THE ROYAL SOCIETY

Received 20 August 2002
Accepted 23 October 2002
Published online 20 January 2003

Inverting the hourglass: quantitative evidence against the phylotypic stage in vertebrate development

Olaf R. P. Bininda-Emonds*, Jonathan E. Jeffery and Michael K. Richardson

Institute of Evolutionary and Ecological Sciences, Leiden University, Kaiserstraat 63, PO Box 9516, 2300 RA Leiden, The Netherlands (bininda@rulsfb.leidenuniv.nl, jeffery@rulsfb.leidenuniv.nl, richardson@rulsfb.leidenuniv.nl)

The concept of a phylotypic stage, when all vertebrate embryos show low phenotypic diversity, is an important cornerstone underlying modern developmental biology. Many theories involving patterns of development, developmental modules, mechanisms of development including developmental integration, and the action of natural selection on embryological stages have been proposed with reference to the phylotypic stage. However, the phylotypic stage has never been precisely defined, or conclusively supported or disproved by comparative quantitative data. We tested the predictions of the 'developmental hourglass' definition of the phylotypic stage quantitatively by looking at the pattern of developmental-timing variation across vertebrates as a whole and within mammals. For both datasets, the results using two different metrics were counter to the predictions of the definition: phenotypic variation between species was highest in the middle of the developmental sequence. This surprising degree of developmental character independence argues against the existence of a phylotypic stage in vertebrates. Instead, we hypothesize that numerous tightly delimited developmental modules exist during the mid-embryonic period. Further, the high level of timing changes (heterochrony) between these modules may be an important evolutionary mechanism giving rise to the diversity of vertebrates. The onus is now clearly on proponents of the phylotypic stage to present both a clear definition of it and quantitative data supporting its existence.

Keywords: phylotypic stage; heterochrony; development; developmental modules; phenotypic variation; vertebrates

1. INTRODUCTION

The idea of a conserved stage in development diagnostic of a particular group of organisms dates back to von Baer, who suggested that many vertebrate species pass through an embryonic stage during which they closely resemble one another phenotypically (reviewed by Gould 1977; Raff 1996; Hall 1997). This stage has been known by a variety of names (see Seidel 1960; Cohen 1967, 1993; Slack *et al.* 1993), but is now commonly referred to as the phylotypic stage (*sensu* Sander 1983). The phylotypic stage is, in principle, applicable to any group of organisms and has been described for insects (Sander 1983), annelids and arthropods (Anderson 1973) and vertebrates (Ballard 1981; Kimmel *et al.* 1995), among others. It has, however, become most closely associated with vertebrates. In particular, Ballard's (1981) pharyngula stage is widely taken to be the vertebrate phylotypic stage, although he intended it only as a hypothetical teaching aid and never equated it with the phylotypic stage.

Today, the phylotypic stage is widely accepted in developmental biology (reviewed in Richardson & Keuck 2002). It has been used to explain the conserved pattern of limb development in tetrapods (Galis *et al.* 2001), general developmental mechanisms and the action of natural selection on developmental stages (Slack *et al.* 1993; Richardson 1999; Galis & Metz 2001; Galis *et al.* 2001), and has been said to represent the link between ontogeny and phylogeny (Hall 1997). However, its existence has also been questioned (Richardson 1995; Richardson *et al.* 1997; Collazo 2000) and, at times, the subject has been an area of intense debate. Despite the role that the phylotypic stage often plays in modern developmental theory, its existence has never been supported or disproved using comparative quantitative data. In part, this reflects the fact that comparative developmental biology is still predominantly a non-quantitative discipline. In this paper, we use two comprehensive datasets of developmental-timing information—one for vertebrates as a whole and one for mammals—to provide the first strong quantitative test of the phylotypic stage.

2. DEFINING THE PHYLOTYPIC STAGE

The phylotypic stage has proved resistant to rigorous testing because it is a nebulous concept with no single explicit definition. Moreover, the definitions are either pattern-based or process-based, confusing things even further. To our knowledge, three main definitions (in a loose sense) of the phylotypic stage exist. The most general definition is that the phylotypic stage is a developmental period of reduced phenotypic divergence (PD) among

[*] Author and address for correspondence: Lehrstuhl für Tierzucht, Technical University of Munich, D-85354 Freising-Weihenstephan, Germany (olaf.bininda@tierzucht.tum.de).

DOI 10.1098/rspb.2002.2242

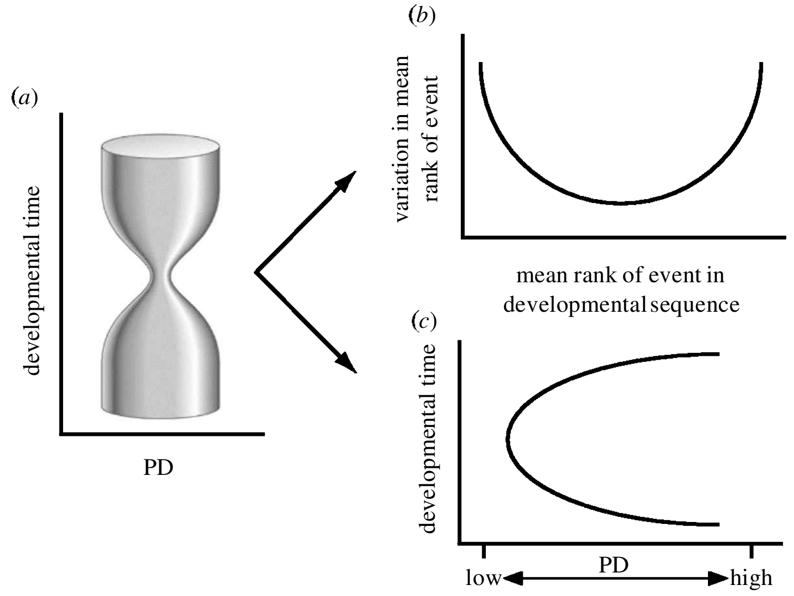
(*a*)
developmental time
PD
(*b*)
variation in mean rank of event
mean rank of event in developmental sequence
(*c*)
developmental time
low
PD
high

Figure 1. Predictions from the hourglass definition of the phylotypic stage. (*a*) In the developmental hourglass, a given group of organisms all pass through an extended phylotypic stage where they closely resemble one another. (*b*,*c*) Two testable predictions derived from the hourglass definition using ranked sequences of developmental events.

a group of organisms (Slack *et al.* 1993; Hall 1996). This pattern-based definition is best visualized in terms of the 'developmental hourglass' model of evolution (Duboule 1994; Raff 1996; figure 1). A second, process-based, definition holds that the phylotypic stage is the period of the most numerous inductive interactions among developmental events (Raff 1996; Galis & Metz 2001; Galis *et al.* 2001). Finally, there have been numerous implicit character-based definitions, relying on key structures (Anderson 1973; Sander 1983; Slack *et al.* 1993; Kimmel *et al.* 1995) or even molecular-expression patterns (Duboule 1994; Yost 1999) held to be present during and to define the phylotypic stage. In vertebrates, these structures are usually held to be the heart, pharyngeal gill arches, a tail bud and paired appendage buds (e.g. Kimmel *et al.* 1995). Galis & Metz (2001) hold the vertebrate phylotypic stage to begin with the onset of neurulation and end with the formation of the last somites. We refer to these definitions as the hourglass, inductive-interaction and key-character definitions, respectively.

These three definitions are not mutually exclusive, nor are they necessarily distinct from one another. The inductive-interaction definition is a possible mechanism that could give rise to the developmental hourglass in the first definition. The use of key characters in the third definition is a crude attempt to quantify the reduced PD of the first definition. The hourglass and inductive-interaction definitions are also explicitly relative in that they imply reduced phenotypic diversity and increased modularity, respectively, at one stage compared with others.

In this paper, we adopt the hourglass definition as our working definition of the phylotypic stage for several reasons. First, directly testing for and quantifying the presence of (numerous) inductive interactions between different molecular pathways would be prohibitively complex (although see Galis & Metz 2001). However, it is possible to rule out the presence of inductive interactions if significant changes in the order in which developmental events occur (i.e. sequence heterochrony) can be established. This is because sequence heterochrony depends on the uncoupling or dissociation of developmental events (Needham 1950). Changes in the timing of development of a tissue will affect its ability to send or respond to an inductive signal. Timing shifts will also affect the ability of linked characters to participate in common signalling pathways. Second, definitions based on key characters are largely untenable (Bininda-Emonds *et al.* 2002). Key characters are highly dependent on character definition (see § 5). More importantly, the choice of key characters is highly subjective and can obscure the fact that many more differences exist within or between other structures. These can be differences in morphology or even in the presence or absence of structures as a result of heterochrony. In a non-quantitative study, heterochrony was shown to exist during the mid-embryonic period (Richardson 1995), which is the accepted time of occurrence of the phylotypic stage.

Instead, using a definition based on reduced PD yields a prediction that can be tested in a quantitative framework. The hourglass definition predicts that species should be least divergent phenotypically in the middle of the developmental sequence. A practical means of testing this prediction is to analyse existing datasets of developmental-timing variation for clues about linkage and phenotypic similarity. This approach is relevant to the phylotypic-stage debate for two reasons. First, sequence heterochronies depend on dissociation of developmental events and therefore provide an index of character linkage (modularity; *sensu* Wagner 1996). Second, timing shifts between organs in different species will reduce phenotypic resemblance at a given developmental stage. Thus, if the phylotypic stage does exist, events in the middle of the developmental sequence (i.e. at the peak of organogenetic activity) should show less variation in timing than ones that occur earlier or later.

3. MATERIAL AND METHODS

(a) *Datasets of developmental-timing variation*

We tested for the existence of the phylotypic stage using two comprehensive datasets of developmental-timing data that we have compiled (see electronic Appendices A and B, available on The Royal Society's Publications Web site, for a list of developmental events). The first dataset (electronic Appendix A) consists of 41 developmental events for a range of 14 vertebrate species (Jeffery *et al.* 2002). The second dataset (electronic Appendix B) consists of an expanded set of 116 events for a more taxonomically restricted sample of 14 mammals plus two amniote outgroups (O. R. P. Bininda-Emonds, J. E. Jeffery and M. K. Richardson, unpublished data). The events were largely developmental transformations, representing the first appearance of a defined morphology (e.g. heart primordia) or morphogenetic movement (e.g. fusion of neural folds) and occur throughout the entire mid-embryonic (organogenetic) period. We concentrated on transformations because they are vital to establishing homology (Wagner 1996). Most developmental events were universal across all species in a dataset (i.e. shared features present in all species).

For both datasets, we transformed the raw developmental-timing data for each species into a ranked developmental sequence (i.e. the first event to occur was given a rank of one, the second a rank of two, and so on) to standardize for chronological age (Bininda-Emonds *et al.* 2002). The sequences were further standardized to account for missing data and artefactual instances of event simultaneity, both of which will 'shorten' the developmental sequence (Bininda-Emonds *et al.* 2002). In each species, events that occurred at the same time were given the same rank and the ranks of all events were adjusted such that they were equally distributed along the entire sequence length, which was fixed at 41 or 116 events (or ranks) depending on the dataset.

(b) *Quantifying PD*

The prediction derived from the hourglass definition provides two sets of hypotheses testable using ranked sequences (Bininda-Emonds *et al.* 2002) of developmental events (figure 1). First, if we examine the mean rank of an event across species, we should find that those events with an intermediate mean rank show the least variation in that mean (figure 1*b*). Second, we can derive a metric that quantifies the amount of PD across species. The mean PD across all events should be lower within the phylotypic stage than outside it (figure 1*c*). In essence, the graph of mean PD over time should recreate the right-hand side of the developmental hourglass.

For the second hypothesis, we divided the entire developmental sequence into periods and recorded whether a given event occurred during that interval. With low PD, a given event will be either mostly present or mostly absent from all species within a given period; with high PD, an event will be about equally present and absent across all species. This idea is easily captured by the following metric:

$$\mathrm{PD}_{\mathrm{event}} = 1 - \left|\frac{n_{\mathrm{present}} - n_{\mathrm{absent}}}{n_{\mathrm{present}} + n_{\mathrm{absent}}}\right|,$$

where n_{present} and n_{absent} are the number of times an event is present or absent, respectively, in a given period across all species. The overall PD for a given period is simply the value of $\mathrm{PD}_{\mathrm{event}}$ averaged over all developmental events.

Because the periods we created are essentially arbitrary, we examined periods of different durations, ranging from one-fifth to one-half of the length of the entire developmental sequence. Further, each period was placed at all points along the developmental sequence in turn in the form of a time-series analysis. For example, given a developmental sequence 116 units long, a period of one-half of the developmental sequence (i.e. 58 units) was placed starting at time 0, 1, 2, ..., 58 units and the value of PD for each placement was calculated. This 'sliding window' analysis will eliminate any artefacts arising from a single subjective placement of a period.

4. RESULTS

For both sets of developmental-timing data, the results are exactly opposite to the predictions in figure 1: variation between species is highest in the middle of the developmental sequence (figure 2). This is true regardless of whether we examined mean ranks or PD. The pattern for mean ranks is highly significant for both the vertebrate (figure 2*a*) and the mammal (figure 2*b*) datasets according to a polynomial regression of order two ($p < 0.0001$), but could arise through a combination of edge effects and event density. Events that are on average closer to the ends of the developmental sequence will show reduced variation in their relative positions (rank) because they can move chiefly in only one direction in the sequence (i.e. towards the middle). By contrast, events in the middle of the sequence are more able to change their relative position through movements in either direction. Similarly, events that occur close in absolute developmental time to many other events (i.e. in an interval of high event density) can change their relative position more easily and to a greater degree than more isolated events. They will therefore possess higher variation across species. However, for those species where event density could be measured because specimens were of a known absolute age (domestic chicken, *Gallus gallus*; mouse, *Mus musculus*; European rabbit, *Oryctolagus cuniculus*; and African clawed toad, *Xenopus laevis*), the highest density is found towards the beginning of the developmental sequence (figure 3), not in the middle where variation in event position is highest. Therefore, event density does not appear to influence our results.

The analysis of PD (figure 2*c*,*d*) avoids the artefacts associated with both edge effects and event density by dispersing them over an extended time interval. Thus, small variations in the exact (relative) timing of an event should not strongly influence the overall pattern, particularly when the developmental period ('window') examined is very large. Again, contrary to the predictions of the inductive-interaction model, phenotypic variation between species was highest in the middle of the developmental sequence for both vertebrates and mammals, regardless of the size of the window used. Therefore, there must be a substantial amount of heterochrony (i.e. changes in developmental timing) present at this time. Unsurprisingly, PD was generally lower among mammals (minimum PD = 0.39) than among the more diverse set of vertebrates examined (minimum PD = 0.54).

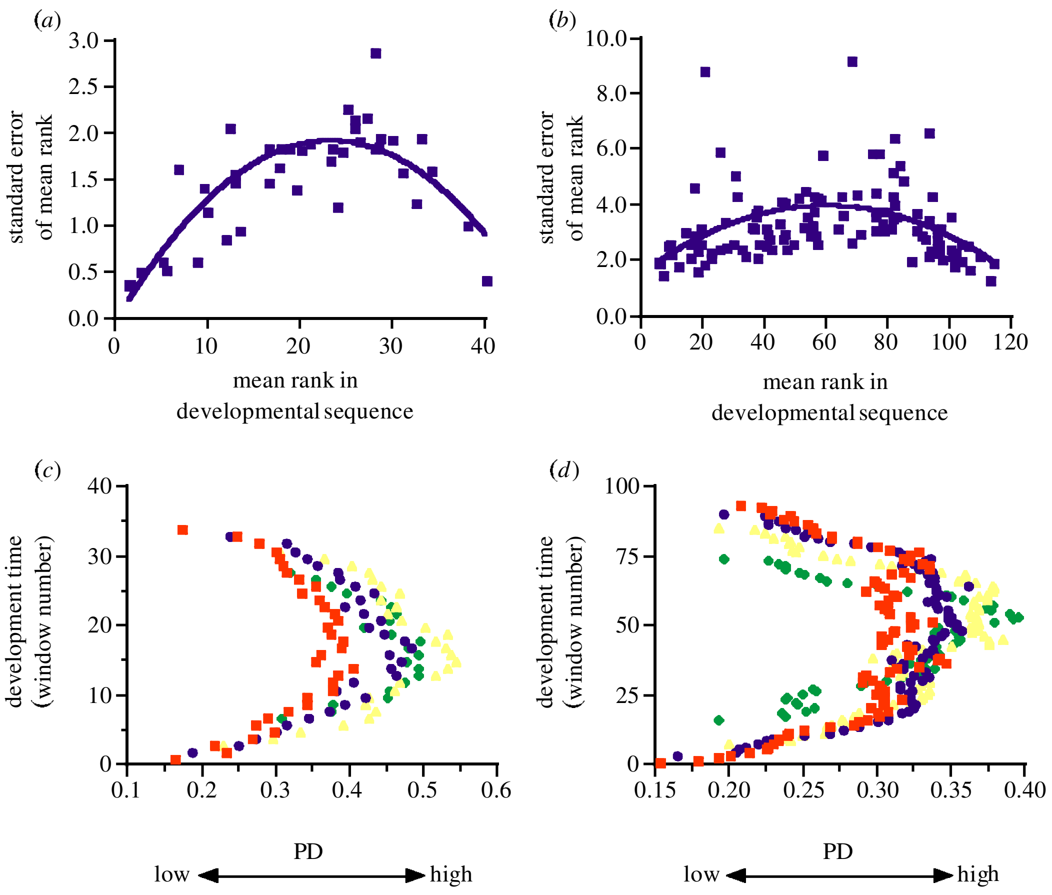

Figure 2. Timing variation during the phylotypic stage. (*a*,*b*) The standard error of the mean rank of a developmental event versus the mean rank. ((*a*) $y = -0.004x^2 + 0.169x - 0.057$, $p < 0.0001$ and (*b*) $y = -0.001x^2 + 0.082x + 1.520$, $p < 0.0001$.) (*c*,*d*) PD across species for a given period during development. Periods in (*c*) and (*d*) are represented as overlapping windows (i.e. composed of events of ranks 1–5, 2–6, 3–7, …) with durations of one-fifth (red squares), one-quarter (blue circles), one-third (yellow triangles) or one-half (green diamonds) of the length of the developmental time span examined. For the data from both vertebrates (*a*,*c*) and mammals (*b*,*d*) the trends run exactly opposite to the predictions in figure 1.

5. DISCUSSION

Support for the hourglass definition of the phylotypic stage derives largely from subjective statements about the overall similarity of embryos of different species, usually based on an examination of pictures of embryos and not from rigorous character-based data analysis. These phenetic statements have two components. First, species are said to be characterized by variation in early (e.g. patterns of cleavage or gastrulation) and late (e.g. feather or hair primordia) developmental features, producing high PD at the two ends of the developmental sequence. Second, few unique features are believed to occur during the intervening period, which involves instead only features that are shared (primitively) among the species. PD is correspondingly reduced in the mid-embryonic phylotypic stage.

In our opinion, much of the difference between these two components stems from issues of character definition. Shared features undoubtedly exist during the mid-embryonic period, but those used to support the phylotypic stage are often defined so coarsely as to obscure potential variation between species. For instance, the statement that vertebrate embryos all possess a heart during the phylotypic stage (Kimmel *et al.* 1995) ignores important variation in how the heart is formed (see Richardson 1995) as well as the existence of heterochrony, which can result in the heart being in different stages of its development when other key characters are all present (which may themselves be at varying stages in their development). Outside the phylotypic stage, character definition becomes more rigorous and exacting. For example, all vertebrate embryos undergo gastrulation, which could be considered the equivalent level of precision to saying that the heart is present. However, in this case, a distinction is made between the structure or event (gastrulation) and the diverse mechanisms giving rise to it (e.g. blastopore, primitive streak), with the emphasis on the latter (Hall 1997). Together with similarly stringent definitions of other events, the end result is an apparent increase in the proportion of unique features and thus higher PD. In essence, much of this discussion boils down to using valid statements of homology, which becomes increasingly difficult at a developmental or embryological level (de Beer 1971; Wagner & Misof 1993; Abouheif *et al.* 1997).

Our methodology does not avoid all the pitfalls mentioned in the preceding paragraph. However, we feel that our character definitions are highly and more uniformly specific and thus equitable among the various events throughout the developmental period examined. Events were also selected to provide comprehensive coverage of the mid-embryonic period for other purposes, without regard to the question of the phylotypic stage. Most

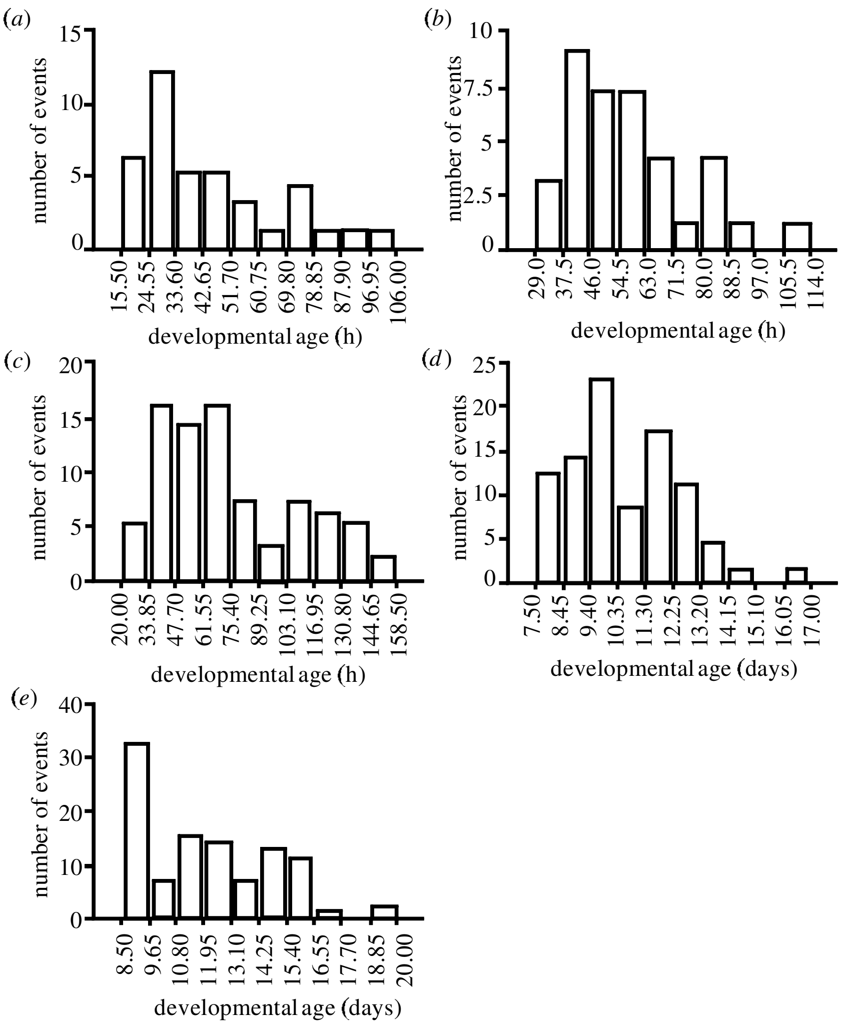

Figure 3. Temporal distribution of developmental events in (*a*) African clawed toad, *Xenopus laevis*; (*b*,*c*) domestic chicken, *Gallus gallus*; (*d*) house mouse, *Mus musculus*; and (*e*) European rabbit, *Oryctolagus cuniculus*. (*a*) and (*b*) use the 41 developmental events of Jeffery *et al.* (2002), while (*c*), (*d*) and (*e*) use the 116 developmental events of O. R. P. Bininda-Emonds, J. E. Jeffery and M. K. Richardson (unpublished data). Owing to heterochrony, the histogram bars are not comparable across species with respect to the events they contain.

importantly, we used these comparative data to test for the existence of the phylotypic stage in an explicitly quantitative framework, thereby avoiding subjective decisions. The results are clear. In both datasets, both metrics showed that PD is actually greatest in the middle of the mid-embryonic period. This strongly contradicts the pattern predicted by the hourglass definition of the phylotypic stage.

This finding results from the unexpectedly high level of heterochrony during the mid-embryonic period, which was suggested previously but without quantitative data (Richardson 1995). The level of heterochrony is actually underestimated by our use of ranked developmental-timing data, which can only indicate relative and not absolute timing changes (Bininda-Emonds *et al.* 2002). Moreover, our findings also argue against the inductive-interaction definition, which holds the phylotypic stage to be regulated more tightly than other stages of development. Because some temporal regulation must be present throughout development (Rougvie 2001), we hypothesize instead that the amount of integration decreases in the mid-embryonic period, with developmental modules becoming more restricted in terms of the characters that comprise them.

This is the opposite conclusion from that reached by Galis & Metz (2001), who observed that application of teratogens during the phylotypic stage caused widespread developmental abnormalities resulting in high mortality. This was not the case outside the phylotypic stage. From these results, they inferred a high level of inductive interactions at this time, which they took to support the existence of the phylotypic stage. However, these observations can be explained equally well by other factors. Their phylotypic stage is extremely broad, spanning virtually the entire organogenetic period, when all the major organ primordia are being specified. This is clearly a crucial point in development when all structures are more susceptible to disturbance than they are during later growth phases (regardless of the level of inductive interactions). Thus, their observations may derive from widespread independent susceptibility rather than widespread inductive interactions. Our results demonstrate that the supposed inductive interactions in the mid-embryonic period can be broken, if only on a small scale. Dramatic alterations are also possible (e.g. delayed onset of limb development in frogs; Richardson 1995; Galis *et al.* 2001), which speaks against a universal phylotypic stage based on this mechanism or on this particular key character (which is not even universal among vertebrates).

Given our results, the onus is now clearly on proponents of the phylotypic stage both to provide a clear definition of it and to support its existence using comparative quantitative data. In the absence of the latter, we argue against the existence of a phylotypic stage (or 'period' *sensu* Richardson 1995) in vertebrates. Instead, the pattern of development more closely resembles a 'spinning top' (see fig. 6*b* in Richardson 1999), with reduced PD at both ends. The high degree of independence we infer for developmental events during the mid-embryonic period may serve as an additional important target for natural selection (Richardson 1999), possibly resulting in macroevolutionary changes. In vertebrates, the independence of mid-embryonic developmental events apparently allows pervasive small-scale timing changes, which may contribute to the corresponding high level of phenotypic diversity across adult vertebrates (Gould 1982).

We thank Wallace Arthur, Lennart Olsson and one anonymous referee for their insightful comments on the manuscript. This work was supported by the van der Leeuw Fonds (The Netherlands) to O.R.P.B.-E. and M.K.R. and by the Leverhulme Trust (UK) to J.E.J.

REFERENCES

Abouheif, E., Akam, M., Dickinson, W. J., Holland, P. W., Meyer, A., Patel, N. H., Raff, R. A., Roth, V. L. & Wray, G. A. 1997 Homology and developmental genes. *Trends Genet.* **13**, 432–433.

Anderson, D. T. 1973 *Embryology and phylogeny in annelids and arthropods*. International series of monographs in pure and applied biology. Division: Zoology. Oxford: Pergamon.

Ballard, W. W. 1981 Morphogenetic movements and fate maps of vertebrates. *Am. Zool.* **21**, 391–399.

Bininda-Emonds, O. R. P., Jeffery, J. E., Coates, M. I. & Richardson, M. K. 2002 From Haeckel to event-pairing: the evolution of developmental sequences. *Theory Biosci.* **121**, 297–320.

Cohen, J. 1967 *Living embryos: an introduction to the study of animal development*. Oxford: Pergamon.

Cohen, J. 1993 Development of the zootype. *Nature* **363**, 307.

Collazo, A. 2000 Developmental variation, homology, and the pharyngula stage. *Syst. Biol.* **49**, 3–18.

de Beer, G. 1971 *Homology: an unsolved problem*. Oxford biology readers. Oxford University Press.

Duboule, D. 1994 Temporal colinearity and the phylotypic progression: a basis for the stability of a vertebrate Bauplan and the evolution of morphologies through heterochrony. *Development* (Suppl.), 135–142.

Galis, F. & Metz, J. A. J. 2001 Testing the vulnerability of the phylotypic stage: on modularity and evolutionary conservation. *J. Exp. Zool.* **291**, 195–204.

Galis, F., van Alphen, J. J. M. & Metz, J. A. J. 2001 Why five fingers? Evolutionary constraints on digit numbers. *Trends Ecol. Evol.* **16**, 637–646.

Gould, S. J. 1977 *Ontogeny and phylogeny*. Cambridge, MA: Belknap.

Gould, S. J. 1982 Change in developmental timing as a mechanism of macroevolution. In *Evolution and development* (ed. J. T. Bonner), pp. 333–346. New York: Springer.

Hall, B. K. 1996 Baupläne, phylotypic stages, and constraint—why are there so few types of animals? *Evol. Biol.* **29**, 251–261.

Hall, B. K. 1997 Phylotypic stage or phantom: is there a highly conserved embryonic stage in vertebrates? *Trends Ecol. Evol.* **12**, 461–463.

Jeffery, J. E., Bininda-Emonds, O. R. P., Coates, M. I. & Richardson, M. K. 2002 Analyzing evolutionary patterns in vertebrate embryonic development. *Evol. Dev.* **4**, 292–302.

Kimmel, C. B., Ballard, W. W., Kimmel, S. R., Ullmann, B. & Schilling, T. F. 1995 Stages of embryonic development of the zebrafish. *Devl Dynam.* **203**, 253–310.

Needham, J. 1950 *Biochemistry and morphogenesis*. Cambridge University Press.

Raff, R. A. 1996 *The shape of life: genes, development, and the evolution of animal form*. University of Chicago Press.

Richardson, M. K. 1995 Heterochrony and the phylotypic period. *Devl Biol.* **172**, 412–421.

Richardson, M. K. 1999 Vertebrate evolution: the developmental origins of adult variation. *BioEssays* **21**, 604–613.

Richardson, M. K. & Keuck, G. 2002 Haeckel's ABC of evolution and development. *Biol. Rev.* **77**, 495–528.

Richardson, M. K., Hanken, J., Gooneratne, M. L., Pieau, C., Raynaud, A., Selwood, L. & Wright, G. M. 1997 There is no highly conserved embryonic stage in the vertebrates: implications for current theories of evolution and development. *Anat. Embryol.* **196**, 91–106.

Rougvie, A. E. 2001 Control of developmental timing in animals. *Nature Rev. Genet.* **2**, 690–701.

Sander, K. 1983 The evolution of patterning mechanisms: gleanings from insect embryogenesis and spermatogenesis. In *Development and evolution* (ed. B. C. Goodwin, N. Holder & C. C. Wylie), pp. 137–159. Cambridge University Press.

Seidel, F. 1960 Körpergrundgestalt und Keimstruktur eine Erörterung über die Grundlagen der vergleichenden und experimentellen Embryologie und deren Gültigkeit bei phylogenetischen Überlegungen. *Zool. Anz.* **164**, 245–305.

Slack, J. M., Holland, P. W. & Graham, C. F. 1993 The zootype and the phylotypic stage. *Nature* **361**, 490–492.

Wagner, G. P. 1996 Homologues, natural kinds and the evolution of modularity. *Am. Zool.* **36**, 36–43.

Wagner, G. P. & Misof, B. Y. 1993 How can a character be developmentally constrained despite variation in developmental pathways? *J. Evol. Biol.* **6**, 449–455.

Yost, H. J. 1999 Diverse initiation in a conserved left–right pathway? *Curr. Opin. Genet. Dev.* **9**, 422–426.

Visit http://www.pubs.royalsoc.ac.uk to see electronic appendices to this paper.