The animal in the genome: comparative genomics and evolution

Richard R. Copley*

Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford OX3 7BN, UK

Comparisons between completely sequenced metazoan genomes have generally emphasized how similar their encoded protein content is, even when the comparison is between phyla. Given the manifest differences between phyla and, in particular, intuitive notions that some animals are more complex than others, this creates something of a paradox. Simplistic explanations have included arguments such as increased numbers of genes; greater numbers of protein products produced through alternative splicing; increased numbers of regulatory non-coding RNAs and increased complexity of the cis-regulatory code. An obvious value of complete genome sequences lies in their ability to provide us with inventories of such components. I examine progress being made in linking genome content to the pattern of animal evolution, and argue that the gap between genomic and phenotypic complexity can only be understood through the totality of interacting components.

Keywords: comparative genomics; evolution; Metazoa; transcription factors; ultraconserved regions

Deus ex machina: A power, event, person, or thing that comes in the nick of time to solve a difficulty; providential interposition…

Oxford English Dictionary

1. INTRODUCTION

Complete genome sequences provide limits to our imaginations. Even just a few years before the human genome was available in rough draft form, it was widely believed to encode at least 50 000 genes (Fields et al. 1994; Nature Genetics Editorial 2000). In contrast, the initial publications estimated 25–40 000 protein-coding genes (Lander et al. 2001; Venter et al. 2001), and since then estimates have generally carried a downward momentum, most recently approaching 20 000 (Goodstadt & Ponting 2006; Pennisi 2007). Although this number is higher than 16 000 or so found in invertebrate chordates (Dehal et al. 2005), it is roughly the same total as the nematode worm Caenorhabditis elegans (Hillier et al. 2005). Whether or not these low numbers of protein-coding genes for vertebrates stand the test of time, the sense of unease surrounding the lack of correlation between organisinal complexity (often measured in numbers of distinct cell types) and protein-coding gene count is evident from the framing of the ‘G-value paradox’ by Hahn & Wray (2002), and the various explanations that have been put forward to ease it, including, for example, miRNAs (Sempere et al. 2006), non-protein-coding DNA (Taft et al. 2007) and alternative splicing (Kim et al. 2007).

Similar gene counts are, of course, a crude measure of biological complexity. There is no reason why two genomes should not encode very different sets of protein-coding genes, but still have similar overall totals. Within the field of animal evolution and the evolution of development (evo–devo), however, the G-value paradox has a particular resonance. Studies in different animal phyla have repeatedly shown the reuse of a core set of developmental genes, the so-called ‘toolkit’ (Carroll et al. 2005), with the HOX genes in particular taking on an iconic significance. Broadly, toolkit genes come from a handful of transcription factor families, defined by the presence of particular structural domains such as the helix-turn-helix (HTH), including the homeobox genes; zinc fingers (ZnFs); leucine zippers and the helix-loop-helix (HLH). As well as transcription factors, there are seven well-conserved pathways responsible for intercellular signalling (Pires-daSilva & Sommer 2003), many of which appear to be present in sponges, the earliest branching clade of animals (Nichols et al. 2006). An extreme interpretation of these data is provided by Davidson (2006): ‘if we focus explicitly on the genes encoding transcription factors, and [...] signalling systems required for developmental spatial regulation, there is almost no qualitative variation among the genomes of bilaterians’.

Given all this, where in the genome do the phenotypic differences between animal taxa arise? The undoubted conservation of the protein-coding developmental genes has, particularly in the evo–devo field with its morphological concerns, focused attention on cis-enhancer elements affecting transcription (Carroll et al. 2005; Davidson 2006; Simpson 2007; Wray 2007), although there are alternative views emphasizing the importance of different kinds of regulatory elements (Alonso & Wilkins 2005) and different protein classes, such as structural genes (Hoekstra & Coyne 2007). As well as the presence of particular genes, the role of gene loss, especially with regard to secondarily simplified organisms such as tunicates and nematodes, is also likely to be of major significance. Below I outline some major themes being...
developed by large-scale genome comparisons, principally of nematodes, insects and vertebrates. My aim is not to present an exhaustive account, but to highlight areas where functionally relevant species-specific differences may arise, within apparently conserved systems. Although I concentrate on the evolution of the systems regulating animal development, this is not to lose sight of the things being regulated: the proteins involved in making nematode cuticles, or asynchronous flight muscles in insects, or the human brain and adaptive immune system, to name but a few, are what make it necessary to evolve those systems.

2. GENE DUPLICATION

Usefully summarizing the differences and similarities between more than 10,000 protein-coding genes from several species at once is not necessarily straightforward. Although pairwise similarities between sequences are easy to compute, they suffer from the imposition of arbitrary cut-offs and are less easy to interpret than measures that explicitly reflect phylogeny. Genes in different species are most obviously interpreted than measures that explicitly reflect phylogeny. Genes in different species are most obviously compared by grouping into sets of orthologues (i.e. genes related by speciation events) and paralogues (genes related by intra-genome duplication events). Closely related species share large numbers of orthologues: 93% of dog (Canis familiaris) and 82% of the marsupial Monodelphis domesticus gene predictions have orthologues in human (Goodstadt et al. 2007). The Linnean hierarchy, however, is not necessarily a good guide of genomic relatedness by this definition of similarity. Within the nematodes, 65% of C. elegans genes share an orthologue with Caenorhabditis briggsae, despite their being from the same genus (Stein et al. 2003). For more distantly related genomes, orthologue counts can drop rapidly. This may be as much a sign of difficulties in reliably assigning gene orthology on a large scale, as a real indication of the extents of the conserved cores.

Paralogues often arise via tandem duplication of genes, giving rise to localized clusters of functionally related genes. As these are the regions where gene content is evolving most rapidly between closely related species, the functions of these genes are of special interest for understanding animal-specific differences. For the most part, for any two closely related vertebrate genomes, the functional classes of genes duplicated in this way are similar—olfaction and chemosensation, reproduction and effectors of the immune response—although the duplications have occurred independently in each lineage (Emes et al. 2003). These large groups of paralogues often show evidence of adaptive evolution in their amino acid sequences, suggesting that new functions have been selected for (Emes et al. 2004a,b). The recurrent nature of duplications within particular functional classes, coupled with the observed diversifying selection suggests that they are a standard adaptive genomic response to environmental challenges. Does similar rapid duplication occur in the kinds of genes, such as transcription factors, that might be implicated in development? A growing number of examples are known. Perhaps most dramatically, in mice a set of 32 tandemly duplicated homeoboxes have arisen from apparently one or two genes in the common ancestor of humans and rodents; they are believed to play a role in germ cell development and embryonic stem cell differentiation (Maclean et al. 2005; Jackson et al. 2006).

ZnF containing transcription factors have undergone independent rounds of gene duplication in insects and tetrapsods. In insects a set of ZnFs is found to co-occur with a ZnF-associated domain (ZAD; Chung et al. 2007); this ZAD class is found in approximately 100 and 150 copies in Drosophila melanogaster and the mosquito Anopheles gambiae, respectively; there is only a single copy in vertebrates (Chung et al. 2007). In D. melanogaster, many are expressed in the female germ line, suggesting a role in oocyte development or embryogenesis (Chung et al. 2007). An analogous story is found with Krüppel-associated box (KRAB) containing Zn fingers in tetrapsods. Successive independent tandem duplication events have occurred in different mammalian lineages, leading to over 400 copies in the human genome (Huntley et al. 2006). The KRAB domain itself appears to have been co-opted from a progenitor sequence conserved throughout euarkyotes (Birtle & Ponting 2006), however, it has evolved so much as to make this similarity difficult to detect; clearly identifiable KRAB domains are specific to tetrapsods. Their functions are largely unknown, and have not been tied to any general aspects of tetrapsod-specific biology. As such, why the family as a whole has expanded is a puzzle.

Nematodes too exhibit lineage-specific expansions of particular transcription factor families, most notably, the nuclear hormone receptors (NHRs). The C. elegans genome encodes 284, far more than the 48 in human and 21 in D. melanogaster. The bulk of these (greater than 200) have arisen from an apparently nematode-specific expansion of a unique gene (Lander et al. 2001; Robinson-Rechavi et al. 2005). Once more, the reasons for such a dramatic lineage-specific expansion of a particular transcription factor family, and any links to taxon-specific biology, are obscure, although it has been speculated that C. elegans relies less on combinatorial re-use of different transcription factors (Antebi 2006). A less dramatic lineage-specific expansion occurs in the case of the T-box-containing transcription factors: there are 21 in C. elegans, with 17 arising from a lineage-specific expansion when compared with D. melanogaster and humans. Ascertaining when and in which taxa these duplications found in C. elegans took place is currently frustrated by a lack of nematode genome sequences—currently, only those of C. elegans and C. briggsae have been published. These T-box genes as a set map to several genomic locations, suggesting that they have arisen over a more protracted time scale than the examples discussed above; some, at least, have known roles in the development of C. elegans (Poole & Hobert 2006).

3. THE INVENTION OF NEW GENES

A number of gene families appear to be metazoan novelities, with no clear sequence similarity to other genes outside the Metazoa, but present in the more basal animal phyla, such as cnidarians and sponges.
Phil. Trans. R. Soc. B

These include key families involved in animal development, such as T-box and SMAD transcription factors, and signalling molecules such as WNTs and fibroblast growth factors (FGFs; Putnam et al. 2007). Was the invention of such families a prerequisite for the evolution of the Metazoa, and were analogous protein inventions required for the evolution of particular taxa, such as insects and vertebrates? Analysis of three-dimensional structures (i.e. the protein fold itself) suggests a more subtle transition than large-scale evolution of new protein folds. In many cases, examination of protein three-dimensional structural similarities shows that these genes have distant homologues in non-metazoan genomes. The MH1 (DNA binding) domain of SMADs, for instance, is probably homologous to a family of homing endonucleases found in all kingdoms of life (Grishin 2001); the T-box shares structural similarities indicative of homology with a variety of other transcription factors, such as STAT DNA-binding factors, suggesting that they too are likely to be homologous (Murzin et al. 1995). In these cases, the novel families are likely to be cases of rapid sequence evolution, accompanying functional shifts, within stem lineages leading to the Metazoa. Sparse sequence sampling of non-fungal and metazoan eukaryotic genomes may contribute to the apparent co-origin of these protein domains with the animals.

As this type of domain evolution is occurring from pre-existing domain types, the process fits within a standard framework of accelerated point mutation and selection for new functions. The invention of the domain type is not a key innovation in itself rather, it can be seen as the extension of functional diversification of subfamilies of the kind that is apparent when comparing more closely related species. The fact that so many new domain types are found to be coincident within the origin of metazoans suggests that the selective pressures giving rise to this kind of accelerated sequence evolution were greater in the metazoan stem lineage.

An example of a more recent domain innovation is found in the Drosophila gene brinker, which plays a key role in the establishment of dorsal-ventral patterning. Although the protein-coding sequence of its DNA-binding domain is well conserved in insects, using current sequence databases it shows no significant sequence similarity to proteins from any other taxa (figure 1), although there is weak (non-significant) similarity to pogo-like transposases, and the structure, which is only folded when complexed with DNA, suggests similarity to various transcription factors (Cordier et al. 2006).

4. EVOLUTION OF TRANSCRIPTION FACTORS: THE ANIMAL IN THE ORTHOLOGUE

Lineage-specific duplication followed by sequence divergence provides one route to species-specific biology, but what scope is there for lineage-specific functional shifts within orthologous genes? In the absence of gene duplication, it is hard to imagine how the DNA specificity of a particular factor might be significantly changed in such a way that it targets new genes, without deleterious consequences. The modular structure of proteins, however, suggests that other routes of functional evolution are available. A protein may have pleiotropic effects, but that is not the same as saying that every amino acid in the protein will be directly involved in all those effects. A recent illustrative example from the hox gene Ultrabithorax, is of an insect-specific 'QA' protein motif, found outside the homeodomain. The region is involved in limb repression; the effects of deleting the motif are strong in some

![Figure 1](http://www.ensembl.org)
tissues but close to undetectable in others (Hittinger et al. 2005). Clearly, changes in the protein-coding sequences of transcription factors, apart from their more obvious DNA-binding residues, must be integrated into our understanding of the evolution of developmental regulation.

The majority of residues in metazoan transcription factors do not fall within regions of well-defined globular structure, with many belonging to so-called ‘intrinsically disordered’ regions—regions that may form a structure when complexed with other macromolecules (Liu et al. 2006; Minezaki et al. 2006). The specific sequences of these regions are typically not obviously conserved between paralogues; because they are unique to particular families they are not covered in domain databases such as SMART and Pfam (Finnila et al. 2006; Letunic et al. 2006). The lack of extreme conservation between distant species has sometimes masked the fact that within closely related species, these regions are conserved. Comparisons of orthologous sequences from closely related genomes (e.g. vertebrates or drosophilids) often show that substantial proportions of these non-domain sequences are undergoing strong purifying selection—they accumulate many more synonymous nucleotide changes than non-synonymous changes—and are thus functional. For the large part, precisely what these biological functions are is unknown; two possibilities, however, suggest themselves. Firstly, they may have relatively uninteresting non-specific effects, such as facilitating folding of the major domain (e.g. by reducing aggregation) or acting as spacers between globular domains. Secondly, and more interestingly from the point of view of animal evolution, they may include short linear peptide motifs that mediate protein–protein interactions (Dyson & Wright 2005; Nedeva & Russell 2005; Nedeva et al. 2005).

There are numerous examples of regulatory motifs found outside of transcription factor domains. Many hox proteins include a YPWM-like hexapeptide motif that interacts with other homeodomain-containing proteins (In der Rieden et al. 2004); Drosophila ftz orthologues have lost this motif but acquired an LXXLL motif coupled to a new role in segmentation (Lohr & Pick 2005); and an N-terminal SSYF-like motif believed to be involved in transcriptional activation is conserved among Hox orthologues and paralogues from different phyla (Tour et al. 2005). Interaction motifs can be coupled with signalling pathways to create cell-type specificity. They can, for instance, be regulated by phosphorylation, such that the phosphorylation status governs what interactions can be made (e.g. Sapkota et al. 2007), or alternative splicing can result in protein–protein interaction motifs being included or excluded from particular cell types, providing additional layers of regulatory complexity that are likely to be species specific (Nedeva & Russell 2005).

The challenge of identifying small regulatory motifs means that their species distributions, and how their presence might produce taxon-specific differences in protein functions, have not been well studied. Examples that tie cleanly to one taxonomic group are less common, but an interesting case has been proposed in bilaterian orthologues of the Brachury gene. These possess an N-terminal motif that is not found in non-bilaterian Metazoa (Marcellini 2006), which instead have a well-defined EHH1-like motif (Copley 2005). The bilaterian motif is believed to be responsible for an interaction with Smad1, and hence to link gastrulation to bilateral pattern formation (Marcellini 2006).

5. ENHANCERS: TRANSCRIPTION FACTOR-BINDING SITES AND ULTRACONSERVED REGIONS

Theoretical considerations have led to an intense focus on transcription factor-binding sites (TFBSs) as a major molecular source of morphological novelty (Wray et al. 2003; Carroll et al. 2005; Davidson 2006; Wray 2007), although see (Hockstra & Coyne 2007) for a critique. Individual TFBSs show rapid turnover in comparisons of closely related genomes, with many being lineage specific (Dermitzakis & Clark 2002; Moses et al. 2006). This dynamic nature may not be revealed in the phenotype—patterns of gene expression may be conserved even though regulatory sequences change at the molecular level (Ludwig et al. 2000; Romano & Wray 2003; Fisher et al. 2006). On the other hand, the gain and the loss of individual TFBSs have been implicated in several recent cases of morphological evolution, in both vertebrates and invertebrates (reviewed in Simpson (2007) and Wray (2007)). The relationship between individual TFBSs and enhancer function is clearly not straightforward, beyond the fact that clustering of individual binding sites can identify some enhancer regions (Markstein et al. 2002). Cases of functional linkages between particular transcription factors have been proposed, for example, between dorsal, twist, Su(H) and an unidentified motif in neurogenic ectoderm formation in Diptera (Markstein et al. 2004), and even a coupling originating prior to the origin of Bilateria, of hairy and E(spl) promoting neural cell fate (Rebeiz et al. 2005).

Comparisons of vertebrate genomes have revealed large regions (more than 100 nucleotides) of extreme conservation of non-coding sequences (conserved non-coding elements (CNEs); Bejerano et al. 2004). These regions are often found near transcription factors and other developmental genes (Sandelin et al. 2004). Outside of the vertebrates, there is evidence for similar regions occurring near developmental genes in flies (Glazov et al. 2005) and nematodes (Vavouri et al. 2007). Although in many cases the conserved regions are even found near orthologous genes, there is no evidence that they are homologous; they appear to have evolved independently in each of the phyla (Vavouri et al. 2007). Experimental evidence from vertebrates shows that many instances have roles as tissue-specific enhancer elements (Woolfe et al. 2005; Pennacchio et al. 2006).

The length and lack of inter-phylum conservation of CNEs is in contrast to individual TFBSs. The DNA specificity of orthologous transcription factors is usually well conserved over large phylogenetic distances, but typical TFBSs are short, of the order of 6–10 nucleotides. An obvious possibility is that longer CNEs are composed of overlapping or adjacent TFBSs. This would suggest a tight packing of transcription factor
proteins on the genomic DNA of these CNEs. There is direct evidence for this: some fragments of highly conserved non-coding sequences are present in crystal structures of transcription factor complexes. An atomic model based on known crystal structures of the interferon-β enhancer, for example, shows 50 consecutive nucleotides in contact with eight different proteins; these nucleotides are well conserved in mammalian species (Panne et al. 2007; see figure 2 for another example). Given that such structures exist, it is not such a leap to imagine 16 proteins binding to 100 nucleotides, or even bigger complexes. This suggests a model where CNE enhancer regions controlling orthologous genes in different phyla are controlled by multiple TFBSs, although not necessarily the same transcription factors or in the same orientation. Moreover, the tight packing of transcription factors on the genomic DNA suggests that the proteins themselves may be co-adapted to interact with each other and aid the cooperative formation of enhancer complexes. Previously, Ruvinsky & Ruvkun (2003) have presented experimental evidence that enhancers and transcription factors co-evolve in this way, with neuronal and muscle-specific enhancer elements from D. melanogaster failing to drive expression in homologous tissue types in C. elegans, and Dover and co-workers have argued for coevolution of bicoid protein and hunchback regulatory regions (McGregor et al. 2001; Shaw et al. 2002).

If protein–protein interactions between transcription factors are often required for the formation of enhancer complexes, close analysis of transcription factor sequence and structure may reveal evidence for co-adaptation of proteins, such as the Hox hexapeptide motif, through which homeotic proteins form complexes with TALE class homeodomains (LaRonde-LeBlanc & Wolberger 2003). We might expect instances of co-adapted transcription factor combinations to be taxon specific, to match the taxon specificity of enhancer sequences.

6. ALTERNATIVE SPLICING

Not all CNEs are associated with enhancer regions. There is good evidence that many are involved in regulating alternative splicing events, including the
alternative splicing of mRNAs of proteins which themselves regulate alternative splicing (Larea et al. 2007; Ni et al. 2007). The presence of highly conserved control elements to regulate alternative splicing indicates that the functional consequences are of importance. Although large very conserved elements may be the exception rather than the rule, detailed comparative analyses have identified smaller conserved motifs regulating alternative splicing, for instance in nematodes (Kabat et al. 2006) and vertebrates (Sorek & Ast 2003; Yeo et al. 2005).

Alternative splicing is often touted as a mechanism by which proteomic complexity is increased. Although early reports suggested that levels of alternative splicing were comparable in vertebrates and invertebrates (Brett et al. 2002), more recent studies suggest that there is indeed more alternative splicing of transcripts in vertebrates (Kim et al. 2007), suggesting a link with increased phenotypic complexity. How relevant is alternative splicing for species-specific biology and morphological differences? Quantitatively, the gene products that appear to be most affected by alternative splicing are typically involved in nervous and immune system function (Modrek et al. 2001). There are, however, ample examples of alternatively spliced transcription factors—as many as 63% of mouse transcription factors have variant exons (Taneri et al. 2004). Although the differences in molecular roles of the alternatively spliced products are often unknown, the genes themselves include developmental classics such as members of Hox, SMAD and T-box families (Fan et al. 2004; Dunn et al. 2005; Noro et al. 2006), although they do not necessarily present obvious morphological correlates (Yoder & Carroll 2006). Alternative splicing of modular proteins is an obvious route through which functions can be changed, by including or excluding particular combinations of domains. In this regard, it is interesting that alternative splicing often affects intrinsically disordered regions outside known protein domains (Romero et al. 2006)—this again points to a critical role for finely tuned protein–protein interactions among transcriptional regulators.

There are few known cases of distant conservation of alternative splice variants of transcription factors; typically, examples are conserved within phyla at best. Widening the search to other classes of gene again suggests that splice variants are not conserved over long periods, although it should be remembered that transcript coverage of most species from which evidence of alternative splicing is obtained is very restricted. Perhaps the best counter-example is currently that of fibroblast growth factor receptor 2 (FGFFR2), where an exon configuration diagnostic of mutually exclusive alternative splicing is found in both vertebrates and the sea urchin Strongylocentrotus purpuratus (Mistry et al. 2003). Examples of orthologous ion channel encoding genes showing similar alternative splicing patterns in D. melanogaster, C. elegans and humans are likely to be cases of parallel evolution (Copley 2004). The shared ability of vertebrates and at least insects and C. elegans to produce alternative transcripts in a regulated manner, but the absence of large numbers of conserved alternative splicing between protostomes and deuterostomes suggests that gene products have become alternatively spliced in parallel between different lineages, while at the same time hinting that the functions performed by alternative splice variants may, over time, be replaced by different genomic solutions.

7. SUMMARY

Although most major classes of protein involved in animal development may be conserved throughout the Metazoa, detailed comparative analysis of these gene types reveals a more dynamic picture, with frequent gene duplication, gene loss, couplings with new motifs and other processes such as alternative splicing and regulation by micro-RNAs, all of which are likely to be important for a full understanding of function. cis-regulatory variation may well be revealed to be quantitatively the most common form of variation between species, but it seems probable that the cumulative effects of multiple cis-regulatory changes will have required that protein networks evolve to accommodate and correctly regulate changed enhancer structures.

Our knowledge of animal evolution and the picture presented here is currently based on a very small sampling of almost exclusively nematode, insect and vertebrate genomes. Although this situation is beginning to change, the fact that many important functional regions, especially those that do not encode proteins, are only revealed by having sets of closely related genome sequences, and that there are 35 or so animal phyla gives some idea of the enormity of the challenges ahead. The rapidly falling costs of genome sequencing do, however, give grounds for optimism.

I thank the organizers and participants of the Novartis Foundation symposium on Animal Evolution, 20 June 2007, for helpful discussions, two anonymous referees for perceptive comments, and the Wellcome Trust for support.

REFERENCES


Fan, W., Huang, X., Chen, C., Gray, J. & Huang, T. 2004 TBX3 and its isoform TBX3+2a are functionally distinctive in inhibition of senescence and are over-expressed in a subset of breast cancer cell lines. *Cancer Res.* 64, 5132–5139. (doi:10.1158/0008-5472.CAN-04-0615)


Phil. Trans. R. Soc. B


