Zoologica Scripta

Invited Review

Comparative genomics and the diversity of life

Casey W. Dunn & Catriona Munro

Submitted: 3 May 2016
Accepted: 11 August 2016
doi:10.1111/zsc.12211

Dunn, C. W. & Munro, C. (2016). Comparative genomics and the diversity of life. — *Zoologica Scripta*, *45*, 5–13.

In the last decade, genomics has come to play a central role in systematics and biodiversity research. In coming years, systematics and phylogenetics will come to play an increasingly important role in genomics. Here, we address the false dichotomy between descriptive- and hypothesis-driven work, discuss the power of descriptive genomics to test questions of broad interest and explore the applications and challenges that arise as comparative genomic analyses come to include more species. Integrated phylogenetic analyses of genome sequences and organism phenotypes across many species will provide a powerful window on genome function that can be used to answer many questions that to date were only tractable in laboratory model systems. Many challenges will arise as the numbers of species in genomic analyses grow by orders of magnitude. In particular, our current nomenclatural systems for describing gene homology (orthology, paralogy and related terms) are breaking down, and the current focus on 'strict orthologs' in many comparative genome analyses will need to be replaced by more holistic approaches that better accommodate gene duplication and loss.

Corresponding author: *Casey W. Dunn, Department of Ecology and Evolutionary Biology, Brown University, 80 Waterman St, Providence, Rhode Island, USA. E-mail: casey_dunn@brown.edu*

Casey W. Dunn, and Catriona Munro, Department of Ecology and Evolutionary Biology, Brown University, 80 Waterman St, Providence, RI, USA. E-mails: casey_dunn@brown.edu, catriona_munro@brown.edu

Introduction

Genomic analyses have recently advanced some of the most important themes of systematics research, including the phylogenetic relationships between species, the understanding of novel phenotypes and adaptive processes in natural populations (Lamichhaney *et al.* 2015; Brawand *et al.* 2014). Systematics and phylogenetics have also influenced genomic work in critical ways, and this impact will be much greater as genome sequences become available for many more individuals across a much greater breadth of species. Initially, the primary influence of systematics on genomics was to inform which genomes to sequence to optimize taxon sampling for particular questions (GIGA Community of Scientists 2014). This has been critical for genomics, but is only the tip of the iceberg of interdisciplinary work to come. In particular, phylogenetic methods provide a natural framework for comparing genomes while explicitly considering the evolutionary processes that produced the observed diversity. This allows for clear articulation of hypotheses about genome evolution and diversity, for example by describing which genome changes occurred along particular branches in the tree of life. Phylogenetic comparative methods also provide a robust approach for testing general hypotheses about genome function, for example by testing predictions that particular evolutionary changes in genome sequence are associated with specific phenotypic changes.

To date, genome biology has been principally focused on model species that can be grown in the laboratory and humans. As genomic approaches are applied to the other >99% of life on Earth, new challenges are faced. Many of the experimental tools presently used to study genome function in the laboratory cannot be applied in the wild, primarily because they require that the study organisms can be cultivated for multiple generations. Variation in wild populations poses technical challenges, but also important new opportunities because it presents standing variation in traits of great interest that can be used to make links between these traits and genome features.

Descriptive biology can both generate and test hypotheses

Most of the genomic data that will be obtained from wild specimens will be descriptive, as experimental manipulation is very difficult in most of these organisms. It is important that we think clearly about what descriptive projects enable. In biology, the term 'descriptive' is often used as a pejorative for studies that do not include experimental manipulations by the investigator and therefore are perceived to not be hypothesis driven (Fig. 1A). But manipulative experiments are not the only way to test hypotheses, and whether a project is experimental or descriptive is unrelated to whether it is hypothesis driven (Fig. 1B). Some of the best descriptive and experimental projects both test and generate further hypotheses.

Descriptive data are among the most powerful resources we have for testing critical hypotheses about the natural world. Many scientific fields, such as astronomy, are based almost entirely on descriptive data. There is broad consensus that Earth goes around the sun and the Universe is expanding, but neither of these hypotheses have been tested through experimental manipulation of the study systems. Initial observations led astronomers to propose these hypotheses, along with others that also explained preliminary data. These hypotheses led to specific predictions that differed from predictions of other hypotheses, and these were further tested with additional descriptive data. It is odd that biologists readily accept hypotheses that have been tested only with descriptive data in other fields, some of them among the greatest successes in science, but downplay the value of descriptive work in Biology.

Experimental approaches are well suited for inducing variation that does not exist in nature, or for controlling the background that the variation exists on. Manipulative experiments have tremendous value for hypothesis testing, and they are also often used to perturb systems in ways that help generate hypotheses and provide information about systems outside of the strict goals of testing particular hypotheses about how systems work. Examples of using experiments to generate hypotheses rather than test-specific hypothesis include, for example, many mutagenic manipulations in forward-genetic screens, or classically, Galvani's application of electricity to frog muscle (Beutler *et al.* 2007; Galvani & Aldini 1792).

Descriptive approaches are especially valuable for testing hypotheses in systems that are at spatial and temporal scales that are not amenable to experimental manipulations, such as astronomy and macroevolution. Descriptive work is also fundamental to generating targeted hypotheses. Additionally, an understanding of the variation in the undisturbed system is critical for interpreting the results of experimental manipulation. Many of the laboratory experimental systems that are widely used today were enabled by foundational descriptive work generations ago that is now largely forgotten and underappreciated. As technical advances enable us to ask new questions in new systems, it is critical to make an initial investment in foundational descriptive work. This will help us formulate the most productive questions and hypotheses, and develop the most effective approaches to answering and testing them.

Tools for identifying associations between genes and phenotypes

One of the central questions in genomics is, which genes influence which phenotypes? Identifying links between genotypes and phenotypes is difficult to do, and must take biological diversity into account to establish how general or specific each link is that is identified in each species. Much of what we have learned in the best studied canonical model organisms applies to a broad diversity of organisms; however, many details are specific to these species and do not describe the biology of other organisms (Bolker 2012).

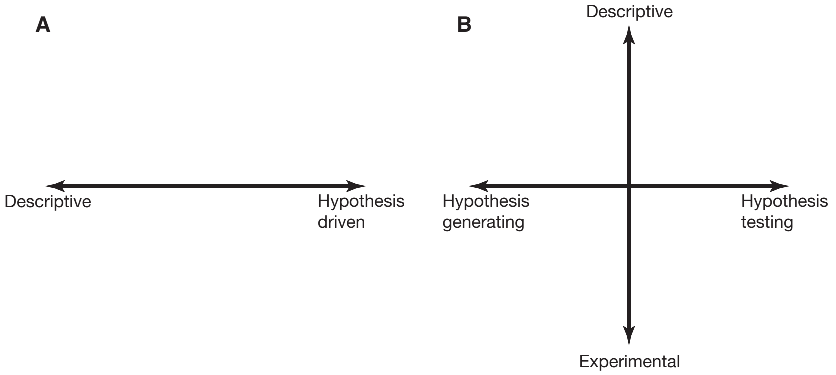

Fig. 1 A. The false dichotomy between descriptive- and hypothesis-driven research, which places these two features at opposite ends of a single spectrum. –B. The degree to which research is experimental (relies on variation induced by the investigator) or descriptive (relies on variation that exists in nature) is orthogonal to the characterization of work as hypothesis generating or hypothesis testing.

This simple fact about the organisms becomes a problem for the study of biology when 'model organisms' are thought of has models of other organisms rather than as organisms with excellent tools for developing models of certain biological processes (Katz 2016). Many biologists are interested in understanding the evolutionary processes that give rise to diversity, and identifying the functional links between genotypes and phenotypes across a wide diversity of species will enable us to identify not only shared traits across clades, but also unique traits and functions (Dunn *et al.* 2015).

For the past hundred years, genetic techniques including mutagenic screens have been highly successful at identifying a number of genes involved in biological pathways that affect phenotypic traits in canonical model organisms (e.g. Winzeler *et al.* 1999). Classic genetic crosses are also a powerful way to survey genomes and identify genetic regions that influence phenotypic changes at the individual, population and species level. These approaches, however, are best applied in species where inbred recombinant lines are possible. Without crosses or the ability to follow pedigrees in wild populations, linkage maps cannot be generated, making quantitative trait locus (QTL) mapping and classic genomewide association studies (GWAS) difficult. Beyond classic methods, new advances in transgenic and genome editing technologies are closing the genotype–phenotype gap in a greater diversity of organisms (Ikmi *et al.* 2014; Perry & Henry 2015). However, these methods also cannot be applied to the vast majority of wild organisms. Fortunately, there are a growing number of tools that can identify associations between genotypes and phenotypes in wild populations.

Population-level approaches

There are multiple well-developed tools for detecting selection in genomic data from wild populations (Akey 2009; Vitti *et al.* 2013; Wray 2013). These can identify specific genes that are associated with selection on particular phenotypes. The implications of these associations can then be further tested by additional observations (e.g. using immunohistochemistry or *in situ* mRNA hybridization), and if possible, through targeted experimental manipulation. Most of these selection-based tools include well-established statistical tests for deviations from neutrality, including gene-based, linkage-disequilibrium-based, and population-differentiation-based models (Vitti *et al.* 2013). Classic gene-based methods include scanning for positive selection through the comparison of non-synonymous (dN) to synonymous (dS) nucleotide substitution rates in protein coding genes to determine regions that may have been under recent selection (McDonald & Kreitman 1991; Stark *et al.* 2007; Yang & Bielawski 2000). A large suite of tests identify regions of strong linkage disequilibrium (LD), which may be indicative of an incomplete selective sweep and positive selection (Hohenlohe *et al.* 2012; Tishkoff *et al.* 2007; Vitti *et al.* 2013). Where two or more populations are sequenced, the most commonly used measure of genetic differentiation between groups is Wright's fixation index (FST) (Lewontin & Krakauer 1973). FST scans and related methods may be used to identify allele frequency variation across the genome to identify outliers between groups with different phenotypic traits (e.g. Jones *et al.* 2012; Shapiro *et al.* 2013).

Trait association studies, where the co-occurrence of a phenotypic trait is found to be statistically associated with one or more loci, may in some cases be used to link traits to phenotypes (approaches include QTL and eQTL mapping, and GWAS). The application of these tools is impeded in wild organisms without an available linkage map, although it may be possible to take advantage of the genome and linkage map of a closely related species (e.g. Dawson *et al.* 2006; Stinchcombe & Hoekstra 2008).

Could phylogenetics be the new genetics?

Phylogenetic comparative approaches are increasingly being applied to answer some of the same questions that have been addressed with classical population genetic approaches (Felsenstein 1988; Hiller *et al.* 2012; Pease *et al.* 2016). In clades with broad taxon sampling of genomes, like mammals, it is already tractable to make associations between phenotype and genotype using a phylogenetic framework to detect molecular convergence (Hiller *et al.* 2012). Phylogenetic methods may also be applied detect selection associated with environmental conditions in phylogenetic- and genomewide association studies or 'PhyloG-WAS' (Pease *et al.* 2016). This approach is limited to situations where clades are not confounded with the environmental condition being measured.

Phylogenetic comparative approaches offer another window that may be used to combine information across species to identify new taxonomically restricted candidate loci that may have an effect on phenotype. Extending phylogenetic methods and tools to multidimensional datasets will further improve the ability to link sequenced genomes to potential phenotypes, for example, by incorporating RNA-seq gene expression data from different cells, tissues and species, and analysing them within a phylogenetic context (Arendt 2008; Brawand *et al.* 2011; Dunn *et al.* 2013a,b; Roux *et al.* 2015).

Sequence homology

As genome sequences become available for a much broader diversity of organisms, and as we move towards incorporating comparative phylogenetic methods, some of the ways that we currently describe patterns and processes of genome evolution will break down. This language worked well for smaller projects, but will not for larger analyses of complex gene trees that incorporate a greater number of species. This gap is particularly problematic for descriptions of gene homology.

The concept of sequence homology is central to the study of genome evolution. Sequences are homologous if they are derived from a shared ancestral sequence (Pearson 2013). Homology is not a statement about sequence similarity or functional similarity, although the term is often misapplied to describe similarity in the literature (as discussed by Gabaldon & Koonin 2013). Homology is a hypothesis about evolutionary history (Wagner 2014). Homologous genes can have very similar sequences, or very different sequences. A set of homologous genes can have very similar functions, or radically different functions. In practice, similarity is used to infer homology under explicit statistical analyses that evaluate the probability that similarity between two sequences is due to shared ancestry as opposed to chance resemblance (Pearson 2013). Similarity is used as a means of inferring homology; however, similarity is not equivalent to homology.

The evolution of homologous sequences is influenced by multiple processes, including speciation, gene duplication events that result in multiple homologous gene copies within the same genome, gene loss and molecular evolutionary processes that change gene sequences. Gene phylogenies are powerful tools for describing the evolution of homologous sequences. The tips of the tree are the homologous gene sequences under consideration, and the root of the tree is their most recent common ancestor. Each internal node in the tree represents a divergence that is due to speciation or duplication.

Many questions in evolutionary genomics require precise language to describe how homologous sequences are related to each other. The most widely used nomenclature annotates each of the tips of the gene tree as orthologs or paralogs (Fitch 1970). Orthologs are sequences whose divergence is due to speciation, and paralogs are those that diverged due to gene duplication events. This terminology can be applied unambiguously when discussing pairs of sequences. The nomenclature also works reasonably well when expanding beyond two sequences when one process dominates over the other. In the extreme cases, for example, it can unambiguously describe strict orthologs sampled one each from multiple species or strict paralogs all sampled from a single species. Problems quickly arise, however, when describing homologs that have a history of both duplication and speciation. These problems are exacerbated as the number of species considered grows. In large complex gene trees, like those regularly encountered in current analyses, the path through the phylogeny between two gene sequences will often include multiple speciation and duplication events. Labelling the tips of the gene trees as orthologs or paralogs cannot fully describe these more complex histories. There have been attempts to address these challenges by expanding the language used to describe genes beyond orthologs and paralogs to include terms such as in-paralogs, out-paralogs and co-orthologs that attempt to capture mixed histories (Sonnhammer & Koonin 2002). Fundamentally, these new terms still have many of the same limitations as the original terms. They cannot fully describe all patterns in gene homology, they often depend on particular reference points in the tree and different evolutionary histories can lead to the same patterns.

The fundamental problem is that these nomenclature systems are attempting to describe attributes of internal nodes in the gene tree (Which internal nodes are speciation events and which are duplication events?) with labels that are applied to the tips of the tree (Which tips are orthologs, which are paralogs and which are variations of the two?). While these problems are minimal for some smaller trees with simple histories, they become far worse for the more complex gene trees that are frequently encountered as analyses include a broader diversity of genomes. Rather than expand the tip-based nomenclature system to refine what we mean by ortholog and paralog, we should instead focus on describing the internal nodes of the gene trees as speciation or duplication events. This is more direct, explicit and clear.

An undue focus on strict orthologs

Many comparative analyses of genomes focus on ‘strict orthologs’, also referred to as ‘single-copy genes’. In these studies, gene families that show evidence of duplication events are actively avoided. This focus is reflected in the many ortholog databases that are available (Nakaya *et al.* 2013) and the frequent reference to the ‘paralogy problem’ in the literature. There are a few reasons for this focus on strict orthologs.

- It is easier to talk about the evolutionary history of strict orthologs, which largely reflect the history of speciation, than it is to consider gene families with many paralogs, where one may need to invoke multiple duplications and losses as well. The focus on orthologs is therefore often imposed as a way to technically simplify analyses.
- Strict orthologs are often presumed to be less prone to molecular evolution processes that could confound topic of interest than are gene families that have many copies. This could reflect, for example, concern that

duplication can modify or relax selection on gene copies through degeneration and complementation (Force *et al.* 1999).
- Some questions, like the phylogenetic analyses of species relationships, are primarily concerned with speciation events in gene trees. Gene families with evidence of duplications are often discarded to focus on speciation events in gene trees.
- The ortholog conjecture (Nehrt *et al.* 2011) is the hypothesis that orthology is a good predictor of conserved function. It is implicitly taken for granted in many analyses and discussions of genome evolution. This expectation of conserved ortholog function is used to apply information on gene function from well-studied organisms to orthologous sequences of poorly studied organisms.

There are, however, problems with each of these points that call into question the motivation for focusing on strict orthologs. Strict orthology is not necessarily an indicator of simpler evolutionary history or processes. Instead, complex histories and processes are hidden by strict orthology because selection restores these genes to single copy after duplication. Just as there has been a conflation of mutation rate (the frequency of genetic changes between parents and offspring) and substitution rates (the rate at which genetic changes become fixed in the population), there is currently a conflation of the rate at which duplicates originate in offspring and the rate at which duplicates become fixed. The duplicate fixation rate is determined by the duplicate origin rate as well as the duplicate loss rate. There is little reason to expect that different genes have different duplicate origin rates. Large-scale patterns in duplicate fixation rate are therefore likely driven in large part by differences in the rate at which duplicates are lost. A growing body of evidence suggests that many more genes occur in single copy than would be expected by chance and that this pattern is driven by selection against duplicates after they arise (De Smet *et al.* 2013). In particular, there are theoretical expectations and now empirical evidence that genes that are usually found in single copy have a higher duplicate loss rate because these genes are prone to dominant negative mutations (De Smet *et al.* 2013). In such genes, a deleterious mutation reduces fitness even in the presence of functional wild-type copies, and duplicates provide more opportunity for such a mutation to arise.

This has important practical implications. It means that we should not think of gene families that tend to have duplicates as outliers with an elevated rate of duplicate origin. Instead, we should think of genes that tend to occur in single copy as having elevated rates of duplicate loss. Focusing on orthologs does not avoid a history of duplication; it hides the duplication that occurred. Efforts to simplify studies of genome evolution by identifying and investigating only strict orthologs may introduce strong biases in many of the patterns and processes that are under study, due in part to uniquely strong selection for reversion to single copy. These biases could be exacerbated by, and mechanistically related to, the lower rates of molecular evolution and higher average expression that are observed among genes with low duplicate fixation rates (De Smet *et al.* 2013; Gout *et al.* 2010). Together, these factors suggest that focusing on strict orthologs can discard many genes with more diverse properties that could be highly relevant to the questions at hand. It could, for example, exclude more rapidly evolving genes that would be highly relevant to recovering difficult to resolve relationships between closely related species.

Many studies, such as phylogenetic analyses of species relationships, are principally interested in studying speciation events in gene trees. Such studies often attempt to isolate speciation events by selecting genes with low duplication fixation rates, *that is* gene families that consist of one member per species. But if differences in the rate of duplicate fixation are driven largely by differences in the rate of duplicate loss, these genes have not been duplicated less – it is just that their history of duplication is quickly erased and is no longer available to the investigator. It is not necessarily the case that every node in a gene tree with one gene sequence per species represents a speciation event. The persistence of some duplicates across speciation events before they are lost could lead to gene tree - species tree incongruence that can mislead the inference of species relationships, just as incomplete lineage sorting of alleles does (Maddison 1997).

Recent analyses suggest that orthology is not necessarily a good predictor of gene function, undermining one of the primary reasons for focusing on orthologs to the exclusion of paralogs. In the limited cases where it has been evaluated, support for the ortholog conjecture has been poor to mixed (Gabaldon & Koonin 2013; Nehrt *et al.* 2011). There are also very interesting and frequent exceptions to the converse conjecture that the same function is performed by orthologous genes in different species (Gabaldon & Koonin 2013; Omelchenko *et al.* 2010). This indicates that orthology may be no better a predictor than homology alone for understanding conserved function. It could be that the evolutionary distance between two genes on a gene tree is alone a good predictor of functional differences, regardless of whether the path on the tree between these sequences transverses speciation and duplicate fixation events or speciation events alone.

All of these issues suggest that the focus on orthologs may have less benefit than is often supposed and can introduce its own problems. Rather than focus on identifying strict orthologs and discarding gene families that have fixed duplicates, evolutionary genomic analyses should take a more holistic approach and broaden their focus to homologs of all sorts. The challenge is that it is then necessary to annotate each node in the gene tree as a speciation or duplication event. Fortunately, the methods and tools for testing these historical hypotheses about speciation and duplication events are rapidly improving.

Identifying speciation and duplication events in gene trees

To better take advantage of and understand the evolution of gene families, it is critical to have tools for inferring which nodes in gene trees are speciation events and which are duplication events. The identification of homologous sequences relies on tools that identify an excess of sequence similarity that suggests shared ancestry (Pearson 2013). Once homologs have been identified, there are two general approaches to identifying speciation and duplication events. The first approach is to use sequence similarity for this step as well. Tools including OMA (Altenhoff *et al.* 2013) and OrthoMCL (Li *et al.* 2003) rely on pairwise comparisons between sequences to identify subsets of sequences with no more than one sequence per species that tend to be more similar to each other than to other homologous sequences. The goal is to isolate subsets of sequences that arose by speciation alone, and not duplication. These similarity-based methods do not attempt to model historical processes, but instead use *ad hoc* criteria to partition genes into putative ortholog sets. These methods are fast, but there is growing concern that they do not perform well. Even in simple cases, pairwise comparisons of similarity have been shown to be poor predictors of orthology (Smith & Pease 2016; Yang & Smith 2014).

The second general approach to identifying speciation and duplication events is to explicitly account for them in a phylogenetic context. There is a growing set of tools that does this. The simplest do not annotate every internal node in the phylogeny. Instead, like OMA, OrthoMCL and related tools, they attempt to identify subtrees of orthologs. The key difference is that they are based on the topology of the gene phylogenies, rather than on sequence similarity. These approaches first build phylogenies of homologous sequences and then identify subtrees in the gene tree that have no more than one sequence per taxon (Ballesteros & Hormiga 2016; Dunn *et al.* 2013a,b; Hejnol *et al.* 2009; Kocot *et al.* 2013; Yang & Smith 2014). These methods differ primarily in the way these subtrees are pruned and filtered. These methods are fast and do not require a species tree ahead of time (in part because they do not attempt to reconcile the subtrees to the species trees). This is particularly advantageous when the species tree is the main goal of the study.

Other tools first infer gene trees and species trees and then reconcile the two by invoking historical hypotheses of speciation, gene duplication and gene loss (Chen *et al.* 2000; Górecki & Eulenstein 2014). This has the advantage of being fast and also providing more detail on the history of gene duplication, but requires having a well resolved species tree ahead of time. Independently inferring gene tree topologies and then reconciling them to a shared species tree has its limitations, however. Poorly supported branches in the gene trees will tend to be incongruent with the species tree, resulting in the inference of a large excess of duplications and losses to reconcile the gene tree topology to the species tree.

The most promising approaches to identifying speciation and duplication nodes in gene trees, and also the most computationally expensive, simultaneously estimate the topologies of the gene trees, the topology of the species trees and the history of gene duplication and loss (Boussau *et al.* 2013; Martins *et al.* 2014; Szollosi *et al.* 2015). These methods do not require a species tree in advance, better account for differences in support across gene trees and species trees and better accommodate uncertainty in previous steps of the analysis (Guang *et al.* 2016). This is an exciting area of methods development that will address many different analysis needs in a single, biologically relevant, explicit framework.

Conclusions

Genomic tools are remarkably complementary to other perspectives, including morphology, functional biology, development and biogeography, and are helping to unify previously independent research programmes in these areas. Now that genome data are less expensive to collect than data on many other organism attributes, genomes will be increasingly useful as a first look at organism biology that helps guide other types of observations. One of the greatest values of genomics for existing research priorities may be to make more informed decisions about how we collect other types of more expensive data. For example, genomes will help us understand which morphological data are most relevant to particular questions. This will drive a resurgence in descriptive biology, both because the genomic data are so interesting and because they will guide the acquisition of other categories of data.

As we move towards the broad application of comparative phylogenetic genomic methods, we need to change the way we talk about central concepts including sequence homology. Annotating genes as orthologs or paralogs is becoming more unwieldy and less informative as gene trees become more complex and better sampled in larger analyses. Instead, the field should focus more on explicit histories of gene evolution, such as gene phylogenies in which internal nodes are annotated according to inferred historical events (including duplication, speciation and loss). This will also help move away from an undue focus on single-copy strict orthologs in comparative genomic analyses. This focus on strict orthologs is often presented as a simplifying step to avoid complexities and potential biases resulting from gene duplication, but it instead may introduce ascertainment biases due to strong selection for these genes to return to single copy.

Acknowledgements

The authors wish to thank Felipe Zapata for reviewing a draft of this manuscript.

References

Akey, J. M. (2009). Constructing genomic maps of positive selection in humans: where do we go from here? *Genome Research*, *19*, 711–722.

Altenhoff, A. M., Gil, M., Gonnet, G. H. & Dessimoz, C. (2013). Inferring Hierarchical Orthologous Groups from Orthologous Gene Pairs. *PLoS ONE*, *8*, e53786.

Arendt, D. (2008). The evolution of cell types in animals: emerging principles from molecular studies. *Nature Reviews Genetics*, *9*, 868–882.

Ballesteros, J. A. & Hormiga, G. (2016). A new orthology assessment method for phylogenomic data: unrooted Phylogenetic Orthology. *Molecular Biology and Evolution*, *33*, 2117–2134.

Beutler, B., Du, X. & Xia, Y. (2007). Precis on forward genetics in mice. *Nature immunology*, *8*, 659–664.

Bolker, J. (2012). Model organisms: there's more to life than rats and flies. *Nature*, *491*, 31–33.

Boussau, B., Szollosi, G. J., Duret, L., Gouy, M., Tannier, E. & Daubin, V. (2013). Genome-scale coestimation of species and gene trees. *Genome Research*, *23*, 323–330.

Brawand, D., Soumillon, M., Necsulea, A., Julien, P., Csárdi, G., Harrigan, P., Weier, M., Liechti, A., Aximu-Petri, A., Kircher, M., Albert, F.W., Zeller, U., Khaitovich, P., Grützner, F., Bergmann, S., Nielsen, R., Pääbo, S. & Kaessmann, H. (2011). The evolution of gene expression levels in mammalian organs. *Nature*, *478*, 343–348.

Brawand, D., Wagner, C. E., Li, Y. I., Malinsky, M., Keller, I., Fan, S., Simakov, O., Ng, A. Y., Lim, Z. W. I, Bezault, E., Turner-Maier, J., Johnson, J., Alcazar, R., Noh, H. J., Russell, P., Aken, B., Alfoldi, J., Amemiya, C., Azzouzi, N., Baroiller, J. - F., Barloy-Hubler, F., Berlin, A., Bloomquist, R., Carleton, K. L., Conte, M. A., D'Cotta, H., Eshel, O., Gaffney, L., Galibert, F., Gante, H. F., Gnerre, S., Greuter, L., Guyon, R., Haddad, N. S., Haerty, W., Harris, R. M., Hofmann, H. A., Hourlier, T., Hulata, G., Jaffe, D. B., Lara, M., Lee, A. P., MacCallum, I., Mwaiko, S., Nikaido, M., Nishihara, H., Ozouf-Costaz, C., Penman, D. J., Przybylski, D., Rakotomanga, M., Renn, S. C. P., Ribeiro, F. J., Ron, M., Salzburger, W., Sanchez-Pulido, L., Santos, M. E., Searle, S., Sharpe, T., Swofford, R., Tan, F. J., Williams, L., Young, S., Yin, S., Okada, N., Kocher, T. D., Miska, E. A., Lander, E. S., Venkatesh, B., Fernald, R. D., Meyer, A., Ponting, C. P., Streelman, J. T. D. D., Lindblad-Toh, K., Seehausen, O. & Di Palma, F. (2014). The genomic substrate for adaptive radiation in African cichlid fish. *Nature*, *513*, 375–381. Available via http://doi.org/10.1038/nature13726

Chen, K., Durand, D. & Farach-Colton, M. (2000). NOTUNG: a program for dating gene duplications and optimizing gene family trees. *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology*, 7, 429–447. Available via http://doi.org/10.1089/106652700750050871

Dawson, D. A., Burke, T., Hansson, B., Pandhal, J., Hale, M. C., Hinten, G. N. & Slate, J. (2006). A predicted microsatellite map of the passerine genome based on chicken–passerine sequence similarity. *Molecular Ecology*, *15*, 1299–1320.

De Smet, R., Adams, K. L., Vandepoele, K., Van Montagu, M. C. E., Maere, S. & Van de Peer, Y. (2013). Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants. *Proceedings of the National Academy of Sciences*, *110*, 2898–2903. Available via http://doi.org/10.1073/pnas.1300127110

Dunn, C. W., Howison, M. & Zapata, F. (2013a). Agalma: an automated phylogenomics workflow. *BMC Bioinformatics*, *14*, 330. Available via http://doi.org/10.1186/1471-2105-14-330

Dunn, C. W., Luo, X. & Wu, Z. (2013b). Phylogenetic analysis of gene expression. *Integrative and Comparative Biology*, *53*, 847–856. ict068.

Dunn, C. W., Leys, S. P. & Haddock, S. H. D. (2015). The hidden biology of sponges and ctenophores. *Trends in Ecology and Evolution*, *30*, 282–291. Available via http://doi.org/10.1016/j.tree.2015.03.003

Felsenstein, J. (1988). Phylogenies and quantitative characters. *Annual Review of Ecology and Systematics.*, *19*, 445–471. Available via http://doi.org/10.2307/2097162

Fitch, W. M. (1970). Distinguishing homologous from analogous proteins. *Systematic Zoology*, *19*, 99–113. Available viahttp://doi.org/10.2307/1563209?ref=search-gateway:a3316b54a0b014c6a41-be406b82fb7ce

Force, A., Lynch, M., Pickett, F. B., Amores, A., Yan, Y. L. & Postlethwait, J. (1999). Preservation of duplicate genes by complementary, degenerative mutations. *Genetics*, *151*, 1531–1545. Available via http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=10101175

Gabaldon, T. & Koonin, E. V. (2013). Functional and evolutionary implications of gene orthology. *Nature Reviews Genetics*, *14*, 360–366. Available via http://doi.org/10.1038/nrg3456

Galvani, L. & Aldini, G.(1792). *De Viribus Electricitatis In Motu Musculari Comentarius Cum Joannis Aldini Dissertatione Et Notis; Accesserunt Epistolae ad animalis Electricitatis Theoriam Pertinentes*. Mutinæ: Apud Societatem Typographicam.

GIGA Community of Scientists. (2014). The Global Invertebrate Genomics Alliance (GIGA): developing community resources to study diverse invertebrate genomes. *Journal of Heredity*, *105*, 1–18. Available via http://doi.org/10.1093/jhered/est084

Górecki, P. & Eulenstein, O. (2014). DrML: probabilistic modeling of gene duplications. *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology*, *21*, 89–98. Available via http://doi.org/10.1089/cmb.2013.0078

Gout, J.-F., Kahn, D. & Duret, L., & Paramecium Post-Genomics Consortium. (2010). The relationship among gene expression, the evolution of gene dosage, and the rate of protein evolution. *PLoS Genetics*, *6*, e1000944. Available via http://doi.org/10.1371/journal.pgen.1000944

Guang, A., Zapata, F., Howison, M., Lawrence, C. E. & Dunn, C. W. (2016). An integrated perspective on phylogenetic workflows. *Trends in Ecology & Evolution*, *31*, 116–126.

Hejnol, A., Obst, M., Stamatakis, A., Ott, M., Rouse, G. W., Edgecombe, G. D., Martinez, P., Baguñà, J., Bailly, X., Jondelius, U., Wiens, M., Müller, W. E. G., Seaver, E., Wheeler, W. C., Martindale, M. Q., Giribet, G. & Dunn, C. W. (2009). Assessing the root of bilaterian animals with scalable phylogenomic methods. *Proceedings. Biological Sciences/The Royal Society*, *276*, 4261–4270. Available via http://doi.org/10.1098/rspb.2009.0896

Hiller, M., Schaar, B. T., Indjeian, V. B., Kingsley, D. M., Hagey, L. R. & Bejerano, G. (2012). A "forward genomics" approach links genotype to phenotype using independent phenotypic losses among related species. *Cell Reports*, *2*, 817–823.

Hohenlohe, P. A., Bassham, S., Currey, M. & Cresko, W. A. (2012). Extensive linkage disequilibrium and parallel adaptive divergence across threespine stickleback genomes. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *367*, 395–408.

Ikmi, A., McKinney, S. A., Delventhal, K. M. & Gibson, M. C. (2014). TALEN and CRISPR/Cas9-mediated genome editing in the early-branching metazoan nematostella vectensis. *Nature Communications*, *5*, 5486.

Jones, F. C., Grabherr, M. G., Chan, Y. F., Russell, P., Mauceli, E., Johnson, J., Swofford, R., Pirun, M., Zody, M. C., White, S., Birney, E., Searle, S., Schmutz, J., Grimwood, J., Dickson, M. C., Myers, R. M., Miller, C. T., Summers, B. R., Knecht, A. K., Brady, S. D., Zhang, H., Pollen, A. A., Howes, T., Amemiya, C. Broad Institute genome sequencing platform & whole genome assembly team, Lander, E. S., Di Palma, F., Lindblad-Toh, K. & Kingsley, D. M. (2012). The genomic basis of adaptive evolution in threespine sticklebacks. *Nature*, *484*, 55–61.

Katz, P. S. (2016). Model organisms' in the light of evolution. *Current Biology*, *26*, R649–R650.

Kocot, K. M., Moroz, L., Citarella, M. & Halanych, K. (2013). PhyloTreePruner: a Phylogenetic Tree-Based Approach for Selection of Orthologous Sequences for Phylogenomics. *Evolutionary Bioinformatics*, *9*, 429–435. Available via http://doi.org/10.4137/EBO.S12813

Lamichhaney, S., Berglund, J., Almén, M. S., Maqbool, K., Grabherr, M., Martinez-Barrio, A., Promerová, M., Rubin, C.-J., Wang, C., Zamani, N., Grant, B. R., Grant, P. R., Webster, M. T. & Andersson, L. (2015). Evolution of Darwin's finches and their beaks revealed by genome sequencing. *Nature*, *518*, 371–375. Available via http://doi.org/10.1038/nature14181

Lewontin, R. & Krakauer, J. (1973). Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics*, *74*, 175–195.

Li, L., Stoeckert, C. & Roos, D. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Research*, *13*, 2178–2189. Available via http://doi.org/10.1101/gr.1224503

Maddison, W. P. (1997). Gene Trees in Species Trees. *Systematic Biology*, *46*, 523–536. Available via http://doi.org/10.1093/sysbio/46.3.523

Martins, L. D. O., Mallo, D. & Posada, D. (2014). A Bayesian Supertree Model for Genome-Wide Species Tree Reconstruction. *Systematic Biology*, *65*, 397–416. syu082. Available via http://doi.org/10.1093/sysbio/syu082

McDonald, J. H. & Kreitman, M., & others. (1991). Adaptive protein evolution at the adh locus in *Drosophila*. *Nature*, *351*, 652–654.

Nakaya, A., Katayama, T., Itoh, M., Hiranuka, K., Kawashima, S., Moriya, Y., Okuda, S., Tanaka, M., Tokimatsu, T., Yamanishi, Y., Yoshizawa, A. C., Kanehisa, M. & Goto, S. (2013). KEGG OC: a large-scale automatic construction of taxonomy-based ortholog clusters. *Nucleic Acids Research*, *41* (Database issue), 353–357. Available via http://doi.org/10.1093/nar/gks1239

Nehrt, N. L., Clark, W. T., Radivojac, P. & Hahn, M. W.(2011). Testing the Ortholog Conjecture with Comparative Functional Genomic Data from Mammals. *PLoS Computational Biology*, 7, e1002073. Available via http://doi.org/10.1371/journal.pcbi.1002073

Omelchenko, M. V., Galperin, M. Y., Wolf, Y. I. & Koonin, E. V. (2010). Non-homologous isofunctional enzymes: a systematic analysis of alternative solutions in enzyme evolution. *Biology Direct*, *5*, 31–20. Available via http://doi.org/10.1186/1745-6150-5-31

Pearson, W. R. (2013). An introduction to sequence similarity ("homology") searching. *Current Protocols in Bioinformatics/Editorial Board, Andreas D. Baxevanis. [et Al.]*, *43*, 3.1.1–3.1.8. *Chapter 3*, Unit 3.1. Available via http://doi.org/10.1002/0471250953.bi0301s42

Pease, J. B., Haak, D. C., Hahn, M. W. & Moyle, L. C. (2016). Phylogenomics reveals three sources of adaptive variation during a rapid radiation. *PLoS Biology*, *14*, e1002379.

Perry, K. J. & Henry, J. Q. (2015). CRISPR/Cas9-mediated genome modification in the mollusc, crepidula fornicata. *Genesis*, *53*, 237–244.

Roux, J., Rosikiewicz, M. & Robinson-Rechavi, M. (2015). What to compare and how: comparative transcriptomics for evo-devo. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*, *324*, 372–382.

Shapiro, M. D., Kronenberg, Z., Li, C., Domyan, E. T., Pan, H., Campbell, M., Tan, H., Huff, C. D., Hu, H., Vickrey, A. I., Nielsen, S. C. A., Stringham, S. A., Hu, H., Willerslev, E., Gilbert, M. T. P., Yandell, M., Zhang, G. & Wang, J. (2013). Genomic diversity and evolution of the head crest in the rock pigeon. *Science*, *339*, 1063–1067. Available via http://doi.org/10.1126/science.1230422

Smith, S. A. & Pease, J. B. (2016). Heterogeneous molecular processes among the causes of how sequence similarity scores can fail to recapitulate phylogeny. *Briefings in Bioinformatics*, *2016*, 1–7. bbw034. Available via http://doi.org/10.1093/bib/bbw034

Sonnhammer, E. L. L. & Koonin, E. V. (2002). Orthology, paralogy and proposed classification for paralog subtypes. *Trends in Genetics: TIG*, *18*, 619–620. Available via http://doi.org/10.1038/nbt749

Stark, A., Lin, M. F., Kheradpour, P., Pedersen, J. S., Parts, L., Carlson, J. W., Crosby, M. A., Rasmussen, M. D., Roy, S., Deoras, A. N., Ruby, J. G., Brennecke, J., Hodges, E., Hinrichs, A. S., Caspi, A., Paten, B., Park, S.-W., Han, M. V., Maeder, M.

L., Polansky, B. J., Robson, B. E., Aerts, S., van Helden, J., Hassan, B., Gilbert, D. G., Eastman, D. A., Rice, M., Weir, M., Hahn, M. W., Park, Y., Dewey, C. N., Pachter, L., Kent, W. J., Haussler, D., Lai, E. C., Bartel, D. P., Hannon, G. J., Kaufman, T. C., Eisen, M. B., Clark, A. G., Smith, D., Celniker, S. E., Gelbart, W. M. & Kellis, M. (2007). Discovery of functional elements in 12 *Drosophila* genomes using evolutionary signatures. *Nature*, *450*, 219–232. Available via http://doi.org/10.1038/nature06340

Stinchcombe, J. & Hoekstra, H. (2008). Combining population genomics and quantitative genetics: finding the genes underlying ecologically important traits. *Heredity*, *100*, 158–170.

Szollosi, G. J., Tannier, E., Daubin, V. & Boussau, B. (2015). The inference of gene trees with species trees. *Systematic Biology*, *64*, e42–e62. Available via http://doi.org/10.1093/sysbio/syu048

Tishkoff, S. A., Reed, F. A., Ranciaro, A., Voight, B. F., Babbitt, C. C. & Silverman, J. S., … others. (2007). Convergent adaptation of human lactase persistence in Africa and Europe. *Nature Genetics*, *39*, 31–40.

Vitti, J. J., Grossman, S. R. & Sabeti, P. C. (2013). Detecting natural selection in genomic data. *Annual Review of Genetics*, *47*, 97–120.

Wagner, G. P. (2014). Homology, Genes, and Evolutionary Innovation. Princeton, NJ: Princeton University Press. Available via http://doi.org/10.2307/j.ctt6wpzfz

Winzeler, E. A., Shoemaker, D. D., Astromoff, A., Liang, H., Anderson, K., Andre, B., Bangham, R., Benito, R., Boeke, J. D., Bussey, H., Chu, A. M., Connelly, C., Davis, K., Dietrich, F., Dow, S. W., El Bakkoury, M., Foury, F., Friend, S. H., Gentalen, E., Giaever, G., Hegemann, J. H., Jones, T., Laub, M., Liao, H., Liebundguth, N., Lockhart, D. J., Lucau-Danila, A., Lussier, M., M'Rabet, N., Menard, P., Mittmann, M., Pai, C., Rebischung, C., Revuelta, J. L., Riles, L., Roberts, C. J., Ross-MacDonald, P., Scherens, B., Snyder, M., Sookhai-Mahadeo, S., Storms, R. K., Véronneau, S., Voet, M., Volckaert, G., Ward, T. R., Wysocki, R., Yen, G. S., Yu, K., Zimmermann, K., Philippsen, P., Johnston, M. & Davis, R. W. (1999). Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science*, *285*, 901–906. Available via http://doi.org/10.1126/science.285.5429.901

Wray, G. A. (2013). Genomics and the Evolution of Phenotypic Traits. *Annual Review of Ecology, Evolution, and Systematics*, *44*, 51–72. Available via http://doi.org/10.1146/annurev-ecolsys-110512-135828

Yang, Z. & Bielawski, J. P. (2000). Statistical methods for detecting molecular adaptation. *Trends in Ecology & Evolution*, *15*, 496–503.

Yang, Y. & Smith, S. A. (2014). Orthology inference in non-model organisms using transcriptomes and low-coverage genomes: improving accuracy and matrix occupancy for phylogenomics. *Molecular Biology and Evolution*, *31*, 245–3092. Available via http://doi.org/10.1093/molbev/msu245