Predictable transcriptome evolution in the convergent and complex bioluminescent organs of squid

M. Sabrina Pankey, Vladimir N. Minin, Greg C. Imholte, Marc A. Suchard, and Todd H. Oakley

*Ecology, Evolution, and Marine Biology Department, University of California, Santa Barbara, CA 93106; Departments of *Statistics and *Biology, University of Washington, Seattle, WA 98195; and Departments of *Biomathematics, *Biostatistics, and *Human Genetics, University of California, Los Angeles, CA 90095

Edited by Sean B. Carroll, University of Wisconsin, Madison, WI, and approved September 23, 2014 (received for review August 28, 2014)

Despite contingency in life’s history, the similarity of evolutionarily convergent traits may represent predictable solutions to common conditions. However, the extent to which overall gene expression levels (transcriptomes) underlying convergent traits are themselves convergent remains largely unexplored. Here, we show strong statistical support for convergent evolutionary origins and massively parallel evolution of the entire transcriptomes in symbiotic bioluminescent organs (bacterial photophores) from two divergent squid species. The gene expression similarities are so strong that regression models of one species’ photophore can predict organ identity of a distantly related photophore from gene expression levels alone. Our results point to widespread parallel changes in gene expression evolution associated with convergent origins of complex organs. Therefore, predictable solutions may drive not only the evolution of novel, complex organs but also the evolution of overall gene expression levels that underlie them.

Unrelated species often evolve predictably similar features when presented separately with the same environmental or biological challenges (1, 2). For example, the camera eyes of cnidarians, cephalopods, and vertebrates represent convergent and adaptive solutions to visual acuity (3), just as the repeated evolution of electric organs in teleost fish provides solutions for communication, navigation, and defense (4). A long-standing question is to what extent convergent molecular changes, either in protein-coding sequences or in gene expression, are associated with convergent phenotypic traits. Recent studies indicate many parallel genetic changes in protein-coding sequences may sometimes be related to convergent phenotypes, even between distantly related species (5). However, published gene expression studies of convergent traits are mainly limited to one or a few genes, and the transcriptomes underlying convergent traits are poorly understood. Although previous transcriptomic comparisons from long-diverged species indicate that largely conserved gene expression profiles characterize homologous traits (6, 7), the potential utility of gene expression suggests that widespread changes in gene expression could readily produce similar phenotypic outcomes over short evolutionary times (8–10), especially if natural selection acts on a highly optimal phenotype. In contrast, if gene expression patterns are evolutionarily conserved, perhaps by historical or developmental constraints, then homologous traits should have highly correlated expression, whereas convergent traits should be less correlated because of their different evolutionary origins (11–14). Depending on the extent to which genes are recruited and/or modulated in parallel, expression levels of genes in convergent traits may be more correlated than expression levels of genes in two dissimilar, nonhomologous traits. We tested this prediction using the photophores of bioluminescent squid.

Bioluminescence and complex associated optical structures have originated frequently in the history of life (15, 16), providing numerous instances of convergent evolution. Although bioluminescent structures are found in many cephalopod species (16), photophores harboring bioluminescent bacteria are known only in five genera of sepiolid squid (including Euprymna), and one genus of loliginid squid (Uroteuthis) (17, 18). The bacterial photophore of Euprymna scolopes has served as a model in the field of symbiosis biology (17), but relatively little is known about photophores of Uroteuthis, beyond the harboring of a similar strain of luminescent Vibionraceae bacteria in an organ bearing a similar morphology to Euprymna’s photophore (19) (SI Appendix, Fig. S1). In addition to using elaborate, eye-like optical devices such as concave reflectors and large crystalline lenses to project emitted light (8–10, 20), these photophores enable E. scolopes to control the intensity of ventrally directed light and to distort the silhouette produced from down-welling light (11, 12, 21, 22). The photophore interior in E. scolopes comprises epithelial crypts capable of mediating bacterial growth and detecting light (15, 16, 19, 23, 24).

Bacterial photophores of squid are evolutionary novelties that may have arisen separately in two squid clades (6, 7, 25), but the details of these evolutionary origins have remained obscure. Although many novelties evolve by modifications to existing structures—such as fishes’ electric organs, which are probably modified muscles—bacterial photophores are not obviously related to any structure in nonluminescent squid that might represent a single, direct evolutionary precursor. The best candidate for a trait with historical relationship to photophores is the accessory nidamental gland (ANG), another bacterial

Significance

Unless there are strong constraints, the probability of complex organs originating multiple times through similar trajectories should be vanishingly small. Here, we report that similar light-producing organs (photophores) evolved separately in two squid species, yet each organ expresses similar genes at comparable levels. Gene expression is so similar that overall expression levels alone can predict organ identity, even in separately evolved traits of squid species separated by tens of millions of years. The striking similarity of expression of hundreds of genes in distinct photophores indicates complex trait evolution may sometimes be more constrained and predictable than expected, either because of internal factors, like a limited array of suitable genetic building blocks, or external factors, like natural selection favoring an optimum.

Author contributions: M.S.P. and T.H.O. conceived the study; M.S.P. and T.H.O. designed research; M.S.P. and T.H.O. performed research; M.S.P. collected specimens, conducted phylogenetic tests, performed qPCR assays, and applied distance and GLM analyses to transcriptome data; M.S.P., V.N.M., G.C.I., and M.A.S. contributed new reagents/analytic tools; V.N.M., G.C.I., M.A.S., and T.H.O. conceived the Bayesian independent origins test; V.N.M., G.C.I., and M.A.S. wrote code for the Bayesian independent origins test; M.S.P. analyzed data; V.N.M. provided advice on GLM analyses; and M.S.P. and T.H.O. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: The data reported in this paper has been deposited in the in National Center for Biotechnology Information Short Read Archive, www.ncbi.nlm.nih.gov/sra (BioProject PRJNA257113).

1To whom correspondence should be addressed. Email: todd.oakley@fiscesi.ucsb.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1416574111/-/DCSupplemental.
symbiotic organ positioned near the photophore in luminous squid (20, 26–28) and also present in nonluminous squid. The ANG consists of a cluster of tubules housing a consortium of bacteria, and was very likely present in the common ancestor of squid (25, 29, 30). The microscopic anatomy, ontogenetic and bacteriogenic properties of the photophore bear striking similarities to those of the ANG (20, 31). Although the photophore does not develop from the ANG, both develop in intimate association with the hindgut–ink sac complex (27, 32). These observations led to the hypothesis that the ANG could have given rise to a second bacteriogenic tubule in the adjacent hindgut–ink sac complex via furbeculation in sepiolid squid including Euprymna (20, 28, 33). Despite similarities with ANG, bacterial photophores appear to be a complex amalgamation of structural and physiological components, which suggests ANG alone cannot be an evolutionary precursor. For instance, photophores share developmental, phototransduction, and optical components with eyes and share immunological properties with skin, gills, and the ANG (17, 22, 23, 34). As such, we posit that—like mammalian nipples, which are also a complex amalgamation of structural features with different evolutionary histories—bacterial photophores of squid belong to a class of novelty termed “combinations of pleiomorphic elements” [*sensu* (35)].

Here, we present strong statistical support for separate origins of bacterial photophores during cephalopod evolution. Furthermore, we observe significantly higher similarity in transcriptomic profiles among these convergent photophores than between photophores and any other tissues, including one hypothesized precursor structure (ANG), suggesting that extensive parallel evolution of gene expression underlies this morphological convergence.

**Results**

**Strong Phylogenetic Support for Separate Origins of Bacterial Bioluminescent Organs in Squid.** Our phylogenetic analyses significantly support separate evolutionary origins (phylogenetic nonhomology) of bioluminescent organs in two distantly related squid species. We found that three separate analyses of character evolution [maximum parsimony, maximum likelihood (ML), and Bayesian inference] produced congruent inferences of at least two separate gains of bacterial photophores during evolution, with at least one gain before the divergence of the Sepiolinae (including Euprymna) and another gain within the Loliginidae clade (including Uroteuthis) (Fig. 1). Alternatively, a hypothesis of photophore homology requires at least eight losses under parsimony. Under a ML framework, we estimated the marginal likelihood of photophore presence or absence at each node in the ML tree topology under a two-rate Markov model (Fig. 1 and *SI Appendix*, Fig. S2), which also recovered significant support for nonhomology of photophores. Here, the scenario in which the most recent common ancestor between *Euprymna* and *Uroteuthis* lacks a photophore fits the data significantly better. Finally, we also found significant support for nonhomology using a Bayesian, continuous-time Markov chain approach to model presence/absence transitions (36) (*SI Appendix, SI Materials and Methods*). Here, we conducted a formal Bayes factor test that returns the ratio of the probability of observing our data under the null hypothesis that photophores were gained at most two times, to the probability of the alternative that photophores were gained at least three times. To account for topological uncertainty, we conducted this test integrating over 1,000 bootstrapped ML topologies. Under a wide variety of different priors on photophore gain and loss rates, the log Bayes factor ranged from $10^{-10}$ to 0.0002.

**Convergent Transcriptomes Are Just as Similar as Homologous Transcriptomes.** Although we recover extensive phylogenetic support for separate origins for photophores in squid, we found overall adult gene expression profiles of photophores to be just as similar in our two target species as overall gene expression of homologous organs of those species (Fig. 2). To compare overall gene expression of different organs, we generated 36 separate Illumina transcriptomes from three mature female individuals from each study species (*E. scolopes* and *U. edulis*). Each of the six animals, we sequenced photophores plus five noncontroversially homologous organs, many of which share physiological or structural similarities to bacterial photophores (eye, brain, gill, skin, ANG). Photophore transcriptomes—even from different species—were statistically more similar to each other than to any other sequenced tissue—including ANG—as evidenced by each of three different distance measures (Spearman rank, cosine, and Bray–Curtis distance; under each measure $P < 0.001$).

Surprisingly, under the cosine and Bray–Curtis measures, the overall gene expression distance between photophores is indistinguishable from the distance between expression levels of any homologous organ of the two species, on average (Fig. 2 and *SI Appendix, Fig. S4*). We next used ordination to quantify the amount of tissue-specific signal contained in the transcriptomes. A scree plot of the principal components extracted from the 36 transcriptomes indicates that the amount of vari-

Pankey et al.

---

**Fig. 1.** Evolutionary reconstructions of bacterial photophores on a ML tree of cephalopods. Bacterial photophores are found in species denoted by red tip symbols. Branch color depicts character state evolution inferred by parsimony. Pie charts report marginal likelihood of character states for selected ancestral nodes under an asymmetrical Markov model. Desaturated pie charts identify nodes where neither state significantly improved the model likelihood. A Bayesian test of independent origin (shown in Inset) also strongly favors hypotheses of at least two gains. (Note: branch lengths not to scale for basal lineages.) Right panels of photophores in *E. scolopes* (Top) and *U. edulis* (Bottom) depict the organ’s ventral positioning on ink sac, the conspicuous bilobed lensed structure, and internal morphology in cross-section. (b.c., bacterial crypt; i.s., ink sac; l, lens; MRCA, most recent common ancestor; r, reflector layer.)
We found transcriptomic signal of organ identity to be so strong that overall gene expression divergence evident between the two species, led us to hypothesize that gene expression profiles of photophores from one species may be sufficient to predict the tissue identity of the phylogenetically nonhomologous photophore transcriptome of another species.

**Convergent Transcriptomes Predictable in Logistic Regression Models.** We found transcriptomic signal of organ identity to be so strong that a multinomial logistic regression model, a type of generalized linear model (GLM), can accurately predict organ identity from knowledge of gene expression levels alone, even in phylogenetically nonhomologous photophores. Widespread parallel changes to gene expression could occur either through numerous independent regulatory changes or through changes in just a few master control genes. In the case of the latter, distance metrics may overestimate similarity between transcriptomes. With this in mind, we implemented an elastic-net penalization during model training, which reduces overparameterization in the GLM—effectively lessening the number of genes (predictor variables) included in the model. This penalization algorithm addresses the issue of colinearity among gene expression levels that may stem from the expression of master control genes such as transcription factors (37). After training and cross-validating a tissue model from each of the 18 transcriptomes from each of our two study species, we could accurately predict tissue identity for the all transcriptomes of the opposite species (all samples: $P < 0.005$). We tested the significance of these predictions by generating a set of 10,000 “null transcriptomes” for each tissue (a null transcriptome was produced by randomly sampling genes without respect to tissue identity). We then attempted to predict tissue identity with the GLM models for each null transcriptome. Regardless of which species was used to train the model, all replicate photophore transcriptomes from the other species were accurately predicted as photophores, with significantly better scores than the null distribution (Fig. 3 and SI Appendix, Fig. S6 A and B, and Table S4).

**Differential Expression in Lens Crystallin, Opsin, and Peroxidase Genes Between Convergent Photophores.** Although expression levels of many genes were indistinguishable in our RNA-seq experiments, we used quantitative PCR (qPCR) on genes previously studied in *E. scolopes* photophore to corroborate and to assess robustness of single-gene estimates of differential expression from RNA-seq experiments. These qPCR tests confirmed significant differences in expression levels of four important candidate loci (*SI Appendix, Fig. S3*) between photophores in adults of a representative sepiolid (*E. scolopes*) and loliginid (*U. edulis*) species.

Using qPCR, we first compared opsin expression. Although *E. scolopes* uses an opsin phototransduction pathway in the light-sensitive epithelial lining that harbors bacterial symbionts (24), we detected no opsin expression in *U. edulis* outside the retina. We next compared halide peroxidase expression, because *E. scolopes* relies on this enzyme to regulate symbiont populations (38). We found that *E. scolopes* expressed nearly 800 times as much halide peroxidase in the photophores as *U. edulis* (Mann–Whitney, $P = 0.048$), suggesting that the *Uroteuthis* photophore may rely on alternate mechanisms to mediate symbiont growth and may be under reduced oxidative...
stress compared with *Euprymna*. Third, we compared expression of lens crystallin genes. Although the lenses of squid eyes are largely composed of S-crystallins (derived from glutathione S-transferases), the photophore of the squid *E. scolopes* uses an unrelated Ω-crystallin (derived from aldehyde dehydrogenase) (39, 40). We found that, although both S- and Ω-crystallins are detectable in bioluminescent organs, *U. edulis* expresses predominantly Ω-crystallins, whereas *E. scolopes* expresses predominantly S-crystallins (Mann–Whitney, P < 0.003). Because the recruitment of Ω-crystallins may be related to adaptation to high levels of oxidative stress in *E. scolopes* bioluminescent organs (41), the dominance of S-crystallins in *U. edulis* organs also suggests reduced oxidative stress compared with *Euprymna*.

**Discussion**

We report very strong statistical phylogenetic evidence that bacterial photophores of two cephalopod species are evolutionary convergent, having originated separately in some loliginids (including *Uroteuthis*) and some sepiolids (including *Euprymna*). At the same time, we show that overall gene expression patterns within the convergent photophores are themselves convergent, to the point that photophore identity is predictable from gene expression models constructed from a convergent photophore. The striking similarity of expression of hundreds of genes in evolutionarily distinct photophores could indicate complex trait evolution may sometimes be more constrained and predictable than expected, either because of internal factors, like developmental constraints, or external factors, like natural selection favoring an optimum. However, there are also important alternatives to consider.

A first alternative to separate origins and massively parallel evolution of gene expression is that bacterial photophores are homologous and were lost numerous times in species that lack them. Parsimony would require eight losses, a number that cannot be evaluated statistically, leading to our Bayesian phylogenetic test. The test quantifies different probabilities to indicate—given our standard discrete state Markov model—the observed data are approximately and conservatively 5,000 times less likely to have arisen from an evolutionary history with less than three gains of photophores than from an evolutionary history with three or more photophore gains (*SI Appendix*, Table S2). Although dependent on the underlying homogeneous model of evolution, this is extraordinarily strong evidence for separate origins of bacterial bioluminescent organs in squid. To put these results in the context of classical t tests, Bayes factors of such absolute magnitude lead to rejection of the null hypothesis with a significance threshold much lower than 0.005 (36).

A second alternative to massively parallel evolution of gene expression accompanying separate origins of photophores is that a homologous precursor trait evolved in parallel in at least two instances into photophores, while maintaining very similar gene expression patterns. Here, much of the similarity in gene expression we observe could be attributable to homology, whereas the differences in expression in a few genes could indicate evolutionary divergence. Some of our results do not support this alternative. Previous morphological and developmental data (20, 27, 28, 31, 32) suggest the best candidate for such an evolutionary precursor of bacterial photophores is the ANG, even though similarities also exist between photophores and other organs like eyes, skin, and gills. If the ANG was the “protophotophore” and the cause of our observed transcriptome sim-

---

**Fig. 3.** Transcriptomes of convergent traits contain predictable signatures of gene expression. *(A)* A regression model fit of *Euprymna* transcriptomes accurately predicts better than chance all three *Uroteuthis* transcriptomes that originated from independent photophores. Gray distributions show model predictions for null transcriptomes (generated by bootstrapping transcript read counts across six tissues). *(B)* A reciprocal model fit of *Uroteuthis* data also predicts which three *Euprymna* transcriptomes belong to photophores. All real photophore transcriptomes scored prediction values ≥ 99.5% of null distribution. Colored tissues identify the six sources of transcriptomes used to train each model. Results for homologous tissue prediction are in *SI Appendix*, Fig. S6 A and B, and Table S4.
illarities between convergent photophores, we would also expect greater transcriptome similarities between each species’ photophore and its own ANG than between the photophores of different species because—as shown for the homologous organs we studied—traits related by evolutionary descent should share greater transcriptomic similarity than unrelated traits (see also refs. 7, 42, and 43). In contrast, we observe significantly greater profile similarity between photophores than between photophores and ANGs (SI Appendix, Fig. S7). In addition, photophores share physiological and developmental features with other organs besides ANG, notably gills and eyes (24, 34, 38, 39). The convergent photophores are also more similar to each other in gene expression than to eye or gill expression. As such, we posit that—like mammalian nipples, which are also a complex amalgamation of structural features with different evolutionary histories—bacterial photophores of squid belong to a class of novelty termed “combinations of pleisiomorphic elements” [sensu (35)]. This origin of a photophore by amalgamation may have occurred at least two times during the history of squid in surprisingly similar ways, leading to predictable similarity in gene expression.

Regardless of the evolutionary significance of the bacteriogenic ANG for the origins of photophores, the gene expression profiles in these two focal species suggest an important role for genic ANG for the origins of photophores, the gene expression ways, leading to predictable similarity in gene expression. At least two times during the history of squid in surprisingly similar “...”—The convergent photophores are also more similar to each other in gene expression than to eye or gill expression. As such, we posit that—like mammalian nipples, which are also a complex amalgamation of structural features with different evolutionary histories—bacterial photophores of squid belong to a class of novelty termed “combinations of pleisiomorphic elements” [sensu (35)]. This origin of a photophore by amalgamation may have occurred at least two times during the history of squid in surprisingly similar ways, leading to predictable similarity in gene expression.

Regardless of the evolutionary significance of the bacteriogenic ANG for the origins of photophores, the gene expression profiles in these two focal species suggest an important role for genic ANG for the origins of photophores, the gene expression ways, leading to predictable similarity in gene expression. At least two times during the history of squid in surprisingly similar ways, leading to predictable similarity in gene expression. A

Transcriptome Sequencing. We prepared samples of all six organs from three individuals from each species. We first extracted RNA from the resulting 36 samples using the RibonPure kit (Ambion) and then prepared libraries for Illumina sequencing on the Hi-Seq platform at University of California, Berkeley, using the TruSeq kit (Illumina).

Phylogenetic Analysis. We estimated the ML topology of 70 cephalopod species in RAxML, version 7.6.8 (50), under a GTR+ΓAMMA model using concatenated nucleotide sequences for up to 13 loci (12S, 16S, 18S, aldolase, ATP synthetase B, COI, cytB, EF1α, histone3α, octopine dehydrogenase, opsin, Pax6) from previously published genes and published or newly generated transcriptome datasets (see SI Appendix, Table S1, for accessions). In particular, species belonging to Sepiolid and Loliginid families were heavily targeted for sequence sampling. The final taxon sample of 70 cephalopods included representatives from all (10 out of 15) sepioïds and all (9 out of 11) loliginids genera for which sequence data are accessible. Sepioïd sampling covered 25 of ~91 nominal species. We scored species as present or absent for bacterial photophores (summarized in ref. 25). We estimated ancestral character states using the ML topology under parsimony in Mesquite (51) and under a two-rate Markov model in the R package “corhMM” (52). To account for phylogenetic uncertainty in cephalopod relationships and to estimate the minimum number of times photophores originated, we used a continuous-time Markov chain model and Bayes factor test on the set of 1,000 RAxML-bootstrapped ML topologies. [The Bayes factor test can be implemented in a new R package (53) (“indorigins”), with additional details on independent origins test in SI Appendix, SI Materials and Methods.] Briefly, these Bayes factors (54) are calculated from the probability of the data conditioned on a null and alternative hypothesis specifying a minimum number (m) of photophore gains occurring on a set of trees:

\[
BF = \frac{Pr(Data | H_m)}{Pr(Data | H_{m+1})}
\]

Quantitative PCR on Candidate Loci. We used qPCR to test whether the relative expression levels differ in U. edulis and E. scolopes for the following genes, which have been previously shown to play important roles in the physiology of the E. scolopes photophore: opsin, cryptochrome1, cryptochrome2, s-crystallin, Ω-crystallin, NFκappaB, and peroxidase. RNA from 10 individuals from each species was isolated from photophores as described earlier for sequencing library preparation. qPCRs were prepared using GoTaq qPCR Master Mix (Promega), and quantification was measured on a Bio-Rad iCycler with a 54 °C annealing temperature. β-Actin serves as a reference gene for comparing expression levels between species. We used a Mann–Whitney U test to determine if log-fold changes were significant between species. Primer sequences are available in SI Appendix, Table S2.

Transcriptome Raw Read Processing. We used FastQC (55) to assess read qualities and identify sequencing artifacts. We then trimmed residual adapters and low-quality end via sliding window using FastX Toolkit (56) until all reads scored above 20 at each position. We identified overrepresented sequences by FastQC removed by TagDust (57). All six read libraries constructed from each individual were pooled for assembly using Trinity (05-2012 release) (58). To remove potential bacterial contamination, any assembled contigs receiving a tBLASTn match (< e-150) to the genomes of Vibrio Fischeri ES114 (GenBank: NC_006840[2-12.0]) or Photobacterium leiognathi (NZ_DF09339936930) were restricted from further analysis. Reads from each tissue library were then mapped back to the individual’s single assembly using Bowtie (59) with the S3EM package (60). The original unaltered reads have been deposited in the National Center for Biotechnology Information Short Read Archive under BioProject PRJNA257113.

Transcriptome Ortholog Clustering. To assign orthology among genes in the six Trinity assemblies, we used an iterative clustering method using CD-HIT software (61). Redundancy within each of the six individual assemblies was reduced by clustering sequences at 100% similarity. Intraspecies assembly clustering was then performed to further cluster sequences at 90% similarity. The longest sequence from each cluster was translated using Trinity’s ORPredictor tool and then used in an interspecific reciprocal BLASTp to identify orthologous sequences.

Quantification of Transcriptome Similarity. Transcriptome saturation was verified by rarefaction analysis using the “vegan” R package (62). Any orthologous gene represented by fewer than three reads in a given library was considered absent in that library. Differences in library size were accounted for by adjusting RSEM’s expected read counts for each library using TMM-normalization.
factors in Edger (63). To prevent artifacts stemming from the separate de novo read assemblies performed on each individual, we only retained orthologs from the analysis in which more than one individual had recovered a minimum of 20 mappable reads across their libraries. Quantifying similarity requires considering the caveats inherent to similarity/distance measures. To ensure that our choice of similarity measure alone did not bias our conclusions, we report results using three different measures. Cosine similarity (an information retrieval metric similar to Pearson’s correlation in that it assesses shared variance but is less sensitive to zeros), Spearman rank correlation coefficient (although similar to Pearson’s), is less sensitive to outliers and more conservative if violating distribution assumptions, and Bray-Curtis similarity (an ecological index that is less influenced by presence of low-count variables) were evaluated across the pairwise tissue comparisons on the log-transformed normalized read count libraries using R packages “lsa,” “bioDist,” and “Fossil.” We evaluated whether the changes in the interspecific similarity across converging tissues (photophores) (7) differs from the similarity between nonhomologous tissues or (ii) differs from the similarity observed between homologous tissues. Because the choice of similarity measure could influence the outcome of significance tests, we conducted all tests with the same three measures as described earlier. We used a two-way test of independence (permutation test) in the “coin” R package to determine significance (64).

Multinomial Logistic Regression Model. To determine whether the gene expression signature of a given tissue is sufficient to accurately predict a sample of the same tissue type in another species, we fit a multinomial logistic regression model for each species using the “glmnet” R package (65), using a hybrid elastic-net penalization of variable coefficients to minimize the number of genes effectively included in the model. For each species, we fit the model using all 18 normalized log-transformed read-count libraries and then used the model to predict the tissue identities of the other species’ libraries. To test whether these model predictions were significantly better than chance, we created a null distribution of read-count data matrices by bootstrapping the read counts of each gene, creating bootstrapped tissue libraries with randomized gene abundances. For each species, we created 10,000 sets of these 18 bootstrapped libraries. These bootstrapped datasets were then subjected to the GLM model fit by the original data from the opposite species. The resulting tissue predictions were used as a null distribution with which we tested the significance of the model predictions on real data.

ACKNOWLEDGMENTS. We are grateful to M. Matsuyama, G. Nishihara, M. J. McFall-Ngai, and the Kewalo Marine Laboratory at the University of Hawaii for facilitating animal collections. We thank P. Weakliem for providing computational resources. We thank D. Stern, S. Haddock, S. Hodges, and T. DeTomaso for manuscript comments. This work was funded by National Science Foundation (NSF) Grants NSF-CNS-0960316 awarded to University of California, Santa Barbara, NSF Scientific Computing Center, NSF Grants DEB-1210673 (to M.S.P. and T.H.O.) and DEB-1146337 (to T.H.O.), and NSF Grants DMS-0856099 and DMS-1264153 and NIH Grant R01-A107034 (to V.N.M. and M.A.S.).