Comparative biogeography, big data and common myths

Alexandre Antonelli

SCI. DAN. B. 6 · TROPICAL PLANT COLLECTIONS
THE ROYAL DANISH ACADEMY OF SCIENCES AND LETTERS · 2017

Abstract

The scientific value of biological collections extends far beyond the naming, classification, and mapping of the world's biodiversity. Molecular phylogenies constructed from voucher specimens, in combination with fossil records, can be used to infer the biogeographical history of lineages, the relation between their diversification and past biotic and abiotic events, and the historical assembly of entire biotas. However, two major challenges remain. First, in order to identify common processes underlying biodiversity patterns, we need to perform analyses under standardised procedures – which I define here as 'comparative biogeography'. This includes data-driven identification and delimitation of biogeographical regions, and the spatial coding of species to estimate range shifts and diversification. Second, we have to learn how to deal with 'big data.' To this end we need to embrace innovative bioinformatic solutions for dealing with errors and biases in public databases, we should make software 'black boxes' more transparent, self-contained and reproducible, and we must increase the engagement of citizens for logging and identifying species. By addressing the common myths about big data and engaging the manifold resources available to us, biodiversity research can move past many standing shortcomings of the field to become a time of huge opportunity.

Key Words: crowd science, evolution, molecular phylogenetics, public databases, species occurrences

Alexandre Antonelli, University of Gothenburg, Gothenburg Global Biodiversity Centre, Box 461, SE 405 30 Göteborg, Sweden; and Gothenburg Botanical Garden, Carl Skottsbergs gata 22A, 413 19 Göteborg, Sweden. Email: alexandre.antonelli@bioenv.gu.se

> 'We are drowning in information, while starving for wisdom.' (Wilson 1999: 294).

There are two striking aspects of the data used in biodiversity research. First, you never know how the information that you gathered today will be used by someone else tomorrow. The type specimen of *Solanum elaeagnifolium* Cav. stored in Madrid is just labelled 'del viaje de los españoles alrededor del mundo' [= from the travels by the Spaniards around the world] (S. Knapp pers. comm.). While that information might have been enough for the purposes of the collector, it drastically reduces the chances of this specimen being useful for more than a handful of research questions. Second, publicly available biodiversity data have increased at a speed and volume similar to that experienced by several other scientific disciplines and society. So both in terms of quantity (breath and

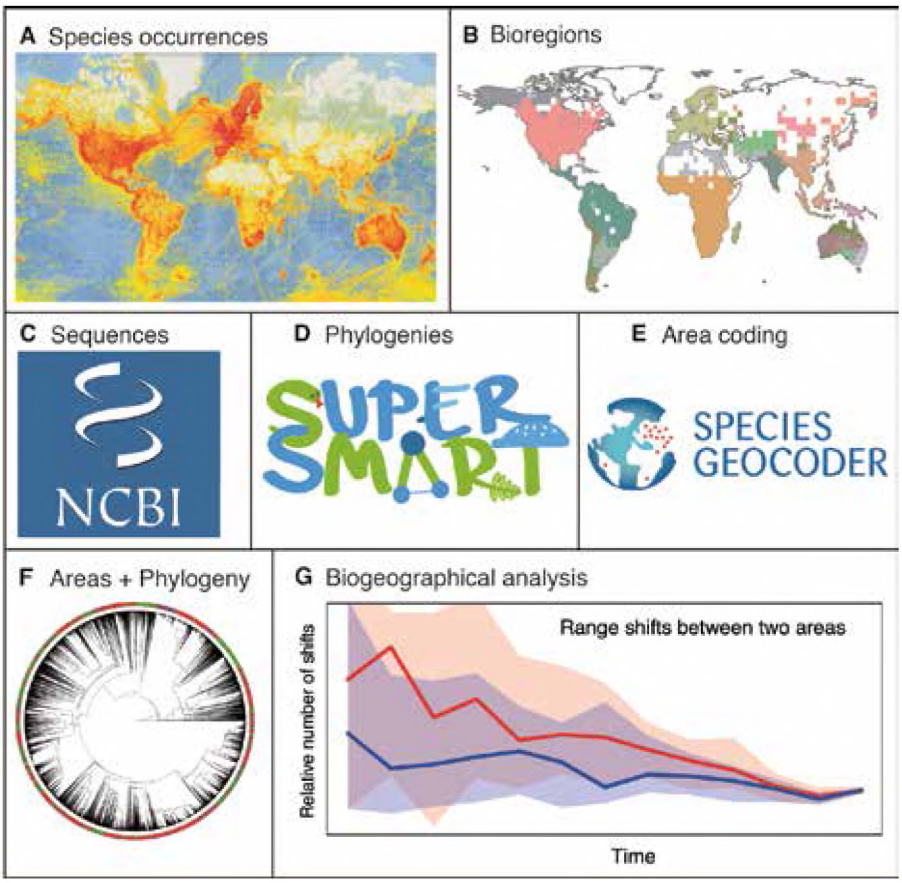

Fig. 1. Schematic workflow of the 'comparative biogeography' approach. First, species occurrence data that have been automatically and/or manually cleaned (A) are clustered into bioregions (B). Molecular sequence data (C) are downloaded, clustered and aligned to produce dated phylogenetic trees with the addition of fossil information (D). Species in the resulting phylogenies are then coded into each of the bioregions delimited (E). The coded phylogeny (F) can then be used by a large number of biogeographical applications, such as ancestral range reconstructions that infer the number of biotic range shifts between any set of areas (G). See the text for more details and references to the packages used.

density) and diversity (taxonomic, genetic, morphological, ecological, spatial, and temporal), biodiversity data are now widely recognised as 'big data' and there is no indication that the rate of data proliferation will cease its rapid outward expansion.

This review is meant to provide an admittedly personal and biased perspective on how to concretely advance some outstanding aspects of biodiversity research, with a focus on the use of species occurrence data and molecular sequences for biogeographical research.

Using Biological Collections in Comparative Biogeography

Fine-scale biogeographical studies based on voucher specimens, such as those investigating the diversification history of a particular genus or clade (e.g., Meseguer *et al.* 2014; Zhang *et al.* 2015; Lagomarsino *et al.* 2016), are essential to our understanding of biogeographical processes. However, in order to identify the factors driving the major patterns of biodiversity observed today, we need to investigate the evolutionary history of many lineages in relation to biotic and abiotic events and the fossil record (e.g., Wesselingh *et al.* 2010; Antonelli & Sanmartín 2011; Favre *et al.* 2014; Linder 2014). Although the term comparative biogeography is not new (Parenti & Ebach 2009), I choose to define it here as "the approach of inferring the evolutionary history of multiple taxa under standardised methods in order to enable direct comparisons and the identification of common processes underlying biodiversity patterns". Unfortunately, the lack of consensus on how to conduct biogeographical studies (albeit beneficial to methodological development) has led to the accumulation of results that cannot be readily

compared. As a practical illustration of how comparative biogeography may be conducted, a short and simple workflow for inferring the geographic history of lineages is provided (Fig. 1) and explained below.

Selecting areas for biogeographical analyses

In most biogeographical methods currently available (e.g., Ree & Smith 2008; Matzke 2014), it is first necessary to decide on a set of discrete operational units (spatial polygons). These could be anything from clearly defined geographical units (such as continents and islands), to regions defined on the basis of their biota (such as biodiversity hotspots, biomes, and ecoregions), units based on geological history (such as the Indian subcontinent, which was long separated from Asia), or a combination of any of these. In many studies, the choice of operational units is not well explained or lacks consistent criteria, and authors do not investigate how arbitrary decisions may influence subsequent results.

One possible solution to standardise operational areas in biogeography is to apply data-driven detection approaches. The idea is not new, as several algorithms have already been proposed based, for instance, on the concept of areas of endemism (Morrone 1994; Linder 2001). However, new methods are able to use massive amounts of geo-referenced species occurrence records that are available, for example, through the Global Biodiversity Information Facility (GBIF; www.gbif.org; Fig. 1A) to identify how species group spatially into higher-level clusters, often called biogeographical regions or simply bioregions. The accuracy of these methods can then be statistically evaluated (Kreft & Jetz 2010). We recently described one such approach – Infomap Bioregions (Fig. 1B) – based on network clustering that appears to outperform similarity-based algorithms (Vilhena & Antonelli 2015; Edler *et al.* 2017).

Assigning species to areas

Once the operational spatial units for biogeographical analysis are decided upon, it is time to assign each species to one or several areas where it occurs. Taking into consideration the rapidly increasing number of geo-referenced specimens now available (over 150 million plant records through GBIF), this task may be extremely time-consuming and error-prone. To facilitate this process, we have developed a software package called SpeciesGeoCoder (Töpel *et al.* 2017), which enables the fast categorisation of species (or population, individuals, etc. – depending on the taxonomic level surveyed) into operational areas (Fig. 1E). These areas may be purely spatial and defined by the borders of a GIS polygon, such as the bioregions delimited in the previous step. However, they can also be defined on the basis of an altitudinal range – such as 'this area under 500 m' – which would facilitate the coding of species occurring in topographically heterogeneous regions. In this way, it would be possible to code species that occur at low, middle, and high elevation zones. Biogeographic analyses would then allow researchers to estimate the origin of the species in each of those zones, for instance along a mountain slope in Borneo (Antonelli 2015; Merckx *et al.* 2015).

Inferring the evolutionary history of lineages and biotas

We are still far from having well-resolved phylogenies for most taxa, as most species have not yet been sequenced. For example, only data for c. 9–11% of all tropical angiosperms with georeferenced species occurrences were available to be assembled into a phylogenetic tree (Zanne *et al.* 2014; Antonelli *et al.* 2015). In addition, most of the molecular sequences produced so far lack any spatial information associated with their records in the National Center for Biotechnology Information database (NCBI/GenBank). For example, only 6.2% of all sequenced species of tetrapods have associated spatial data in GenBank, despite being such a charismatic clade (Gratton *et al.* 2017).

Rapidly decreasing sequencing costs and new sequencing technologies should soon ameliorate this issue. In plants, sequencing coverage may be greatly increased in the coming years by massive sequencing of herbarium samples (Bakker *et al.* 2016; Bakker *et al.*

2017). Large initiatives targeting the sequencing of full genomes and/or transcriptomes of birds (Zhang *et al.* 2014), insects (Misof *et al.* 2014), and vertebrates (Haussler *et al.* 2009) will further increase taxonomic coverage across the tree of life. However, it is important to consider that even if research labs worldwide would together manage to sequence all species, without proper co-ordination of efforts and synthesis of results we would still be unable to address most questions in comparative biogeography. This is because researchers are now using widely disparate sets of sequence regions for phylogenetic inference, different methods for estimating divergence times, and fossils of very different quality and informativeness for calibrating molecular phylogenies. This is a fundamental issue that cannot be solved by synergistic initiatives such as the Open Tree of Life (Hinchliff *et al.* 2015), which essentially produces a mega-tree by grafting together smaller trees from various studies. We recently proposed a complementary approach (Antonelli *et al.* 2015), which allows the inference of phylogenies based on all suitable sequences publicly available. Our initiative, termed the Self-Updating Platform for Estimating Rates of Speciation and Migration, Ages and Relationships of Taxa (SUPERSMART) is designed as a user-friendly platform for producing dated phylogenies for any taxon (Fig. 1D). It differs from so-called 'supertree' and 'supermatrix' approaches (e.g., Haeseler 2012) by performing phylogenetic estimation in a stepwise way. First, SUPERSMART produces higher-level backbone trees based on a set of common sequence regions, which are typically slowly evolving and well represented taxonomically. Then, the package expands internal nodes to comprise all suitable species and sequences, calculating their relationship and divergence times under coalescent methods. Finally, it grafts all species-level phylogenies back to the backbone tree, thus resulting in very large species-level dated phylogenies for all sequenced taxa. Once the species in a phylogeny are area-coded (Fig. 1F), it becomes relatively straightforward to perform any phylogeny-based biogeographical analysis (e.g., Fig. 1G).

Dealing with Big Data

The examples provided above for biogeographical research depict some of the drastic changes that are affecting the research landscape in biodiversity. These include a rapid increase in the volume of data (and the associated biases and errors), an urge to use biological collections to address a plethora of new questions as compared to earlier uses, and an increased complexity associated with understanding, choosing and applying new methods. In order to further advance biodiversity research, it is therefore crucial to re-evaluate some of the most prominent myths in our community regarding this recent onslaught of easily accessible data.

Myth 1: Public databases are useless for research

There is general and widespread scepticism concerning the use of publicly available databases - such as GBIF and GenBank – for 'serious' research. Just as many researchers today refer to Wikipedia for a rapid check on a term or event, public databases are often seen as a place where researchers can do a quick check of what is available, but no more. One critical limitation of public data is that the resolution (e.g., spatial, taxonomic, genetic, temporal) might not be appropriate to the study question. But even when the available biodiversity and molecular data at first sight seem appropriate, they may contain numerous errors and biases (Valkiunas *et al.* 2008; Mendonça *et al.* 2011; Nilsson *et al.* 2012; Meyer *et al.* 2015; Meyer *et al.* 2016). Fortunately many of these can now be properly identified and quantified, and analytical tools are being developed to further increase their usefulness in research.

'Errors' can often be identified by algorithms. For instance, when dealing with species occurrence data, SpeciesGeoCoder can flag terrestrial species with occurrences in the sea, those with coordinates outside of the countries in which they were recorded, those located significantly farther away from the rest, and those with occurrences in country centroids, among other anomalies. The package biogeo (Robertson *et*

al. 2016) goes a step further, providing suggestions to correct entries after approval by the user. When time or resources are scarce, researchers can thus focus their initial attention on such automatically identified potential errors. Similarly, when constructing molecular datasets for phylogenetic inference, SUPERSMART will not include those sequences that are significantly different to other sequences for the same taxon, thus reducing the inclusion of wrongly identified or spuriously sequenced specimens.

'Biases' are equally amenable to automated detection and quantification. For instance, taxonomic biases are straightforward to calculate by counting how many sequences and occurrence records are available in public databases in relation to the expected diversity in a clade. Collection biases can also be calculated by computing the distance between each geo-referenced record and the nearest road, river, city, national park, etc. (Fithian *et al.* 2015; see also https://github.com/azizka/sampbias).

Most biogeographical methods now allow for the incorporation of biases, such as incomplete species sampling, in diversification rate analyses (Cusimano *et al.* 2012; Rabosky 2014). Others allow the user to set absolute and relative thresholds to account for undetected errors. For instance, in SpeciesGeoCoder the user can set a minimum of 10 records (or a certain percentage) before a species is coded as present in an area. By testing multiple threshold levels, it becomes easy to assess the robustness of results in relation to varying degrees of error in the data used (Antonelli *et al.* 2015).

Myth 2: Humans outperform analytical 'black boxes'

Many biologists do not trust computers. Some systematists prefer to spend days checking and adjusting molecular sequence alignments by hand, despising those who run a sequence alignment software for a few seconds. Some will fight tooth and nail for the use of parsimony over likelihood-based methods (Editors 2016), largely because parsimony is easier for most of us to grasp. Botanists will rarely trust computer programs for the identification of species, despite the fact that trained software has been shown under pilot tests to identify species based on leaves (Durgante *et al.* 2013; Wilf *et al.* 2016) and pollen grains (Punyasena *et al.* 2012; Riley *et al.* 2015) at similar or even higher success rates than humans. We must be open to novel bioinformatic and technical developments, in particular when dealing with the time-consuming analysis of increasingly large data volumes (García-Roselló *et al.* 2015; Maldonado *et al.* 2015).

Besides saving us time and often improving quality, software also has the ability to make biodiversity research more explicit and reproducible. Ideally, researchers should provide all input and output files for their analyses, as well as the settings they used, in supplementary material to articles, or in permanent open repositories. However, given the wide range of tools currently available, even when such data are provided it may still be nearly impossible to reproduce a published study, due to the continuous update of versions in the software and operational platforms used.

Biodiversity research is not an exception to the increasing concern for the reproducibility of studies, which is a growing problem for all sciences (e.g., Buck 2015; Open Science Collaboration 2015). Self-contained, integrative analytical platforms - sometimes unfavourably denoted 'black boxes' - have the potential to solve the general issue of reproducibility, when combined with long-term data storage, and proper reporting of analytical methods and settings. The common fear of using such platforms is that researchers loose control over what is done. To tackle this valid criticism, it is essential to make such 'boxes' less black and more transparent (Borregaard & Hart 2016), but still retaining their 'boxiness' – i.e., containing all software dependencies in one file that can be properly version-tracked and re-run. To this end, collaboration between biologists and bioinformaticians is key.

Myth 3: Citizen identifications cannot be trusted

Systematists have long held a monopoly over taxonomic knowledge, and the word 'amateur' (s/he who loves) is too often used as someone inept. This atti-

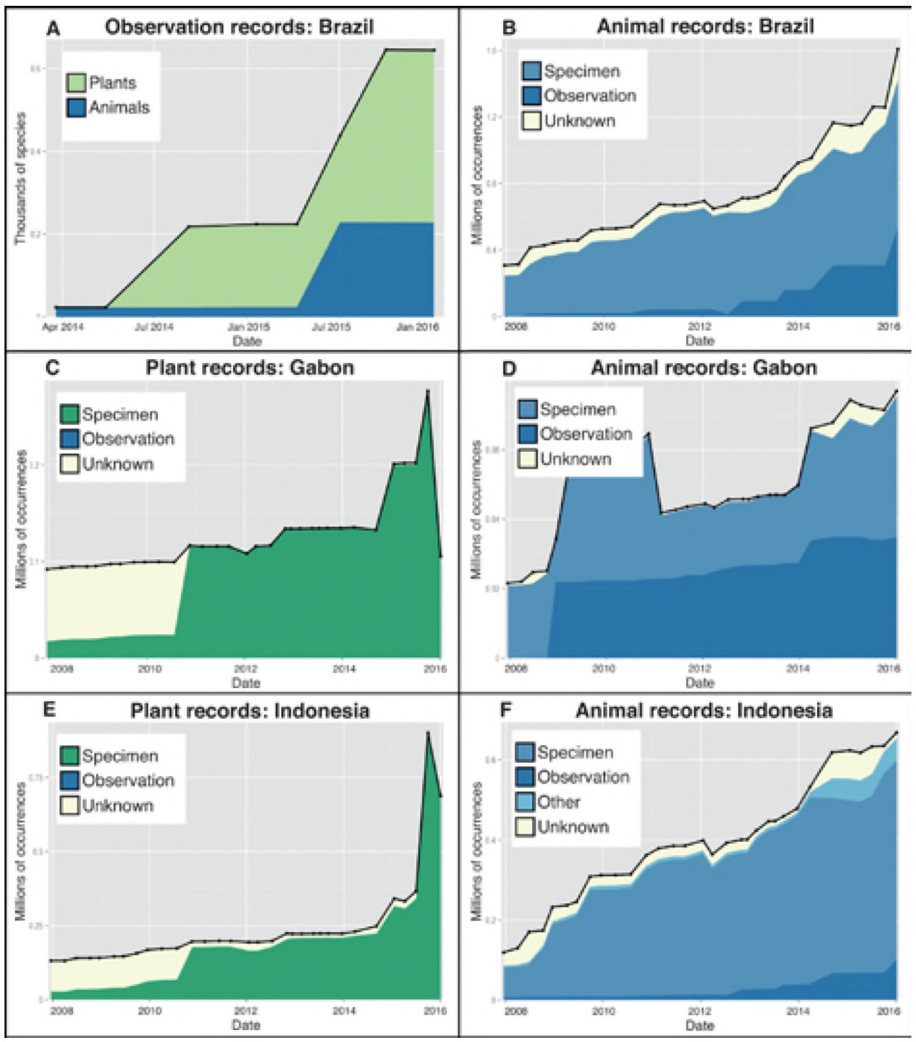

Fig. 2. The increase in observation records in GBIF. The graphs show temporal trends in the rate of new records being made accessible through the Global Biodiversity Information Facility (GBIF) for three tropical countries: Brazil (A–B), Gabon (C–D) and Indonesia (E–F). (A) exemplifies the differences between animal and plants records, while (B–F) highlight the differences between specimens and observation records. Observations are records that are not associated with a voucher specimen, such as geo-referenced photographs from citizens and records from ecological monitoring projects. (Downloaded and adapted from http://www.gbif.org/analytics on 29 March, 2016).

tude is seen not only among scientists, but also governmental agencies. In some countries, people lacking a PhD degree cannot apply for collection permits. While we complain about the lack of funding for biodiversity research, outside our ivory towers are potentially billions of people who could contribute valuable data for biodiversity research and conservation. It is now time to seize the opportunity given by the smartphone era and the increased appeal of outdoor activities in order to engage citizens in locating, identifying, and sharing information about species.

Several mobile applications already facilitate the logging and identification of species, such as iSpot, iNaturalist, Project Noah, Plant-o-Matic, and Map of Life. It is clear that certain species are not as amenable to identification through smartphone photos as others (e.g., when micro-morphological characters are required for reliable identifications). However, despite general scepticism that citizens cannot be trusted to identify specimens, this is likely not a major concern. In a recent assessment, citizens were able to correctly identify about 92% of all sightings recorded on iSpot, and more than half of these within an hour of submission (Silvertown *et al.* 2015).

Thanks to the recent engagement of citizens and scientists alike, the number of observation records – i.e., those sightings not associated with a museum

specimen, but often containing georeferenced images and other locality data – is increasing rapidly in many databases, such as iNaturalist (over 2.3 million observations to date). However, there are still large differences among countries and between plants and animals, which can be seen by comparing the statistics in GBIF for three tropical countries (Fig. 2). In Brazil, there are currently about twice as many recorded observations for plants as for animals (Fig. 2A), although most records entering GBIF are still specimen-based (Fig. 2B). Both in Gabon (Fig. 2C-D) and Indonesia (Fig. 2E-F), only observations of animals are accessible through GBIF. Interestingly, the number of observation records for animals in Gabon is about the same as the number of digitized museum specimens (Fig. 2D). However, it should be noted that many of these observations were recorded during research projects and not by citizens, as just a few citizen science databases (such as eBird and iNaturalist) currently provide data to GBIF.

Species observations do not replace the fundamental role of voucher specimens for biodiversity research, a role that is correctly demonstrated throughout this volume and in the literature (e.g., Rocha *et al.* 2014). However, engaging citizens in recording novel sightings of species will greatly complement biological collections and increase our knowledge on the distribution of species - thus decreasing the 'Wallacean shortfall' (Hortal *et al.* 2015). This knowledge will in turn improve our studies on the biogeographical history of taxa, the future distribution of species based on ecological niche modelling, and the conservation status of species, among many other applications. In addition, this could also lead to a general increase in the societal understanding and appreciation for biodiversity at large.

Conclusions and Outlook

Most people outside academia have no idea about how little we actually know about the natural world. For instance, few realize that we still have not described nearly about 86% of all terrestrial and about 91% of all marine species (Mora *et al.* 2011), and of the described ones only about 20% have been sequenced or barcoded (Hinchliff *et al.* 2015).

We have not even picked the low-hanging fruit. Until very recently, it was widely assumed that the clash between the North and South American continents through the Panama Isthmus took place 3.5 million years ago – an event with major significance for the biotic interchange between these continents, global ocean circulation, and climate. It therefore came as a surprise that the connection probably took place approximately 10 million years earlier, as was shown by biogeographical analyses based on biological collections (Bacon *et al.* 2015) and backed up by new geological evidence (Montes *et al.* 2015). Similarly, 150 years after the first classifications of the world's biogeographical regions (Sclater 1858; Wallace 1876), we are still at a point where the simple use of different methods on the same dataset of species distributions can produce large discrepancies in results and different delimitations of the world's biogeographical regions (Holt *et al.* 2013; Vilhena & Antonelli 2015). Clearly, many controversies remain to be settled and major discoveries to be made.

I hope to have convinced you that there is an exciting and prosperous future for many upcoming generations of biodiversity scientists. Public databases, open bioinformatic tools, and increased data sharing hold the potential to greatly advance biodiversity and biogeographical research (Wen *et al.* 2013; Borregaard & Hart 2016; Poisot *et al.* 2016). However, the promising prospects are accompanied by big challenges. These include our ability to thrive in an age where data are big, full of gaps and biases, and (still) poorly synthetized and integrated. In addition, we need to increase openness while safeguarding intellectual property – e.g., by providing data providers and software developers the adequate credit for their work. And we must, or course, still strongly encourage, support and reward fieldwork and the generation of novel biodiversity data. Bioinformatic solutions can do a great deal of the 'dirty work' for us, so that we can focus our limited time and resources on making sense out of noise, improving algorithms to do what we want them to, and en-

gaging the scientific community and the general public to join forces in the understanding and protection of biodiversity.

Acknowledgements

I wish to thank Henrik Balslev and Ib Friis for organising this symposium and volume, my colleagues and students who developed much of the software mentioned in this paper and helped to form my thinking on this topic, Allison Perrigo and two anonymous reviewers for constructive comments on this paper, biodiversity scientists around the world for generously contributing data to open repositories, and my funding agencies for support (the European Research Council under the European Union's Seventh Framework Programme [FP/2007-2013, ERC Grant Agreement n. 331024], the Wallenberg Foundation for a Wallenberg Academy Fellowship, and the Swedish Research Council [2015-04857]).

References

Antonelli, A. (2015). Multiple origins of mountain life. *Nature* 524: 300-301.

Antonelli, A., Hettling, H., Condamine, F.L., Vos, K., Nielsson, R.H., Nilsson, R.H., Sanderson, J., Sauquet, H., Scharn, R., Silvestro, D., Töpel, M. Bacon C.D., Oxelman, B. & Vos, R.A. (2017). Towards a Self-Updating Platform for Estimating Rates of Speciation and Migration, Ages, and Relationship of Taxa. *Systematic Biology* 66(2): 152-166.

Antonelli, A. & Sanmartín, I. (2011). Why are there so many plant species in the Neotropics? *Taxon* 60: 403-414.

Antonelli, A., Zizka, A., Silvestro, D., Scharn, R., Cascales-Miñana, B. & Bacon, C.D. (2015). An engine for global plant diversity: Highest evolutionary turnover and emigration in the American tropics. *Frontiers in Genetics* 6. http://dx.doi.org/10.3389/fgene.2015.00130 (accessed 6 April, 2016)..

Bacon, C.D., Silvestro, D., Jaramillo, C., Smith, B.T., Chakrabarty, P. & Antonelli, A. (2015). Biological evidence supports an early and complex emergence of the Isthmus of Panama. *Proceedings of the National Academy of Sciences* 112: 6110-6115.

Bakker, F.T., Lei, D., Yu, J., Mohammadin, S., Wei, Z., Kerke, S., Gravendeel, B., Nieuwenhuis, M., Staats, M. & Alquezar-Planas, D.E. (2016). Herbarium genomics: Plastome sequence assembly from a range of herbarium specimens using an Iterative Organelle Genome Assembly pipeline. *Biological Journal of the Linnean Society* 117: 33-43.

Bakker, F.T., Lei, D. & Holmer, R. (2017). Herbarium genomics, skimming and plastome sequencing. *Scientia Danica, B (Biologica)* 6: 271-283.

Borregaard, M.K. & Hart, E.M. (2016). Towards a more reproducible ecology. *Ecography* 39(4): 349-353.

Buck, S. (2015). Solving reproducibility. *Science* 348: 1403-1403.

Cusimano, N., Stadler, T. & Renner, S.S. (2012). A new method for handling missing species in diversification analysis applicable to randomly or nonrandomly sampled phylogenies. *Systematic Biology* 61: 785-792.

Durgante, F.M., Higuchi, N., Almeida, A. & Vicentini, A. (2013). Species Spectral Signature: Discriminating closely related plant species in the Amazon with Near-Infrared Leaf-Spectroscopy. *Forest Ecology and Management* 291: 240-248.

Editors, T. (2016). Editorial. *Cladistics* 32: 1.

Edler, D., Guedes, T., Zizka, A., Rosvall, M. & Antonelli, A. (2017). Infomap Bioregions: Interactive mapping of biogeographical regions from species distributions. *Systematic Biology* 66(2): 197-204.

Favre, A., Päckert, M., Pauls, S.U., Jähnig, S.C., Uhl, D., Michalak, I. & Muellner-Riehl, A.N. (2014). The role of the uplift of the Qinghai-Tibetan Plateau for the evolution of Tibetan biotas. *Biological Reviews* 90 (1): 236-253.

Fithian, W., Elith, J., Hastie, T. & Keith, D.A. (2015). Bias correction in species distribution models: Pooling survey and collection data for multiple species. *Methods in Ecology and Evolution* 6: 424-438.

García-Roselló, E., Guisande, C., Manjarrés-Hernández, A., González-Dacosta, J., Heine, J., Pelayo-Villamil, P., González-Vilas, L., Vari, R.P., Vaamonde, A., Granado-Lorencio, C. & Lobo, J.M. (2015). Can we derive macroecological patterns from primary Global Biodiversity Information Facility data? *Global Ecology and Biogeography* 24: 335-347.

Gratton, P., Marta, S., Bocksberger, G., Winter, M., Trucchi, E. & Kuehl, H. (2017). A world of sequences. Can we use georeferenced nucleotide databases for an automated phylogeography? *Journal of Biogeography* 44: 475-486. doi: 10.1111/jbi.12786.

Haeseler, A. von (2012). Do we still need supertrees? *BMC Biology* 10(1): 13.

Haussler, D., O'Brien, S.J., Ryder, O.A., Barker, F.K., Clamp, M., Crawford, A.J., Hanner, R., Hanotte, O., Johnson, W.E. & McGuire, J.A. (2009). Genome 10K: A proposal to obtain whole-genome sequence for 10 000 vertebrate species. *Journal of Heredity* 100: 659-674.

Hinchliff, C.E., Smith, S.A., Allman, J.F., Burleigh, J.G., Chaudhary, R., Coghill, L.M., Crandall, K.A., Deng, J., Drew, B.T., Gazis, R., Gude, K., Hibbett, D.S., Katz, L.A., Laughinghouse, H.D., McTavish, E.J., Midford, P.E., Owen, C.L., Ree, R.H., Rees, J.A., Soltis, D.E., Williams, T. & Cranston, K.A. (2015). Synthesis of phylogeny and taxonomy into a comprehensive tree of life. *Proceedings of the National Academy of Sciences* 112: 12764-12769.

Holt, B.G., Lessard, J.-P., Borregaard, M.K., Fritz, S.A., Araújo, M.B., Dimitrov, D., Fabre, P.-H., Graham, C.H., Graves, G.R., Jønsson, K.A., Nogués-Bravo, D., Wang, Z., Whittaker, R.J., Fjeldså, J. & Rahbek, C. (2013). An Update of Wallace's Zoogeographic Regions of the World. *Science* 339: 74-78.

Hortal, J., de Bello, F., Diniz-Filho, J.A.F., Lewinsohn, T.M., Lobo, J.M. & Ladle, R.J. (2015). Seven shortfalls that beset large-scale knowledge on biodiversity. *Annual Review of Ecology, Evolution, and Systematics* 46: 523-549.

Kreft, H. & Jetz, W. (2010). A framework for delineating biogeographical regions based on species distributions. *Journal of Biogeography* 37: 2029-2053.

Lagomarsino, L.P., Condamine, F.L., Antonelli, A., Mulch, A. & Davis, C.C. (2016). The abiotic and biotic drivers of rapid diversification in Andean bellflowers (Campanulaceae). *New Phytologist* 210(4): 1430-1442. http://onlinelibrary.wiley.com/doi/10.1111/nph.13920/full (accessed 6.4.2016).

Linder, H.P. (2001). On areas of endemism, with an example from the African Restionaceae. *Systematic Biology* 50(6): 892-912.

Linder, H.P. (2014). The evolution of African plant diversity. *Frontiers in Ecology and Evolution* 2. http://dx.doi.org/10.3389/fevo.2014.00038 (accessed 6 April, 2016).

Maldonado, C., Molina, C.I., Zizka, A., Persson, C., Taylor, C.M., Albán, J., Chilquillo, E., Rønsted, N. & Antonelli, A. (2015). Estimating species diversity and distribution in the era of Big Data: To what extent can we trust public databases? *Global Ecology and Biogeography* 24 (8):973-984 .

Matzke, N.J. (2014). Model selection in historical biogeography reveals that founder-event speciation is a crucial process in island clades. *Systematic Biology* 63(6): 951-970.

Mendonça, R.S. de, Navia, D., Diniz, I.R., Auger, P. & Navajas, M. (2011). A critical review on some closely related species of *Tetranychus* sensu stricto (Acari: Tetranychidae) in the public DNA sequences databases. *Experimental and Applied Acarology* 55: 1-23.

Merckx, V.S.F.T., Hendriks, K.P., Beentjes, K.K., Mennes, C.B., Becking, L.E., Peijnenburg, K.T.C.A., Afendy, A., Arumugam, N., de Boer, H., Biun, A., Buang, M.M., Chen, P.-P., Chung, A.Y.C., Dow, R., Feijen, F.A.A., Feijen, H., Feijen-van Soest, C., Geml, J., Geurts, R., Gravendeel, B., Hovenkamp, P., Imbun, P., Ipor, I., Janssens, S.B., Jocque, M., Kappes, H., Khoo, E., Koomen, P., Lens, F., Majapun, R.J., Morgado, L.N., Neupane, S., Nieser, N., Pereira, J.T., Rahman, H., Sabran, S., Sawang, A., Schwallier, R.M., Shim, P.-S., Smit, H., Sol, N., Spait, M., Stech, M., Stokvis, F., Sugau, J.B., Suleiman, M., Sumail, S., Thomas, D.C., van Tol, J., Tuh, F.Y.Y., Yahya, B.E., Nais, J., Repin, R., Lakim, M. & Schilthuizen, M. (2015). Evolution of endemism on a young tropical mountain. *Nature* 524: 347-350.

Meseguer, A.S., Lobo, J.M., Ree, R., Beerling, D.J. & Sanmartín, I. (2014). Integrating fossils, phylogenies, and niche models into biogeography to reveal ancient evolutionary history: The case of *Hypericum* (Hypericaceae). *Systematic Biology* 64(2): 215-232.

Meyer, C., Kreft, H., Guralnick, R. & Jetz, W. (2015). Global priorities for an effective information basis of biodiversity distributions. *Nature Communications* **6** http://www.nature.com/ncomms/2015/150907/ncomms9221/full/ncomms9221.html (accessed 6 April, 2016).

Meyer, C., Weigelt, P. & Kreft, H. (2016). Multidimensional biases, gaps and uncertainties in global plant occurrence information. *Ecology Letters* 19(8): 992-1006.

Misof, B., Liu, S., Meusemann, K., Peters, R.S., Donath, A., Mayer, C., Frandsen, P. B., Ware, J., Flouri, T., Beutel, R.G., Niehuis, O., Petersen, M., Izquierdo-Carrasco, F., Wappler, T., Rust, J., Aberer, A.J., Aspöck, U., Aspöck, H., Bartel, D., Blanke, A., Berger, S., Böhm, A., Buckley, T.R., Calcott, B., Chen, J., Friedrich, F., Fukui, M., Fujita, M., Greve, C., Grobe, P., Gu, S., Huang, Y., Jermiin, L.S., Kawahara, A.Y., Krogmann, L., Kubiak, M., Lanfear, R., Letsch, H., Li, Y., Li, Z., Li, J., Lu, H., Machida, R., Mashimo, Y., Kapli, P., McKenna, D.D., Meng, G., Nakagaki, Y., Navarrete-Heredia, J.L., Ott, M., Ou, Y., Pass, G., Podsiad-

lowski, L., Pohl, H., von Reumont, B.M., Schütte, K., Sekiya, K., Shimizu, S., Slipinski, A., Stamatakis, A., Song, W., Su, X., Szucsich, N.U., Tan, M., Tan, X., Tang, M., Tang, J., Timelthaler, G., Tomizuka, S., Trautwein, M., Tong, X., Uchifune, T., Walzl, M.G., Wiegmann, B.M., Wilbrandt, J., Wipfler, B., Wong, T.K.F., Wu, Q., Wu, G., Xie, Y., Yang, S., Yang, Q., Yeates, D.K., Yoshizawa, K., Zhang, Q., Zhang, R., Zhang, W., Zhang, Y., Zhao, J., Zhou, C., Zhou, L., Ziesmann, T., Zou, S., Li, Y., Xu, X., Zhang, Y., Yang, H., Wang, J., Wang, J., Kjer, K.M. & Zhou, X. (2014). Phylogenomics resolves the timing and pattern of insect evolution. *Science* 346: 763–767.

Montes, C., Cardona, A., Jaramillo, C., Pardo, A., Silva, J.C., Valencia, V., Ayala, C., Pérez-Angel, L.C., Rodriguez-Parra, L.A., Ramirez, V. & Niño, H. (2015). Middle Miocene closure of the Central American Seaway. *Science* 348: 226–229.

Mora, C., Tittensor, D.P., Adl, S., Simpson, A.G.B. & Worm, B. (2011). How many species are there on earth and in the ocean? *PLoS Biol* 9: e1001127. http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1001127 (accessed 6.4.2916).

Morrone, J.J. (1994). On the identification of areas of endemism. *Systematic Biology* 43: 438–441.

Nilsson, R.H., Tedersoo, L., Abarenkov, K., Ryberg, M., Kristiansson, E., Hartmann, M., Schoch, C.L., Nylander, J.A., Bergsten, J. & Porter, T.M. (2012). Five simple guidelines for establishing basic authenticity and reliability of newly generated fungal ITS sequences. *MycoKeys* 4: 37–63.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science* 349: 943.

Parenti, L.R. & Ebach, M.C. (2009). *Comparative Biogeography: Discovering and Classifying Biogeographical Patterns of a Dynamic Earth*. University of California Press, Berkley and Los Angeles.

Poisot, T., Gravel, D., Leroux, S., Wood, S.A., Fortin, M.-J., Baiser, B., Cirtwill, A. R., Araújo, M.B. & Stouffer, D.B. (2016). Synthetic datasets and community tools for the rapid testing of ecological hypotheses. *Ecography* 39(4): 402–408.

Punyasena, S.W., Tcheng, D.K., Wesseln, C. & Mueller, P.G. (2012). Classifying black and white spruce pollen using layered machine learning. *New Phytologist* 196: 937–944.

Rabosky, D.L. (2014). Automatic detection of key innovations, rate shifts, and diversity-dependence on phylogenetic trees. *PLoS ONE* 9: e89543. http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0089543 (accessed 6 April, 2016).

Ree, R.H. & Smith, S.A. (2008). Maximum likelihood inference of geographic range evolution by dispersal, local extinction, and cladogenesis. *Systematic Biology* 57: 4–14.

Riley, K.C., Woodard, J.P., Hwang, G.M. & Punyasena, S.W. (2015). Progress towards establishing collection standards for semi-automated pollen classification in forensic geo-historical location applications. *Review of Palaeobotany and Palynology* 221: 117–127.

Robertson, M.P., Visser, V. & Hui, C. (2016). Biogeo: an R package for assessing and improving data quality of occurrence record datasets. *Ecography* 39(4), 394–401.

Rocha, L.A., Aleixo, A., Allen, G., Almeda, F., Baldwin, C., Barclay, M., Bates, J. M., Bauer, A., Benzoni, F. & Berns, C. (2014). Specimen collection: An essential tool. *Science* 344: 814–815.

Sclater, P.L. (1858). On the general geographical distribution of the members of the class Aves. *Journal of the Proceedings of the Linnean Society of London. Zoology* 2: 130–136.

Silvertown, J., Harvey, M., Greenwood, R., Dodd, M., Rosewell, J., Rebelo, T., Ansine, J. & McConway, K. (2015). Crowdsourcing the identification of organisms: A case-study of iSpot. *ZooKeys* 480: 125–146. doi: 10.3897/zookeys.480.8803 (accesssed 6 April, 2016).

Töpel, M., Zizka, A., Calió, M.F., Scharn, R., Silvestro, D. & Antonelli, A. (2017). SpeciesGeoCoder: Fast categorisation of species occurrences for analyses of biodiversity, biogeography, ecology and evolution. *Systematic Biology* 66(2): 145-151.

Valkiunas, G., Atkinson, C.T., Bensch, S., Sehga, R.N.M. & Ricklefs, R.E. (2008). Parasite misidentifications in GenBank: How to minimize their number? *Trends in Parasitology* 24: 247–248.

Vilhena, D.A. & Antonelli, A. (2015). A network approach for identifying and delimiting biogeographical regions. *Nature Communications* 6. http://www.nature.com/ncomms/2015/150424/ncomms7848/full/ncomms7848.html (accesssed 6 April, 2016).

von Haeseler, A. (2012). Do we still need supertrees? *BMC Biology* 10(1): 13 (accesssed 6 April, 2016).

Wallace, A.R. (1876). *The Geographical Distribution of Animals: With a Study of the Relations of Living and Extinct Faunas as Elucidating the Past Changes of the Earth's Surface:* In Two Volumes. Macmillan & Co., London.

Wen, J., Ree, R.H., Ickert-Bond, S.M., Nie, Z. & Funk, V. (2013). Biogeography: Where do we go from here? *Taxon* 62(5): 912–927.

Wesselingh, F.P., Hoorn, C., Kroonenberg, S.B., Antonelli, A., Lundberg, J.G., Vonhof, H.B. & Hooghiemstra, H. (2010). On the origin of Amazonian landscapes and biodiversity: A synthesis. *In*: C. Hoorn & F.P. Wesselingh (eds.), *Amazonia, Landscape and Species Evolution*, 1st edition. Blackwell publishing, Hoboken. Pp. 421–431.

Wilf, P., Zhang, S., Chikkerur, S., Little, S.A., Wing, S.L. & Serre, T. (2016). Computer vision cracks the leaf code. *Proceedings of the National Academy of Sciences* 113: 3305–3310.

Wilson, E.O. (1999). *Consilience: The Unity of Knowledge*. Random House Digital, Inc., New York.

Zanne, A.E., Tank, D.C., Cornwell, W.K., Eastman, J.M., Smith, S.A., FitzJohn, R.G., McGlinn, D.J., O'Meara, B.C., Moles, A.T., Reich, P.B., Royer, D.L., Soltis, D.E., Stevens, P.F., Westoby, M., Wright, I.J., Aarssen, L., Bertin, R.I., Calaminus, A., Govaerts, R., Hemmings, F., Leishman, M.R., Oleksyn, J., Soltis, P.S., Swenson, N.G., Warman, L. & Beaulieu, J.M. (2014). Three keys to the radiation of angiosperms into freezing environments. *Nature* 506: 89–92.

Zhang, G., Jarvis, E.D. & Gilbert, M.T.P. (2014). A flock of genomes. *Science* 346: 1308–1309.

Zhang, M.-L., Temirbayeva, K., Sanderson, S.C. & Chen, X. (2015). Young dispersal of xerophil *Nitraria* lineages in intercontinental disjunctions of the Old World. *Sci Rep* 5: 13840. Doi: 10.1038/srep13840 (accessed 6 April, 2016).