MOLECULAR ECOLOGY

Molecular Ecology (2015)

NEWS AND VIEWS

OPINION

Not the time or the place: the missing spatio-temporal link in publicly available genetic data

LISA C. POPE,* LIBBY LIGGINS,†‡ JUDE KEYSE,* SILVIA B CARVALHO§ and CYNTHIA RIGINOS*

*School of Biological Sciences, The University of Queensland, Brisbane, Qld 4072, Australia, †Allan Wilson Centre for Molecular Ecology and Evolution, New Zealand Institute for Advanced Study, Institute of Natural and Mathematical Sciences, Massey University, Auckland 0745, New Zealand, ‡Auckland War Memorial Museum, Tāmaki Paenga Hira, Auckland 1142, New Zealand, §CIBIO/InBIO – Centro de Investigação em Biodiversidade e Recursos Genéticos da Universidade do Porto, R. Padre Armando Quintas, 4485-661 Vairão, Portugal

Abstract

Genetic data are being generated at unprecedented rates. Policies of many journals, institutions and funding bodies aim to ensure that these data are publicly archived so that published results are reproducible. Additionally, publicly archived data can be 'repurposed' to address new questions in the future. In 2011, along with other leading journals in ecology and evolution, *Molecular Ecology* implemented mandatory public data archiving (the Joint Data Archiving Policy). To evaluate the effect of this policy, we assessed the genetic, spatial and temporal data archived for 419 data sets from 289 articles in *Molecular Ecology* from 2009 to 2013. We then determined whether archived data could be used to reproduce analyses as presented in the manuscript. We found that the journal's mandatory archiving policy has had a substantial positive impact, increasing genetic data archiving from 49 (pre-2011) to 98% (2011–present). However, 31% of publicly archived genetic data sets could not be recreated based on information supplied in either the manuscript or public archives, with incomplete data or inconsistent codes linking genetic data and metadata as the primary reasons. While the majority of articles did provide some geographic information, 40% did not provide this information as geographic coordinates. Furthermore, a large proportion of articles did not contain any information regarding date of sampling (40%). Although the inclusion of spatio-temporal data does require an increase in effort, we argue that the enduring value of publicly accessible genetic data to the molecular ecology field is greatly compromised when such metadata are not archived alongside genetic data.

Keywords: biological ontology, data accessibility, metadata, reproducibility, reuse, standards

Received 4 January 2015; revision received 7 May 2015; accepted 22 May 2015

Correspondence: Lisa C. Pope, Fax: +61 7 334 67646; E-mail: l.pope@uq.edu.au

Introduction

Molecular ecology is a rapidly growing field, and genetic data are being generated at an exponential rate (Kodama *et al.* 2012; Parr *et al.* 2012). Reliable archiving and public access to such data are essential to allow the reproducibility of published research to be assessed, a central tenet of science. Furthermore, data archives with public access can support the application of new statistical approaches, syntheses across studies, and allow the 'repurposing' of data, that is enabling researchers to address questions that differ from those for which the data were originally collected (Sidlauskas *et al.* 2010; Stoltzfus *et al.* 2012).

Molecular ecology encompasses a broad range of topics, illustrated by the sections of this journal. Questions aligning with different topics can often be addressed using the same genetic markers (see Fig. S1, Supporting information), providing extensive opportunities for genetic data to be repurposed in this field. Examples of data repurposing include the construction of megaphylogenies (e.g. the open tree of life project – http://blog.opentreeoflife.org/), delineating genetic 'hot spots' (e.g. Vandergast *et al.* 2008; Wood *et al.* 2012), testing the generality of the central-margin hypothesis (Eckert *et al.* 2008) and predicting the spread of invasive species (Gaither *et al.* 2013), to name but a few. The future value of spatio-temporal genetic data to investigating questions such as the impact of climate change, the ongoing biodiversity crisis and disease spread is incalculable.

In 2011, *Molecular Ecology* entered into the Joint Data Archiving Policy (http://datadryad.org/pages/jdap) motivated by low voluntary rates of public data archiving among contributors (Rausher *et al.* 2010); the JDAP calls for published studies to be reproducible and to facilitate data reuse (see Box 1). Several 'best-practice guides' and recommendations for the provisioning of genetic data and metadata have been contributed (e.g. Leebens-Mack *et al.* 2006; Whitlock 2011; White *et al.* 2013; Cranston *et al.* 2014). Multilocus genotypes and/or DNA sequences identifiable to the level of individual (i.e. individual-based genetic data) are preferable for assessing both the reproducibility of

Box 1. The Joint Data Archiving Policy (JDAP) describes a requirement that data supporting publications be publicly available. This policy was adopted in a joint and coordinated fashion by many leading journals in the field of evolution in 2011 (Rausher *et al.* 2010), and JDAP has since been adopted by additional journals across various disciplines.

Molecular ecology policy on data archiving

Data are important products of the scientific enterprise, and they should be preserved and usable for decades in the future. As such, *Molecular Ecology* requires authors to archive the data supporting their results and conclusions along with sufficient details so that a third party can interpret them correctly. Studies with exemplary data and code archiving are more valuable for future research and, all else being equal, will be given higher priority for publication. Data should be archived in an appropriate public archive, such as GenBank, Gene Expression Omnibus, TreeBASE, Dryad, the Knowledge Network for Biocomplexity, and your own institutional or funder repository, or as Supporting Information on the *Molecular Ecology* website. The utility of archived data is greatly enhanced when the scripts and input files used in the analyses are also made available. Given that scripts may be a mix of proprietary and freely available code, their deposition is not compulsory, but we nonetheless strongly encourage authors to make these scripts available whenever possible. As discussed by Whitlock *et al.* (2010), accurate interpretation of data will likely 'require a short additional text document, with details specifying the meaning of each column in the data set. The preparation of such shareable data sets will be easiest if these files are prepared as part of the data analysis phase of the preparation of the paper, rather than after acceptance of a manuscript'. For additional guidelines on data deposition best practice, please visit http://datadryad.org/depositing. Data must be publicly available at the time of publication. Embargos may be granted in exceptional instances at the discretion of the Managing Editors. Exemptions to this policy may also be granted, especially for sensitive information such as human subject data or the location of endangered species.

published research and reuse (discussed in Whitlock 2011). It is more difficult, however, to specify which metadata may be relevant for the reproducibility of a study and/or future repurposing of genetic data. Nonetheless, spatiotemporal information, such as the location and time of genetic sampling, is of central importance to most ecological and evolutionary studies, and their inclusion is certain to expand the scope for future data reuse or repurpose.

Despite the universal nature of geographic and temporal information, there is often no requirement for these metadata to be associated with genetic data by existing public databases, journals or institutions. For example, since 2005, DNA sequences submitted to the National Centre for Biotechnology Information (NCBI) have been encouraged, but not required, to contain geographic information. Voluntary inclusion of this information appears limited, with fewer than 7% of sequences for 'barcoding genes' submitted to NCBI since mid-2011 containing geographic coordinates (Marques *et al.* 2013). However, some data archives do require such metadata to be deposited and linked to genetic data (e.g. Metagenomics Analysis Server, http://press.igsb.anl.gov/mg-rast/metadata-in-mg-rast/, requires latitude and longitude, but not time; the NCBI Bioproject requires both spatial and temporal information, Dugan *et al.* 2014). How biological databases should cross-communicate and how genetic and biodiversity ontologies can support such information exchange has been an active topic of discussion and implementation. For example, Gene Ontology (GO) standards promote the exchange of information among the GO Consortium, including FlyBase, WormBase, J Craig Venter Institute and Mouse Genome Informatics (http://geneontology.org/page/go-consortium-contributors-list). Similarly, metadata standards such as the Darwin Core underpin biodiversity databases such as the Global Biodiversity Information Facility (GBIF) and the Ocean Biogeographic Information System (OBIS). The need for more inclusive standards that encompass all aspects of biodiversity, including genetic biodiversity, is recognized, and such standards are under development (Yilmaz *et al.* 2011; Walls *et al.* 2014).

At present, there are no universal genetic or biodiversity databases to suit the variety of studies published in *Molecular Ecology*, and NCBI and DataDryad are currently the most used archive facilities for this journal (based on studies reviewed herein). NCBI has provided an immense resource to molecular ecologists via their restricted entry format and active data quality-checking facilities, enabling uniform and programmatic data retrieval. On the other hand, DataDryad provides a platform that allows the user to provide any data in any format. While this flexibility is extremely useful, it becomes difficult to assess exactly which data have been archived without an informative, accompanying text or key. Furthermore, although geographic and temporal information is sometimes contained within a publication, differences between reference codes used in metadata and genetic data files can render this information useless. While it is possible to contact the original authors to obtain data, this approach was found to have low success in other areas of genetic research (Magee *et al.* 2014).

Increasingly, the importance of public data archiving is being recognized by funding bodies (e.g. National Institutes of Health 2003, National Science Foundation, Natural

Sciences and Engineering Research Council of Canada, National Environment Research Council of the UK, The Austrian Science Fund, the Deutsche Forschungsgemeinschaft in Germany, and Australian Research Council 2013), universities and journals (Moore *et al.* 2010; Fairbairn 2011; Baker 2013; Lin & Strasser 2014). A dramatic increase in public archiving of genetic data has resulted from these institutional policies (Vines *et al.* 2013). However, the importance of archiving associated metadata is less recognized. Consequently, while great progress has been made towards the public availability of genetic data, the lack of emphasis on provision of associated information, such as geographic location and time of sampling, may impede our ability to fully reproduce such studies or use their genetic data in new ways (Anonymous, 2008).

Given that *Molecular Ecology* was one of the first journals in evolution and ecology to adopt a mandatory public data access policy, here we gauge the impact of this policy on public archiving of genetic data. Then, given the importance of geographic and temporal information to many ecological and evolutionary studies, we evaluate the extent to which spatio-temporal data associated with genetic data are being made publicly available by the molecular ecology community. To do this, we examined manuscripts from 20 issues of *Molecular Ecology* from 2009 to 2013. For these articles, we determined the following: Have genetic data been made publicly available? Could the analyses presented be reproduced based on the total information made publicly available? Has geographic or temporal information been provided and at what scale? What is the scope for repurposing the associated data for future studies?

Methods

We examined all original articles containing newly generated genetic data from 20 issues of *Molecular Ecology* in the 2009 to 2013 period (two issues from July, and two from December each year), a total of 289 articles. Many articles utilized multiple markers, which were often archived differently. From 289 articles, we obtained 419 genetic data sets (one data set for each different genetic marker used in each manuscript) for which we assessed public archiving rates as well as the ability to 'reproduce' analyses based on the provisioning of genetic data and spatial and temporal sampling information.

We defined articles as having 'publicly available' genetic data if *any* genetic data were lodged in a public repository (e.g. NCBI, DataDryad) or provided as supplementary material on the journal website; thus, our criterion for publically available data was very lenient. We searched the text of each article for reference to public data archives and, for articles published after 2011, utilized the 'Data Accessibility' section. We recorded the following: genetic marker type, type of genetic data archived (e.g. individual, population level), and the location of the genetic data if it had been archived.

To address the 'reproducibility' of an article, we assessed whether the genetic data could be recreated and whether sufficient metadata had been provided such that *all* the analyses presented in that article could be reproduced. As different public archiving practices are often used for different genetic marker types (e.g. sequences versus microsatellites), and rates of public archiving have varied for different genetic marker types over time, genetic data were considered at the level of data set. For articles that included links or references to publically available genetic data files, we started by examining these files and applied a simple set of criteria to gauge whether it would be possible to *recreate* the original genetic data set(s): (i) reference codes used to identify individuals in the publically archived genetic data had to match those used in the manuscript or a linking file, and (ii) individual genetic data could be determined. For multilocus nuclear markers, if only summary allele frequencies were provided, the data set was not deemed recreatable as full genotypes with linkage relationships could not be inferred. For DNA sequence-based studies where only unique sequences were archived, data sets were only considered recreatable if haplotype frequencies and sample sizes were provided, allowing individual genetic data to be recreated. Haplotype information provided non-numerically, such as pie graphs, was not considered sufficient for reconstructing genotypes. For phylogenetic studies, a single sequence per species was sufficient to meet the recreatable genetic data set criterion. These criteria were selected based on common problems we had personally experienced when attempting to reanalyse data sets. We were conservative when designating a data set as not 'recreatable', and if there was any doubt, we assumed that the archived data could be used to recreate a data set.

The ability to recreate the relevant genetic data set(s) (using data set criteria described previously) was deemed essential for an article to be considered 'reproducible'. In some cases, article 'reproducibility' only required that the genetic data set(s) were recreatable, for instance when spatial and temporal information was irrelevant to the study objectives. In other cases, spatial and/or temporal information of an appropriate scale (i.e. metadata) was also required for the analyses to be reproducible. If these metadata were not provided at a sufficiently accurate scale to allow the presented analyses to be performed, the article was not classed as reproducible. Because we did not recreate the actual presented analyses, our assessment of complete article-level reproducibility is certain to be upwardly biased.

We examined the number of articles for which spatio-temporal metadata were provided, and assessed the precision of these data. We excluded a small number of articles for which it could be argued that geography was not relevant (e.g. laboratory/methodological/within-population studies); thus, 252 articles were examined for spatio-temporal metadata. As with genetic data, all publicly available materials, including the text of the article, supplementary text and publically archived data, were searched to determine whether geographic or temporal information was provided. When geographic information was present, we categorized its level of precision:

1 Where geographic information was provided as text only, we defined this as two categories: 'locality' and 'region'. Text was classified as 'region' if the area specified was 'large', for example ocean, country, state, region, or province; otherwise, it was classed as 'locality' (town, district, etc.).
2 Where coordinates were provided (latitude, longitude or UTM), we defined their precision using three categories: ~ > 100 km (degrees only); ~ 1 – 100 km (degrees up to two decimal places, or minutes); and ~ < 1 km (remainder).
3 Where geographic information was provided using an undefined coordinate system, if only a map was provided with no text, this was categorized as 'other'.

Where a record was provided of the time of sampling, we categorized the precision as: year range, year, and more accurate than a year (< year). In the majority of articles, the same spatio-temporal information applied to all data sets within the article. In the small number of cases where spatio-temporal information differed between data sets (9 of 252), the more precise spatio-temporal data were used.

Finally, to examine the potential to 'repurpose' archived data, we combined information on publicly archived genetic data sets that provided linking codes and individual genetic data ('recreatable' genetic data sets), with information on spatio-temporal metadata, for articles published post-JDAP. We plotted the number of recreatable data sets for which geographic and temporal information was provided at various levels of precision, providing an indication of the extent to which genetic data sets and accompanying spatio-temporal data are available for 'repurposing'.

Results

Many articles contained multiple data sets and used more than one public database to store their genetic data. Nearly half of the articles examined stored some data in NCBI (147/289; 47%). The majority of DNA sequence data sets were stored in NCBI (133/156; 85%), and 45% of these included some kind of geographic information in NCBI itself (60 of 156 sequence data sets), although this was generally 'country'. The second most commonly used public database was DataDryad (112/289; 39%).

Public archiving of data increased greatly over the 5-year period examined, starting at 49% and ending at 98% (based on 289 articles, Fig. 1a). This gain was primarily due to increased public archiving of non-DNA sequence data such as microsatellite and SNP genotypes (Fig. S2, Supporting information). While public archiving rates improved over time, increasing the overall proportion of genetic data sets that could be 'recreated', other issues remained steady: in addition to data sets with no public archiving (72/419; 12%), 19% failed to provide individual level genetic data (79/419) and 10% did not provide a link between codes used in the manuscript and those used in the archived data (43/419), with some studies failing to provide both codes and individual genotypes (Fig. 1b). In summary, 31% of genetic data sets that have nominally been publically archived could not be recreated (106/347). Articles evaluated as presenting completely reproducible analyses also increased over time, although again the proportion was not as great as might be expected given public data archiving rates (18–72%; Fig. 1a). Restricting consideration to only those articles that included public data archiving (242 articles), only 41% (100) presented fully reproducible analyses.

In contrast to the gains in genetic data archiving, the provisioning of geographic and temporal data changed little from 2009 to 2013 (Fig. 1c, d). All articles for which geography was deemed relevant provided geographic information of some kind. However, over a third of articles provided geographic information as text only (90/252, 36%), with 18% describing geography in the text at a regional-level only (ocean, country, state, region or province; 45/252). Only 60% of articles provided geographic coordinates (151/252). There has been an increase in the level of precision of geographic coordinates when provided (<1 km increased from 29 to 46%); however, the overall rate of latitude and longitude reporting has remained steady (Fig. 1c). Similarly, reporting of time of sampling remained fairly constant (Fig. 1d). Around 40% of articles did not provide any temporal information (100/252), and many provided only a range of years (50/252, 20%). Thus, only 40% of articles (102/252) reported year of sampling (or greater precision).

For genetic data sets from 2011 onwards that were able to be recreated (178 from a total of 228 geographically relevant data sets, Fig. 2a), Fig. 2(b) illustrates the varying levels of precision of archived spatio-temporal metadata. The proportion of data sets available for repurposing will vary depending on the spatio-temporal needs of the new study. If temporal information is not required and if authors are willing to use locality text information, in addition to geographic coordinates, a large proportion of recreatable data sets could be reused (83%; 148/178). However, if latitude and longitude are required, fewer data sets are repurposable (64%; 115/178), and if latitude and longitude along with year of sampling or better are desired, a much smaller pool of data sets are available for repurposing (21%; 35/178).

Discussion

Policies mandating public data archiving have clearly increased archiving of genetic data, as shown in Fig. 1 (see also Vines *et al.* 2013). These developments in *Molecular Ecology* align with a sociological shift towards data sharing in ecology and evolution (e.g. Jones *et al.* 2006; Poisot *et al.* 2013; and discussed in Constable *et al.* 2010). In general, researchers in the fields of molecular ecology, phylogenetics and genomics have accepted this practice as fundamental to the requirement of reproducibility in science. Several institutions exist with the intention of making molecular genetic data publically accessible (e.g. NCBI, EBI, datadryad.org, http://www.free-the-data.org/ etc.); however,

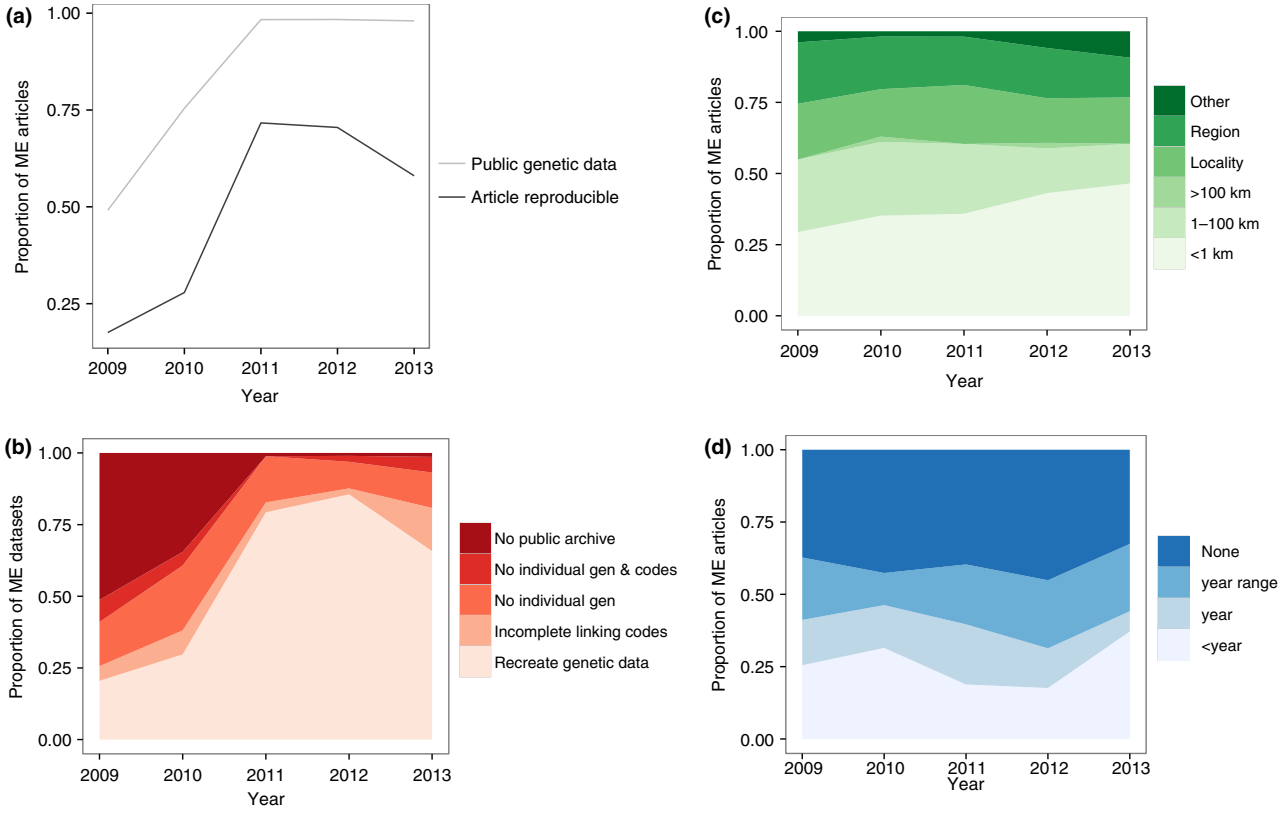

Fig. 1 Public availability of genetic data and associated spatio-temporal information. Results are based on all articles in both July and December issues for the year indicated (289 articles). (a) Rates of public data archiving and 'reproducibility' over time. 'Public genetic data' refers to the deposition of any kind of genetic data in a public database or publication. 'Article reproducible' refers to articles for which sufficient information was provided such that all analyses presented in the original article could be reproduced (see text for a full definition). (b) Rates of genetic data set archiving over time (n = 419). Genetic data sets were assessed as 'recreatable', and if not, the reason for this was classed as either: no public archive of genetic data, incomplete linking codes, no individual genetic data, or both no linking codes and no individual genetic data. c) Precision of geographic information in articles over time. For articles where geography was relevant (i.e. not captive bred/experimental, a total of 252 articles), we determined with what precision geographic information was made available: 'nothing/other', text only regional, text only locality, or coordinates > 100 km, 1–100 km, and <1 km precision. d) Precision of time of sampling information over time. Articles where time of sampling was relevant (252 articles) were graded on the amount of information provided: nothing, a range of years, year, and greater precision than a year.

data utility and linkages to other biodiversity databases are limited by submission formats and ontologies (Yilmaz *et al.* 2011; Walls *et al.* 2014). We argue that in order for molecular ecological data to be truly accessible to the public, at a minimum, individual genotypes should be recoverable and linked to geographic and temporal information. Our study indicates that voluntary rates of supplying this information could be substantively improved (post-JDAP introduction in 2011, 21% of genetic data sets could not be recreated, 45% of data sets provided no temporal information, and 40% no geographic coordinates). Both the JDAP and the *Molecular Ecology* policy on data archiving (see Box 1) emphasize that all data supporting the publications be available (not just genetic data files), which is consistent with our opinion. Thus, we suspect that shortcomings in full implementation stem from misinterpretation of these data archiving policies, difficulties in cross-referencing without clear standards or appropriately structured databases, unintentional oversights of busy people, and poor (self-) regulation of the field. Undoubtedly, we are also personally guilty of inadequate data archiving.

There are many reasons why spatio-temporal metadata may not be associated with genetic data. The location of samples may not be deemed relevant, such as for captively reared or artificially selected organisms. In other cases, the original time and place of genetic sampling will be unknown. This might occur where samples have been 'inherited' from previous projects, deposited in museums with locality unknown, or collected in such a manner that a precise locality cannot be determined (e.g. markets). Occasionally, locations of endangered species or sites of archaeological importance might be withheld from public release (see Rausher *et al.* 2010).

In other cases, the place and time of genetic sampling are known, but these metadata are not publically archived. Data submission can be a lengthy process, and fast ways to

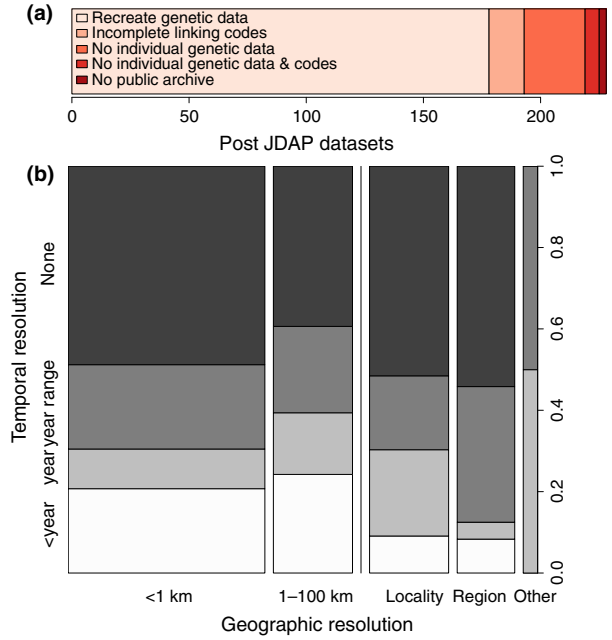

Fig. 2 Public archive rates, spatio-temporal data availability and opportunities for 'repurposing' of data sets (from 228 articles for which geography was relevant to the article objective published post-JDAP): (a) This bar represents the number of data sets with: no publically archived metadata (3/228), insufficient information to recreate individual genetic data (26/228), did provide individual genetic data, but without codes to link all the provided data together to recreate genetic data (16/228), both no linking codes and no individual genetic data (6/228). (b) The remaining 'recreatable' data sets ($n = 178$) are presented in terms of the precision of the available spatio-temporal data to indicate the data available for repurposing based on a researcher's requirements. Categories used are described in Fig. 1.

include metadata are often not obvious. Many popular population genetic data formats have no method for appending metadata (e.g. Arlequin, GENEPOP, Nexus and STRUCTURE formats). Some formats do, however, allow the inclusion of spatial information (GENALEX, Peakall & Smouse 2006; geneGIS, Dick *et al.* 2014; TESS, Chen *et al.* 2007). Unfortunately, for many of the loci employed in molecular ecology studies (especially microsatellites and AFLPs), there are no standard data repositories; thus, extra care is required in preparing archived files for these data. Additionally, because manuscript acceptance is typically decoupled from public data submission, often changes in the reference codes linking genetic data to metadata creep in during the revision process.

To improve the standards of public genetic data and spatio-temporal metadata in our field will require the effort of all parties: authors, reviewers, journals, institutions, public data repositories and the *Molecular Ecology* community as a whole. Based on the results from this study, we recommend that best practice for genetic data archiving for most *Molecular Ecology* studies (consistent with the JDAP) should include the following: (i) genetic data files that present individual genotypes, (ii) unified reference codes identifying individuals across any archived data sets from a single publication, (iii) year (ideally date) of sampling, and (iv) sampling locations provided as geographic coordinates.

In particular, we stress the need for higher community standards regarding geographic reporting with the expectation that spatial information be provided as georeferenced coordinates (presently missing in 40% of examined articles). The best practice for spatial data should include both a text description of the locality and geographic coordinates (including a description of the system used), as several location names are shared worldwide (e.g. Bird Rock), and/or only locally known (e.g. Bob's corner; discussed in a Anonymous 2008). In some locations, the use and precision of GPS is limited due to signal weakness and/or disturbance (i.e. underwater, under forest canopy, little satellite coverage). However, in such situations, geographic coordinates can be complemented with an estimate of spatial uncertainty. Tools to facilitate the estimation of uncertainty are available (e.g. GeoLocate, http://www.museum.tulane.edu/geolocate/) and are already incorporated in record keeping protocols for other forms of biological data (e.g. VertNet, Constable *et al.* 2010).

Preferably, genetic data should be deposited in appropriate repositories, rather than as supplemental files, which have been shown to decay over time (Evangelou *et al.* 2005; Anderson *et al.* 2006). Structured repositories with controlled ontologies can be efficiently queried and searched by end–users, and there are growing efforts to link genetic and other biodiversity databases via shared ontologies (see Introduction), preserving long-term value to the field. Where possible, spatial and temporal information should accompany database submissions; for example, these data can be included in NCBI records. In the short term, however, many *Molecular Ecology* data will not find an obvious home in a structured repository, and thus, flexible methods of data archiving (such as DataDryad) are extremely valuable. We recommend that authors prepare files in line with the recommendations listed above (utilizing commonly used genotype based files, consistent codes, date and location of sampling) along with an overarching readme file (see Whitlock *et al.* 2010) and review these files at the time of final submission. A quick check that these minimal elements are available could be undertaken by handling editors (based on our reviews of studies here, we found that well-prepared files can be summarily checked in a few minutes).

The last 5 years have shown a massive increase in the public archiving of genetic data. Despite these positive developments, many of the studies published in *Molecular Ecology* today are not reproducible, a central tenet of public archiving. *Molecular Ecology* represents one of the leaders in the call for essential data archiving, so this situation is likely worse for journals without clear and enforced data access policies. Additionally, many studies do not include geographic coordinates, or even year of sampling, restricting the future reuse of these genetic data. We advocate 'a

higher expectation for the quality and quantity of descriptive data' (Field 2008). How this is best achieved is open to debate. Whether higher rates and quality of spatio-temporal data can be achieved through raised awareness and standards, without explicit mandates, remains to be seen. We do know that careful archiving of genetic data with associated spatio-temporal data *now* will result in a much more valuable legacy for future research. To fully realize the future potential of this data legacy, there should now be a greater push to link spatio-temporal metadata to genetic data and to develop standards and repositories that facilitate data deposition, curation and searchability.

Acknowledgements

LCP was funded by a University of Queensland Women's Postdoctoral Research Fellowship. LL was funded by a New Zealand Allan Wilson Centre for Molecular Ecology and Evolution Postdoctoral Research Fellowship. SBC was funded by a postdoctoral grant from Fundacão para a Ciência e Tecnologia (FCT) (SFRH/BPD/74423/2010), and through the project PTDC/BIA-BIC/118624/2010-FCOMP-01-0124-FEDER-019676, supported by FEDER funds through the Operational Programme for Competitiveness Factors – COMPETE and by National Funds through FCT. The authors thank J. Sheehan with assistance in retrieving information from manuscripts.

References

Anderson NR, Tarczy-Hornoch P, Bumgarner RE (2006) On the persistence of supplementary resources in biomedical publications. *BMC Bioinformatics*, **7**, 260.

Anonymous (2008) A place for everything. *Nature*, **453**, 2.

Australian Research Council (2013) *ARC Open Access Policy - Version 2013.1*. Australian Research Council AG, Canberra, ACT, Australia.

Baker CS (2013) Journal of heredity adopts joint data archiving policy. *Journal of Heredity*, **104**, 1.

Chen C, Durand E, Forbes F, François O (2007) Bayesian clustering algorithms ascertaining spatial population structure: a new computer program and a comparison study. *Molecular Ecology Notes*, **7**, 747–756.

Constable H, Guralnick R, Wieczorek J *et al.* (2010) VertNet: a new model for biodiversity data sharing. *PLoS Biology*, **8**, e1000309.

Cranston K, Harmon LJ, O'Leary MA, Lisle C (2014) Best practices for data sharing in phylogenetic research. *PLOS Currents Tree of Life*, **6**. doi: 10.1371/currents.tol.bf01eff4a6b60ca4825c69293dc59645.

Dick DM, Walbridge S, Wright DJ *et al.* (2014) geneGIS: Geoanalytical Tools and Arc Marine Customization for Individual-Based Genetic Records. *Transactions in GIS*, **18**, 324–350.

Dugan VG, Emrich SJ, Giraldo-Calderon GI *et al.* (2014) Standardized metadata for human pathogen/vector genomic sequences. *PLoS ONE*, **9**, e99979.

Eckert CG, Samis KE, Lougheed SC (2008) Genetic variation across species' geographical ranges: the central-marginal hypothesis and beyond. *Molecular Ecology*, **17**, 1170–1188.

Evangelou E, Trikalinos TA, Ioannidis JP (2005) Unavailability of online supplementary scientific information from articles published in major journals. *FASEB Journal*, **19**, 1943–1944.

Fairbairn DJ (2011) The advent of mandatory data archiving. *Evolution*, **65**, 1–2.

Field D (2008) Working together to put molecules on the map. *Nature*, **453**, 978.

Gaither MR, Bowen BW, Toonen RJ (2013) Population structure in the native range predicts the spread of introduced marine species. *Proceedings of the Royal Society B: Biological Sciences*, **280**, 20130409.

Jones MB, Schildhauer MP, Reichman OJ, Bowers S (2006) The new bioinformatics: integrating ecological data from the gene to the biosphere. *Annual Review of Ecology and Systematics*, **37**, 519–544.

Kodama Y, Shumway M, Leinonen R, International Nucleotide Sequence Database (2012) The Sequence Read Archive: explosive growth of sequencing data. *Nucleic Acids Research* **40**, D54–D56.

Leebens-Mack J, Vision T, Brenner E *et al.* (2006) Taking the first steps towards a standard for reporting on phylogenies: minimum information about a phylogenetic analysis (MIAPA). *OMICS: A Journal of Integrative Biology*, **10**, 231–237.

Lin J, Strasser C (2014) Recommendations for the role of publishers in access to data. *PLoS Biology*, **12**, e1001975.

Magee AF, May MR, Moore BR (2014) The dawn of open access to phylogenetic data. *PLoS ONE*, **9**, e110268.

Marques AC, Maronna MM, Collins AG (2013) Putting GenBank data on the map. *Science*, **341**, 1341.

Moore AJ, McPeek MA, Rausher MD, Rieseberg L, Whitlock MC (2010) The need for archiving data in evolutionary biology. *Journal of Evolutionary Biology*, **23**, 659–660.

National Institutes of Health (2003) *NOT-OD-03-032: Final NIH Statement on Sharing Research Data*. Department of Health and Human Services National Institutes of Health Office of Extramural Research, Rockville, Maryland.

Parr CS, Guralnick R, Cellinese N, Page RD (2012) Evolutionary informatics: unifying knowledge about the diversity of life. *Trends in Ecology and Evolution*, **27**, 94–103.

Peakall ROD, Smouse PE (2006) Genalex 6: genetic analysis in Excel. Population genetic software for teaching and research. *Molecular Ecology Notes*, **6**, 288–295.

Poisot T, Mounce R, Gravel D (2013) Moving toward a sustainable ecological science: don't let data go to waste!. *Ideas in Ecology and Evolution*, **6**, 11–19.

Rausher MD, McPeek MA, Moore AJ, Rieseberg L, Whitlock MC (2010) Data archiving. *Evolution*, **64**, 603–604.

Sidlauskas B, Ganapathy G, Hazkani-Covo E *et al.* (2010) Linking big: the continuing promise of evolutionary synthesis. *Evolution*, **64**, 871–880.

Stoltzfus A, O'Meara B, Whiteacre J *et al.* (2012) Sharing and re-use of phylogenetic trees (and associated data) to facilitate synthesis. *BMC Research Notes*, **5**, 574.

Vandergast AG, Bohonak AJ, Hathaway SA, Boys J, Fisher RN (2008) Are hotspots of evolutionary potential adequately protected in southern California? *Biological Conservation*, **141**, 1648–1664.

Vines TH, Andrew RL, Bock DG *et al.* (2013) Mandated data archiving greatly improves access to research data. *FASEB Journal*, **27**, 1304–1308.

Walls RL, Deck J, Guralnick R *et al.* (2014) Semantics in support of biodiversity knowledge discovery: an introduction to the biological collections ontology and related ontologies. *PLoS ONE*, **9**, e89606.

White EP, Baldridge E, Brym ZT *et al.* (2013) Nine simple ways to make it easier to (re)use your data. *Ideas in Ecology and Evolution*, **6**, 1–10.

Whitlock MC (2011) Data archiving in ecology and evolution: best practices. *Trends in Ecology and Evolution*, **26**, 61–65.

Whitlock MC, McPeek MA, Rausher MD, Rieseberg L, Moore AJ (2010) Data archiving. *The American Naturalist*, **175**, 145–146.

Wood DA, Vandergast AG, Barr KR *et al.* (2012) Comparative phylogeography reveals deep lineages and regional evolutionary hotspots in the Mojave and Sonoran Deserts. *Diversity and Distributions*, **19**, 722–737.

Yilmaz P, Kottmann R, Field D *et al.* (2011) Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. *Nature Biotechnology*, **29**, 415–420.

All authors were involved in formulating the ideas for this manuscript, data entry and writing the manuscript. CR instigated the manuscript and performed the R analyses on 'reproducible articles'. LP performed the majority of analyses and coordinated the final text. All the authors are interested in comparative spatial genetics and use repurposed data in their research. LCP, JK and CR are usually found at -27.498 153.012, LL at -36.734 174.703 and SCB at 41.181 -8.602 (WGS84).

doi: 10.1111/mec.13254

Data accessibility

An Excel spreadsheet of all the *Molecular Ecology* articles and data sets included in our analyses, along with a readme file, R script and R data file, are available on Dryad doi:10.5061/dryad.kg943.2

Supporting information

Additional supporting information may be found in the online version of this article.

Fig. S1 Use of different marker types across *Molecular Ecology* sections.

Fig. S2. The number of data sets without publically archived data, as a function of genetic marker type.