The evolution of animal genomes
Casey W Dunn¹ and Joseph F Ryan²,³

Genome sequences are now available for hundreds of species sampled across the animal phylogeny, bringing key features of animal genome evolution into sharper focus. The field of animal evolutionary genomics has focused on identifying and classifying the diversity genomic features, reconstructing the history of evolutionary changes in animal genomes, and testing hypotheses about the evolutionary relationships of animals. The grand challenges moving forward are to connect evolutionary changes in genomes with particular evolutionary changes in phenotypes, and to determine which changes are driven by selection. This will require far greater genome sampling both across and within species, extensive phenotype data, a well resolved animal phylogeny, and advances in comparative methods.

Addresses
¹ Department of Ecology and Evolutionary Biology, Brown University, 80 Waterman St., Providence, RI 02906, USA
² Whitney Laboratory for Marine Bioscience, University of Florida, 9505 Ocean Shore Blvd., St Augustine, FL 32080, USA
³ Department of Biology, University of Florida, Gainesville, FL 32611, USA

Corresponding author: Dunn, Casey W (casey.dunn@brown.edu)

Current Opinion in Genetics & Development 2015, 35:25–32
This review comes from a themed issue on Genomes and evolution
Edited by Antonis Rokas and Pamela S Soltis
For a complete overview see the Issue and the Editorial
Available online 9th September 2015
http://dx.doi.org/10.1016/j.gde.2015.08.006
0959-437X/© 2015 Elsevier Ltd. All rights reserved.

The goals of animal evolutionary genomics
Four main goals drive much of the work on animal evolutionary genomics: firstly, to reconstruct the history of animal genome evolution, secondly, to identify evolutionary changes in genomes which contributed to evolutionary changes in phenotypes, thirdly, to understand the evolutionary processes (including selection and neutral drift) that led to genome change, and finally to use the evolution of genomes as a proxy for understanding other historical patterns such as phylogeny and demography. This review examines progress towards these goals, provide examples of biological insights that have already been achieved, and discuss priorities for future work. It focuses on macroevolution of nuclear genomes.

We are still at a very early stage of the genomic revolution, not unlike the state of computing in the late nineties [1]. Genome sequences of various levels of quality are now available for a few hundred animal species thanks to the concerted efforts of large collaborative projects [2–4], more recent efforts by smaller groups [5–7], and even genome projects undertaken largely by independent laboratories [8,9]. At least 212 animal genomes have been published (Figure 1) and many others are publicly available in various stages of assembly and annotation. This sampling is still very biased towards animals with small genomes that can be bred in the laboratory and those from a small set of phylogenetic clades. 83% of published genomes belong to vertebrates and arthropods (Figure 1). Even so, the breadth of sampling is much better than it was a few years ago, with genomes now available for distantly related species that span the root of the animal tree.

All the goals outlined above are advanced by improved taxonomic sampling [10]. Despite progress sequencing a diversity of animal genomes, there is still a long way to go. Genome sequences have been published for less than 0.015% of the roughly 1.5 million described animal species (Figure 1). Many published genome sequences are of poor quality due to the challenges of repetitive DNA, heterozygosity, and contamination [11–13], and some important clades of animals remain entirely unsampled. Improving quality and sampling is a critical and tractable goal, and there are multiple efforts underway to achieve this (e.g., [14*]). As sequencing and analysis methods improve, the limiting steps for gaining biological insight from animal genomes will shift from sequencing expenses and general analysis expertise to the taxon-specific expertise required to collect specimens and interpret results, and the costs associated with obtaining wild animals across the globe. At present, there are 167,802 animal species with molecular sequence data in NCBI (Figure 1), about 11.5% of described species (and about 6% of the animal species thought to live on Earth). These are the species that we know can be collected, identified, and brought back to the lab at current funding levels. To an optimistic first approximation, the taxon sampling of genomes in the next couple decades will be similar to the current sampling of animals for which at least one gene sequence is available.

Reconstructing the history of animal genome evolution
As animal genome sampling has improved, the results have surprised many investigators. It was postulated that animals differ greatly in complexity, and it was expected that gene number would be correlated with these differences [15]. Instead, genome after genome has revealed
that, while genomes were radically transformed along the branch that gave rise to animals, there is far less diversity within animals in terms of gene number and other genome features than expected [5,16–24,25**,26]. The genomes of animals that are sometimes erroneously referred to as ‘higher,’ such as vertebrates, are remarkably similar to those that were characterized as ‘lower’, such as sea anemones and sponges. This has been considered a paradox [27–29], but it says far more about our conceptions of animals and complexity than it does about the biology of the animals themselves. It is based on biases that lead us to greatly underestimate the intricacy of many poorly studied animals [30] and chauvinism that leads us to overestimate vertebrate complexity [28*], and presupposes that there are a small number of common types of genome changes (such as acquisition of new protein...
coding genes) that underlie particular phenotypic changes.

In order to understand the evolutionary history of animal genome diversity, it is helpful to first identify the basic features of the genome of the most recent common ancestor of all animals. By analyzing a broad sampling of genomes from animals and their close relatives in a phylogenetic context, it is possible to infer a minimal set of genes that were present in this organism. The current availability of genomic data from animals and animal outgroups is sparse, but does span the root of the animal tree (since genome sequences are available for both sponges and eukaryotes, this holds regardless of which is the sister group to all other animals [16,17,31–34]). Based on the gene inventories described in [17], the most recent common ancestor of all animals had at least 6289 gene families that are also shared with outgroups and 2141 gene families that appear to be animal-specific. This number will be revised upward as taxon sampling improves, and may be revised downward if improved analysis methods detect previously overlooked homology. Though taxon sampling is still minimal, some interesting qualitative patterns have emerged regarding the composition of the genome of the most recent common ancestor of all animals (Table 1). For example, while many cell adhesion genes in animals are shared with outgroups, many cell signalling genes are animal-specific [19**].

The genome of the most recent common animal ancestor gave rise to the genomes of all living animals through many types of evolutionary changes. These include whole genome duplication events, changes in synteny [35*], gene family evolution (through the origin, duplication, and loss of genes) [36*], coding sequence evolution (through site insertion, deletion and substitution), modifications of regulatory regions, horizontal gene transfer [11], expansion and loss of transposable elements, and modification of local repeats (including satellite DNA, simple sequence repeats, and tandem repeats). The genomic features affected by these processes include introns [37], centromeres, telomeres, sex chromosomes, transposable elements, sequence-dependent variation in chromatin-organization, gene regulatory elements, as well as microRNAs and other noncoding RNAs.

Perhaps the most conspicuous of genome changes are changes in genome size [38]. Despite the relative stability in protein-coding gene content, animal genome sizes are highly variable and these changes are thought to be largely driven by nonadaptive processes [39]. In at least some clades, genome size correlates with cell size, cell division rate, resting metabolic rate, average adult size and therefore is potentially impacted by selection [40].

While gene inventory is more conserved across animals than had been expected, there are many genes that are found only in particular animal subclades [41]. These taxon-restricted ‘orphan’ genes may be formed de novo from noncoding sequence or alternatively may be highly derived genes that no longer show a signature of homology at the sequence level [42,43]. At least some of these genes have the ability to quickly become essential [44], but it is not yet clear if there are general patterns in the association of the origin of novel genes and phenotype evolution [41,45].

Connecting genome evolution to phenotype evolution

Not all changes in genomes result in changes in phenotypes — there is a ‘many-to-one’ mapping between genomes and phenotypes [46]. This has sweeping implications for some of the most basic questions in evolutionary genomics — What genomic differences across the animal phylogeny underlie the diversity of animal phenotypes? What is the phenotypic impact of particular genomic differences? What phenotypes are evolutionarily accessible? Because there are so many genomic differences and so many of them have no impact on phenotypes it is extremely difficult to make causative associations between the two. The most common strategy is to extrapolate experimental results from laboratory model organisms to other species of interest (i.e., the candidate gene approach), but this can introduce considerable bias since it is largely blind to novel features in nonmodel species [30]. Another approach is to collect enough comparisons between genotypes and phenotypes to make statistically significant associations between particular evolutionary changes in the genome and particular evolutionary changes in the phenotype [47*], whether at the population or macroevolutionary scale. This requires genomic data from a large number of organisms that are variable in the phenotypic trait of interest, and many independent evolutionary changes in the trait of interest. Even so, it has already proven tractable for specific traits in clades such as mammals [48] where taxon sampling is already quite good and there is much experimental evidence regarding gene function in some species. This approach will be more powerful as genome and phenotype sampling improve. Wider adoption of these approaches will also hinge on developing analysis methods that can account for between species variation as well as within speciation variation [49], and approaches that can extend phylogenetic comparative methods to high dimensional data [50]. The incorporation of functional genomic data sampled across a broad diversity of species, including gene expression data [50,51], will also greatly assist in understanding the phenotypic implications of genomic changes.

Identifying the evolutionary processes that shape genome evolution

Not all changes in phenotype result in a change in fitness — there is a ‘many-to-one’ mapping between phenotypes
and fitness. Because genotype to phenotype mapping is also many-to-one (as discussed above), it is extraordinarily difficult to understand whether differences in genome sequences result in differences in phenotypes that in turn result in fitness differences that can be acted on by natural selection. While there has been a historical bias to invoke adaptationist arguments for genome variation, others have powerfully argued that most genome variation is neutral (i.e., has no impact on fitness) [52].

<table>
<thead>
<tr>
<th>Clade</th>
<th>Gene</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metazoa</td>
<td>LIM classes (ABLIM, Enigma, LHX, LMO7, Mical)</td>
<td>[65]</td>
</tr>
<tr>
<td>Metazoa</td>
<td>HD: ANTP, PRD, POU, LIM, SINE classes</td>
<td>[17]</td>
</tr>
<tr>
<td>Metazoa</td>
<td>Sox (B, C, E, F)</td>
<td>[66]</td>
</tr>
<tr>
<td>Metazoa</td>
<td>Class I Fox TFs: D, G, L2</td>
<td>[36*]</td>
</tr>
<tr>
<td>Metazoa</td>
<td>bHLH: Group A (E12/E47, ProtoAtonal), E.</td>
<td>[36*]</td>
</tr>
<tr>
<td>Metazoa</td>
<td>bZIP: Jun, Fox, XBP1, MAF, Nfe2</td>
<td>[36*]</td>
</tr>
<tr>
<td>Metazoa</td>
<td>Ets, Smad, NR, Interferon Regulatory Factor, MADF, AP-2, Doublesex</td>
<td>[36*]</td>
</tr>
<tr>
<td>Metazoa</td>
<td>T-box: Eomes, Tbx2/3, Tbx8, TbxPor, Tbx1/15/20</td>
<td>[36*]</td>
</tr>
<tr>
<td>Metazoa</td>
<td>Wnt</td>
<td>[67]</td>
</tr>
<tr>
<td>Metazoa</td>
<td>LGR</td>
<td>[68]</td>
</tr>
<tr>
<td>Metazoa</td>
<td>Shaker K+ channels (Shaker)</td>
<td>[69]</td>
</tr>
<tr>
<td>Metazoa</td>
<td>EAG K+ channels</td>
<td>[70]</td>
</tr>
<tr>
<td>Porifera + Parahoxozoa</td>
<td>Immune genes: IRF, MyD88, TLR2</td>
<td>[71]</td>
</tr>
<tr>
<td>Porifera + Parahoxozoa</td>
<td>Axin</td>
<td>[67]</td>
</tr>
<tr>
<td>Porifera + Parahoxozoa</td>
<td>iGluR: (AMPA)</td>
<td>[71]</td>
</tr>
<tr>
<td>Porifera + Parahoxozoa</td>
<td>ADAR1</td>
<td>[72]</td>
</tr>
<tr>
<td>Porifera + Parahoxozoa</td>
<td>Nuclear receptors with zinc fingers</td>
<td>[73]</td>
</tr>
<tr>
<td>Porifera + Parahoxozoa</td>
<td>homeodomain + paired domain</td>
<td>[17]</td>
</tr>
<tr>
<td>Porifera + Parahoxozoa</td>
<td>Drosha, Pasha, microRNAs</td>
<td>[74]</td>
</tr>
<tr>
<td>Parahoxozoa</td>
<td>HD: HNF, PRD (SS0 and K50), Hox</td>
<td>[17]</td>
</tr>
<tr>
<td>Parahoxozoa</td>
<td>Perlecan/HSPG2</td>
<td>[75]</td>
</tr>
<tr>
<td>Parahoxozoa</td>
<td>LIM: LMO class</td>
<td>[65]</td>
</tr>
<tr>
<td>Parahoxozoa</td>
<td>RHD: NFAT and Rel</td>
<td>[36*]</td>
</tr>
<tr>
<td>Parahoxozoa</td>
<td>GCM</td>
<td>[36*]</td>
</tr>
<tr>
<td>Parahoxozoa</td>
<td>Tbx6</td>
<td>[36*]</td>
</tr>
<tr>
<td>Parahoxozoa</td>
<td>Class I Fox TFs: A, B, C, E, Q1, Q2</td>
<td>[36*]</td>
</tr>
<tr>
<td>Parahoxozoa</td>
<td>bHLH: Group A (Achaete Scute, Paraxis, Twist, Net, Mesp), Group C (Bmal, Hairy), Group D</td>
<td>[36*]</td>
</tr>
<tr>
<td>Parahoxozoa</td>
<td>TGF-beta: Noggin, Follistatin</td>
<td>[76]</td>
</tr>
<tr>
<td>Parahoxozoa</td>
<td>Contactin, Neurexin</td>
<td>[77]</td>
</tr>
<tr>
<td>Parahoxozoa</td>
<td>BMP-antagonist</td>
<td>[68]</td>
</tr>
<tr>
<td>Parahoxozoa</td>
<td>Shaker K+ channels (Shab, Shaw)</td>
<td>[69]</td>
</tr>
<tr>
<td>Parahoxozoa</td>
<td>EAG K+ channels (Eag, Elk,Erg)</td>
<td>[70]</td>
</tr>
<tr>
<td>Parahoxozoa</td>
<td>Gustatory receptor-like</td>
<td>[78]</td>
</tr>
<tr>
<td>Planuloozoa</td>
<td>Wnt: Wnt2, Wnt3, Wnt4, Wnt5, Wnt7, Wnt11, Wnt16, DKK</td>
<td>[67]</td>
</tr>
<tr>
<td>Planuloozoa</td>
<td>Nuclear receptors: NR2F, NR1/4, NR2B, NR2C/D/E</td>
<td>[73]</td>
</tr>
<tr>
<td>Planuloozoa</td>
<td>TGF-beta: Chordin</td>
<td>[76]</td>
</tr>
<tr>
<td>Planuloozoa</td>
<td>Erbin, CASK, Neurologin</td>
<td>[17]</td>
</tr>
<tr>
<td>Planuloozoa</td>
<td>mir-100</td>
<td>[74]</td>
</tr>
<tr>
<td>Planuloozoa</td>
<td>PDGF</td>
<td>[68]</td>
</tr>
<tr>
<td>Planuloozoa</td>
<td>Shaker K+ channels (Shal)</td>
<td>[69]</td>
</tr>
<tr>
<td>Planuloozoa</td>
<td>Cyclooxygenase</td>
<td>[79]</td>
</tr>
<tr>
<td>Bilateria</td>
<td>bZIP: BACH, B-ATF, Atf3</td>
<td>[36*]</td>
</tr>
<tr>
<td>Bilateria</td>
<td>ClassI Fox TFs: F, H,I,L1</td>
<td>[36*]</td>
</tr>
<tr>
<td>Bilateria</td>
<td>HD: CUT, PROS, ZF</td>
<td>[36*]</td>
</tr>
<tr>
<td>Bilateria</td>
<td>HMGbox: Sox-like genes</td>
<td>[66]</td>
</tr>
<tr>
<td>Bilateria</td>
<td>bHLH: Group A (MyoD, NeuroD, Neurogenin, MyoR), Group B (SRC)</td>
<td>[36*]</td>
</tr>
<tr>
<td>Bilateria</td>
<td>Stargazin</td>
<td>[17]</td>
</tr>
</tbody>
</table>

There are several ways to investigate whether genome variation is the result of neutral evolution or has been shaped by natural selection. First, one can measure the impact of genotype variation on phenotype variation and phenotype variation on fitness. This is exceptionally difficult, though, as most of these steps are difficult and very labor intensive. Second, one can search directly for signatures of selection in the genome, such as selective sweeps, phylogenetic footprints, or deviations in the
nonsynonymous mutation rate [53], without investigating the phenotypes that mediated the selection [54].

**Genome evolution as a proxy**

Many evolutionary biologists who collect and analyze genome data are not primarily motivated by the desire to study genome evolution — they seek to use the evolutionary history of the genomes as a proxy to understand other aspects of evolution. Population geneticists use genomes to infer demographic events in the history of populations, such as migration and ancestral population sizes [55]. Phylogenetic systematics now uses genome-scale data to infer the relationships between species.

The first animal phylogenies based entirely on genome data [56] supported relationships, including the Coelomata hypothesis, that were later shown with the addition of more data and improved analysis methods to be artefacts [33,57]. This reinforces how important broad taxonomic sampling is for interpreting genome data. Even if an investigator is interested in taxonomically restricted questions, such as genome evolution in a particular clade, there is great value in considering those questions in the context of genome data from additional species. Most phylogenetic analyses based on genome data now also include transcriptome sampling [33,57], which can be a less expensive way to expand taxon sampling. The vast majority of studies that use genomic data to investigate phylogenetic relationships consider only the molecular sequence evolution of protein coding genes. There is also considerable interest and great promise in tapping other types of genomic data, such as gene gain and loss and non-coding sequences, to understand species relationships [58].

**Conclusions**

The real power in animal evolutionary genomics will not be in what we learn about genomes themselves, but how we connect this information to phenotypic diversity and the historical processes that led to the diversity of genomes we see today. As genome sequences become available from a greater number of species, phylogenetic comparative methods [49,59] will become increasingly important for interpreting the biological implications of genome diversity and leveraging the diversity of genomes across species to understand genome function within species. They will first need to be retooled to work effectively with high dimensional data, though. There are many phenotype changes and many more genome changes along each branch of the animal tree. Just because a genome change is evolutionarily coincident with a particular phenotype change does not mean it is causative. It could have no phenotypic impact at all, or be related to one of the many other phenotypic changes along the same phylogenetic branch. Unfortunately, such spurious associations are still a mainstay of animal genome papers that purport to identify genes specific to phenotypic novelties [60,61]. One of the most common problems with such conclusions is the failure to account for multiple independent tests.

The concept of ‘a genome’ for a given animal species is an invention of convenience — there is no such thing. Species are sets of individuals with many different genomes. Genomes can even vary between cells within individuals due to programmed genome rearrangement [62] and somatic mutation. As the field matures, many evolutionary questions will require sampling genomes at all these scales (multiple cells from multiple individuals of multiple species).

The field of evolutionary animal genomics needs to stop looking for relationships between vague notions of genome complexity and vague notions of phenotype complexity, and instead focus on understanding the evolution of genes and phenotypes on a trait-by-trait basis. There is no single axis of organism complexity, each animal has a mix of traits that could be characterized as simple or complex. In addition, our sense of differences in complexity across animals is inflated by biases that lead us to focus on complex traits in well-studied animals and miss complex novel features in poorly studied animals [30]. There is also reason to question the expected differences in genomes. Because there is a many-to-one mapping between genomes, phenotypes, and fitness, there are many different changes in genomes that can lead to equivalent phenotype changes. Instead of seeking a small number of genome evolutionary patterns that explain sweeping patterns in animal diversity, such as the gain of novel protein coding genes or micro-RNAs, we should be prepared to identify and explain very different genotype evolution stories that underlie every phenotypic change.

**Acknowledgements**

Thanks to Felipe Zapata and Stefan Siebert for providing feedback on this manuscript. This work was supported in part by the National Science Foundation Alan T Waterman Award to CWD.

**References and recommended reading**

Papers of particular interest, published within the period of review, have been highlighted as:

- of special interest
- of outstanding interest


This review highlights regions of the animal tree that are undersampled and provides detailed justification for generating genomic sequences from animals within these lineages.


This genome paper greatly clarified the understanding of genome evolution in Holozoa (the clade that includes animals) and provided important insight into the changes in genome composition that preceded the radiation of animals.


This is the first animal genome to be sequenced outside of Bilateria. It revealed that many genome features were much older than previously thought, and that many of the differences between vertebrate and ecdysozoan genomes were due to reductions in ecdysozoa, not expansions in vertebrates.


A concise summary of the paradox, or lack thereof, between the number of protein coding genes and morphological complexity.


A detailed recent review on comparative animal genomics.


A detailed review of many papers that seek to find the phylogenetic distribution of animal transcription factors. The work provides insight into evolutionary forces that shaped the repertoire of transcription factors in animals, how gene regulatory networks influenced the evolution of transcription factors, and what forces led to the extraordinary regulatory capabilities of animal transcription factors.


47. Wray GA: Genomics and the evolution of phenotypic traits. • Annu Rev Ecol Evol Syst 2013, 44:51-72. Review shows how genomic approaches are being applied to long-standing issues in evolutionary genetics and how these approaches are expanding the field.


77. Ganot P, Zoccola D, Tambutté E, Voolstra CR, Aranda M, Allemand D, Tambutté S: Structural molecular components of
