BIOINFORMATICS DISCOVERY NOTE

Vol. 28 no. 24 2012, pages 3178–3181
doi:10.1093/bioinformatics/bts623

Genome analysis

Advance Access publication October 18, 2012

Lengthening of 3′UTR increases with morphological complexity in animal evolution

Cho-Yi Chen[1,2], Shui-Tein Chen[2], Hsueh-Fen Juan[1,*] and Hsuan-Cheng Huang[3,*]

[1]Genome and Systems Biology Degree Program, Department of Life Science, Institute of Molecular and Cellular Biology, Graduate Institute of Biomedical Electronics and Bioinformatics, National Taiwan University, Taipei 116, [2]Institute of Biological Chemistry, Academia Sinica, Taipei 115 and [3]Institute of Biomedical Informatics, Center for Systems and Synthetic Biology, National Yang-Ming University, Taipei 112, Taiwan

Associate Editor: Martin Bishop

ABSTRACT

Motivation: Evolutionary expansion of gene regulatory circuits seems to boost morphological complexity. However, the expansion patterns and the quantification relationships have not yet been identified. In this study, we focus on the regulatory circuits at the post-transcriptional level, investigating whether and how this principle may apply.

Results: By analysing the structure of mRNA transcripts in multiple metazoan species, we observed a striking exponential correlation between the length of 3′ untranslated regions (3′UTR) and morphological complexity as measured by the number of cell types in each organism. Cellular diversity was similarly associated with the accumulation of microRNA genes and their putative targets. We propose that the lengthening of 3′UTRs together with a commensurate exponential expansion in post-transcriptional regulatory circuits can contribute to the emergence of new cell types during animal evolution.

Contact: yukijuan@ntu.edu.tw or hsuancheng@ym.edu.tw.

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on August 6, 2012; revised on October 10, 2012; accepted on October 16, 2012

1 INTRODUCTION

For the past decade of genome exploration, research into the evolution of organismal diversity and morphological complexity has shifted its focus from raw genome size to gene regulatory complexity (Carroll, 2001; Vogel and Chothia, 2006; Tan *et al.*, 2009). Multicellular organisms may contain a variety of widely differing and specialized cell types. The morphological complexity, thus, can be defined as the number of distinct cell types within a species (Valentine *et al.*, 1994; Vogel and Chothia, 2006). Generation of new cell types may require exploring and reshaping the expression landscape that relies on expansion or rewiring of extant regulatory circuits, leading to a variety of cell-specific expression patterns. Complex gene regulatory circuits may allow a greater variety of potential combinations of gene expression, thereby facilitating the generation of novel cell types. This principle has been studied mainly at the transcriptional level, that is, the regulatory circuits of transcription factors (TFs) and corresponding *cis*-regulatory elements on DNA (Carroll, 2001). However, recent studies have shown that the post-transcriptional gene regulation is believed to be widespread (Bartel, 2009; De Mulder and Berezikov, 2010; Ji *et al.*, 2009). Also, extensive co-regulation and crosstalk between the transcriptional and the post-transcriptional levels have been suggested (Chen *et al.*, 2011; Cui *et al.*, 2007; Hobert, 2008; Lin *et al.*, 2012). Thus, it is of great interest to examine whether and to what extent this principle may also apply at the post-transcriptional level.

*To whom correspondence should be addressed.

2 MATERIALS AND METHODS

2.1 Collection of mature mRNA transcripts

Mature mRNA transcript and untranslated region (UTR) sequences were downloaded from RefGene table through UCSC Genome Browser (Rhead *et al.*, 2010) and from UTRdb (Grillo *et al.*, 2010). Fifteen organisms, for which the 5′ and 3′UTRs of mature mRNA transcripts have been well characterized and annotated (RefSeq), were used in this study, *Caenorhabditis elegans*, *Apis mellifera* (honey bee), *Anopheles gambiae*, *Drosophila melanogaster* (fruit fly), *Ciona intestinalis*, *Danio rerio* (zebrafish), *Takifugu rubripes* (Fugu rubripes), *Xenopus tropicalis* (western clawed frog), *Gallus gallus* (chicken), *Canis familiaris* (dog), *Bos taurus* (cattle), *Rattus norvegicus* (Norway rat), *Mus musculus* (house mouse), *Pan troglodytes* (chimpanzee) and *Homo sapiens* (human). *Saccharomyces cerevisiae* (yeast) transcript data, derived from RNA sequencing, were referenced from Nagalakshmi *et al.* (2008). A common set of orthologous genes were selected from NCBI's HomoloGene database (Supplementary Methods). Regression analysis was performed to study the relationships between the region length and the organismal complexity (Fig. 1A and Supplementary Fig. S1) based on four different models, linear, exponential, logarithm and power models (Supplementary Methods). The exponential model gave the best performance in most cases and was reported thereinafter.

2.2 Putative TFs and microRNA genes in organisms

The number of predicted TFs was obtained from the DBD 2.0. These putative TFs all contain sequence-specific DNA-binding domains (Wilson *et al.*, 2008). The number of microRNA (miRNA) genes that have been identified in each organism was derived from miRBase release 18 (Griffiths-Jones *et al.*, 2008). Seventeen metazoan species that have reported miRNA genes were used in this study, *Caenorhabditis briggsae*, *C.elegans*, *A.gambiae*, *A.mellifera*, *D.melanogaster*, *C.intestinalis*, *D.rerio*, *T.rubripes*, *Tetradon nigroviridis*, *X.tropicalis*, *G.gallus*, *C.familiaris*, *B.taurus*, *R.norvegicus*, *M.musculus*, *P.troglodytes* and *H.sapiens* (Fig. 1B and Supplementary Fig. S2).

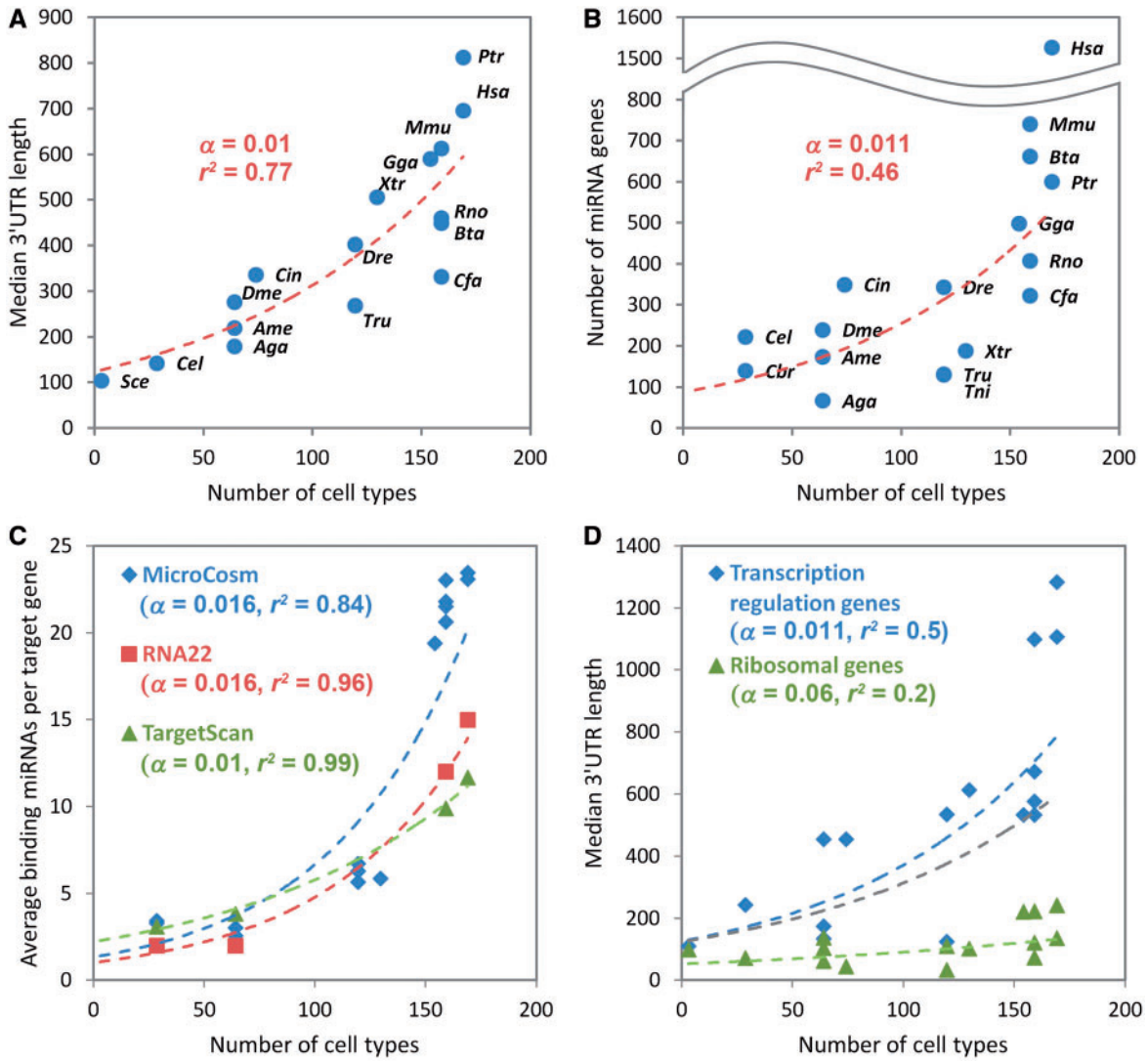

Fig. 1. Exponential correlation between miRNA-mediated regulation and morphological complexity. (**A**) A strong exponential correlation exists between the median length of 3′UTRs and the morphological complexity among 15 metazoan species, as measured by distinct cell types (Vogel and Chothia, 2006). Budding yeast (*S.cerevisiae*) is included for comparison as a unicellular eukaryote. The dashed line indicates a single exponential fit, together with the fitting parameter α and the coefficient of determination r^2. (**B**) The number of miRNA genes in each genome and (**C**) miRNA binding complexity (average numbers of putative binding miRNAs per target gene) correlate exponentially with morphological complexity. See also Supplementary Figure S3 for species labels. (**D**) The growth profiles of median 3′UTR length for transcription regulation genes (GO: 0006355) and ribosomal genes (GO: 0005840) show different trends. Budding yeast is included for comparison. The gray dash line, showing the global trend, is adapted from (A) for comparison

2.3 Complexity of miRNA regulation on target genes

The putative miRNA binding sites among target genes were obtained from MicroCosm Targets 5 (formerly miRBase Targets) (Griffiths-Jones *et al.*, 2008), TargetScan 5.1 (Friedman *et al.*, 2009) and RNA22 (Miranda *et al.*, 2006). In the TargetScan and RNA22 databases, predictions were only available for worm (*C.elegans*), fly (*D.melanogaster*), mouse (*M.musculus*) and human (*H.sapiens*), whereas predictions for 15 species were available in the MicroCosm database (Supplementary Fig. S3). The miRNA binding complexity for each organism was defined as the average numbers of putative binding miRNAs per target gene. Although the three databases used different algorithms to predict miRNA binding sites (Bartel, 2009), their results still displayed a close correlation between miRNA binding complexity and morphological complexity (Fig. 1C).

2.4 Functional annotation

Organism functional annotation categorized by Gene Ontology (GO) was obtained from UniProtKB (Magrane and Consortium, 2011) and DAVID Bioinformatics Resources 6.7 (Huang *et al.*, 2009). Typical GO categories of miRNA targets and anti-targets were according to Stark *et al.* (2005).

3 RESULTS AND DISCUSSION

After analysing the structure of genomic mRNA transcripts from a variety of multicellular animals, we observed a striking exponential correlation (Pearson's $r = 0.88$, $P = 1.69 \times 10^{-5}$) between the typical length of the 3′UTR and morphological complexity as measured by the number of cell types (Fig. 1A). Conversely, there was no significant correlation observed between morphological complexity and 5′UTR length (Supplementary Fig. S1a) or coding sequence length (Supplementary Fig. S1b).

Considering the possibility that the observed variations in the 3′UTR length could be merely because of the implications of gene gain and loss in evolution, we selected a common set of orthologous genes among these organisms and repeated the same analysis on these new subjects, obtaining a similar pattern (Supplementary Fig. S4) as shown before, which indicates that the exponential lengthening of 3′UTR can be observed not only in a genome-wide scale but also in the broadly conserved gene subset.

As animal miRNAs mainly target the *cis*-regulatory elements in 3′UTR, this difference may be because of miRNA-based post-transcriptional regulation (Mazumder *et al*., 2003). Indeed, miRNAs have been suggested as a practical phylogenetic marker for metazoa (Kosik, 2009; Wheeler *et al*., 2009). Recent studies also showed that miRNAs play a crucial role in the evolution of 3′UTR and the establishment of tissue identity (De Mulder and Berezikov, 2010; Farh *et al*., 2005; Kosik, 2009; Shkumatava *et al*., 2009; Stark *et al*., 2005). The likelihood that a transcript will be targeted by miRNA may increase with 3′UTR length because 3′UTR expansion can yield additional miRNA binding sites (Ji *et al*., 2009; Supplementary Figs S5 and S6). Furthermore, longer 3′UTRs may include multiple polyadenylation signals, and thus produce several alternative transcript isoforms with different combinations of miRNA binding sites, leading to a large increase in regulatory complexity (Ji *et al*., 2009; Majoros and Ohler, 2007; Mangone *et al*., 2010). Ji *et al*. (2009) also reported the observation of progressive lengthening of 3′UTR during mouse embryonic development, together suggesting that lengthening of 3′UTR may play essential roles in both evolution and developmental processes.

The number of miRNA genes in the genome was also found to grow exponentially with morphological complexity (Pearson's $r = 0.68$, $P = 2.79 \times 10^{-3}$; Fig. 1B). The co-expansion of *trans*-regulators (miRNA genes) and potential *cis*-elements (length of 3′UTR) suggests a possible evolutionary increase in the complexity of post-transcriptional regulation. Indeed, we found that the miRNA binding complexity in 3′UTR, as predicted by different algorithms, also strongly correlated with morphological complexity (Fig. 1C and Supplementary Fig. S3).

In addition, the typical miRNA target and anti-target categories have a clear divergence in 3′UTR length patterns (Fig. 1D). Transcriptional regulation genes, known as a typical class of miRNA targets (Enright *et al*., 2003; Stark *et al*., 2005), grew exponentially in 3′UTR length as organismal complexity increased (Pearson's $r = 0.71$, $P = 3.07 \times 10^{-3}$). In sharp contrast to this, ribosomal genes, a class of housekeeping genes and miRNA anti-targets (Stark *et al*., 2005), had relatively short 3′UTRs in all model organisms and no significant correlation. These results reinforce the argument that the lengthening of 3′UTR is highly associated with the expansion of miRNA regulatory programs.

Previous research has shown that the metazoan genome is continuously acquiring novel miRNA families (Chen and Rajewsky, 2007; Griffiths-Jones *et al*., 2008; Kosik, 2009). A novel miRNA would have an initially genome-wide stochastic regulatory effect on all potential targets, and then the natural selection drives the preservation or avoidance of these miRNA binding sites (Chen and Rajewsky, 2007). On the other hand, the increase in 3′UTR length could be an effective strategy to stochastically accommodate more miRNA binding sites in 3′UTR and boost the complexity of miRNA regulation combinations, thus expanding the expression space for natural selection to explore. The selection process may finally result in a cross-the-board lengthening for 3′UTR, but still with some flexibility. For example, housekeeping genes tend to avoid miRNA binding sites and maintain their 3′UTR length relatively short, so that they may perform their household tasks without unneeded interference (Stark *et al*., 2005).

Although the association between 3′UTR lengthening and increasing of miRNA regulatory complexity could be observed within and across multiple species (Supplementary Figs S5 and S6), the cause–effect relationship between the two factors is still in a tangle, unless the temporal precedence could be firstly demonstrated by comprehensive genome and transcriptome reconstruction of the last common ancestors in the phylogenic trees of animals. We suspect that the expansion of miRNA regulatory circuits may provide additional advantages for organisms, such as strengthening the gene expression buffering and anti-fluctuation ability (Li *et al*., 2009; Stark *et al*., 2005), which in turn help the organisms to explore their genotype space in races with other species or environment changes. However, the possibility of opposite direction of selection force cannot be ruled out; in other words, some undiscovered factors that directly promote the 3′UTR lengthening along with increasing miRNA binding sites as its side effect could exist beyond our knowledge.

A similar exponential correlation of morphological complexity was demonstrated in the number of TFs (Pearson's $r = 0.70$, $P = 1.81 \times 10^{-3}$; Supplementary Fig. S2). This suggests that the expansion of gene regulators occurred and co-evolved at both the transcriptional and post-transcriptional levels. Indeed, the gene regulatory circuits at both levels were found highly analogous and extensively cross-wired to each other, allowing for even more complicated regulation of cellular processes and homeostasis (Cui *et al*., 2007; Hobert, 2008; Lin *et al*., 2012). Given the role of miRNAs in stabilizing gene expression and conferring robustness to gene regulatory networks (De Mulder and Berezikov, 2010; Hornstein and Shomron, 2006; Li *et al*., 2009; Stark *et al*., 2005; Tsang *et al*., 2007), miRNAs are likely to play an important role in establishing stable expression landscapes and maintaining tissue identities in cooperation with TFs, together contributing to increase of morphological complexity.

Interestingly, all the regulation-related variables (L) examined earlier in the text showed similar exponential growth against the number of cell types (n):

$$L(n) \sim e^{\alpha n} \quad (1)$$

or equivalently:

$$dL/dn \sim \alpha L \quad (2)$$

with a factor α ranging from 0.01 to 0.016 (Fig. 1A–D). It indicated that the generation of a new cell type involves an ~1% increase in regulatory complexity. This also implied that the cost of generating a new cell type was ~1% additional redundancy in the regulatory circuits. The redundancy may provide robustness and have the ability to accumulate variation to explore new genotype space, leading to evolutionary innovations (Wagner, 2011).

The emergence of morphological complexity seems to have its multi-origins and different evolutionary paths in many aspects of cellular processes and molecular systems. For example, in addition to the expansion of TF/miRNA regulatory networks, cell signalling networks also play a role in promoting organismal complexity (Li *et al*., 2012). On the other hand, regulatory elements in 5′UTR do not seem to be a critical factor to the increased organismal complexity in metazoan (Chen *et al*., 2012). It is of great interest to elucidate what factors and to what extents these factors contributed to the increased morphological complexity.

4 CONCLUSIONS

Our findings align with recent studies showing that protein domain recombination and *cis*-regulatory element recombination may also have a major role in evolution (Paixão and Azevedo, 2010; Peisajovich *et al*., 2010). As gene duplication and recombination are combinatorial processes that rely on existing genetic materials (Olson, 2006), it follows the principle 'complexity will breed complexity', leading to a non-linear expansion in post-transcriptional interactions. Based on our empirical evidence of a strong exponential trend between the development of new cell types and the expansion of post-transcriptional regulatory circuits, we, therefore, propose that 3′UTR lengthening can contribute to greater morphological complexity in animals, and this process was driven by natural selection.

ACKNOWLEDGEMENTS

The authors thank Ben-Yang Liao and Feng-Chi Chen for helpful discussion.

Funding: National Science Council, Taiwan (NSC 99-2621-B-010-001-MY3 and 002-005-MY3); National Health Research Institutes (NHRI-EX98-9819PI).

Conflict of Interest: none declared.

REFERENCES

Bartel,D.P. (2009) MicroRNAs: target recognition and regulatory functions. *Cell*, **136**, 215–233.

Carroll,S.B. (2001) Chance and necessity: the evolution of morphological complexity and diversity. *Nature*, **409**, 1102–1109.

Chen,C.H. *et al*. (2012) The plausible reason why the length of 5′ untranslated region is unrelated to organismal complexity. *BMC Res. Notes*, **4**, 312.

Chen,C.Y. *et al*. (2011) Coregulation of transcription factors and microRNAs in human transcriptional regulatory network. *BMC Bioinformatics*, **12**, S41.

Chen,K. and Rajewsky,N. (2007) The evolution of gene regulation by transcription factors and microRNAs. *Nat. Rev. Genet*., **8**, 93–103.

Cui,Q. *et al*. (2007) MicroRNAs preferentially target the genes with high transcriptional regulation complexity. *Biochem. Biophys. Res. Commun*., **352**, 733–738.

De Mulder,K. and Berezikov,E. (2010) Tracing the evolution of tissue identity with microRNAs. *Genome Biol*., **11**, 111.

Enright,A. *et al*. (2003) MicroRNA targets in Drosophila. *Genome Biol*., **5**, R1.

Farh,K.K. *et al*. (2005) The widespread impact of mammalian microRNAs on mRNA repression and evolution. *Science*, **310**, 1817–1821.

Friedman,R.C. *et al*. (2009) Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res*., **19**, 92–105.

Griffiths-Jones,S. *et al*. (2008) miRBase: tools for microRNA genomics. *Nucleic Acids Res*., **36**, D154–D158.

Grillo,G. *et al*. (2010) UTRdb and UTRsite (RELEASE 2010): a collection of sequences and regulatory motifs of the untranslated regions of eukaryotic mRNAs. *Nucleic Acids Res*., **38**, D75–D80.

Hobert,O. (2008) Gene regulation by transcription factors and microRNAs. *Science*, **319**, 1785–1786.

Hornstein,E. and Shomron,N. (2006) Canalization of development by microRNAs. *Nat. Genet*., **38** (Suppl.), S20–24.

Huang,D.W. *et al*. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc*., **4**, 44–57.

Ji,Z. *et al*. (2009) Progressive lengthening of 3′ untranslated regions of mRNAs by alternative polyadenylation during mouse embryonic development. *Proc. Natl. Acad. Sci. USA*, **106**, 7028–7033.

Kosik,K.S. (2009) MicroRNAs tell an evo-devo story. *Nat. Rev. Neurosci*., **10**, 754–759.

Li,L. *et al*. (2012) The human phosphotyrosine signaling network: evolution and hotspots of hijacking in cancer. *Genome Res*., **22**, 1222–1230.

Li,X. *et al*. (2009) A microRNA imparts robustness against environmental fluctuation during development. *Cell*, **137**, 273–282.

Lin,C.C. *et al*. (2012) Crosstalk between transcription factors and microRNAs in human protein interaction network. *BMC Syst. Biol*., **6**, 18.

Magrane,M. and Consortium,U. (2011) UniProt knowledgebase: a hub of integrated protein data. *Database*, **2011**, bar009.

Majoros,W. and Ohler,U. (2007) Spatial preferences of microRNA targets in 3′ untranslated regions. *BMC Genomics*, **8**, 152.

Mangone,M. *et al*. (2010) The landscape of C. elegans 3′UTRs. *Science*, **329**, 432–435.

Mazumder,B. *et al*. (2003) Translational control by the 3′-UTR: the ends specify the means. *Trends Biochem. Sci*., **28**, 91–98.

Miranda,K. *et al*. (2006) A pattern-based method for the identification of microRNA binding sites and their corresponding heteroduplexes. *Cell*, **126**, 1203–1217.

Nagalakshmi,U. *et al*. (2008) The transcriptional landscape of the yeast genome defined by RNA Sequencing. *Science*, **320**, 1344–1349.

Olson,E.N. (2006) Gene regulatory networks in the evolution and development of the heart. *Science*, **313**, 1922–1927.

Paixão,T. and Azevedo,R.B. (2010) Redundancy and the evolution of cis-regulatory element multiplicity. *PLoS Comput. Biol*., **6**, e1000848.

Peisajovich,S.G. *et al*. (2010) Rapid diversification of cell signaling phenotypes by modular domain recombination. *Science*, **328**, 368–372.

Rhead,B. *et al*. (2010) The UCSC genome browser database: update 2010. *Nucleic Acids Res*., **38**, D613.

Shkumatava,A. *et al*. (2009) Coherent but overlapping expression of microRNAs and their targets during vertebrate development. *Genes Dev*., **23**, 466–481.

Stark,A. *et al*. (2005) Animal microRNAs confer robustness to gene expression and have a significant impact on 3′UTR evolution. *Cell*, **123**, 1133–1146.

Tan,C.S. *et al*. (2009) Positive selection of tyrosine loss in metazoan evolution. *Science*, **325**, 1686–1688.

Tsang,J. *et al*. (2007) MicroRNA-mediated feedback and feedforward loops are recurrent network motifs in mammals. *Mol. Cell*, **26**, 753–767.

Valentine,J.W. *et al*. (1994) Morphological complexity increase in metazoans. *Paleobiology*, **20**, 131–142.

Vogel,C. and Chothia,C. (2006) Protein family expansions and biological complexity. *PLoS Comput. Biol*., **2**, e48.

Wagner,A. (2011) The molecular origins of evolutionary innovations. *Trends Genet*., **27**, 397–410.

Wheeler,B.M. *et al*. (2009) The deep evolution of metazoan microRNAs. *Evol. Dev*., **11**, 50–68.

Wilson,D. *et al*. (2008) DBD—taxonomically broad transcription factor predictions: new content and functionality. *Nucleic Acids Res*., **36**, D88–D92.