Assessment of modularity in the urodele skull: An exploratory analysis using ossification sequence data.

Michel Laurin

Assessment of modularity in the urodele skull: an exploratory analysis using ossification sequence data

By

Michel Laurin[*1]

Sorbonne Universités, CR2P, CNRS/MNHN/UPMC, Muséum National d'Histoire Naturelle, Bâtiment de Géologie. Case Postale 48, 57 rue Cuvier, 75005 Paris, France

Number of text figures: 10

Number of tables: 5

Number of Supplementary On-line Materials: 2

Abbreviated title: Exploratory analysis of modularity in urodeles

[*] Correspondence to: Michel Laurin, Sorbonne Universités, CR2P, CNRS/MNHN/UPMC, Muséum National d'Histoire Naturelle, Bâtiment de Géologie. Case Postale 48, 57 rue Cuvier, 75005 Paris, France. E-mail: michel.laurin@upmc.fr or laurin@mnhn.fr

[1] This research was funded by the CNRS and French Ministry of Research through the operating grant to the CR2P.

Abstract

The potential presence of developmental modules is studied in the urodele skull using several classical statistical methods that have not previously been used in this context. Principal Component Analysis (PCA) of ossification sequence data on 21 bones in 21 extant urodele species suggests the presence of up to four developmental modules, but examination of statistically significant correlations using Phylogenetic Independent Contrasts (PIC) and correcting for multiple tests using the False Discovery Rate suggests the presence of only two modules of uneven size and of two bones that may not be part of these modules. Thus, PCA does not appear to be a reliable method to investigate modularity; direct investigation of statistically significant correlations using PIC or other phylogeny-informed methods is recommended. A binomial test of the distribution of significant correlations between characters shows significant heterogeneity, which suggests that modularity is indeed present in the data. A cluster analysis gives inconsistent results that apparently do not reflect developmental modules. The data include a phylogenetic signal, as shown by a permutation-based test with squared change parsimony, but this is detectable only when the whole matrix is analyzed, and a plot of the tree onto developmental space through Evolutionary Principal Component Analysis shows that homoplasy is pervasive. Evolutionary rates between characters vary about 90-fold. Canonical Variates Analyses suggest that obligatorily neotenic urodeles may be discriminated from other urodeles on the basis of cranial ossification sequence data.

Key words: modularity, evo-devo, ossification sequences, event heterochrony, developmental modules

Text

Introduction

The fashionable concept of modularity, the phenomenon that creates releases covariation between sets of characters, has been studied by biologists from genes to colonies (Schlosser and Wagner, ’04). Modularity may have played a major role in evolution by relaxing some constraints and increasing evolvability, perhaps in a manner vaguely analogous with compartmentation, which acts at another (molecular) level (Kirchner and Gerhart ‘98). Modularity arises from processes that creating sets of characters that are more strongly correlated to each other than to elements of other sets. These process may allow selection to optimize separately sets of characters, thus allowing organisms to reach a fitness value that would be much harder to achieve if all characters were tightly linked to each other (Wagner et al., ’07). In the context of evo-devo tackled below, processes generating developmental modules should release developmental constraints and facilitate the appearance of heterochronies that may improve the fitness of individuals.

Several methods have been used to assess the presence and significance of developmental modules. Some, like clustering methods, have the drawback that linked elements in each cluster may exhibit similarity, but not necessarily covariation (Goswami, ’06: 274). In the context of an event heterochrony, such clusters might simply lump events occurring at similar times in the ontogeny, but not necessarily those that are linked developmentally and whose timing changes are correlated in evolution. Other methods, such as the RV coefficient, as used by Klingenberg (’09) and Santana and Lofgren (’13) are based on assessing the magnitude of covariation, and are probably more appropriate to detect or assess the statistical significance of presumed developmental clusters. Covariation-based methods have been used with

morphometric data (e.g. Goswami, ’06; Goswami and Polly, ’10; Santana and Lofgren, ’13), but not, to my knowledge, in the context of event heterochrony.

Most previous studies of developmental modules using developmental sequences (the type of data typically used in sequence heterochrony analyses) have tested the validity of modules proposed on the basis of other data. Following Poe (’04), they have assumed that elements of a developmental module would show correlation in sequence position (e.g., Goswami, ’07; Goswami et al., ’09; Koyabu et al., 2011). For instance, Goswami (’07) tested the hypothesis that the six phenotypic modules previously found by Ackermann and Cheverud (’04) and Goswami (’06) in the therian skull were also developmental modules. To assess this, she compared rank correlation of various characters within and between modules for various pairs of taxa; the null hypothesis predicts that rank correlation between events is independent of the modules (i.e. there are no modules, or the modules have been erroneously delimited). This can be rejected if correlations within modules are greater than in 95% of the sets of randomly chosen events (irrespective of which module they belong to). Goswami (’07) could not reject the null hypothesis, which implies that the phenotypic modules reflecting adult morphology are not developmental modules.

Given our preliminary knowledge of developmental modules, a method that would allow us to detect and delimit the modules (rather than assess pre-defined modules) should be useful. This was pointed out recently by Koyabu et al. (’11: 16) when they stated that “Development of analytical tools to detect modularity directly from sequence heterochrony data shall provide new avenues to understand the role of modularity in the evolution of vertebrate cranial diversity.” Actually, statistical methods that could do this were developed long ago, but have not been used in that context yet, to my knowledge. Namely, a Principal Component Analysis (PCA thereafter) can be used to visualize covariation between all characters of a dataset. Below, I give an example of how this method can be used in this context, apply it to a dataset

of cranial ossification sequence in lissamphibians previously used by Germain and Laurin (’09) using a correlation matrix, and discuss the advantages and drawbacks of this method.

The composition of developmental modules yielded by PCA should be viewed as preliminary hypotheses to be tested because PCA is a parametric method that assumes normality of the data, and it typically ignores phylogenetic relationships. Developmental sequences, even when standardized to be analyzed quantitatively (Germain and Laurin, ’09) are probably not normally distributed. This is not a problem if PCA is used only as an exploratory step to delimit potential modules, but it would be a problem if PCA or another parametric method were used to assess the statistical significance of the correlation between events. More generally, parametric methods can be used to describe covariation even if the data do not follow a normal distribution, but statistical significance cannot be assessed because the probabilities yielded by parametric tests are wrong, in these circumstances. This problem can be solved by using another method to verify that characters within modules are fairly tightly correlated; phylogenetic independent contrasts (PIC; Felsenstein, ’85) are ideally suited for this, if the contrasts are adequately standardized. Some data cannot be analyzed by PIC because this method assumes that the data have evolved according to a Brownian motion model on the reference tree, and for these simpler, more robust, non-parametric methods can be used. Among these, pairwise comparisons (Read and Nee, ’95; Maddison, ’00) are especially interesting because they incorporate the phylogeny into the analysis.

In this paper, I use several classical statistical methods to assess the presence of developmental modules in the urodele skulls. More specifically, I try to determine if integration is sufficiently heterogeneous to infer that modularity is plausibly present, and I try to determine the number and composition of developmental modules. I also assess the relative performance of two methods to delimit modules (PCA and assessment of the statistical significance of observed correlations between all pairs of characters). In addition, I test the

hypothesis that neoteny results in changes in cranial ossification sequences, and I assess the relative evolutionary rates of bone ossification sequence positions. The methods used in this paper fall into several families, the main ones being ordination to carry out exploratory analyses, and regressions to assess the statistical significance of observed covariation. These methods are for the most part frequently used, which is advantageous because this means that their statistical properties are well-understood, but they have not, to my knowledge, been previously used in this context.

Methods

The reference phylogeny (Fig. 1) used by Germain and Laurin (’09) to analyze urodele ossification sequences using continuous analysis had been taken from Wiens et al. (05). This tree was incompletely resolved tree because Wiens et al. (05) did not include all terminal taxa used by Germain and Laurin (’09), and at the time, we were not able to resolve the phylogeny within *Ambystoma* using other references. This part of the tree was resolved using Pyron and Wiens (’11).

The data, consisting in the position of 21 bones in the ossification sequence of 21 urodele terminal taxa (Tables 1, 2 of SOM 1) were taken from Germain and Laurin (’09), who had themselves collected it from older literature (see Germain and Laurin, ’09 for references). The three characters with the most missing data (out of the 24 analyzed by Germain and Laurin, ’09) were removed because some analytical methods used here do not cope well with missing data. The remaining 21 characters were first analyzed through a PCA to examine the patterns of co-variation between characters, in the hope that the position of characters on PC axes 1, 2, and eventually 3 (depending on the amount of variance explained by this axis) might reveal developmental modules. Each module would appear as tightly grouped, nearly parallel vectors

(because characters of a module would co-vary strongly with each other). PCAs and all classical statistical tests were performed using Statistica 6, distributed by StatSoft France.

To verify if the developmental modules suggested by PCA were plausible, the correlation between individual characters was assessed through a Phylogenetic Independent Contrast analysis (PIC hereafter). This method, initially proposed by Felsenstein (1985), has now become the most widely-used comparative method, and it is implemented in the user-friendly PDAP:PDTREE module for Mesquite (Midford et al., '10), which I used for these analyses. The mathematically equivalent PGLS (Phylogenetic Generalized Least Squares) method (Martin and Hansen '97) is also available either in the commercial software NTSYS (Rohlf '11) or from command-driven software, such as BayesTraits (Pagel and Meade '06). For this analysis, the data were incorporated into a Mesquite file (SOM 2). Pearson product-moment correlations were also obtained, for comparison purposes.

To determine if the PIC analysis could be performed, I verified that the contrasts were adequately standardized using the four tests available in the PDAP:PDTREE module. To determine if PIC regressions were preferable over classical, non-phylogenetic linear regressions, I assessed the presence of a phylogenetic signal by comparing the squared length of each character on the timetree to a population of 10000 trees on which character value had been permuted, as I suggested earlier (Laurin, '04). This test was performed in Mesquite (Maddison and Maddison, '11). Several analytical methods, such as continuous analysis (Germain and Laurin, '09) or methods based on discrete data, like event pairing (Jeffery et al., '02) and methods derived from it (Jeffery et al., '05), as well as PGi (Parsimov-based Genetic inference; Harrison and Larsson, '08), implicitly or explicitly rest on the presence of a phylogenetic signal in the data. To determine the suitability of these data to these methods, I also assessed the phylogenetic signal on the original sequences, which were analyzed through

ordered parsimony, the appropriate approach when data represent discretization of underlying continuous data (Grand et al., ’13).

The obtained modules are also compared with a phenogram obtained from Statistica 6. This is useful to determine if modules assessed through character covariation differ substantially from those obtained from character similarity.

To assess the probability that the perceived modules simply reflect a random distribution of significant correlations between elements, a binomial test was performed to assess if the significant correlations were randomly distributed between elements. Under the null hypothesis of no modularity, the number of significant correlations between elements should correspond to a binomial distribution with the following parameters: n (number of tests) = 20 (because each of the 21 bones is examined for correlations with 20 other bones), and p (the probability of each test being significant) = number of significant correlations/number of comparisons made. This is akin to assuming that correlations are drawn randomly from a sample, with replacement. In this case, there are 210 tested correlations (21*20/2) because the matrix is symmetrical and the diagonal is by definition 1. This concludes the essential steps of the module detection analysis, whose workflow is summarized in Figure 2.

Given that timetrees are still not available for all taxa, I checked if the classical (non-phylogenetic) character variance could be used to infer approximate evolutionary rates. For this, I checked for the correlation between evolutionary rates computed through PIC in Mesquite’s PDAP: PDTREE module, character variance, and I determined how much variance of the evolutionary rates could be explained by these other (non-phylogenetic) data.

A Canonical Variates Analysis was performed in Mesquite to determine if neotenic, facultatively neotenic, and metamorphic taxa could be segregated on the basis of their ossification sequence. In Mesquite, the data for this analysis cannot have missing values, so

all characters with missing data were deleted, but all taxa were retained. The resulting matrix has only 11 characters (out of the 21 initially present). To check for statistical validity of this analysis, it was repeated with fewer (six) bones, to increase the ratio taxon number/character number.

An Evolutionary PCA (EPCA) was performed in the Rhetenor package for Mesquite (Dyreson and Maddison, ’06), using the same 11 characters without missing data, to show the distribution of the taxa into developmental space (i.e. developmental variation visualized using PCA or EPCA). This is a PCA that takes the phylogeny into consideration by selecting axes that maximize evolutionary variance, rather than static (non-phylogenetic) variance. This should give more even weight to sister-clades. This could be important here because the sample includes far more species in Diadectosalamandroidea than in Cryptobranchoidea. This analysis is mostly used to visualize homoplasy.

A complementary analysis was done with the complete dataset, with values completed with PhyloPars (Bruggeman et al., ’09), to compare the character loadings on EPCA axes 1 and 2 with character loadings on PCA axes 1 and 2, and with the modules redelimited using PIC. PhyloPars was designed to estimate missing values using both character correlations (evaluated through PIC) and the phylogenetic signal (the closest relatives determine part of the estimated values). This combined approach should give the best educated guess we can make about missing values, and a cross-validation analysis indicates how much better the results (mean bias and mean error) are than using only the nearest neighbor or the mean to estimate missing values. The versions of the matrix (both unscaled and scaled) with estimated missing values from PhyloPars is made available in SOM 2 and in the scaled version is also in Table 3 of SOM 1 (PhyloPars estimates are in the red cells). However, these estimates are not new data, so they were used only in the analyses mentioned above that could not accommodate missing data.

Results

Character evolution and phylogenetic signal

The standardization of contrasts on the tree is adequate for all but two characters (1, coronoid, and 4, dentary, failing one of the four tests each), at least if correction for multiple testing is done for this (Table 4 of SOM 1), following a suggestion from Canoville and Laurin (’10). Thus, PIC can be used to assess character correlation for this dataset (with caution to results pertaining to the coronoid and dentary).

The phylogenetic signal of the standardized data (for continuous analysis and most analyses presented in this paper) analyzed through squared-change parsimony looks rather weak when each character is examined individually, with a minority of characters returning probabilities below the classical 0.05 threshold without correction for multiple testing; after such corrections, none would be significant (Table 2). However, analysis of the whole matrix of standardized data yields significant results ($p = 0.0008$). Thus, PIC are adequate to assess correlations between characters, and this has the additional advantage of providing a more independent test of the modules delimited by PCA (to avoid circularity).

PCA and developmental modules

The first two Principal Component (PC hereafter) axes account for more than 50% of the total character variance (Table 2). They also provide a clearer delimitation of prospective modules than PC axis three. Therefore, only PC axes one and two will be considered below. Character loadings onto PC axes 1 and 2 suggest that up to four developmental modules may occur in the urodele skull (Fig. 3A), which will be named according to the highlighting color in the figures and SOM 1, to facilitate discussion.

The validity of these potential modules is assessed by looking at the strength and statistical significance of character covariation assessed through PIC. Several PIC regressions between pairs of characters yielded probabilities below the usual 0.05 threshold (Table 5 of SOM 1), but given that 210 tests were made, corrections for multiple testing are required. After FDR (False Discovery Rate) corrections (Benjamini and Hochberg, '95; Curran-Everett, 2000), only 45 probabilities remain significant (at $p < 0.01$), of which 18 are within the presumed four developmental modules. A chi-square test (Table 6 of SOM 1) indicates that this proportion is significant ($p < 0.0163$). However, comparison of the intra- vs. inter-module coefficients of determination (R^2) fails to show that intra-module explained variance exceeds inter-module variances (Tables 7 and 8 in SOM 1), perhaps because the modules require revision (see below). This was tested through a U Mann-Whitney test because the R^2 values deviate strongly from a normal distribution and have unequal variances (tested through Statistica 6, from StatSoft France), which would have prevented use of a Student's t test.

One module ("blue module" hereafter), the weakest according to the correlation test (Tables 5 and 7 of SOM 1), is composed of quadrate, coronoid and prearticular. It thus corresponds with the posterior part of the mandible and the quadrate, with which the mandible articulates, but includes dermal and visceral (endochondral) elements. This module is the only one among the four potential modules suggested by PCA not to be supported by at least one significant correlation after correction for multiple testing, so it is probably artifactual. Its characters load positively on the first PC axis.

A second ("pink") module, which loads positively on the second PC axis, and slightly negatively on the first axis, includes bones of the dermal skull roof (nasal, prefrontal), the maxilla, parts of the palate (vomer, palatine, and pterygoid), in addition to the orbitosphenoid, which connects skull roof and palate. This module is supported by two significant

correlations, between vomer and palatine, and between maxilla and prefrontal. Both pairs of bones are composed of bones in topological connexion.

A third ("green") module, which loads negatively on both first and second PC axes, includes much of the skull roof (frontal, parietal, squamosal), the premaxilla, the parasphenoid (all of which are dermal), as well several endochondral elements (opisthotic and stapes). This module is the most supported by the PIC analysis, with fifteen significant correlations after corrections for multiple testing; four of these concern the dentary and are unreliable because standardization of the contrasts of this character is suboptimal, but eleven significant results after FDR correction are still by far the strongest result from this preliminary analysis. Most of it is topologically connected and is composed of part of the braincase and of the dermal bones most intimately associated with it. However, the premaxilla is disconnected from these other bones. Yet, its presence in this module is supported by four significant correlations, with the squamosal, frontal, parietal, and parasphenoid (Tables 5 and 7 of SOM 1).

The fourth and last ("yellow") module, which loads negatively on the first PC axis, includes part of the neurocranium (exoccipital and prootic), but also, surprisingly, the septomaxilla, which is located far from the rest of this module. The existence of this module is further supported by a significant correlation found between exoccipital and prootic, but not with the septomaxilla, which instead shows significant correlation (even after FDR correction) with the vomer ($p = 0.002$). The septomaxilla might thus not belong to this module (see below).

The Evolutionary PCA (EPCA) performed on the completed dataset (with missing values estimated by PhyloPars) yields fairly different results (Fig. 3B). The four modules suggested by the classical PCA (with missing values ignored) cannot be recognized. Differences from the classical PCA result partly from the missing values estimated by PhyloPars and the resulting re-scaling of the other values; PCA performed on this completed dataset (Table 3 of

SOM 1) does not allow recognizing the four modules (Fig. 1 of SOM 1). However, this is only part of the picture because PCAs performed by Jorn Bruggeman on the phylogenetic correlation matrix from these data yielded different results, on which the four modules could not be recognized either (results not shown).

Redelimination of developmental modules based on PIC

Given the discrepancies obtained by PCA (with some missing values) and EPCA (with missing values estimated by PhyloPars), the composition of modules suggested by examination of scores on PC axes 1 and 2 can at best be viewed as a first hypothesis that needs to be tested and refined or refuted by examination of correlations between pairs of characters. Below, the results from classical PCA will be emphasized because they suggest possible modularity, whereas no such modularity is suggested by EPCA. An examination of correlations between characters through PIC shows that the results described above by PCA indeed appear to require major revision.

First, both small (blue and yellow) modules seem to be artifactual. To start with the yellow module, the prootic and exoccipital (both from the yellow module) are significantly correlated both with each other and with several bones of the green module, especially with the squamosal and parietal (Table 5 of SOM 1). Finally, the septomaxilla (yellow module) does not appear to be correlated with the exoccipital and prootic ($p > 0.4$ in both cases), even though scores on PC axes 1 and 2 suggest that these three bones pertain to the yellow module. On the contrary, the septomaxilla is significantly correlated with the vomer, which belongs to the pink developmental module. Thus, the yellow module does not appear to exist; a more optimal solution is to place the septomaxilla in the pink module and the exoccipital and prootic in the green module (Fig. 4).

The blue module may similarly be artifactual. The most convincing result concerning the existence of this module is a correlation between the prearticular and quadrate ($p = 0.033$; not significant according to FDR), but the prearticular appears to be more correlated with six elements of the green module (p comprised between $7.6*10^{-4}$ and 0.01). Thus, the case for reassigning the prearticular to the green module is strong. The quadrate can also convincingly be reassigned to that module. It shows five significant correlations (p ranging from $3.7 * 10^{-4}$ to 0.01) with elements of the green module, and one weaker correlation ($p = 0.01$) with the maxilla of the pink module (Table 5 of SOM 1). Thus, the quadrate is tentatively reassigned to the green module. Finally, the coronoid, which showed no correlation with other elements of the blue module, is weakly correlated with the palatine ($p = 0.020$; not significant after FDR) of the pink module according to PIC, although this result is not reliable because contrasts for the coronoid could not be adequately standardized. Two additional methods, pairwise comparisons (Read and Nee, '95; Maddison, '00) and a sign test using PIC failed to yield any significant results for the coronoid (Table 9 of SOM 1), although two are marginally non-significant, before correction for multiple tests. Thus, the three elements of the blue module can be reassigned to the pink and green module, which results in only two modules (Table 10 of SOM 1 and Table 3).

Of the two large modules, all elements initially assigned to the green module seem to belong there, but five elements initially assigned to the pink module, the orbitosphneoid, nasal, maxilla, prefrontal, and possibly the pterygoid should probably be reassigned to the green module. There is no significant correlation between these five bones and any other bone of the pink module (after FDR). However, the orbitosphenoid is significantly correlated with four bones of the green module, and the nasal, maxilla and prefrontal, with one of them (after FDR, in both cases). The case of the pteryoid is not as clear because it shows at least three weak correlations ($0.01 < p < 0.05$; not significant after FDR) with elements of the green

module. Thus, I tentatively reassign these five bones to that module (Table 3, and Table 10 of SOM 1). This leaves only four elements in the pink module, whose existence is supported by only two significant correlations after FDR corrections are applied (Fig. 5). These correlations are between vomer, septomaxilla and palatine. The coronoid might be part of this module, but this is based on weak evidence of a possible link with the palatine ($p = 0.02$, not significant after FDR) and remains tentative. Thus, this module makes up the anterior part of the palate (vomer and palatine) and includes a small bone located near its dorsal surface (septomaxilla).

A Fisher's exact test and a U Mann-Whitney test (Tables 11 and 12 in SOM 1) of the null hypothesis that inter-and intra-module coefficients of determination are the same is rejected with great confidence ($p < 0.0001$). To conclude, classical (non-phylogenetic) PC scores do not appear to provide a delimitation of modules developmental modules; these should indeed be based on examination of detailed pairwise phylogenetic regressions.

Examination of the (non-phylogenetic) Pearson product-moment correlation coefficients (Table 13 of SOM 1) generally yields congruent results, with most significant results after FDR correction (bold red type in the appendix table) occurring within the reassessed modules. However, the Pearson product-moment correlation is probably less reliable with these data because these include a phylogenetic signal, and because there are strong deviations from a normal distribution (but conversely, contrasts for most characters are adequately standardized). Thus, a more detailed comparison with PIC results is not necessary.

A binomial test of the hypothesis that the significant correlations are randomly distributed between bones (Table 14 of SOM 1) is rejected ($p = 0.03798$). This is supported by a visual examination of the distribution of the number of significant correlations per bones; both extremes of the range have much greater observed frequencies than predicted by the binomial law (Fig. 6).

Adding PhyloPars estimates on PIC does not seem to greatly alter the pattern of statistically significant correlations (Table 15 of SOM 1). These estimates should be reasonably reliable because cross-validation analyses indicate that PhyloPars estimates have on average 17 times less error than the simpler average and nearest neighbor modles (Table 16 of SOM 1). Similarly, artifact checks suggest that standardization of the contrasts is similar, with two characters displaying significant artifacts (Table 17 of SOM 1). Probabilities obtained with and without missing values estimated from PhyloPars are strongly correlated ($p<10^{-5}$) according to a Spearman rank correlation and Kendall's Tau (Table 18 of SOM 1).

Similarity and developmental modules

The phenogram (Fig. 7) yields a nested hierarchy of characters that differs substantially from the two modules suggested here (Table 3). The pink module is broken into three clusters, and the green one, into a large "paraphyletic" group. The hierarchical structure of the tree does not suggest the presence of more than two modules, a large one encompassing pterygoid, prootic, exoccipital, parasphenoid, parietal, frontal, squamosal, premaxilla, dentary, palatine, and vomer, and a small module that would include nasal and maxilla. The latter (small) phenetic module is supported by a Pearson product-moment correlation ($p < 0.001$) but not by PIC ($p = 0.86$). According to PIC (Table 5 of SOM 1), the maxilla is more strongly correlated with the prefrontal ($p = 2.5 * 10^{-7}$), but the nasal appears to co-vary strongly only with the frontal (Tables 5 and 10 of SOM 1). The less reliable Person product-moment correlation (Table 13 of SOM 1) does support a grouping of nasal with the maxilla, but also with the vomer ($0 < 0.001$ in both cases), a bone from the pink module. The large phenetic module is also problematic because pterygoid appears to co-vary with other bones neither according to PIC nor to the Pearson correlation. Similarly, the vomer does not co-vary with the pterygoid, prootic, or parasphenoid according to PIC. Conversely, several robust results according to

PIC, such as between the maxilla and prefrontal (see above), between the opisthotic and stapes ($p = 4*10^{-5}$), or stapes and prootic ($p = 9*10^{-5}$) are not reflected in the phenogram.

Evolutionary rates and character variance

Evolutionary rates (Table 4) differ a bit from those estimated from Germain and Laurin (’09), which is explained by the fact that two taxa with much missing data were removed, and that a polytomy was resolved using a recent phylogeny. Non-phylogenetic character variance explains 75% of the variance in evolutionary rates, which means that a rough estimate of relative (but not absolute) evolutionary rates can be obtained easily from character variance. However, PCs 1 and 2, which account for 50% of the total (non-phylogenetic) variance are not significantly correlated with evolutionary rates (PC axes 1 and 2 explain less than 12% of variance in evolutionary rates, and this is marginally non-significant).

Position of taxa in developmental space

Projection of the 21 urodele taxa on conventional PC axes one and two shows that most taxa form a fairly elongated cluster along PC1, with the notable exception of the obligatorily neotenic *Necturus maculosus* (Fig. 8). Of the three other obligatorily neotenic taxa, *Andrias japonicus* is isolated from other taxa by a combination of negative scoring on PC1 and positive scoring on PC2. The two other obligatorily neotenic taxa, *Siren intermedia* and *Amphiuma means*, are near the edges of the main cluster. All of the facultatively neotenic taxa are part of this main cluster, although *Lissotriton vulgaris* is at an extremity of this cluster.

The tree plot into EPCA (Evolutionary PCA) space (along axes one and two) based on the 11 characters without missing data similarly shows that most taxa are spread along axis one, although the taxa are more scattered along axis two than in the conventional PCA, and the outliers are not restricted to the obligatorily neotenic taxa (Fig. 9). In addition, the plot shows that there is much developmental homoplasy, with many branches crossing each other, and

many small clades (such as Hynobiidae and Salamandridae) widely scattered into developmental space. As expected from mathematical reasons, the root of the tree (Urodela) is located approximately in the center of the data points.

As expected, the Canonical Variates Analysis results in better separation between the three groups defined on the basis of life history (Fig. 10A). The Canonical Variates Analysis (in Mesquite) still shows a broad overlap between facultatively neotenic from metamorphic taxa, which are partly separated along axis two. This broad overlap between facultatively neotenic and metamorphic urodeles is not surprising because for at least some of the former, such as *Notophthalmus viridescens*, there is no difference in cranial ossification sequence between neotenic and metamorphic populations (Reilly, ’86); in several other facultatively neotenic taxa and some obligatorily neotenic ones in which neoteny developed in the recent geological past, development simply stops at a stage comparable to a late larval to metamorphic stage of closely related populations or taxa (Rose, ’03). Obligatorily neotenic urodeles are well-segregated along Canonical Axis one, but a greater taxonomic sample would be required to reach firm conclusions on this point. However, given that this is the most interesting result of this analysis, it will be exposed in greater detail. According to this preliminary analysis (Table 5), Canonical Variate Axis one has, approximately in decreasing absolute value of character loadings, negative loadings of the pterygoid and exocipital, positive loadings of squamosal, dentary, parietal, and prootic (to mention only the bones that have a loading of at least 0.25 on this axis). Given that neotenic taxa have a more negative score than other taxa on this axis, this means that the pterygoid and exoccipital ossify later, whereas the squamosal, dentary, parietal, and prootic ossify earlier than in other taxa, and examination of the original data confirms this. However, these results cannot be currently validated by a PIC analysis, perhaps because of the small taxonomic sample size (Table 19 of SOM 1).

Canonical Variates Analysis based on the six characters with highest scoring on canonical axes one and two still gave fairly good separation between the ontogenetic groups, with obligatory neotenes barely overlapping with the two other categories, and the facultative neotenes partially separated from metamorphic taxa (Fig. 10B). However, reducing the character sampling further to five yielded much worse results (not shown).

Discussion

Phylogentic signal, evolutionary rate, and developmental modularity

The presence of a phylogenetic signal in the standardized urodele cranial ossification sequence data, when the whole matrix is analyzed with squared-change parsimony, validates the use of character optimization to infer character history for these urodele ossification sequences. This lends some additional support to the conclusions drawn by Germain and Laurin (’09) on the basis of continuous analysis and possibly on the basis of Parsimov.

The phylogenetic signal should routinely be assessed in comparative studies on modularity and evo-devo in general, but that is not yet established practice in the field (e.g., Jeffery et al., ’05; Smirthwaite et al. ’07; Hugi et al., ’12; Hautier et al., ‘13). Several methods in evo-devo implicitly rely on the presence of a phylogenetic signal, such as event paring (with, or without PARSIMOV), PGi, or the approach, first proposed by Poe (’04) and recently used in other studies (e.g. Goswami, ’07) that tests modularity by comparing the rank correlation of sets of events within and between modules in pairs of taxa. In Poe’s (’04) method, several comparisons involve internal nodes for which the sequence had to be inferred from its descendants. If no phylogenetic signal existed in these data, inferred nodal sequences would be unreliable and alternate methods could be used, either conventional (non-phylogenetic) correlations between terminal taxa, or comparisons of extant sister-taxa only, as in Maddison’s (’00) pairwise comparisons.

Evolutionary rates are best assessed, for quantitative characters, by PIC over timetrees. A rough estimate of relative evolutionary rate can be obtained from character variance, if timetrees are unavailable, but this is clearly suboptimal because this approach is very imprecise. For instance, there is a roughly 90-fold difference in evolutionary rates between the bone most constrained in its position in the developmental sequence across all taxa here investigated (dentary, $2.6*10^{-5}$ standardized sequence position per lineage per Ma) and the least constrained (stapes, $2.3*10^{-3}$). The non-phylogenetic variances between these characters varies about 294-fold (Table 4). PC scores on the first few axes should not be used because they represent only part of the variance, and the latter is unevenly distributed among PC axes. In any case, simple non-phylogenetic variance is easier to compute than PC scores, so the latter have no advantage in this context.

Evolutionary rates need to be routinely assessed in heterochrony and modularity studies, but this is not common practice. Thus, Goswami et al. (2009: 192) invoked similarity in the sequences (which implies low evolutionary rates) to explain that only 27 of the 510 tests yielded significant results. Similarly, Wilson (’13) noted that “...emerging from these studies is a general picture of conservatism in cranial ossification sequences among placental mammals...”. These suggestions might be correct, but they need to be validated by a quantification of evolutionary rates to determine if these rates are lower for these datasets than for other kinds of characters, because otherwise, there is no way to tell that there is more conservatism in mammalian cranial ossification sequences than for any other character or taxon set. Currently, comparisons of evolutionary rates between studies is hampered by the fact that rates have been assessed using a variety of metrics, many of which cannot be compared between datasets because they are influenced by the number of taxa and the evolutionary distance between taxa. For instance, Smith (’97) reported that 163 event pairs out of 378 (43%) were invariable across the sampled therian mammals. This measure cannot

be meaningfully compared with the rates reported here (in amount of standardized sequence position change per lineage per million years) because the units of change differ between studies and because the number of event pairs that can be expected to be invariant in a dataset, for a given evolutionary rate, should decrease with taxonomic sample size and with evolutionary distance between taxa. Worse, even if we knew the number of most parsimonious changes implied by the event pair data of Smith (’97), that would still not be comparable to another dataset because the number of changes expected increases with the number of sampled taxa, the mean evolutionary distance between these taxa (in Ma or in another unit), and the number of characters. Clearly, a standardization effort using an evolutionary rate metric that can be comparable across studies is desirable in this field. Ideally, this metric should yield a number (for discrete characters) or amount of change (for continuous characters) per character, per lineage, per Ma. The metric used here is thus one of the suitable metrics (for continuous data). The equivalent for discrete characters has occasionally been used, in other contexts (e.g. Meslin et al., 2012: table 2).

Quantitative exploratory analyses of modularity

The above analyses show that standard developmental sequence data can be productively analyzed through methods typically employed for continuous data, provided that they are standardized by methods such as that used for continuous analysis. However, data on absolute time or even on a proxy of time (developmental stages or size) at which events occur should be even more useful for such approaches, and such data might be more normally distributed than event sequence data. I hope that in the future, developmental biologists will report more frequently such data, at least when they produce them. In several recent studies, such data were usually discarded and converted into event sequences (e.g., Harrington et al., ’13: 346), presumably because these studies focused on methods that could not use the more quantitative data, although such data were reported in more descriptive literature (e.g. Rose, ’03: 1719).

When discussing modules sought using morphometric data, Klingenberg (’09) argued that “Because developmental interactions are tissue bound, it is sensible to require that modules should be spatially contiguous.” This is a reasonable assumption if there is good reason to believe that processes such as the diffusion of signalling factors affect simultaneously several elements or structures, which would form a developmental module (Klingenberg, ’13). Bone growth typically occurs along its edges, so except in the earliest phases of bone ossification, growth typically occurs along sutures, which is also expected to generate integration between adjacent elements, at least once the sutures have formed (this may not occur in the timing of bone ossification because bone primordial are typically not connected to each other, initially). However, other developmental processes could generate spatially discontinuous modules. A possible example is the fore and hind paired appendages of gnathostomes, which may form a module in the context of event heterochrony, at least in placental mammals (Goswami et al., ’09), perhaps because the same set of genes, such as hoxa and hoxd 9–13 genes (Ahn and Ho, ’08) are involved in their patterning. The urodele cranial ossification sequence data analyzed above seem to support the hypothesis that developmental modules are typically composed of spatially contiguous elements. However, several statistically significant correlations were found between fairly distant elements, such as the prearticular and the frontal, or the quadrate and the premaxilla (Fig. 5). These elements appear to be part of a large module, but evidence for integration within this module is obviously not limited to correlations between anatomically-connected elements.

Operationally, the new method involving PCA and, more importantly, an assessment of the correlations through phylogeny-informed analyses such as PIC differs from the established technique, Poe’s (’04) method of testing modularity using Kendall’s τ, in the way taxa and characters are handled. Poe’s (’04) method requires computing correlations between multiple pairs of taxa and comparing the correlation between elements within a module with

correlations between random sets of elements to assess statistical significance. For large datasets (including several terminal taxa), this implies a large number of comparisons. This is a drawback given that the composition of developmental modules is not expected to vary greatly, at least between closely related taxa; Goswami et al. (’09) found evidence of differences between marsupials and placentals, but these clades diverged in the Mesozoic, about 150–190 Ma ago (Meredith et al., ’11). The method proposed here can use all the data (all characters in all taxa) in a single operation, using classical parametric statistical methods (like PCA and PIC) implemented in standard software, which may, or may not, incorporate phylogenetic data, because Mesquite (Maddison and Maddison, ’11) can perform a phylogenetic PCA (EPCA). This is made possible by using the standardization procedure employed in continuous analysis (Germain and Laurin, ’09; Laurin and Germain, ’11), and the methods used here can be seen as an extension of the continuous analysis for developmental modular analyses. The proposed method can also be used to test modularity in subsets of taxa, simply by pruning the data matrix to retain only the relevant taxa, but it does not require testing all possible taxon partitions. However, as seen above, the new method also requires assessing correlations between all possible pairs of characters; this can easily be done in a non-phylogenetic analysis with conventional statistical software like Statistica, but doing this through PIC is more time-consuming because in Mesquite, this is done for each pair of characters separately (although exporting the standardized contrasts to statistical software would allow this to be done in a single operation). Other software, like PhyloPars, can produce phylogenetic correlation matrices, and this could be used to perform EPCA or detect modules directly. Alternatively, a script for R for multi-variate PGLS (Rohlf, ’01) was written by Outomuro et al. (’13).

The most fundamental difference between both methods is that Poe’s (’04) method was developed to assess the statistical support for developmental modules whose number and

composition is hypothesized a priori or based on other types of data. On the contrary, the new method proposed here is aimed at assessing the potential presence of developmental modules independently of any a priori hypothesis. Thus, both methods are complementary, and hopefully, both will contribute to progress in this field. Exploratory analyses are extremely important in biology, because the number of conceivable hypotheses is great. In any case, phylogenetics (Rogers and Swofford, ’98) and more generally comparative biology continue to extensively use exploratory analyses, as recent scientific meetings demonstrate (e.g. Laurin, ‘13). Even in studies of integration and modularity, exploratory analyses have been used, along with tests of pre-conceived hypotheses. Thus, cluster analyses were used by Goswami (2006: fig 3) to detect and delimit modules (though I don’t recommend this, based on the above results and theoretical considerations).

This deep difference in approach might be especially important if previous work on other types of data (genetic or morphometric, for instance) suggested the presence of modules (let’s designate them A and B) whose composition differed from that of developmental modules (A’ and B’). If the composition of the two kinds of modules overlapped sufficiently (suppose, for instance, 90% overlap between A and A’, and about the same overlap between B and B’), Poe’s (’04) method might return significant results, and we would conclude, erroneously, that the developmental data supports the existence of the genetic or morphometric modules. In the same situation, the method developed here should show that the two sets of modules differ slightly in composition. However, these expectations based on theoretical considerations and on the results obtained on the single empirical dataset analyzed here will have to be validated by additional studies on other datasets. Such extensive tests, however, are beyond the scope of this exploratory study.

Drawbacks of the methods tested here are mainly twofold. First, examination of character loadings on the first PC axes is probably not a reliable method. In this first empirical study,

the modules suggested by PCA could not be recognized using EPCA or PCA performed on phylogenetic correlations established by PhyloPars, and examination of significant correlations using PIC resulted in several changes in module composition and number. This is possibly because patterns of covariance suggested by classical statistics differ from those of phylogeny-informed analyses like PIC. The usefulness of PCA-based method to detect and delimit developmental modules will need to be assessed using additional empirical and simulated datasets, but these findings support the claim that such methods, like clustering methods, yield incongruent results (Goswami and Polly, ‘10). Fortunately, examining statistically significant correlations using PIC seems to yield far more coherent results (Table 17 of SOM 1). Some software such as PhyloPars (Bruggeman et al., ’09) can provide directly the matrix of phylogenetic correlations between all pairs of characters, and this could be used directly to assess the developmental modules. Here, I did not develop an automated method using such correlation matrices, but the procedure consisting in building a network of significantly correlated elements (Fig. 5) is operationally fairly simple and can be implemented using existing software. However, given the great number of statistical tests involved, corrections for multiple tests (done here through FDR) are essential. This method should not be confused with network model analysis (Esteve-Altava et al., ’11, ‘13), which has recently been used to analyze tetrapod skulls (Esteve-Altava and Rasskin-Gutman, ’14). Network model analysis could of course be applied to the data used here, although such a study would be beyond the scope of this paper. My results demonstrate some correlations (Fig. 5) between bones without direct connections (Fig. 4), such as between the premaxilla and several other bones (frontal, parietal, squamosal, quadrate, parasphenoid, and dentary), but this is compatible with modules composed of elements that are not all directly connected to each other.

Second, as pointed out by Klingenberg (’13: 47) in a slightly different context (clustering analyses), some methods can produce clusters even when none are present, and if these are validated using similar methods on the same data, the reasoning would largely be circular. Here, I attempted to minimize this potential circularity problem by initially defining the modules based on the first two (non-phylogenetic) PC axes, which represent only 50% of the variance, and validating them using a phylogeny-informed analysis that considers all of the variance (PIC). However, rather than a clear validation, this comparison demonstrated the need to reassess the number and composition of the developmental modules. More useful in this context is the test that the significant correlations are randomly distributed between elements (using a binomial distribution). The fact that this hypothesis can be rejected ($p = 0.03798$; Table 14 of SOM 1) suggests that some parts of the skull are more integrated than others, although it does not indicate the number of modules present or their composition. This is reflected by the much greater observed frequencies of bones at both ends of the distribution than predicted by the binomial distribution.

PCA on the urodele cranial ossification dataset suggested the presence of four modules, but examination of PIC results suggests the presence of a single, very large, well-supported (green) module, and of a much smaller (pink) module composed of three to four bones (Fig. 5). The small module appears to make up the anterior part of the palate (vomer, palatine) and surrounding area (septomaxilla). Two bones, the coronoid and pterygoid evolve more freely and they display no significant correlations with bones from the two modules, according to the FDR analysis (Fig. 5). Thus, the palate of the urodele skull appears largely uncoupled developmentally from the rest of the skull. The uncoupling of coronoid and pterygoid from both modules does not appear to be linked to their evolutionary rates because other bones change in relative timing faster or slower (Table 4). Most of the bones uncoupled from the main (green) module are presumably involved in biting (except for the septomaxilla),

although the dentary, prearticular and quadrate, which should share this function, are part of the green module.

The probability of the null hypothesis that the coefficients of determination (which are about twice as large intra- than inter-module) are randomly distributed with regard to module composition is so low (about 10^{-4}; Tables 11 and 12 of SOM 1) that the two presumed developmental modules (green and pink) are unlikely to represent a mere statistical artifact. Other methods and data types (molecular, morphometric, etc.) should be used in the future to reassess modularity in the urodele skull and verify these preliminary results. More ossification sequences, and more informative data (with actual developmental times, developmental stages, or body size at which events occur) would also help clarify these issues. The methods used here are more suitable to quantitative data, and PhyloPars had problems calculating phylogenetic correlations (Bruggeman, personal communication, January 21, 2014), probably because the data are far from normally distributed and not very abundant. Part of the problems reported here may simply reflect the fact that the work had to be carried out from sequence data. More detailed data is often available in the literature, but it was not incorporated into the recent compilations, such as the one used here (Rose, 2003).

Obviously, phenetic (distance-based) methods are unsuitable to discover evolutionary modules. The resulting clusters of similar characters can differ substantially from modules based on co-variation, as is the case here. Fortunately, phenetic methods have not been widely used in this context.

Modularity in development

The reasonably strong evidence found here for the presence of two developmental modules of very unequal size in the urodele skull is encouraging because most studies on developmental modules failed to find statistical support for them. Thus, Goswami (’07) failed to find

developmental support for six phenotypic modules in therian mammals, and Klingenberg and Marugán-Lobón (‘13) failed to support the presence of two modules (facial and braincase) in the avian skull.

In the above discussion, modularity has been considered only between mutually exclusive sets of elements, but this may not be the correct model to express the correlations that prevail in development. Modules might conceivably be nested, or some elements might link two modules (be correlated with spatially adjacent elements of two different modules), or some structures (like the urodele skull) might be composed of a set of tightly integrated elements (like the green module) and include other less integrated elements. Testing such hypotheses may require more sophisticated analytical methods and probably finer data (on more taxa, and incorporating absolute developmental time or a proxy, rather than sequences). The role and form of modularity in development should thus remain a fruitful research topic for many years, and the present study barely scratches the surface of what could be accomplished in the future, but I hope that this modest contribution will foster additional work in this field.

Heterochrony and missing bones in urodeles

Several PhyloPars estimates place missing values at the very end of the sequence, after ossification of the last bones documented in the sequence (Table 3 of SOM 1). This is logical to the extent that several of these bones (such as the prefrontal in sirenids, or the orbitosphenoid, nasal, and septomaxilla in *Necturus*) are absent in the relevant taxon, perhaps as a result of paedomorphosis truncating the developmental sequence before appearance of these bones. In this respect, the PhyloPars estimates placing missing bones in the middle (e.g. the maxilla in *Necturus*) or at the very beginning of the sequence (e.g. the coronoid in *Amphiuma*) are more puzzling.

The fact that obligatorily neotenic taxa can be better discriminated than facultatively neotenic taxa from metamorphic taxa by a canonical analysis may not relate only to the developmental constraints imposed by the retention of a flexible life history. The obligatorily neotenic taxa have apparently been neotenic for a long time, according to the fossil record of these taxa. Thus, cryptobranchids appear to have been neotenic for a long time. The affinities of the possible earliest stem-cryptobranchids are still disputed (Carroll and Zheng, ’12; Marjanović and Laurin, ‘14), but these forms, such as *Jeholotriton* from the Bathonian (Middle Jurassic) were already neotenic. More recent, undisputed cryptobranchids such as *Cryptobranchus saskatchewanensis* from the Thanetian (56–59 Ma) were probably already neotenic, given their great morphological similarity with extant *Cryptobranchus* (Naylor, ’81). The oldest known stem-sirenid, *Kababisha humarensis* from the Cenomanian (94–100 Ma) and the oldest known amphiumid, *Proamphiuma cretacea* from the late Maastrichtian or early Paleocene (Gardner, ’03), were already neotenic (Evans et al. 1996). The oldest known (Paleocene) proteid, *Necturus krausei* (Naylor, ’78), is known only from a few vertebrae, but both its suggested affinities within the proteid crown and faunal associations with other aquatic urodeles suggest that it was paedomorphic too (Gardner, ’03). By contrast, facultatively neotenic urodeles necessarily represent very recent heterochronic events because they diverged from their closest extant metamorphic relatives mostly in the Neogene (Zhang et al., ‘08; Marjanović and Laurin, ‘14), and their neoteny is not completely genetically fixed, given that they can metamorphose either under some environmental signals (like seasonal drying of the water bodies they inhabit) or hormonal treatment. Thus, they are expected to differ little from their metamorphic relatives in their ossification sequences.

Acknowledgments

I thank Jorn Bruggeman for help with the use of PhyloPars, and the organizers of the symposium, Laura Wilson and Ingmar Werneburg, for funding to attend the meeting, for editing this special issue, and for comments on the paper. Borja Esteve-Altava also provided many constructive comments, and a less enthusiastic anonymous referee forced me to tighten somewhat the argumentation. This work was financed by the CNRS and the French ministry of research through the recurring grant to the CR2P. I have no conflict of interest to declare.

Literature cited

Ahn D, Ho RK. 2008. Tri-phasic expression of posterior Hox genes during development of pectoral fins in zebrafish: Implications for the evolution of vertebrate paired appendages. Develop Biol 322:220–233.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B–Stat Methodol 57:289–300.

Bruggeman J, Heringa J, Brand BW. 2009. PhyloPars: estimation of missing parameter values using phylogeny. Nucleic Acids Res 37:W179–W184.

Canoville A, Laurin M. 2010. Evolution of humeral microanatomy and lifestyle in amniotes, and some comments on paleobiological inferences. Biol J Linn Soc 100:384–406.

Carroll RL, Zheng A. 2012. A neotenic salamander, *Jeholotriton paradoxus*, from the Daohugou Beds in Inner Mongolia. Zool J Linn Soc 164:659–668.

Curran-Everett D. 2000. Multiple comparisons: philosophies and illustrations. Am J Physiol Regulatory Integrative Comp Physiol 279:1–8.

Duellman WE, Trueb L. 1986. Biology of Amphibians. New York: McGraw-Hill.

Dyreson E, Maddison WP. 2006. Rhetenor package for morphometrics. Version 1.11. http://mesquiteproject.org/mesquite0.98/mesquite/rhetenor/rhetenor.html

Esteve-Altava B, Marugán-Lobón J, Botella H, Rasskin-Gutman D. 2011. Network Models in Anatomical Systems. J Anthrop Sci 89:175–184.

Esteve-Altava B, Marugán-Lobón J, Botella H, Bastir M, Rasskin-Gutman D. 2013. Grist for Riedl's mill: a network model perspective on the integration and modularity of the human skull. J Exp Zool B (Mol Dev Evol) 320:8.

Esteve-Altava B, Rasskin-Gutman D. 2014. Theoretical morphology of tetrapod skull

networks. C R Palevol 13:41–50.

Evans SE, Milner AR, Werner C. 1996. Sirenid salamanders and a gymnophionan amphibian from the Cretaceous of the Sudan. Palaeontology 39:77–95.

Felsenstein J. 1985. Phylogenies and the comparative method. Am Nat 125:1-15.

Gardner JD. 2003. The fossil salamander *Proamphiuma cretacea* Estes (Caudata: Amphiumidae) and relationships within the Amphiumidae. J Vert Paleont 23:769-782.

Germain D, Laurin M. 2009. Evolution of ossification sequences in salamanders and urodele origins assessed through event-pairing and new methods. Evol Dev 11:170–190.

Goswami A. 2006. Cranial modularity shifts during mammalian evolution. Am Nat 168:270–280.

Goswami A. 2007. Cranial modularity and sequence heterochrony in mammals. Evol Dev 9:290–298.

Goswami A, Polly PD. 2010. Methods for studying morphological integration and modularity. In: Alroy J, Hunt G editors. Quantitative methods in paleobiology: The Paleontological Society. p. 213–243.

Goswami A, Weisbecker V, Sánchez-Villagra M. 2009. Developmental modularity and the marsupial–placental dichotomy. J Exp Zool B (Mol Dev Evol) 312B:186–195.

Grand A, Corvez A, Duque Velez LM, Laurin M. 2013. Phylogenetic inference using discrete characters: performance of ordered and unordered parsimony and of three-item statements. Biol J Linn Soc 110:914–930.

Harrington SM, Harrison LB, Sheil CA. 2013. Ossification sequence heterochrony among amphibians. Evol Dev 15:344–364.

Harrison LB, Larsson HCE. 2008. Estimating evolution of temporal sequence changes: a

practical approach to inferring ancestral developmental sequences and sequence heterochrony. Syst Biol 57:378–387.

Hautier L, Bennett NC, Viljoen H, Howard L, Milinkovitch MC, Tzika AC, Goswami A, Asher RJ. 2013. Patterns of ossification in southern vs. northern placental mammals. Evolution 67:1994–2010.

Hugi J, Hutchinson MN, Koyabu D, Sánchez-Villagra MR. 2012. Heterochronic shifts in the ossification sequences of surface- and subsurface-dwelling skinks are correlated with the degree of limb reduction. Zoology 115:188–198.

Jeffery JE, Richardson MK, Coates MI, Bininda-Emonds ORP. 2002. Analyzing developmental sequences within a phylogenetic framework. Syst Biol 51:478-491.

Jeffery JE, Bininda-Emonds ORP, Coates MI, Richardson MK. 2005. A new technique for identifying sequence heterochrony. Syst Biol 54:230-240.

Josse S, Moreau T, Laurin M. 2006. Stratigraphic tools for Mesquite. http://mesquiteproject.org/packages/stratigraphicTools/

Kirschner M, Gerhart J. 1998. Evolvability. Proc Natl Acad Sci USA. 95(15): 8420–8427.

Klingenberg CP. 2009. Morphometric integration and modularity in configurations of landmarks: tools for evaluating a priori hypotheses. Evol Dev 11:405–421.

Klingenberg CP. 2013. Cranial integration and modularity: insights into evolution and development from morphometric data. Hystrix 24:43–58.

Klingenberg CP, Marugán-Lobón J. 2013. Evolutionary covariation in geometric morphometric data: analyzing integration, modularity, and allometry in a phylogenetic context. Syst Biol 62:591–610.

Koyabu D, Endo H, Mitgutsch C, Suwa G, Catania KC, Zollikofer CP, Oda S-i, Koyasu K, Ando M, Sánchez-Villagra MR. 2011. Heterochrony and developmental modularity of cranial osteogenesis in lipotyphlan mammals. EvoDevo 2:1–18.

Laurin M. 2004. The evolution of body size, Cope's rule and the origin of amniotes. Syst Biol 53:594-622.

Laurin M. 2013. Systematics beyond phylogenetics/La systématique au-delà de la phylogénétique. C R Palevol 12:327–331.

Laurin M, Germain D. 2011. Developmental characters in phylogenetic inference and their absolute timing information. Syst Biol 60:630–644.

Maddison WP. 2000. Testing character correlation using pairwise comparisons on a phylogeny. J Theor Biol 202:195–204.

Maddison WP, Maddison DR. 2011. Mesquite: a modular system for evolutionary analysis. Version 2.75. http://mesquiteproject.org

Marjanović D, Laurin M. 2013. An updated palaeontological timetree of lissamphibians, with comments on the anatomy of Jurassic crown-group salamanders (Urodela). Hist Biol. (Early View version). DOI: 10.1080/08912963.2013.797972

Martins EP, Hansen TF. 1997. Phylogenies and the comparative method: a general approach to incorporating phylogenetic information into the analysis of interspecific data. Am Nat 149:646–667.

Meredith RW, Janečka JE, Gatesy J, Ryder OA, Fisher CA, Teeling EC, Goodbla A, Eizirik E, Simão TLL, Stadler T, Rabosky DL, Honeycutt RL, Flynn JJ, Ingram CM, Steiner C, Williams TL, Robinson TJ, Burk-Herrick A, Westerman M, Ayoub NA, Springer MS, Murphy WJ. 2011. Impacts of the Cretaceous Terrestrial Revolution and KPg Extinction on Mammal Diversification. Science 334:521–524.

Meslin C, Mugnier S, Callebaut I, Laurin M, Pascal G, Poupon A, Goudet G, Monget P. 2012. Evolution of genes involved in gamete interaction: evidence for positive selection, duplications and losses in vertebrates. PLoS ONE 7 (9): e44548.

Midford P, Garland TJ, Maddison WP. 2010. PDAP Package for Mesquite. Version 1.16.

http://mesquiteproject.org/pdap_mesquite/index.html

Naylor BG. 1978. The earliest known *Necturus* (Amphibia, Urodela), from the Paleocene Ravenscrag formation of Saskatchewan. J Herpetol 12:565–569.

Naylor BG. 1981. Cryptobranchid salamanders from the Paleocene and Miocene of Saskatchewan. Copeia 1981:76–86.

Outomuro D, Adams DC, Johansson F. 2013. The evolution of wing shape in ornamented-winged damselflies (Calopterygidae, Odonata). Evol Biol 40:300–309.

Pagel M, Meade A. 2006. BayesTraits. http://www.evolution.rdg.ac.uk/BayesTraits.html

Poe S. 2004. A test for patterns of modularity in sequences of developmental events. Evolution 58:1852–1855.

Pyron RA, Wiens JJ. 2011. A large-scale phylogeny of Amphibia including over 2800 species, and a revised classification of extant frogs, salamanders, and caecilians. Mol Phyl Evol 61:543–583.

Read AF, Nee S. 1995. Inference from binary comparative data. J Theor Biol 173:99–108.

Reilly SM. 1986. Ontogeny of cranial ossification in the eastern newt, *Notophthalmus viridescens* (Caudata: Salamandridae), and its relationship to metamorphosis and neoteny. J Morph 188:315-326.

Rogers JS, Swofford DL. 1998. A fast method for approximating maximum likelihoods of phylogenetic trees from nucleotide sequences. Syst Biol 47:77-89.

Rohlf FJ. 2001. Comparative methods for the analysis of continuous variables: geometric interpretations. Evolution 55:2143–2160.

Rohlf FJ. 2011. NTSYSpc, Numerical Taxonomy System. Setauket, NY: Exeter Software. http://www.exetersoftware.com/cat/ntsyspc/ntsyspc.html

Rose CS. 2003. The developmental morphology of salamanders skulls. In: Heatwole H, Davies M editors. Amphibian Biology. Chipping Norton: Surrey Beatty & Sons. p.

1684–1781.

Santana SE, Lofgren SE. 2013. Does nasal echolocation influence the modularity of the mammal skull? J Evol Biol 26:2520–2526.

Schlosser G, Wagner GP. 2004. Modularity in Development and Evolution. Chicago: University of Chicago Press.

Smirthwaite JJ, Rundle SD, Bininda-Emonds ORP, Spicer JI. 2007. An integrative approach identifies developmental sequence heterochronies in freshwater basommatophoran snails. Evol Dev 9:122-130.

Smith KK. 1997. Comparative patterns of craniofacial development in eutherian and metatherian mammals. Evolution 51:1663–1678.

Wiens JJ, Bonett RM, Chippindale PT. 2005. Ontogeny discombobulates phylogeny: paedomorphosis and higher-level salamander relationships. Syst Biol 54:91-110.

Wilson LAB. 2013. Cranial Suture Closure Patterns in Sciuridae: Heterochrony and Modularity. J Mammal Evol. (Early View version).

Zhang P, Papenfuss TJ, Wake MH, Qu L, Wake DB. 2008. Phylogeny and biogeography of the family Salamandridae (Amphibia: Caudata) inferred from complete mitochondrial genomes. Mol Phyl Evol 49:586–597.

Figure legends

Figure 1. Timetree of urodeles used for the phylogeny-informed analyses. The topology is updated from Germain and Laurin (’09) by incorporating data from Pyron and Wiens (’11) to resolve the phylogeny within *Ambystoma*. The tree is modified from the output from Mesquite (Maddison and Maddison, ’11), with the geological timescale superimposed using the StratigraphicTools (Josse et al., ’06).

Figure 2. Work flow schema outlining the proposed module detection procedure.

Figure 3. Projection of the characters (position of bones in ossification sequences) on the first two Principal Component axes of: A, classical (non-phylogenetic) PCA performed in Statistica on all 21 characters; B, EPCA performed in Mesquite on the 11 characters without missing data. The four detected modules (most obvious in part A) are indicated by a color code. Abbreviations: art, articular; cor, coronoid; d, dentary; exo, exoccipital; fr, frontal; mx, maxilla; nas, nasal; op, opisthotic; os, orbitosphenoid; pal, palatine; ps, parasphenoid; par, parietal; pra, prearticular; prf, prefrontal; pmx, premaxilla; pro, prootic; pt, pterygoid; q, quadrate; smx, septomaxilla; sq, squamosal; st, stapes; vo, vomer.

Figure 4. Proposed developmental modules in urodeles shown on a skull of *Cryptobranchus allegheniensis* (not represented in the dataset, but closely related and similar to *Andrias japonicus*) on the basis of scores on PC axes 1 and 2. Redrawn from Carroll and Holmes (’80: fig. 5). Abbreviations as in Figure 2. The asterisk (*) indicates that the septomaxilla is absent in cryptobranchids (Duellman and Trueb, ’86: 498), but the area where it is located in other tetrapods has been colored to show to which module this element belongs in the urodeles that possess it. The pterygoid is shown in light green because its inclusion into this module is equivocal.

Figure 5. Statistically significant correlations between elements. These result from PIC analysis and take into consideration corrections for multiple tests using the FDR. Elements are in a circular arrangement that takes into consideration (as far as possible) their topological relationships. Intra-module correlations are represented in the color of the module; inter-module correlations are in black. Note that assignment of the pterygoid and coronoid to modules is not supported by any significant correlation after FDR, but weaker evidence suggests a possible attribution to these modules. Also note that the pink module might be composed of two sub-modules.

Figure 6. Distribution of the numbers of significant correlations per bone. The light blue (grey) boxes represent the numbers of bones (y axis) for which the number of significant correlations indicated in the x axis was found. The thick red (grey) line represents the number of bones for each number of correlations predicted by the binomial law. Note the great discrepancy reflecting the concentration of significant relationships between fewer bones than predicted.

Figure 7. Phenogram of the characters (ossification times of the 21 bones) produced by Statistica 6. The similarity (distance)-based groups differ strongly from the covariation-based modules; compare with Figures 2, 3, and Table 3. Several bones are so similar that the topology of the tree cannot be read, but does not matter because all of these (pterygoid, prootic, exoccipital, parasphenoid, parietal, frontal, squamosal, premaxilla, dentary, palatine, and vomer) are so close to each other that they would have to be considered a single module. Abbreviations: lpm, large phenetic module; spm, small phenetic module. For other (bone) abbreviations, see legend of Figure 3.

Figure 8. Taxa projected onto PC axes 1 and 2 (redrawn from a screen shot in Statistica). This is based on the complete dataset (21 characters). To make the figure more legible, only the

genus name is given when a single species per genus is included here, unless the data point is located in an uncluttered part of the graph. For species of *Ambystoma*, the genus name is abbreviated and the specific epithet is given, because several species of this genus are included. The circles of metamorphic taxa are in black, those of facultatively neotenic taxa, in yellow (light grey in the print version), and those of neotenic taxa, in red (dark grey in the print version). Complete species names, in alphabetical order, are as follows (see also Table 1 of SOM 1): *Ambystoma texanum, Ambystoma talpoideum, Ambystoma tigrinum, Ambystoma maculatum, Ambystoma mexicanum, Amphiuma means, Andrias japonicus, Dicamptodon tenebrosus, Eurycea bislineata, Gyrinophilus prophyriticus, Hemidactylium scutatum, Lissotriton vulgaris, Necturus maculosus, Notophthalmus viridescens, Onychodactylus japonicus, Pleurodeles waltl, Ranodon sibiricus, Rhyacotriton cascadeum, Salamandra salamandra, Salamandrella keyserlingii,* and *Siren intermedia*.

Figure 9. Tree (topology only) projected onto morphospace, modeled by an Evolutionary PCA. Redrawn from a screen shot in Mesquite. Based on the 11 characters without missing data. The circles of metamorphic taxa are in black, those of facultatively neotenic taxa are in yellow (light grey in the print version), and those of neotenic taxa are red (dark grey in the print version). Uncertainty in nodal state inferences are represented by circles with two colors.

Figure 10. Taxa projected onto Canonical Variates Axes 1 and 2 (redrawn from a screen shot in Mesquite). A, Based on the pruned dataset (11 characters without missing data). B, based on a subset of six bones (squamosal, parasphenoid, frontal, pterygoid, exoccipital, and prootic) that had the greatest scores on canonical axes 1 and 2. For taxon names, see legend of Fig. 4. The circles of metamorphic taxa are in black, those of facultatively neotenic taxa, in yellow (light grey in the print version), and those of neotenic taxa, in red (dark grey in the print version).

Supplementary on-line materials. These can be downloaded from the journal's web site and in Dryad (http://datadryad.org).

SOM 1. Data and various analyses (Excel format).

SOM 2. Mesquite Nexus file incorporating the data (ossification sequences, both "raw" and standardized, and subsets of the latter used for the EPCA and Canonical Variates Analyses.