Supplementary Methods. Testing phylogenetic independent origin hypotheses using Bayes factors

Abstract

To assess support for the hypothesis that photophores originated more than once, we developed a Bayes factor test in which we compare the prior and posterior probabilities of the observed data under two opposing hypotheses (that the number of gains required is either less than or greater/equal to M). To approximate these prior and posterior probabilities we developed a computational method which uses MCMC to account for phylogenetic uncertainty and uncertainty in rates of gain and loss. Here we describe the model assumptions and computational details, which have been implemented in the R package `indorigin` available at `https://github.com/vnminin/indorigin`.

Modeling assumptions

We start with a binary character (e.g. absence/presence of a morphological trait) measured in n species. We collect these measurements into vector $\mathbf{y} = (y_1, \ldots, y_n)$, where each $y_i \in \{0, 1\}$. Suppose that the evolutionary relationship among the above species can be described by a phylogeny τ, which includes branch lengths. We assume that the binary character had evolved along this phylogeny according to a two-state continuous-time Markov chain (CTMC) with an infinitesimal rate matrix

$$\mathbf{\Lambda} = \begin{pmatrix} -\lambda_{01} & \lambda_{01} \\ \lambda_{10} & -\lambda_{10} \end{pmatrix}.$$

We also assume that we have another set of data $\mathbf{x}$, molecular and/or morphological, collected from the same species. In principle, we can set up an evolutionary model for this second data set, with evolutionary model parameters $\boldsymbol{\theta}$ (e.g. substitution matrix, rate heterogeneity parameters) and then approximate the posterior distribution of all model parameters conditional on all available data:

$$\Pr(\tau, \boldsymbol{\theta}, \lambda_{01}, \lambda_{10} \mid \mathbf{x}, \mathbf{y}) \propto \Pr(\mathbf{x} \mid \tau, \boldsymbol{\theta}) \Pr(\mathbf{y} \mid \tau, \lambda_{01}, \lambda_{10}) \Pr(\tau) \Pr(\boldsymbol{\theta}) \Pr(\lambda_{01}) \Pr(\lambda_{10}), \tag{1}$$

where we assume that *a priori* $\lambda_{01} \sim \text{Gamma}(\alpha_{01}, \beta_{01})$ and $\lambda_{10} \sim \text{Gamma}(\alpha_{10}, \beta_{10})$, with the rest of the priors left unspecified for generality. However, in practice the contribution of the data vector $\mathbf{y}$ to phylogenetic estimation is negligible when compared to the contribution of the data matrix $\mathbf{x}$. Therefore, we take a two-stage approach, where we first approximate the posterior distribution

$$\Pr(\tau, \boldsymbol{\theta} \mid \mathbf{x}) \propto \Pr(\mathbf{x} \mid \tau, \boldsymbol{\theta}) \Pr(\tau) \Pr(\boldsymbol{\theta}) \Pr(\lambda_{01}) \Pr(\lambda_{10})$$

via Markov chain Monte Carlo (MCMC). This produces the posterior sample of K phylogenies, $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_K)$. This sample can also be generated via a bootstrap procedure within the maximum likelihood analysis. Next, we form an *approximate* posterior distribution

$$\begin{aligned} \widetilde{\Pr}(\lambda_{01}, \lambda_{10} \mid \mathbf{x}, \mathbf{y}) &= \int_{\tau} \Pr(\lambda_{01}, \lambda_{10} \mid \tau, \mathbf{y}) \Pr(\tau \mid \mathbf{x}) \mathrm{d}\tau \\ &\propto \int_{\tau} \Pr(\mathbf{y} \mid \tau, \lambda_{01}, \lambda_{10}) \Pr(\lambda_{01}) \Pr(\lambda_{10}) \Pr(\tau \mid \mathbf{x}) \mathrm{d}\tau \\ &\approx \left[\sum_{k=1}^{K} \Pr(\mathbf{y} \mid \tau_k, \lambda_{01}, \lambda_{10}) \right] \Pr(\lambda_{01}) \Pr(\lambda_{10}). \end{aligned} \tag{2}$$

that helps us estimate the rates of gain and loss of the trait, λ_{01} and λ_{10}, appropriately accounting for phylogenetic uncertainty. The approximate posterior (2) has only two parameters and therefore can be approximated by multiple numerical procedures, including deterministic integration techniques, such as Gaussian quadrature. We implement a MCMC algorithm that targets posterior (2), but plan to experiment with deterministic integration in the future.

So far our modeling assumptions and approximations follow standard practices in statistical phylogenetics as applied to macroevolution. For example, one could use software packages BayesTraits [Pagel et al., 2004] or Mr.Bayes [Ronquist et al., 2012], among many others, to approximate the posterior distributions (1) or (2). The main novelty of our methodology, explained in the next section, comes from the way we use these posteriors to devise a principled method for testing hypotheses about the number of gains and losses of the trait of interest.

Hypotheses and their Bayes factors

Let N_{01} be the number of gains and let N_{10} be the number of losses. Conservatively, in this work we assume that the root of the phylogenetic tree relating the species under study is in state 1. This means that the parsimony score for the number of gains associated with vector $\mathbf{y}$ and *any* phylogeny is 0, because under our assumption about the root any binary vector can be generated with only trait losses, even though such an evolutionary trajectory may be very unlikely.

We fix a nonnegative threshold m and formulate an *independent origin hypothesis* associated with this threshold as

$$\mathrm{H}_0 : N_{01} \leq M,$$

with the corresponding alternative

$$\mathrm{H}_a : N_{01} > M.$$

This means that our null hypothesis is that the trait was gained at most $M + 1$ times — we add one because we know that the trait was gained at least once. For example, using $M = 0$ corresponds to testing the null hypothesis that the trait was gained only once some time prior to the time of the most recent common ancestor of the species under study. We use a Bayes factor test [Kass and Raftery, 1995] to compare the above two hypotheses:

$$\mathrm{BF}_M = \frac{\Pr(\mathbf{y} \mid N_{01} \leq M)}{\Pr(\mathbf{y} \mid N_{01} > M)} = \frac{\Pr(N_{01} \leq M \mid \mathbf{y})/\Pr(N_{01} \leq M)}{\Pr(N_{01} > M \mid \mathbf{y})/\Pr(N_{01} > M)}, \tag{3}$$

where $\Pr(N_{01} \leq M \mid \mathbf{y})$ and $\Pr(N_{01} > M \mid \mathbf{y})$ are the posterior probabilities of the null and alternative hypotheses, and $\Pr(N_{01} \leq M)$ and $\Pr(N_{01} > M)$ are the corresponding prior probabilities. We explain how we compute these probabilities in the next section.

Computational details

We approximate the posterior (2) by a MCMC algorithm that starts with arbitrary initial values $\lambda_{01}^{(0)}$, $\lambda_{01}^{(0)}$ and at each iteration $l \geq 1$ repeats the following steps:

1. Sample uniformly at random a tree index k from the set $\{1, \ldots, K\}$ and set the current tree $\tau^{(l)} = \tau_k$.

2. Conditional on the phylogeny and the gain and loss rates from the previous iteration, draw a realization of the full evolutionary trajectory (also known as stochastic mapping [Nielsen, 2002]) on phylogeny $\tau^{(l)}$ using the uniformization method [Lartillot, 2006] and record the following missing data summaries: $N_{01}^{(l)}$, $N_{10}^{(l)}$, defined as before, and $t_0^{(l)}$, $t_1^{(l)}$ — total times the trait spent in state 0 and 1 respectively.

3. Draw new values of gain and loss rates from their full conditionals:

$$\lambda_{01}^{(l)} \sim \text{Gamma}(N_{01}^{(l)} + \alpha_{01}, t_0^{(l)} + \beta_{01}),$$
$$\lambda_{10}^{(l)} \sim \text{Gamma}(N_{10}^{(l)} + \alpha_{10}, t_1^{(l)} + \beta_{10}).$$

Advantages of using the above Gibbs sampling algorithm are: a) no tuning is required and b) augmenting the state space with latent variables, N_{01}, N_{10}, t_0, t_1, and sampling these latent variables efficiently yield rapid convergence of the MCMC, in our experience.

The last important computational issue is computing prior and posterior probabilities needed to compute the Bayes factor (3). Consider computing the posterior probability $\Pr(N_{01} \leq M \mid \mathbf{y})$ — a surprisingly nontrivial task, as it turns out. For example, the most straightforward approximation of this probability from our MCMC output is

$$\Pr\left(N_{01} \leq M \mid \mathbf{y}\right) \approx \frac{1}{L} \sum_{l=1}^{L} 1_{\{N_{01}^{(l)} \leq M\}},$$

where $1_{\{\}}$ is an indicator function. This approximation has substantial Monte Carlo error, a result of the large variance of N_{01}, which makes using this approximation infeasible for Bayes factor calculations, especially when $\Pr(N_{01} \leq M \mid \mathbf{y})$ is close to 0 or to 1. Alternatively, a better approximation can be formed as follows:

$$\Pr(N_{01} \leq M \mid \mathbf{y}) \approx \frac{1}{L} \sum_{l=1}^{L} \Pr(N_{01}^{(l)} \leq M \mid \mathbf{y}, \lambda_{01}^{(l)}, \lambda_{10}^{(l)}, \tau^{(l)}), \tag{4}$$

where $\Pr(N_{01} \leq M \mid \mathbf{y}, \lambda_{01}, \lambda_{10}, \tau)$ is the posterior probability of at most m jumps on a fixed tree τ, assuming known gain and loss rates, λ_{01} and λ_{10}. To compute the last posterior probability, we first compute $\Pr(N_{01} = m \mid \mathbf{y}, \lambda_{01}, \lambda_{10}, \tau)$ for $m = 0, \ldots, M$ and then sum these probabilities to obtain the desired quantity.

Computing $\Pr(N_{01} = m \mid \mathbf{y}, \lambda_{01}, \lambda_{10}, \tau)$ can be accomplished by combining analytic results of Minin and Suchard [2008] and a dynamic programming algorithm of Siepel et al. [2006]. We further extend the analytic results of Minin and Suchard [2008] with an alternate representation of the two-state model solution to improve the numerical stability of our calculations. We compute the prior probability of at most M jumps using an approximation analogous to formula (4), with the exception of averaging over independent draws from priors of λ_{01} and λ_{01}, and over uniform draws of candidate phylogenies $\tau_1, \ldots, \tau_K$.

Implementation and illustrations

Software implementing the above procedure is available in the form of an open-source R package `indorigin` (`https://github.com/vnminin/indorigin`). To install the package install the `devtools` package and then install `indorigin` using 'install_github' command:

```r
## install.packages("devtools") # uncomment if "devtoos" is not installed
## install_github("vnminin/indorigin") # uncomment or copy and paste into R terminal
library(indorigin)

## Loading required package:  Rcpp
## Loading required package:  RcppArmadillo
## Loading required package:  testthat
```

Notice that installing from github requires installing the package from source. To learn about package installation see `http://cran.r-project.org/doc/manuals/R-admin.html`.

Simulated data

Let's simulate a tree and fast/slow evolving binary traits on this tree.

```r
library(diversitree) # diversitree is only needed for simulations

## Loading required package:  deSolve
## Loading required package:  ape
## Loading required package:  subplex
## Loading required package:  methods

set.seed(3245)

## Simulate a tree
phy<-NULL
while(is.null(phy)){
  phy <- tree.bd(c(.1, .03), max.taxa=50)
}

## Simulate FAST EVOLVING 0/1 characters on this tree
states1 <- sim.character(phy, c(.03, .1), x0=1, model="mk2")

## Simulate SLOW EVOLVING 0/1 characters on this tree
states2 <- sim.character(phy, c(.001, .01), x0=1, model="mk2")
```

First, we analyze the data simulated under the fast evolving trait regime. In this case, the Bayes factor strongly rejects the hypothesis that there were 0 gains of the trait.

```r
## run the independet origin analysis on the simulated data.
## Notice that the first argument must be a list of trees even if you are
## supplying one tree. Hence, c(phy) command.
testIndOriginResults1 = testIndOrigin(inputTrees=c(phy), traitData=states1,
initLambda01=.01, initLambda10=.01, priorAlpha01=1, priorBeta01=10,
priorAlpha10=1, priorBeta10=10, mcmcSize=2100, mcmcBurnin=100,
mcmcSubsample=1, mcSize=10000)

## pre-processing trees and trait data
```

```r
## plot the tree with the simulated histories at all nodes
par(mfrow=c(1,2))
plot(phy, show.tip.label=FALSE, no.margin=TRUE)
col <- c("#004165", "#eaab00")
tiplabels(col=col[states1+1], pch=19, adj=1)
nodelabels(col=col[attr(states1, "node.state")+1], pch=19)

plot(phy, show.tip.label=FALSE, no.margin=TRUE)
col <- c("#004165", "#eaab00")
tiplabels(col=col[states2+1], pch=19, adj=1)
nodelabels(col=col[attr(states2, "node.state")+1], pch=19)
```

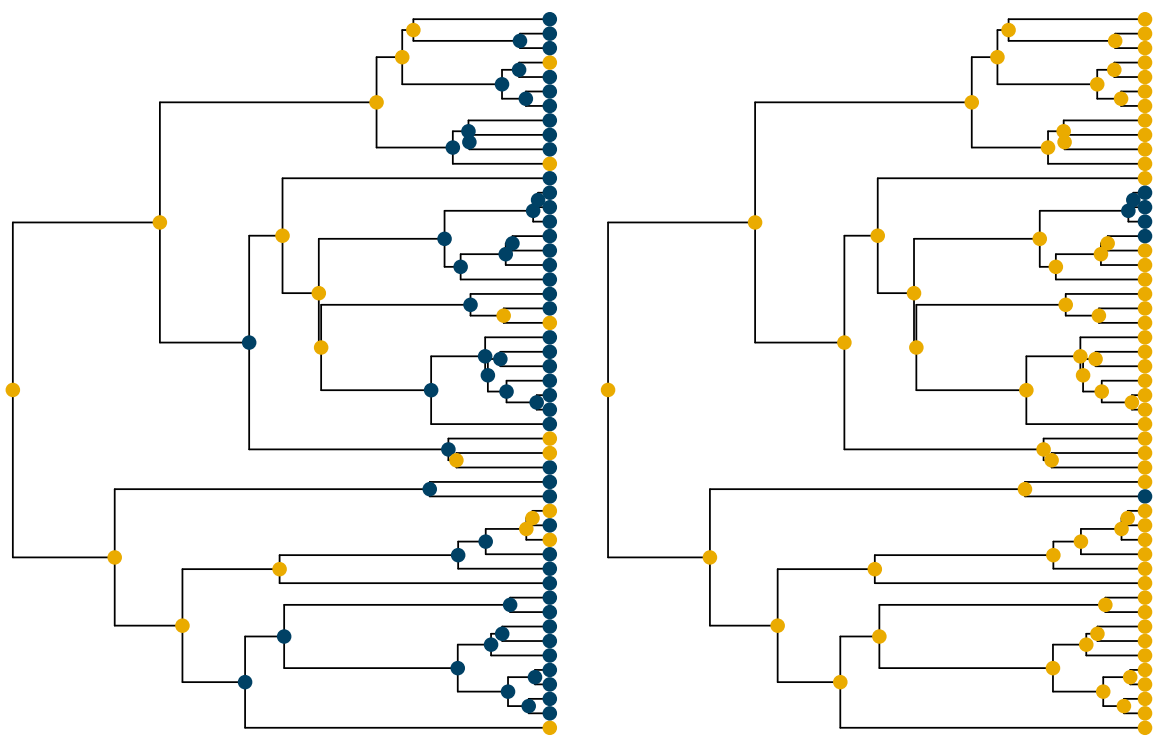


Figure 1: Fast (left figure) and slow (right figure) evolving binary traits with true internal node states plotted.

```
## running Gibbs sampler
## Computing posterior probabilities
## Computing prior probabilities

getBF(testIndOriginResults1)

## BF for N01<=0     log10(BF)   2xlog_e(BF)
##     5.173e-06    -5.286e+00    -2.434e+01
```

When we perform a similar analysis for the slow evolving trait, the Bayes factor supports the hypothesis of 0 gains, but the support is very weak. This is expected, because data generated under the slow evolving model have very little information about the gain/loss rates, so there is a lot of uncertainty about these rates.

```
testIndOriginResults2 = testIndOrigin(inputTrees=c(phy), traitData=states2,
initLambda01=.01, initLambda10=0.1, priorAlpha01=1, priorBeta01=10,
priorAlpha10=1, priorBeta10=10, mcmcSize=2100, mcmcBurnin=100,
mcmcSubsample=1, mcSize=10000)

## pre-processing trees and trait data
## running Gibbs sampler
## Computing posterior probabilities
## Computing prior probabilities

getBF(testIndOriginResults2)

## BF for N01<=0     log10(BF)   2xlog_e(BF)
##       1.11587       0.04761       0.21927
```

Photophores data

First, we are going to load 1000 phylogenies of 70 cephalopod species and a corresponding vector of trait values (presence/absence of photophores).

```
library(ape)

cephalopodTrees = read.tree("BLsonboots70.phy")
tree.num = length(cephalopodTrees)
cephalopodTraits = read.csv("BLspecies70.csv", header=FALSE)

# a little massaging to get trait data into a vector format
tip.num = dim(cephalopodTraits)[1]
cephalopodTraitVec = numeric(tip.num)
tipNames = as.character(cephalopodTraits[,1])
cephalopodTraitVec = cephalopodTraits[,2]
names(cephalopodTraitVec) = tipNames

# run the analysis
```

```
cephalopodIndOriginResults = testIndOrigin(inputTrees=cephalopodTrees,
traitData=cephalopodTraitVec,initLambda01=.01, initLambda10=0.1,
priorAlpha01=1, priorBeta01=100, priorAlpha10=1, priorBeta10=10,
mcmcSize=1100, mcmcBurnin=100, mcmcSubsample=1, mcSize=1000, testThreshold=2)

## pre-processing trees and trait data
## running Gibbs sampler
## Computing posterior probabilities
## Computing prior probabilities

# get the Bayes factor
getBF(cephalopodIndOriginResults)

## BF for N01<=2     log10(BF)   2xlog_e(BF)
##     3.722e-06    -5.429e+00    -2.500e+01
```

The above results reproduce the Bayes factors in the 7th row of the SI Table 2. To reproduce the rest of the rows, one can change 'priorBeta01', 'priorBeta10', and 'testThreshold' parameters to manipulate the priors and the hypotheses. Note that we kept the number of MCMC iterations low, so it is possible to reproduce all of the above examples quickly. You should increase this number when attempting your own analyses.

References

R.E. Kass and A.E. Raftery. Bayes factors. Journal of the American Statistical Association, 90:773–795, 1995.

N. Lartillot. Conjugate Gibbs sampling for Bayesian phylogenetic models. Journal of Computational Biology, 13:1701–1722, 2006.

V.N. Minin and M.A. Suchard. Counting labeled transitions in continuous-time Markov models of evolution. Journal of Mathematical Biology, 56:391–412, 2008.

R. Nielsen. Mapping mutations on phylogenies. Systematic Biology, 51:729–739, 2002.

M. Pagel, A. Meade, and D. Barker. Bayesian estimation of ancestral character states on phylogenies. Systematic Biology, 53:673–684, 2004.

F. Ronquist, M. Teslenko, P. van der Mark, D.L. Ayres, A. Darling, S. Höhna, B. Larget, L. Liu, M.A. Suchard, and J.P. Huelsenbeck. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Systematic Biology, 61:539–542, 2012.

A. Siepel, K.S. Pollard, and D. Haussler. New methods for detecting lineage-specific selection. In Proceedings of the 10th International Conference on Research in Computational Molecular Biology, pages 190–205, 2006.

SI Table 1. GenBank sequence identifiers (GI numbers) for loci used in analysis listed alphabetically. Changes to the species name used in this manuscript are listed in bold. Newly generated sequences indicated by yellow cells.

GenBank Name	12S	16S	18S	ald3	ATPsyth	COI	cytb	ef1a	H3	odh	opsin	pax
Afrololigo mercatoris		93004770				194474586					194474656	
Alloteuthis africana		194474470				194474508					194474600	
Alloteuthis media		194474490				194474550					194474634	
Alloteuthis subulata		194474505				194474580					194474650	
Argonauta nodosa	45510911	45510935	49482067			4003405	76359244		50346987	45510951	45511050	
Chiroteuthis calyx		209969953	209969999			209970103			209970211			
Loligo gahi **= Doryteuthis gahi**		93004761				5678658						
Loligo opalescens **= Doryteuthis opalescens**	2402666124	2402666123				2402666121	2402666122			12055096		1778016
Loligo pealei **= Doryteuthis pealeii**		56567271	34369176			18026439	EF423109_1		38607257		AY450853_1	
Loligo plei **= Doryteuthis plei**		93004765				14120087						
Eledone cirrhosa	48994420	18073265	49482072							48994473	48994575	48994529
Enteroctopus dofleini	62005876	45510940				62084159	239735795	JF927838_1		45510965	45510966	45511018
Euprymna berryi	34542045	14161379				13195587						
Euprymna hyllebergi	34542048	34542074				34542132						
Euprymna morsei	62005906	34542070										
Euprymna scolopes	34542046	34542072			JF927843_1	34542128		JF927842_1		48994346	48994378	21667880
Euprymna tasmanica	34542047	34542073				34542130					48994587	48994541
Graneledone verrucosa	45510922	82561543	49482073			4003433			50346991	45510973	158828841	45511024
Heterololigo bleekeri	97906254	97906253				979062511	97906252					
Heteroteuthis hawaiiensis	34542063	34542089	49482077			34542160			50346997	48994344	48994376	48994406
Idiosepius paradoxus	62005903	62005892				62084191						AB716344_1
Loligo forbesi	45510930	498938				5678661				45510987	45511086	45511040
Loligo reynaudii		93004766				5678665						
Loligo vulgaris	77157317	498939				5678656				13561066		
Loliolus beka		384228941				384228950						
Loliolus japonica		93004768				5678666	145244494					
Loliolus uyii		325073373				330426895						
Lolliguncula brevis	48762898	93004769				5678655				48994332	48994364	48994394
Lolliguncula diomedeae		209969959	209969984			209970073			209970169			
Nautilus macromphalus	944906654	9449066533	17385427			911773921	944906652					
Nautilus pompilius	48994439	38607021	34369177	JF927850_1		18026437		JX036488_1				48994567
Nautilus scrobiculatus		571333	18026322									
Neorossia caroli		117960047										
Octopus bimaculoides	45510917	18076178		JF927853_1		15421827		JF927854_1		45510961	45511062	45511014
Octopus vulgaris	537938874	537938873				537938871	537938872			HM104284_1	116829804	HM104274_1
Rondeletiola minor	34542060	34542086				34542154						
Rossia bipapillata	34542059	34542085										
Rossia pacifica	62005904	62005893				5353806				48994342	48994374	48994404
Rossia palpebrosa		209969908	49482078			4003471			50346999			
Semirossia tenera	38374163	38374164				38374165						
Sepia officinalis	892552844	892552843	49482076	FO202690.1	FO162309_1	892552841	892552842	FO202892_1	50346995	13561068	315075492	121495688
Sepiadarium austrinum	48994423	48994450								48994479	48994581	48994535

SI Table 1. GenBank sequence identifiers (GI numbers) for loci used in analysis listed alphabetically. Changes to the species name used in this manuscript are listed in bold. Newly generated sequences indicated by yellow cells.

GenBank Name	12S	16S	18S	ald3	ATPsyth	COI	cytb	ef1a	H3	odh	opsin	pax
Sepiadarium kochi	62005907	34542087				38154317						
Sepietta neglecta	34542056	34542082				34542148						
Sepietta obscura	34542057	34542083				34542150						
Sepietta oweniana	34542058	34542084				34542152						
Sepietta sp.		498954										
Sepiola affinis	34542050	209969909	49482079			34542136			50347001			
Sepiola atlantica	34542055	34542081				34542146						
Sepiola birostrata	34542049	34542075				34542134						
Sepiola intermedia	34542052	34542078				34542140						
Sepiola ligulata	34542051	34542077				34542138						
Sepiola robusta	34542053	34542079				34542142						
Sepiola rondeleti	34542054	34542080				34542144						
Sepiolina nipponensis	34542062	34542088				34542158						
Sepioloidea lineolata	48994422	48994449				4003477				48994477	48994579	48994533
Sepioteuthis australis	48762899	93004772				4003479				48994334	48994366	48994396
Sepioteuthis lessoniana	892554074	892554073	49482085			890008521	892554072		50347013	48994336	48994368	48994398
Sepioteuthis sepioidea		93004775				5678651						
Stauroteuthis gilchristi	45510909	45510933								45510947	45511046	45510998
Stauroteuthis syrtensis		82622206	49482062			4003483			50346978			
Stoloteuthis leucoptera	34542064	209969910	49482080			4003485			50347003			
Taonius pavo	AY616959	209969961	209969992			209970121			209970184			
Uroteuthis chinensis		209969912	23450949			28207579			50347009			
Uroteuthis duvauceli		3618168				5678659						
Uroteuthis etheridgei		93004779										
Uroteuthis edulis		3618173				169247791						
Uroteuthis noctiluca	34542041	34542065				34542116						
Uroteuthis sp. JMS 2004	48762901	48762912								48994338	48994370	48994400
Vampyroteuthis infernalis	1531248574	1531248573	34369180			1531248571	1531248572		50346982	45510945	45511044	45510996

Legend

Black text GI number for public data

sequence generated in this study

Nautilus pompilius:
sequence assembled from available short-read datasets:
Accessions: SRR330442; SRR108979; DRR001114; DRR001111

Octopus vulgaris:
sequence assembled from available short-read datasets:
Accesssions: SRR1946; SRR108980

Idiosepius_paradoxus:
sequence assembled from available short-read datasets:
Accessions DRR001110; DRR001113:

Hypotheses compared for Bayes Factor test	Priors on rates of gain:loss	Prior Probability on H0	BF	log10(BF)	2xlog_e(BF)	Posterior Probablity on H0
HA: Ngains >=2 H0: Ngains <=1	1:100	0.9999944	1.08E-06	-5.97	-27.47	0.162277
	1:10	*0.9995186*	*3.29E-06*	*-5.48*	*-25.25*	*0.00677985*
	1:1	0.9817623	6.23E-06	-5.21	-23.97	0.00033508
	10:1	0.9981345	3.54E-12	-11.45	-52.73	1.89E-09
	100:1	0.9998135	5.63E-19	-18.25	-84.04	3.02E-15
HA: Ngains >=3 H0: Ngains <=2	1:100	1	3.05E-08	-7.52	-34.61	0.4422237
	1:10	*0.9999708*	*3.82E-06*	*-5.42*	*-24.95*	*0.115802*
	1:1	0.993742	2.07E-04	-3.68	-16.96	0.031836
	10:1	0.9993716	5.04E-07	-6.30	-29.00	0.00080095
	100:1	0.9999358	9.45E-11	-10.02	-46.17	1.47E-06
HA: Ngains >=4 H0: Ngains <=3	1:100	1	6.17E-09	-8.21	-37.81	0.9644096
	1:10	*0.9999982*	*1.12E-05*	*-4.95*	*-22.79*	*0.865057*
	1:1	0.9977474	4.82E-03	-2.32	-10.67	0.6810581
	10:1	0.9997861	1.31E-04	-3.88	-17.87	0.3806625
	100:1	0.9999785	8.03E-07	-6.10	-28.07	0.0360379
HA: Ngains >=5 H0: Ngains <=4	1:100	1	3.37E-09	-8.47	-39.02	0.9995157
	1:10	*0.9999999*	*9.63E-06*	*-5.02*	*-23.10*	*0.991713*
	1:1	0.9992335	7.95E-03	-2.10	-9.67	0.912042
	10:1	0.9999263	1.09E-04	-3.96	-18.25	0.5968848
	100:1	0.9999926	8.37E-07	-6.08	-27.99	0.1018062

SI Table 2. Results of Test of Independent Origins.
The Bayes Factor test results for MCMC runs under 4 different null-alternative hypothesis pairs. For each hypothesis test, the rates of gain and loss were varied in the prior parameters, ranging from 1:100 (losses occur, on average, 100 times more often than gains) to 100:1 (gains 100 times more frequent). Earlier ML analysis under a 2-rate model estimated losses 10 more likely than gains (italized rows). Bayes Factors strongly and consistenly favored hypotheses of at least 2 or 3 gains across rate priors tested. Posterior probabilities indicate that these null hypotheses are least favored under priors which increase the relative rate of loss.

		U. edulis			*E. scolopes*		
Control genes	*actin*	*Ue_actinRT_F*	**56**	CACCGCCGAGAGAGAAATTG	*Es_actinRT_F*	**56.4**	ATGTTCCCCGGTATTGCTGA
		Ue_actinRT_R	**55.9**	CCTGTTCGAAGTCAAGAGCG	*Es_actinRT_R*	**56.3**	CGCCGATCCAGACAGAGTAT
		amplicon(bp)		***71***			***115***
	18s	*RT_Sepiola_18s-F*	**53.7**	CGTTTTCCTCGATCAAGAGC	*RT_Sepiola_18s-F*	**53.7**	CGTTTTCCTCGATCAAGAGC
		RT_Sepiola_18s-R	**54.8**	CATCGTTTACGGTCGGAACT	*RT_Sepiola_18s-R*	**54.8**	CATCGTTTACGGTCGGAACT
		amplicon(bp)		***77***			***77***
Photo-detection	*opsin*	*Ue_ops_F*	**56.1**	GGGCTATCGGCCCTATCT	*RT_Es_Op_F*	**54.8**	CGAAGCATATGAGCCACAGA
		Ue_ops_R	**54.9**	AATGTTGGATCGTGTAGCTGTATC	*RT_Es_Op_R*	**55.7**	CCGATAGCCCATAGGACAGA
		amplicon(bp)		***107***			***76***
	loph-opsin				*Es_lophops_F*		CTCTCAATCAGCACGCTAACA
					Es_lophops_R		GCCCAGAAGATGGCATACAC
		amplicon(bp)				**140**	
	cry1	*Ue_cry1_F*	**53**	TGCTTTCGAAAAAGCCTTACA	*Es_cry1_F*	**54**	TGTATGGCATGAGGATGGATAG
		Ue_cry1_R	**53.2**	GGACGAATTCACCATTCCTATC	*Es_cry1_R*	**54.4**	TTCCACGGACAAAACAGATACTT
		amplicon(bp)		***86***			***97***
	cry2	*Ue_cry2_F*	**55.9**	GACTGGTTTCCCCTGGATAGA	*Es_cry2_F*	**56.1**	GACTGGCTTCCCTTGGATAGA
		Ue_cry2_R	**54.3**	CCAAAGATCACCTCTGGTTAAGA	*Es_cry2_R*	**54.3**	CCAAAGATCACCTCTGGTTAAGA
		amplicon(bp)		***112***			***112***
Lens proteins	*s-crystallin*	*Ue_ScrystRT_F*	**55.8**	TGGACATGATGAGGTGTGACT	*Es_ScrystRT_F*	**56.1**	GGTACTTGGCCCGTGAATTC
		Ue_ScrystRT_R	**55.9**	TCCGTTCTTCCAGTGGTAGTAC	*Es_ScrystRT_R*	**55.8**	GAAGCGTCCGTTCTTTTCGT
		amplicon(bp)		***86***			***134***
	o-crystallin	*Ocrys_Ue_F*	**54.9**	TTGAACCAACCGTCTTCTCC	*Ocrys_Es_F*	**52.6**	GCGGAAAGAGCAATTTGAAG
		Ocrys_Ue_R	**55.1**	GCCATTCCATAGTCGGTGTT	*Ocrys_Es_R*	**54.8**	ACCGGACCAATGAGATTGAC
		amplicon(bp)		***142***			***146***
Immuno/ symbiosis proteins	*NFkappaB*	*nfkb_Ue_F*	**54.9**	TGTGAAACCTGAGCTTCCTG	*nfkb_Es_F*	**56.3**	GAAGCTGCTGGTTGTCCTTC
		nfkb_Ue_R	**56.8**	TTTCTTCTGGTGGTGGTGCT	*nfkb_Es_R*	**55.4**	GGATGTTGCTGCCTGAATCT
		amplicon(bp)		***110***			***66***
	peroxidase	*Ue_peroxRT_F*	**55.1**	GGGTGACCGATTCTGGTATG	*Es_peroxRT_F*	**54.7**	TTCCGAAGATGACGCTAACC
		Ue_peroxRT_R	**54**	TCCCTGGATTTGTTGGATGT	*Es_peroxRT_R*	**56.2**	TCAGAAACACCACCAGTCCA
		amplicon(bp)		***127***			***81***
	TLR						

SI Table 3. Primers used for relative quantitative PCR in *E.scolopes* and *U. edulis*.

Euprymna scolopes transcriptomes

	Prediction score of each *Euprymna* sample under *Uroteuthis* GLM transcriptome model of tissue types					
	ang	brain	eye	gill	phot	skin
Es1_skin	0.0268	0.0047	0.007	0.0339	0.0024	0.9253
Es1_phot	0.1393	0.0286	0.0267	0.0878	0.6291	0.0885
Es1_gill	0.0192	0.0115	0.0083	0.9463	0.0018	0.0129
Es1_eye	0.0015	0.0018	0.9876	0.0004	0.0005	0.0083
Es1_brain	0.0046	0.975	0.0065	0.0039	0.0013	0.0086
Es1_ang	0.8589	0.0093	0.0133	0.024	0.0036	0.0911
Es2_skin	0.0066	0.0027	0.0086	0.0226	0.005	0.9545
Es2_phot	0.0514	0.0234	0.0574	0.0862	0.5309	0.2508
Es2_gill	0.0065	0.011	0.0267	0.8537	0.0052	0.0968
Es2_eye	0.0031	0.006	0.979	0.0013	0.0014	0.0092
Es2_brain	0.0029	0.9792	0.0075	0.0028	0.0013	0.0063
Es2_ang	0.476	0.0087	0.019	0.0182	0.0051	0.473
Es3_skin	0.0119	0.0077	0.0082	0.0467	0.0051	0.9205
Es3_phot	0.0879	0.015	0.0155	0.0675	0.6509	0.1632
Es3_gill	0.0074	0.0114	0.0067	0.9583	0.0012	0.015
Es3_eye	0.0007	0.0014	0.9956	0.0004	0.0003	0.0016
Es3_brain	0.0072	0.9543	0.0096	0.0047	0.002	0.0221
Es3_ang	0.8466	0.0142	0.0116	0.0181	0.0076	0.1019

	Pvalues: ***probability of observed prediction on 10000 bootstrapped transcriptomes***					
	ang	brain	eye	gill	phot	skin
Es1_skin	0.8982	0.9996	0.9972	0.8731	0.9832	0.0033
Es1_phot	0.3399	0.9164	0.9239	0.5982	0.0004	0.8854
Es1_gill	0.945	0.9924	0.9953	0	0.9907	0.9966
Es1_eye	0.9999	0.9999	0	1	0.9999	0.9985
Es1_brain	0.998	0	0.9975	0.9999	0.996	0.9984
Es1_ang	0.0001	0.9959	0.9806	0.9254	0.9576	0.8807
Es2_skin	0.996	0.9999	0.9946	0.9317	0.9224	0.0001
Es2_phot	0.7427	0.9462	0.7473	0.6051	0.0021	0.6038
Es2_gill	0.996	0.9931	0.9239	0.0002	0.9185	0.8708
Es2_eye	0.9991	0.9992	0	0.9999	0.9945	0.9983
Es2_brain	0.9994	0	0.9964	0.9998	0.9965	0.999
Es2_ang	0.0244	0.9965	0.9583	0.9536	0.9208	0.3106
Es3_skin	0.9799	0.9979	0.9955	0.7983	0.9218	0.0043
Es3_phot	0.5348	0.9835	0.9727	0.6932	0.0003	0.7499
Es3_gill	0.9945	0.9924	0.9974	0	0.9966	0.9957
Es3_eye	0.9999	0.9999	0	0.9999	0.9999	0.9999
Es3_brain	0.9949	0	0.9921	0.9981	0.9884	0.9903
Es3_ang	0.0002	0.9851	0.9862	0.954	0.8551	0.8605

Uroteuthis edulis transcriptomes

	Prediction score of each *Uroteuthis* sample under *Euprymna* GLM transcriptome model tissue types					
	ang	brain	eye	gill	phot	skin
Ue1_skin	0.082	0.3002	0.2957	0.0106	0.0137	0.2978
Ue1_phot	0.103	0.4288	0.1332	0.1359	0.1823	0.0168
Ue1_gill	0.034	0.0904	0.0925	0.7113	0.013	0.0587
Ue1_eye	0.0032	0.0124	0.9764	0.0025	0.0038	0.0018
Ue1_brain	0.0019	0.9772	0.0155	0.0015	0.0026	0.0013
Ue1_ang	0.4838	0.1889	0.1223	0.0519	0.1359	0.0172
Ue2_skin	0.0229	0.735	0.1749	0.0057	0.0114	0.05
Ue2_phot	0.1177	0.2616	0.1532	0.1039	0.3197	0.0439
Ue2_gill	0.0448	0.166	0.1148	0.5228	0.0037	0.1479
Ue2_eye	0.0016	0.0127	0.9797	0.0024	0.001	0.0026
Ue2_brain	0.0018	0.9775	0.0167	0.0019	0.0016	0.0005
Ue2_ang	0.7395	0.0811	0.0554	0.0867	0.029	0.0082
Ue3_skin	0.0415	0.3851	0.2809	0.0123	0.0626	0.2175
Ue3_phot	0.1344	0.3099	0.1455	0.1137	0.2031	0.0936
Ue3_gill	0.0527	0.0998	0.0677	0.6855	0.0131	0.0812
Ue3_eye	0.0117	0.0472	0.9288	0.0046	0.0013	0.0064
Ue3_brain	0.0021	0.9552	0.0282	0.0028	0.011	0.0007
Ue3_ang	0.7156	0.082	0.0427	0.0667	0.0526	0.0404

	Pvalues: ***probability of observed prediction on 10000 bootstrapped transcriptomes***					
	ang	brain	eye	gill	phot	skin
Ue1_skin	0.3901	0.792	0.4654	0.9117	0.8103	0.0036
Ue1_phot	0.2608	0.4612	0.9265	0.189	0.0447	0.8616
Ue1_gill	0.8792	0.9989	0.9861	0.0001	0.8241	0.2503
Ue1_eye	0.9999	0.9999	0	0.9974	0.9915	0.9999
Ue1_brain	0.9999	0	0.9999	0.9998	0.997	0.9999
Ue1_ang	0	0.9654	0.9482	0.5947	0.1024	0.8543
Ue2_skin	0.9658	0.0131	0.8246	0.9686	0.8619	0.3249
Ue2_phot	0.1933	0.869	0.8832	0.2894	0.0109	0.3926
Ue2_gill	0.762	0.981	0.959	0.0018	0.9922	0.0337
Ue2_eye	0.9999	0.9999	0	0.9975	0.9999	0.9998
Ue2_brain	0.9999	0	0.9999	0.9989	0.9995	0.9999
Ue2_ang	0	0.9993	0.9997	0.3648	0.5422	0.9827
Ue3_skin	0.7982	0.5761	0.5039	0.8966	0.2826	0.0109
Ue3_phot	0.1375	0.7691	0.9004	0.2536	0.0412	0.1057
Ue3_gill	0.6747	0.9982	0.9981	0.0001	0.8226	0.1399
Ue3_eye	0.9979	0.9999	0	0.9807	0.9997	0.9937
Ue3_brain	0.9999	0	0.9999	0.9965	0.8709	0.9999
Ue3_ang	0	0.9992	0.9999	0.4855	0.3341	0.4358

SI Table 4. Results of multinomial logistic regression test.

18 transcriptome libraries from each species were bootstrapped to generate 10000 'null transcriptomes'.

After the regression model was fit and cross-validated using U. edulis, the 18 real E scolopes transcriptomes, and the 10000 null E. scolopes datasets were predicted under the model.

The same approach was repeated to create a GLM of Euprymna transcriptomes and test the fit of real Uroteuthis data and generated null data.

Proportional support for each tissue category shown in top panels, and correspond to attached plots.

Exact p-values in lower tables represent the proportion of the null distributions in which larger predictions are observed.

P-values in red denote samples whose prediction score fell outside 95% of the null distribution for the tissue type.

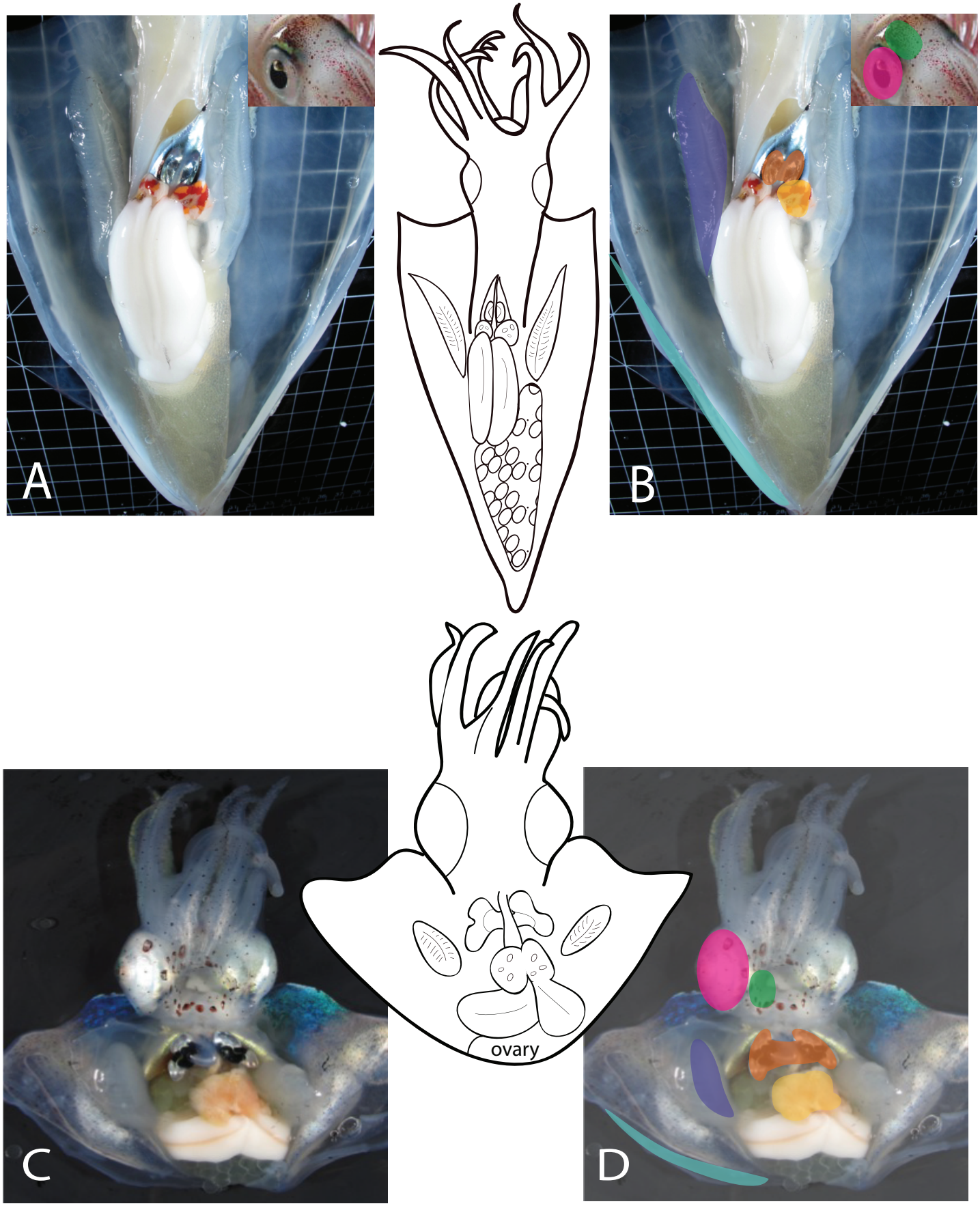


Figure S1. ***Uroteuthis*** **and** ***Euprymna*** **tissues.**
Ventral dissection of *Uroteuthis edulis* (top row A,B) and *Euprymna scolopes* (bottom row C, D) showing organs sampled for transcriptome analyses. Photophores shaded (in orange) in right panels (B,D) , along with homologous organs: eyes (pink), brain (green), ANG (yellow), gill (purple), skin (blue). Cartoons display simplified anatomy.

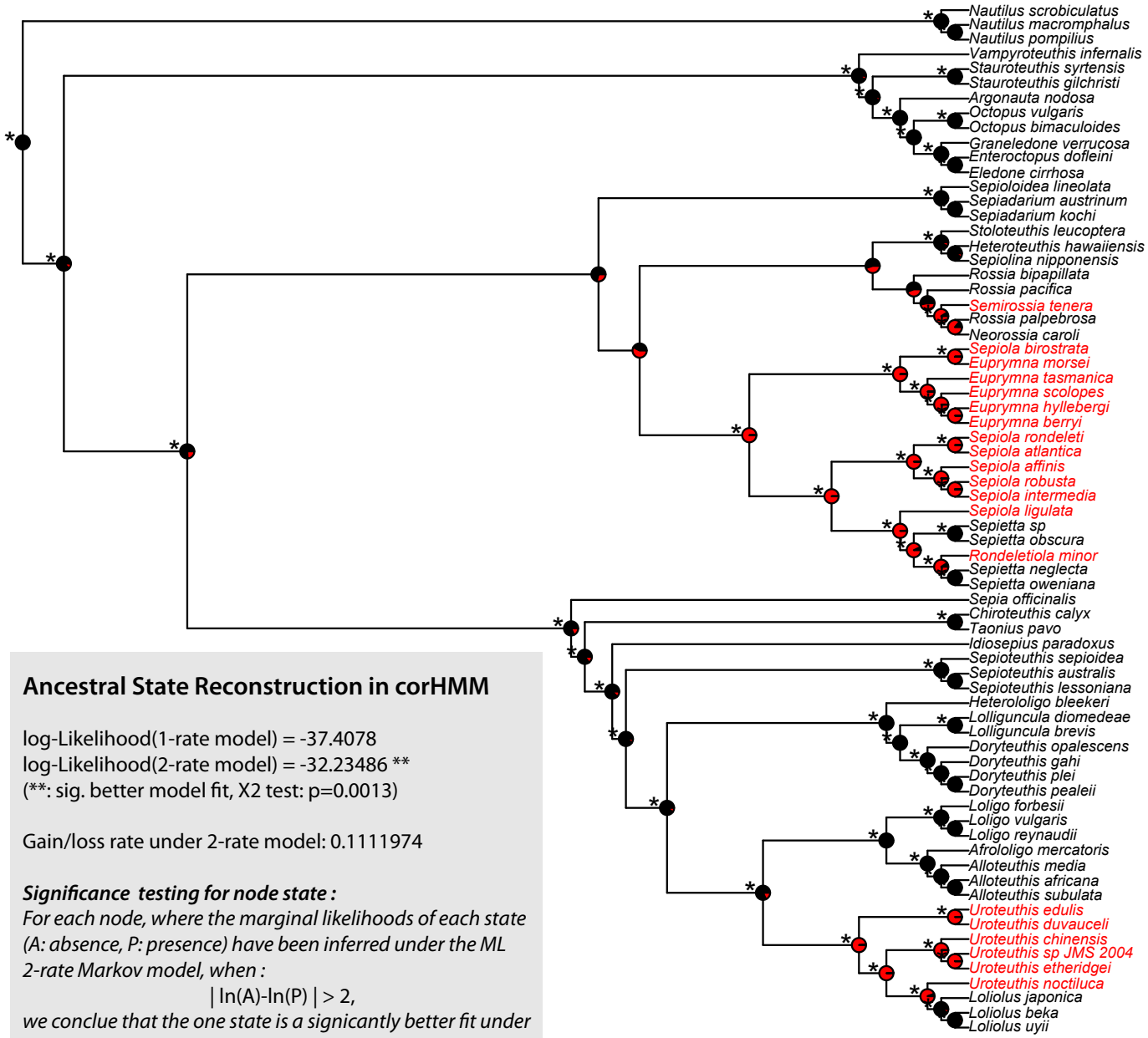


Figure S2. Marginal likelihoods for photophore presence (red) or absence (black) under 2-rate Markov model at ancestral nodes of ML topology. Nodes at which one state signficantly improved the fit of the model over the other state are indicated by (*).

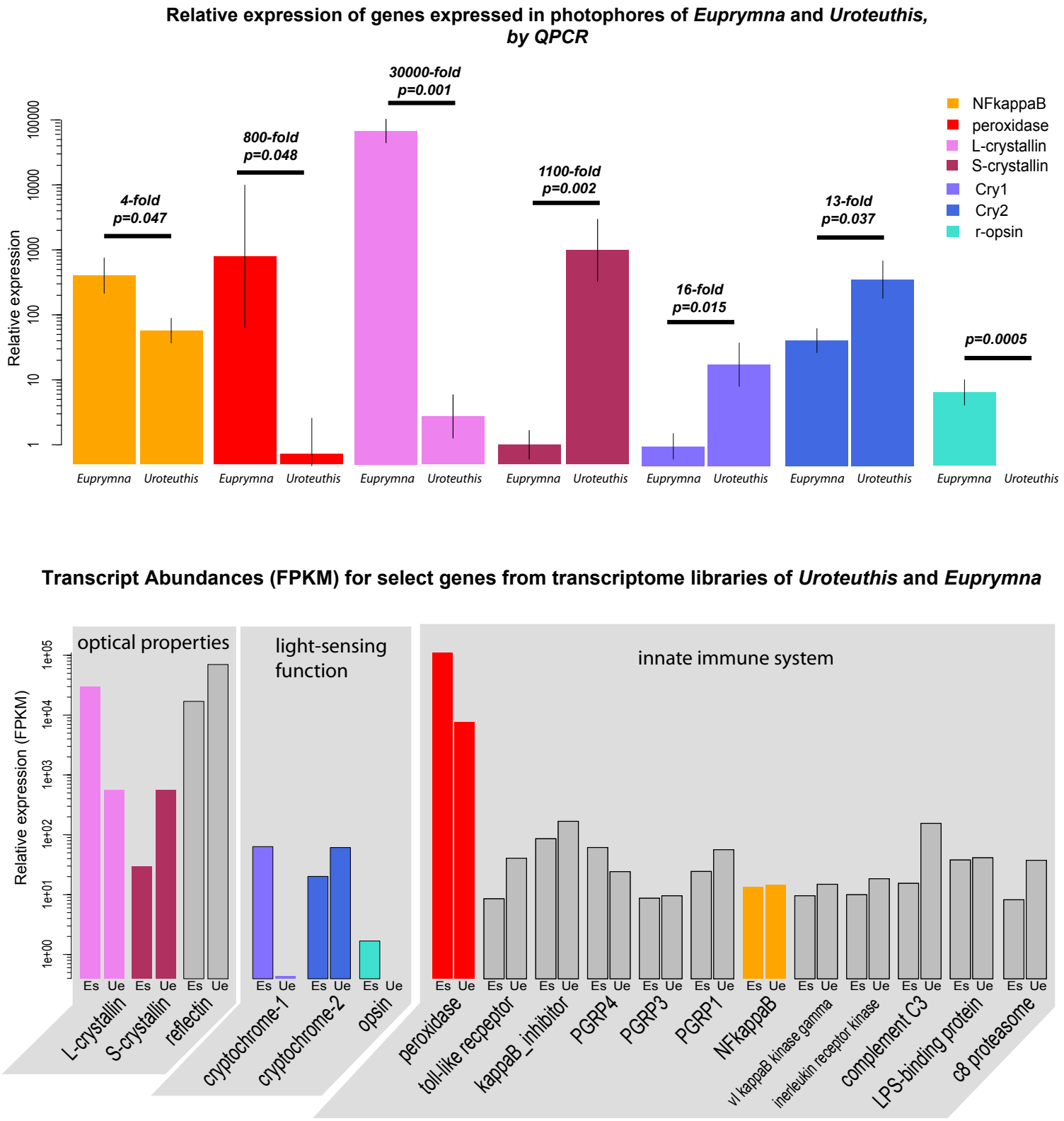


Figure S3. Relative expression of genes expressed in photophores of Euprymna and Uroteuthis.
Top panel: Mean expression levels for L-crystallin, S-crystallin, opsin, and peroxidase in qPCR assays, standardized by actin. Fold-abundance difference, S.E. bars and p-values from Wilcoxon Rank- sum test indicated.
Lower panel: Mean normalized transcript abundances (FPKM) for genes identified in photophore transcriptome libraries (each n=3). Genes grouped by putative functional categories. Only genes in color were assayed for expression differences via qPCR.

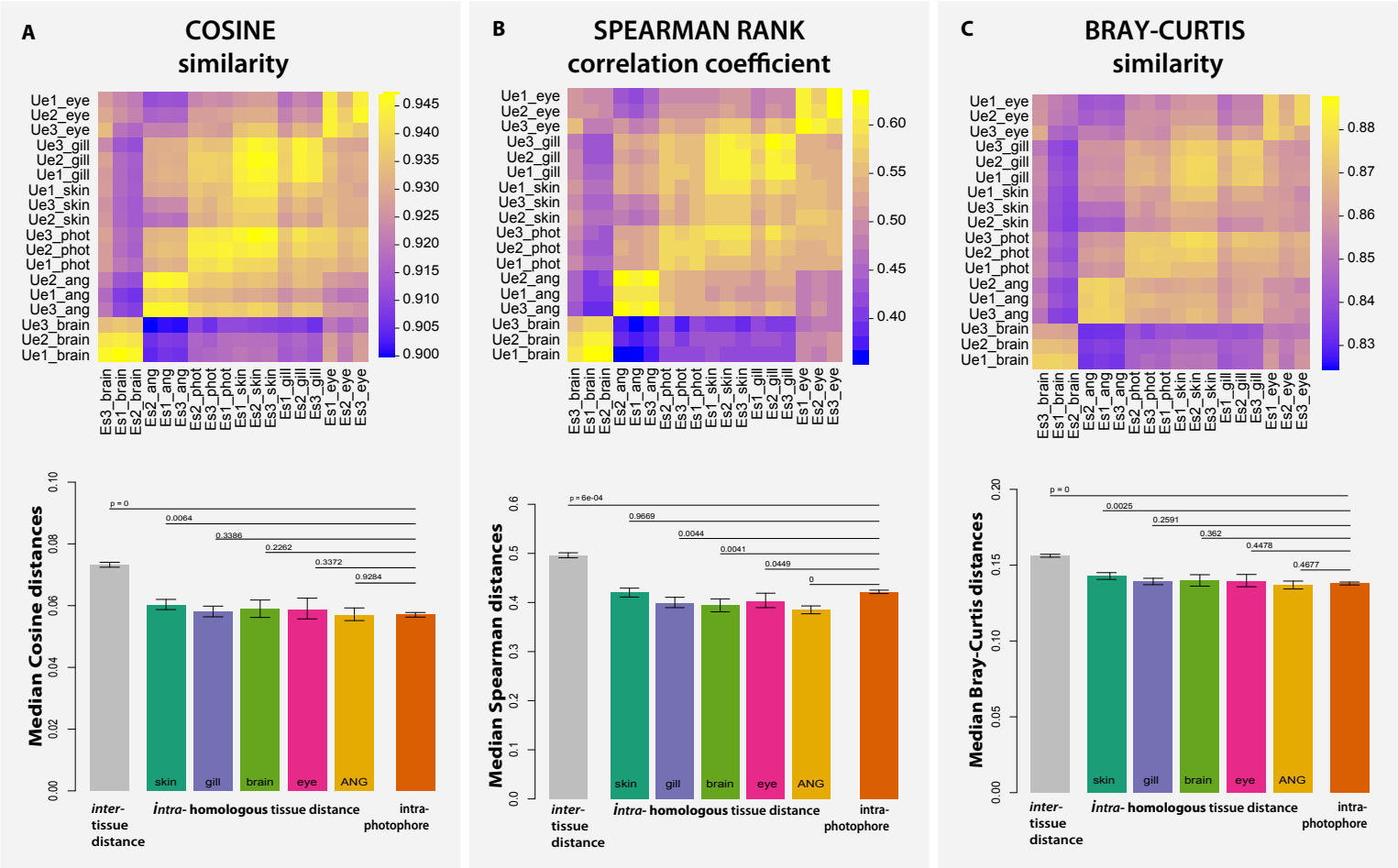


Figure S4. Distances between transcriptomes as measured under (A) Cosine distance, (B) Spearman Rank distance, and (C) Bray-Curtis distance. Upper panel heatmaps depict similarity between the 18 sequenced libraries from each species, ranging from most similar (yellow) to least similar (blue). Lower panel barplots depict the median dissimilarity measured between tissues. Under all 3 distance measures, photophores from Euprymna and Uroteuthis (orange) are less distant (more similar) to each other than expected given non-homologous tissues' distances (grey).

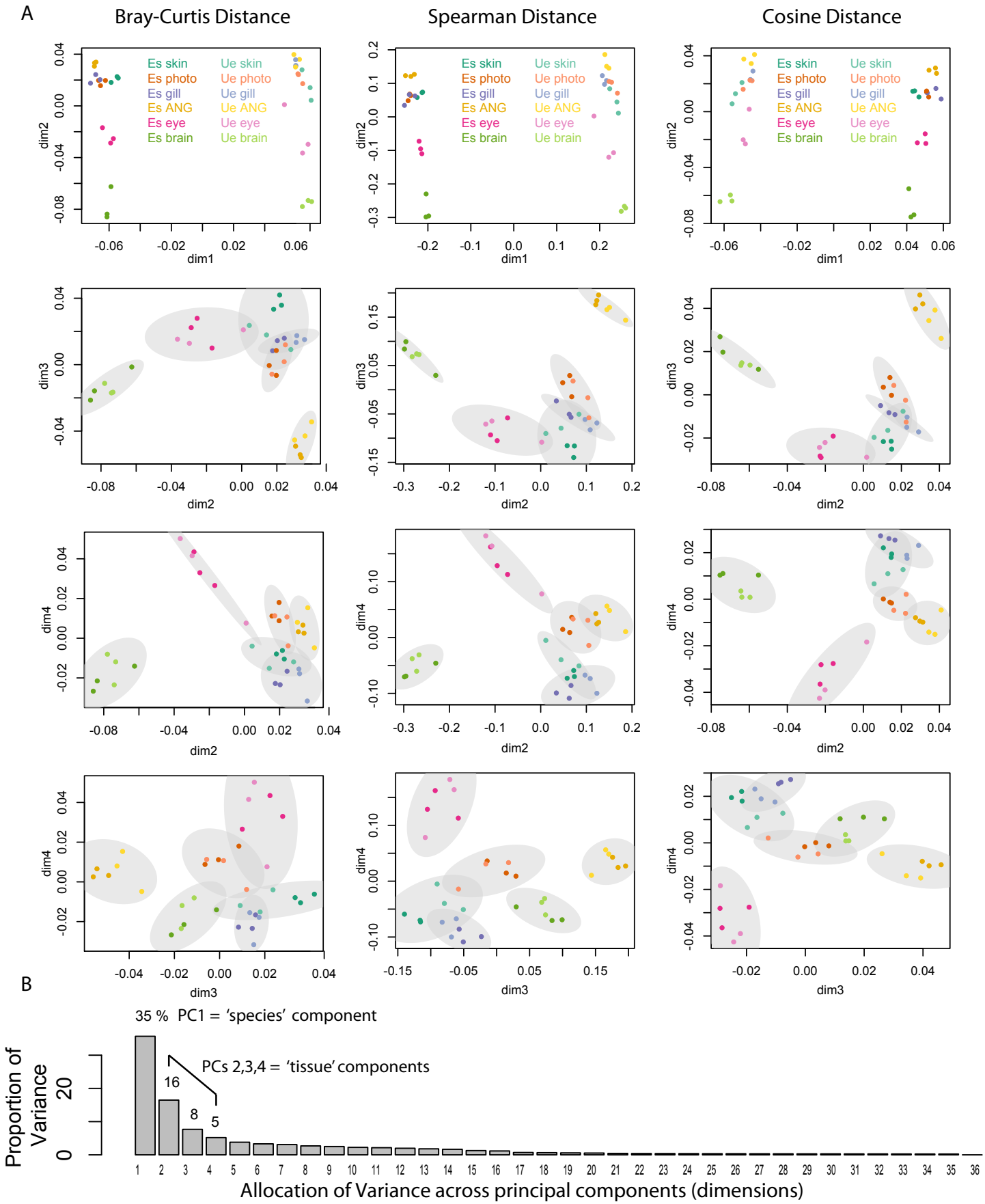
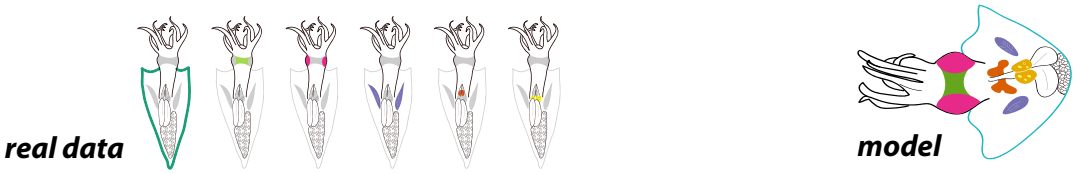


Figure S5. Ordination of latent structure in transcriptomes distinguishes species and tissue signals. A, Non-metric Multidimensional scaling of 36 transcriptomes (2 species, 6 tissues, 3 replicates each) using 3 differentmeasure of distance. For each measure, d imentions 2, 3, and 4 capture variance in gene expression which are shared between tissue type in both species. Dimension 1 capture variance explained by species differences. B, Scree plot showing proportion of variance in the 36-transcriptome dataset captured by each dimension (principal component). Gene expression differences due to species (Dimension 1) accounts for the largest proportion of variation while tissue (additively dimensions 2, 3, 4) accounts nearly the same proportion.

Figure S 6A. *Uroteuthis* transcriptomes tested against *Euprymna* GLM



Transcriptome data from *Uroteuthis* (18 samples; 3 from each tissue type) were predicted under a GLM fitted by 18 *Euprymna* transcriptomes from corresponding tissues. Circles denote the prediction scores for each of the 18 *Uroteuthis* transcriptomes. Filled points represent scores which fell outside of 95% of the null distribution . Nul distributions for the predction scores for each tissue type were generated by testing 10000 bootstrapped *Uroteuthis* libraries against the same *Euprymna* model.

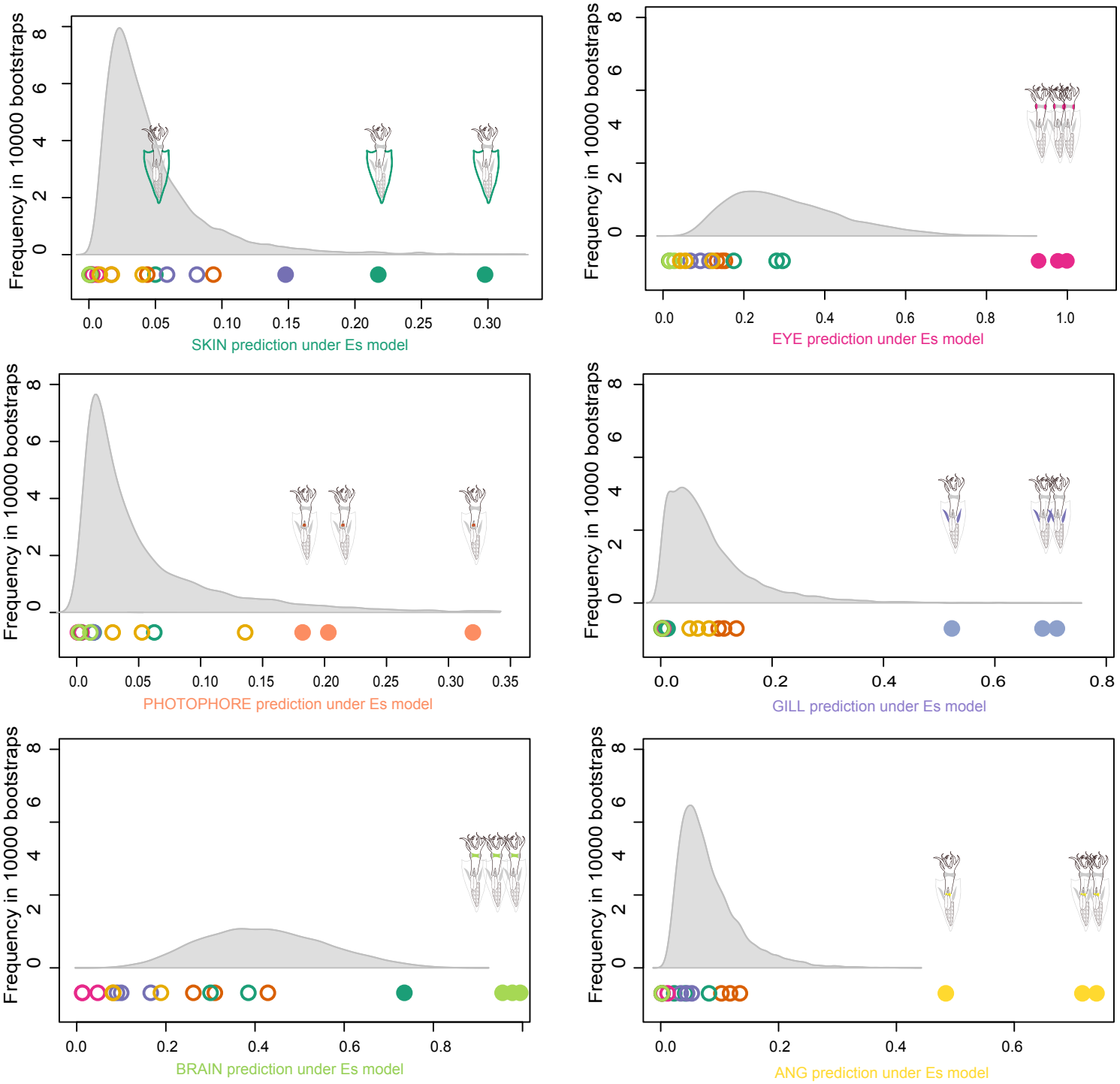
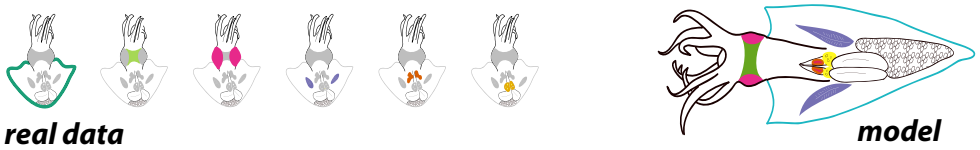
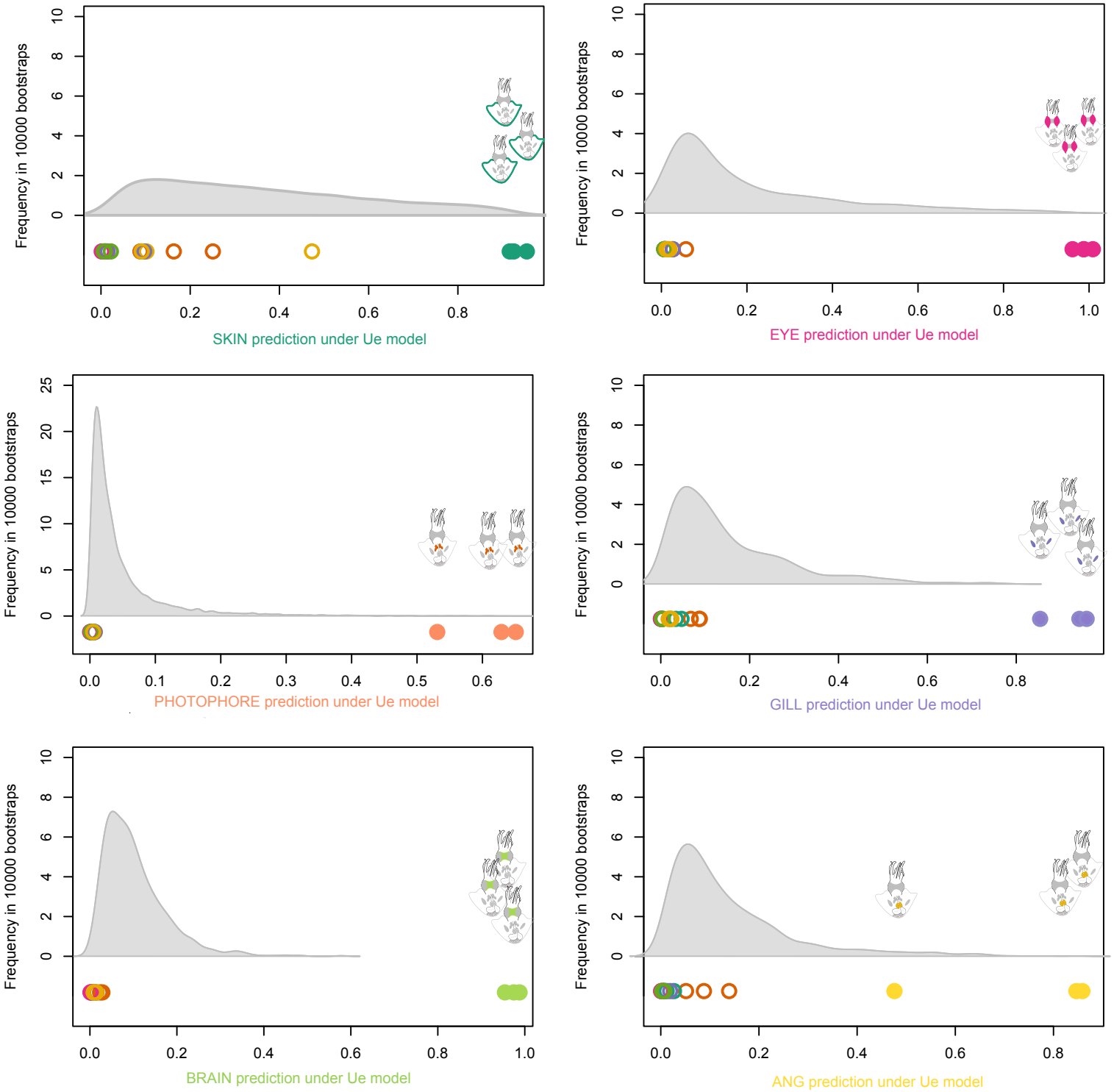


Figure S 6B. *Euprymna* transcriptomes tested against *Uroteuthis* GLM



Transcriptome data from *Euprymna* (18 samples; 3 from each tissue type) were predicted under a GLM fitted by 18 *Uroteuthis* transcriptomes from corresponding tissues. Circles denote the prediction scores for each of the 18 *Euprymna* transcriptomes. Filled points represent scores which fell outside of 95% of the null distribution . Nul distributions for the predction scores for each tissue type were generated by testing 10000 bootstrapped *Euprymna* libraries against the same *Uroteuthis* model.



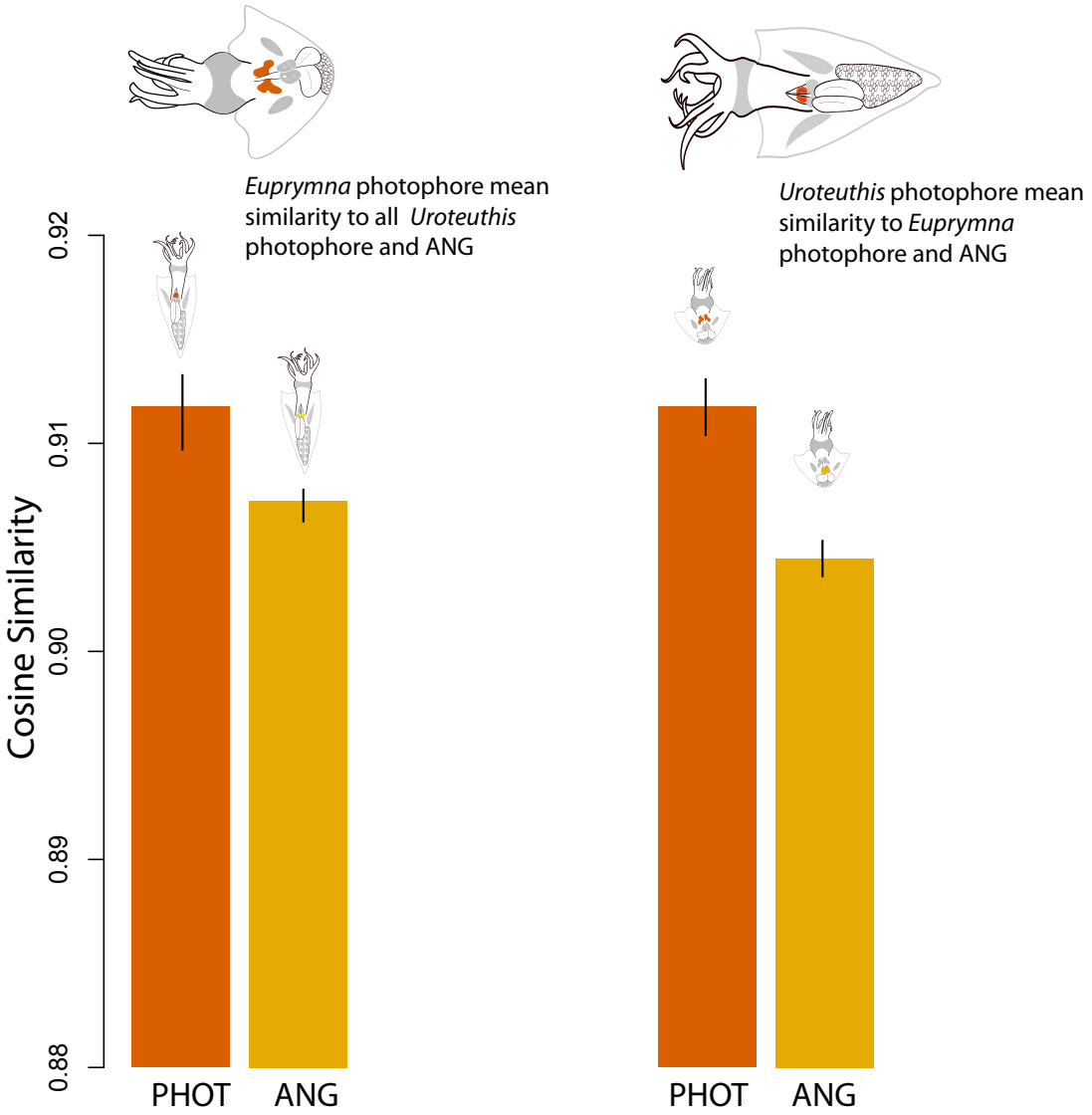


Figure S 7. Photophore transcriptomes share greater similarity with other photophores than photophores do with accessory nidamental glands. Bars represent mean cosine similarities between photophores or between photophores and ANGs. Error bars depict 95% confidence intervals estimated by 500 bootstrap replicates.