Zootaxa 3754 (2): 195–200
www.mapress.com/zootaxa/

Article

ISSN 1175-5326 (print edition)
ZOOTAXA
ISSN 1175-5334 (online edition)

http://dx.doi.org/10.11646/zootaxa.3754.2.8
http://zoobank.org/urn:lsid:zoobank.org:pub:08FD1323-600E-457E-805D-4CB0BA16C1E7

Correcting the disconnect between phylogenetics and biodiversity informatics

JOSEPH T. MILLER & GARRY JOLLEY-ROGERS
Centre for Australian National Biodiversity Research, CSIRO Plant Industry, Canberra, ACT 2601 Australia.
E-mail: joe@acaciamulga.net, garry.jolleyrogers@gmail.com

Abstract

Rich collections of biodiversity data are now synthesized in publically available databases and phylogenetic knowledge now provides a sound understanding of the origin of organisms and their place in the tree of life. However, these knowledge bases are poorly linked, leading to underutilization or worse, an incorrect understanding of biodiversity because there is poor evolutionary context. We address this problem by integrating biodiversity information aggregated from many sources onto phylogenetic trees. PhyloJIVE connects biodiversity and phylogeny knowledge bases by providing an integrated evolutionary view of biodiversity data which in turn can improve biodiversity research and the conservation decision making process. Biodiversity science must assert the centrality of evolution to provide effective data to counteract global change and biodiversity loss.

Key words: Taxonomy, phylogeny, distribution, biodiversity, database

The biodiversity research community has divergent agendas. One side, the biological collection informatics community, is driven by large-scale collections-based projects, exemplified by the *Global Biodiversity Information Facility* (GBIF), *Encyclopedia of Life* (EOL), *Integrated Digitized Biocollections* (iDigBio), *Atlas of Living Australia* (ALA) (Beaman *et al.* 2007; Matsunaga *et al*. 2013) and online floras and faunas that describe biodiversity by species, taxonomies and localities. For example, a full description of the authoritative taxonomy, distribution and ecology, including images and literature references, is available at EOL for *Chelonoidis nigra* Abingdonii, the Galapagos giant tortoise, and tens of thousands of other species. The other side, the evolution and phylogenetics community, driven largely by the National Science Foundation and academia, is building the tree of life (Open Tree of Life 2013; APGIII 2009; NSF 2013), describing biodiversity through clades, diversifications and distributions. For example, the Angiosperm Phylogeny Group provides a synthesis of the evolutionary relationships (APGIII 2009) of the angiosperms and the NSF Assembling the Tree of Life (ATOL) initiative (Choumane *et al.* 2000) seeks to construct an evolutionary history for all major lineages of life.

However, biodiversity is neither a list of taxonomic entities with morphological and geographic attributes nor solely a phylogeny with divergence dates. Comparative biology is predicated on the expectation that closely related organisms share traits—such as morphology, ecology, biogeography, disease resistance, ecosystem services—that are not common in more distantly related organisms. The rich biodiversity data available through GBIF is underutilized because it is not integrated with the tree of life: only by applying the analytical power that comes from a phylogeny can we understand the relationships of biodiversity data (Mishler 2010).

While much work remains to accomplish the individual goals of each community, a synthesis is badly needed that will make integrated biodiversity data the foundation for scientifically based biodiversity-management decisions (van der Linde & Houle 2008; Varón *et al.* 2010; Parr *et al.* 2012; Jonathan *et al.* 2013). As Cracraft (2002) noted *"the ability to search multiple databases using the nodes of a phylogenetic tree may be the single most important contribution of systematics to conservation and sustainable use of biodiversity"*. Viewed through phylogeny, the evolutionary context of morphological, spatial and ecological data becomes clear and accessible to scientists and non-scientists alike.

The global biodiversity crisis, exacerbated by climate change, necessitates that science be better integrated so that decisions are made with the best available data provided by all communities. However, taxonomy and

Accepted by M. Ebach: 5 Dec. 2013; published: 14 Jan. 2014

phylogeny are not always aligned, and phylogenetic analyses have identified numerous examples of non-monophyletic taxa (e.g. Miller & Bayer 2001; Fukami *et al.* 2004; Mast & Thiele 2007). This decouples the evolutionary history from the distributions, morphology and species interactions in the biodiversity databases, thereby lessening the applied value of our knowledge.

Decisions made on the basis of data not integrated with phylogeny are potentially harmful to conservation. The disconnect is clearly exemplified for biota on the World Heritage-listed Great Barrier Reef (GBR). Endemism of corals at the generic level is different between the Atlantic (low) and Indo-Pacific (high) regions, and endemism is used to describe diversity in the Great Barrier Reef (SEWPAC 2013). However, in 2004, a phylogenetic study of corals indicated that many coral genera shared between the two regions are non-monophyletic, rendering these metrics uninformative (Fukami *et al.* 2004). Unfortunately, this phylogenetic knowledge is not reflected in the classifications used by the biodiversity collection informatics community, which often uses polyphyletic generic names. Distribution maps from the ALA and EOL data suggest that natural monophyletic lineages occur with broad distributions in both the Atlantic and Indo-Pacific regions (Fig 1). However, it has been known to science for nine years that the genera are not monophyletic and therefore assumptions of biogeography, trait evolution and potential adaption to climate change are invalid.

Figure 1.

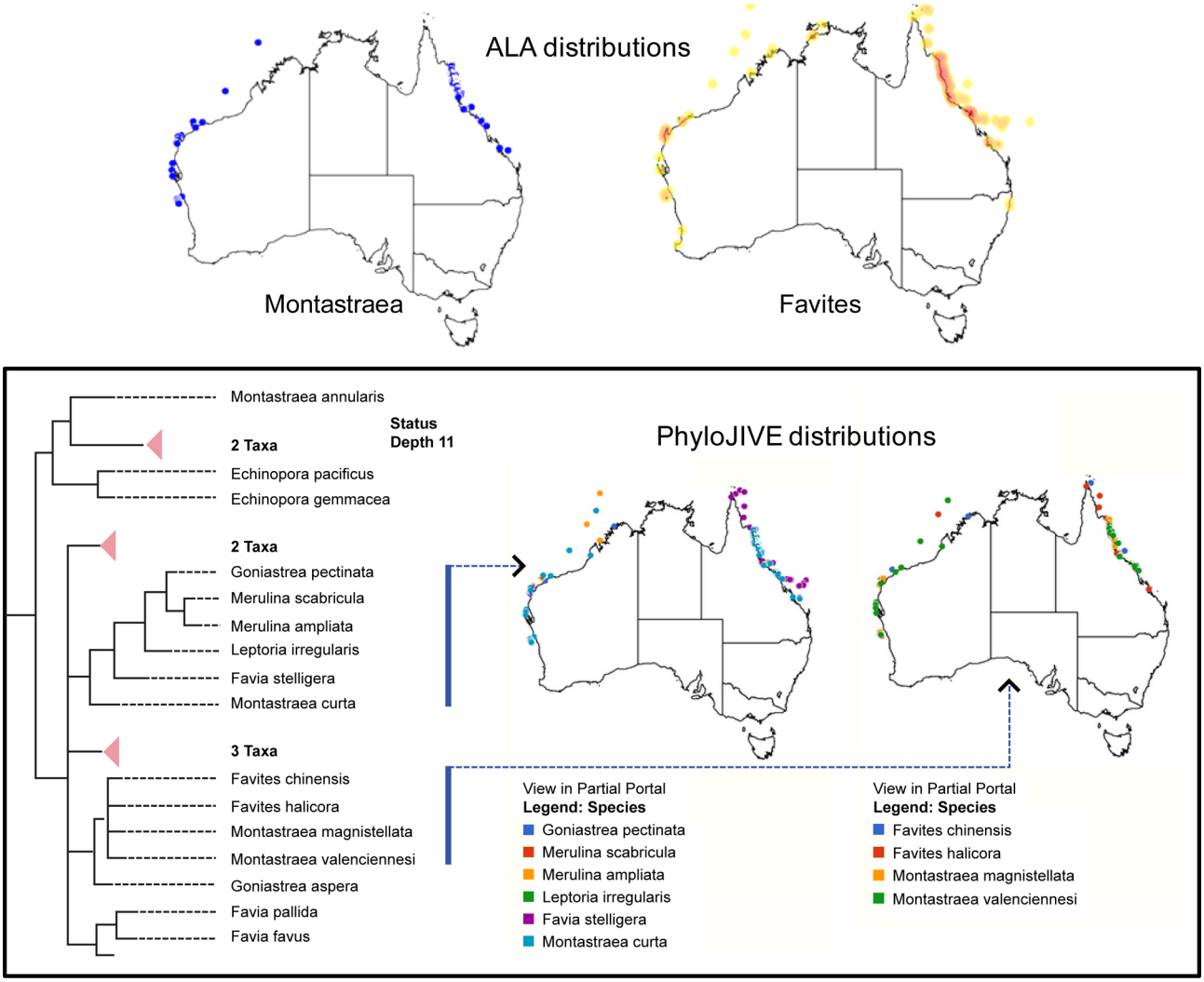

FIGURE 1. Top. Distribution of two coral genera (*Montastraea* and *Favites*) in the waters off the northern Australian coast with data drawn from the Atlas of Living Australia. Bottom. Distributions of two clades of corals as inferred from a phylogeny of the corals (Fukami *et al.* 2004). The left panel is the tree in which two separate clades are mapped on the right. The maps are drawn from individual species distributions from the Atlas of Living Australia.[i] Distribution of six coral species including *Montaserea curta* (light blue), and (ii) Distribution of four species of coral including two species of *Montastrea*. The distributions of *M. curta* and *M. valenciennesi* are broadly sympatric, but the genus is non-monophlytic, which suggests evolutionary convergence.

Despite the great quantity of taxonomic, phylogenetic and biogeographic knowledge of corals, the complete picture, which includes pervasive morphological convergence (Fukami *et al.* 2004), is not presented to non-expert viewers of coral biodiversity, like these authors, or more importantly, to conservation planners. Indeed, heritage assessments of the Great Barrier Reef (SEWPAC 2013) refer to the high level of endemic taxa in the Great Barrier Reef without mentioning the phylogenetic context to support the uniqueness that is implied by endemism.

To meet the need to link phylogenies to other biodiversity information, we have developed PhyloJIVE (Phylogeny Javascript Information Visualiser and Explorer), a web-based application that integrates biodiversity information aggregated from many sources with phylogenetic trees (Jolley-Rogers *et al.* Under Revision). To demonstrate its utility, we constructed a phylogenetic tree of corals based on published data (Fukami *et al.* 2004) and uploaded it to the instance of PhyloJIVE at the ALA. The viewer immediately sees that the genus *Montastraea* de Blainville is not monophyletic, that is, all species of this genus do not a have a common most recent ancestor, and viewers can interrogate any terminal or node for a taxon distribution map (Fig 1). These maps are directly sourced live via web services from the ALA georeferenced database. By investigating a node containing *Montastraea valenciennesi* Milne Edwards & Haime and *Favites halicora* Ehrenberg, a distribution map appears simultaneously mapping all four species in the clade. A separate clade contains another *Montastraea* species, which indicates the complex nature of the distributions of these corals. The user quickly views the non-monophyly of *Montastraea* and its relationships with other coral genera, and the convergent evolution is clear. Viewed through PhyloJIVE, conservation planners can identify the evolutionary context of the endemic corals. Since the diversity is more complicated, due to convergence, than that presented by the ALA, decisions on GBR conservation could be based on a finer geographic scale of phylogenetic diversity.

The growing catalogue of online biodiversity data sources has reached a point that integration can inform research, and evolutionary perspectives can improve the quality of biodiversity data. There are approximately 2.5 billion biological specimens in the world's natural history collections (OECD 1999; Beaman *et al.* 2007). Currently, GBIF holds records for 86.5 million preserved or fossil specimens (GBIF 2013). Even though less than 4% of all collected specimens are available digitally from the GBIF, it is nonetheless a substantial resource which includes historical as well as current records that are essential for research in taxonomy, systematics and ecology.

The data provided by GBIF is indexed by species or taxon names and is presented in a classification hierarchy from kingdom to species. While not explicitly an evolutionary framework, each classification essentially is a backbone "tree" with polytomies at most nodes (Page 2006, 2008, 2012). In principle, these classification schemes should be consistent with current phylogenetic evidence. In practice, classification schemes vary and many are currently in use. They reflect opinions, organizing principles for collections, refined phylogenetic methods, and cultural inertia due to the significant effort and expense to implement change (Page 2006, 2008).

Additionally, the generation of large phylogenies of hundreds or thousands of terminals makes it difficult, if not impossible, to interrogate phylogenetic trees to identify testable hypotheses or even understand the complexity of the allied data. Data analysis programs are available that map single characters on a phylogeny. Aggregators such as EOL and ALA typically provide data such as distribution maps and morphology characters one species page at a time. In order to visualize these connected characters in a phylogenetic framework, a tool such as PhyloJIVE is required to directly integrate data for multiple characters and maps simultaneously on the tree.

Knowledge of evolutionary relationships is essential for framing and interpreting many pressing problems in biology, such as food security, disease control and climate change (Shaffer *et al.* 1998; Cracraft 2002; Suarez and Tsutsui 2004; Chapman 2005; Grytnes & Romdal 2008; Pinto *et al.* 2010). The biodiversity informatics and the phylogenetic communities have been focused, and rightly so, on the monumental tasks of generating, organizing and presenting biodiversity data. Now is the time to develop tools that explicitly integrate phylogenies and biodiversity data.

We envision an integrated system that uses a tool based on PhyloJIVE (Fig. 2) to seamlessly integrate the tree of life with the spatial and trait data currently maintained in the biodiversity informatics community. Future advances in web service protocols and text mining will underpin a fundamental shift in biodiversity informatics data, imparting immediate integrated evolutionary-based biodiversity data for informed decisions to manage biodiversity.

Biodiversity scientists, or naturalists as they were known, were the pre-eminent biologists of the seventeenth and eighteenth centuries. At a time when medicine amounted to bleeding the ill and organized plant breeding did not exist, naturalists such as Linnaeus, Darwin, Banks and Humboldt, and their indigenous colleagues explored the

world, described biodiversity and identified evolution by means of natural selection as the dominant biological force. Since that time, and especially over the past 50 years, medical science and genomics have overtaken biodiversity research in their impact on society. It is time for biodiversity science to follow the lead of physics and assert evolution's centrality. Physicists acknowledge the power of the fundamental forces, such as gravity, incorporating them into every calculation. To successfully conserve our dwindling global biodiversity which we rely upon economically and aesthetically for sustenance and pleasure, biology must integrate evolution into every assessment of biodiversity.

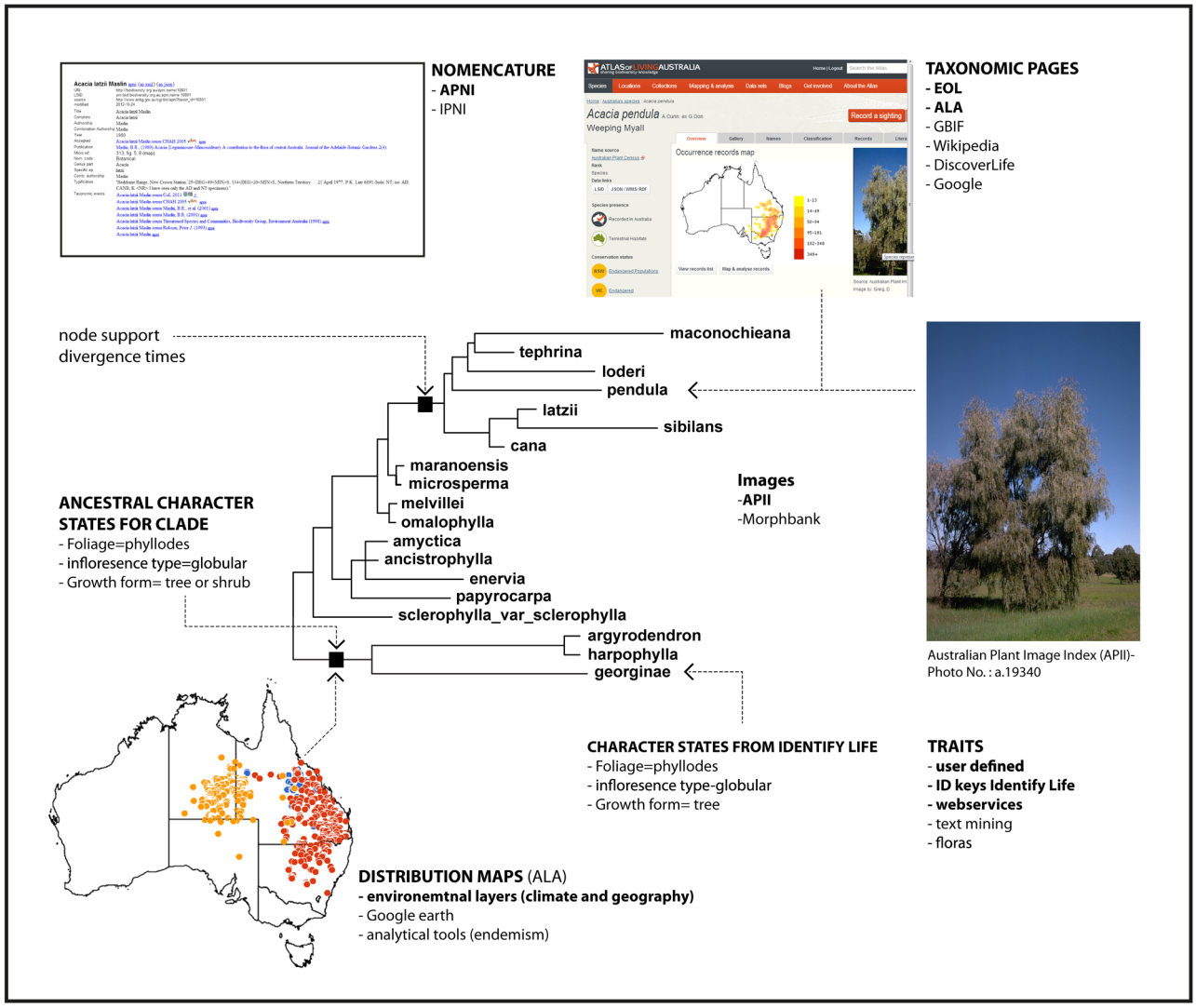

FIGURE 2. Biodiversity resources available through PhyloJIVE. Bold text indicates web services and external web pages currently available.

Acknowledgments

PhyloJIVE is available through the Atlas of Living Australia portal (http://phylojive.ala.org.au/treeViewer/show/Corals). We wish to acknowledge the Taxonomy Research & Information Network (TRIN), the Centre for Australian National Biodiversity Research (CANBR) and the Atlas of Living Australia (ALA) for supporting this project. We thank Temi Varghese, Paul Harvey, and Nick dos Remedios for programming PhyloJIVE. Executables, source code, sample data, documentation and screencasts are available at http://phylojive.ala.org.au/.

References

APGIII (2009) An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III. *Botanical Journal of the Linnean Society*, 161, 105–121.
http://dx.doi.org/10.1111/j.1095-8339.2009.00996.x

Beaman, R., Macklin, J.A., Donoghue, M.J. & Hanken, J. (2007) Overcoming the digitization bottleneck in natural history collections. A summary report on a workshop held 7–9 September 2006 at Harvard University. Available from: http://www.etaxonomy.org/wiki/images/b/b3/Harvard_data_capture_wkshp_rpt_2006.pdf. (accessed 23 December 2013)

Chapman, A.D. (2005) Uses of Primary Species-Occurrence Data, version 1.0. Global Biodiversity Information Facility, Copenhagen.

Choumane, W., Winter, P., Weigand, F. & Kahl, G. (2000) Conservation and variability of sequence-tagged microsatellite sites (STMSs) from chickpea (*Cicer aerietinum* L.) within the genus *Cicer*. *Theoretical and Applied Genetics*, 101, 269–278.
http://dx.doi.org/10.1007/s001220051479

Cracraft, J. (2002) The Seven Great Questions of Systematic Biology: An Essential Foundation for Conservation and the Sustainable Use of Biodiversity. *Annals of the Missouri Botanical Garden*, 89, 157–144.
http://dx.doi.org/10.2307/3298558

Fukami, H., Budd, A.F., Paulay, G., Solé-Cava, A., Chen, C.A., Iwao, K. & Knowlton, N. (2004) Conventional taxonomy obscures deep divergence between Pacific and Atlantic corals. *Nature*, 427, 832–835.
http://dx.doi.org/10.1038/nature02339

GBIF (2013) http://www.gbif.org/occurrence/ (accessed 2 December 2013)

Grytnes, J.-A. & Romdal, T.S. (2008) Using Museum Collections to Estimate Diversity Patterns along Geographical Gradients. *Folia Geobotanica*, 43, 357–369.
http://dx.doi.org/10.1007/s12224-008-9017-6

Jolley-Rogers, G., Varghese, T., Harvey, P., dos Remedios, N. & Miller, J.T. (Under Revision) PhyloJIVE: Integrating biodiversity data with the Tree of Life. *Bioinformatics*.

Jonathan, B.L., Stevan, J.A., Gill, B., Brodie, E.D., Hibbett, D., Hoekstra, H.E., Mindell, D.P., Monteiro, A., Moritz, C., Orr, H.A., Petrov, D.A., Renner, S.S., Ricklefs, R.E., Soltis, P.S. & Turner, T.L. (2013) Evolutionary Biology for the 21st Century. *PLoS Biology*, 11, e1001466.
http://dx.doi.org/10.1371/journal.pbio.1001466

Mast, A.R. & Thiele, K. (2007) The transfer of *Dryandra* R.Br. to *Banksia* L.f. (Proteaceae). *Austalian Systematic Botany*, 20, 63–71.
http://dx.doi.org/10.1071/sb06016

Matsunaga, A., Thompson, A., Figueiredo, R.J., Gremain-Aubrey, C.C., Collins, M., Beaman, R.S., MacFadden, B.J., Riccardi, G., Soltis, P.S., Page, L.M. & Fortis, J.A.B. (2013) *A Computational and Storage-Cloud for Integration of Biodiversity. Proceedings of the 2013 IEEE 9th International Conference on e-Science*. Beijing, IEEE, China, pp 78–87.
http://dx.doi.org/10.1109/eScience.2013.48

Miller, J.T. & Bayer, R.J. (2001) Molecular phylogenetics of *Acacia* (Fabaceae : Mimosoideae) based on the chloroplast matK coding sequence and flanking trnK intron spacer regions. *American Journal of Botany*, 88, 697–705.
http://dx.doi.org/10.2307/2657071

Mishler, B.D. (2010) Species are not uniquely real biological entities. *In*: Ayala, F. & Arp, R. (Eds.), *Contemporary Debates in Philosophy of Biology*. Wiley-Blackwell, pp. 110–122.

NSF (2013) *Assembling the Tree of Life (ATOL)*. National Science Foundation, Arlington VA, 10 pp.

OECD (1999) *Final Report of the megascience forum working group on biological informatics*. OECD, Paris, 74 pp.

Open Tree of Life (2013) Available from: http://opentreeoflife.org/ (accessed 2 December 2013)

Page, R. (2006) Taxonomic Names, Metadata, and the Semantic Web. *Biodiversity Informatics*, 3, 1–15.

Page, R.D. (2008) Biodiversity informatics: the challenge of linking data and the role of shared identifiers. *Briefings in Bioinformatics*, 9, 345–354.
http://dx.doi.org/10.1093/bib/bbn022

Page, R.D. (2012) Space, time, form: viewing the Tree of Life. *Trends in Ecology and Evolution,* 27, 113–120.
http://dx.doi.org/10.1016/j.tree.2011.12.002

Parr, C.S., Guralnick, R., Cellinese, N. & Page, R.D.M. (2012) Evolutionary informatics: unifying knowledge about the diversity of life. *Trends in Ecology and Evolution*, 27, 94–103.
http://dx.doi.org/10.1016/j.tree.2011.11.001

Pinto, C.M., Baxter, B.D., Hanson, J.D., Mendez-Harclerode, F.M., Suchecki, J.R., Grijalva, M.J., Fulhorst, C.F. & Bradley, R.D. (2010) Using museum collections to detect pathogens. *Emerging Infectious Diseases*, 16, 356–7.
http://dx.doi.org/10.3201/eid1602.090998

SEWPAC (2013) Statement of Outstanding Universal Value of the Great Barrier Reef. Department of Sustainability, Environment, Water, Population and Communities, Canberra, Australia. Available from: http://www.environment.gov.au/heritage/places/world/great-barrier-reef/values.html (accessed 9 September 2013)

Shaffer, H.B., Fisher, R.N. & Davidson, C. (1998) The role of natural history collections in documenting species declines. *Trends in Ecology and Evolution*, 13, 27–30.
http://dx.doi.org/10.1016/s0169-5347(97)01177-4

Suarez, A.V. & Tsutsui, N.D. (2004) The Value of Museum Collections for Research and Society. *BioScience*, 54, 66–74.
http://dx.doi.org/10.1641/0006-3568(2004)054[0066:tvomcf]2.0.co;2

Van der Linde, K. & Houle, D. (2008) A supertree analysis and literature review of the genus *Drosophila* and closely related genera (Diptera, Drosophilidae). *Insect Systematics and Evolution*, 39, 241–267.
http://dx.doi.org/10.1163/187631208788784237

Varón, A., Vinh, L.S. & Wheeler, W.C. (2010) POY version 4: phylogenetic analysis using dynamic homologies. *Cladistics*, 26, 72–85.
http://dx.doi.org/10.1111/j.1096-0031.2009.00282.x