MOLECULAR ECOLOGY

Molecular Ecology (2015) **24**, 1164–1171

NEWS AND VIEWS

OPINION

The evolution of phylogeographic data sets

RYAN C. GARRICK,* ISABEL A. S. BONATELLI,† CHAZ HYSENI,* ARIADNA MORALES,‡ TARA A. PELLETIER,‡ MANOLO F. PEREZ,† EDWIN RICE,‡ JORDAN D. SATLER,‡ REBECCA E. SYMULA,* MARIA TEREZA C. THOMÉ§ and BRYAN C. CARSTENS‡

**Department of Biology, University of Mississippi, Oxford, MS 38677, USA, †Departamento de Biologia, Universidade Federal de São Carlos, Campus Sorocaba, Caixa Postal 18052780, Sorocaba, São Paulo, Brazil, ‡Department of Evolution, Ecology and Organismal Biology, The Ohio State University, 318 W. 12th Avenue, Columbus, OH 43210-1293, USA, §Departamento de Zoologia, Instituto de Biociências, UNESP – Univ Estadual Paulista, Campus Rio Claro, Caixa Postal 19913506-900, Rio Claro, São Paulo, Brazil*

Empirical phylogeographic studies have progressively sampled greater numbers of loci over time, in part motivated by theoretical papers showing that estimates of key demographic parameters improve as the number of loci increases. Recently, next-generation sequencing has been applied to questions about organismal history, with the promise of revolutionizing the field. However, no systematic assessment of how phylogeographic data sets have changed over time with respect to overall size and information content has been performed. Here, we quantify the changing nature of these genetic data sets over the past 20 years, focusing on papers published in *Molecular Ecology*. We found that the number of independent loci, the total number of alleles sampled and the total number of single nucleotide polymorphisms (SNPs) per data set has improved over time, with particularly dramatic increases within the past 5 years. Interestingly, uniparentally inherited organellar markers (e.g. animal mitochondrial and plant chloroplast DNA) continue to represent an important component of phylogeographic data. Single-species studies (cf. comparative studies) that focus on vertebrates (particularly fish and to some extent, birds) represent the gold standard of phylogeographic data collection. Based on the current trajectory seen in our survey data, forecast modelling indicates that the median number of SNPs per data set for studies published by the end of the year 2016 may approach ~20 000. This survey provides baseline information for understanding the evolution of phylogeographic data sets and underscores the fact that development of analytical methods for handling very large genetic data sets will be critical for facilitating growth of the field.

Keywords: DNA sequences, information content, phylogeography, sampling, single nucleotide polymorphisms, temporal trends

Received 31 October 2014; revised 21 January 2015; accepted 4 February 2015

Correspondence: Ryan C. Garrick, Fax: +1 662-915-5144; E-mail: rgarrick@olemiss.edu and Bryan C. Carstens, Fax: +1 614-292-2030; E-mail: carstens.12@osu.edu

Introduction

Phylogeographers have been working to collect multilocus data ever since a series of theoretical papers pertinent to the discipline demonstrated that estimates of key demographic parameters improve as the number of loci increases (e.g. Edwards & Beerli 2000; Hey & Nielsen 2004; Felsenstein 2006; Carling & Brumfield 2007). Recent improvements in DNA sequencing technology have led to platforms with greater speed, resolution and/or output (e.g. Margulies *et al.* 2005; Bentley *et al.* 2008; Rothberg *et al.* 2011) when compared to the traditional Sanger method. These technological advances, together with the development of general-purpose protocols for discovering and screening many DNA sequence polymorphisms arrayed across a species' genome (e.g. Baird *et al.* 2008; Kerstens *et al.* 2009; Faircloth *et al.* 2012; Peterson *et al.* 2012), are transforming the field of phylogeography to one that is no longer data limited. Investigations concerned with reconstructing long-term population history generally require large numbers of sampled alleles (i.e. many individuals and populations), across multiple loci, to adequately characterize levels of diversity and spatial genetic structuring (McCormack *et al.* 2013), and these data are now in reach. However, for moderately recent timescales (e.g. at or near the Last Glacial Maximum), it is clear that the number of individuals sampled is also important (Robinson *et al.* 2014a) and that the ideal ratio of loci to individuals depends on the question at hand (Maddison & Knowles 2006). Furthermore, given that phylogeographic studies often focus on organisms for which few or no genomic resources exist (e.g. Avise 2000), the new technical developments enable detailed investigations of non-model species and expand the complexity and scope of questions that can be addressed. For example, sampling of genomewide DNA sequence polymorphisms strengthens

our ability to distinguish between recent vs. historical migration and admixture (Pool & Nielsen 2009), and yields previously unattainable insights into both neutral and selective processes (Gompert *et al.* 2014).

Even prior to the recent application of next-generation sequencing in phylogeography, a transformation in the basic approaches used to draw historical inferences from molecular data was underway (Brito & Edwards 2009; Hickerson *et al.* 2010). Awareness of the limitations of single locus data sets had grown during the early- to mid-2000s (Hare 2001; Arbogast *et al.* 2002; Brumfield *et al.* 2003). For one, the possibility of undetected selection influencing historical demographic inferences was particularly problematic when studies relied on data from a single protein-coding locus. Indeed, the potential for selective sweeps to affect animal mitochondrial DNA has been raised repeatedly over the past decade (albeit not without contention; Bazin *et al.* 2006), and evidence for non-neutrality of plant chloroplast DNA is now also emerging (Bock *et al.* 2014). Perhaps most alarming was the realization that inherent locus-to-locus stochasticity in levels of polymorphism and the extent of lineage sorting can be extreme, such that a single locus represents just one of many possible realizations of a given demographic history (Knowles 2004). These concerns were partly addressed by the development of several coalescent-based analytical methods that provided an opportunity to integrate across loci when estimating population parameters and testing a priori hypotheses (Beerli & Felsenstein 2001; Hey & Nielsen 2004). These analytical advances, together with the concurrent development of transferrable single copy nuclear markers (Jarman *et al.* 2002; Backström *et al.* 2008; Spinks *et al.* 2010), promoted a scaling-up of the number of loci included in empirical data sets (Beheregaray 2008). However, while it appears that the number of loci assayed in phylogeographic studies has increased over the years (but see Turchetto-Zolet *et al.* 2012), we know of no systematic assessment of how such data sets have changed with respect to overall size and information content. As a corollary, the extent to which animal mitochondrial DNA (mtDNA) and plant chloroplast DNA (cpDNA) still play central roles also remains unquantified.

In this paper, we aim to encourage discussion about the current state and future directions of phylogeography by examining temporal changes in the composition of genetic data sets on which historical inferences are based. To do this, we surveyed a representative sample of empirical studies published over the past two decades. We concentrated on papers in *Molecular Ecology* as this journal has been a leading venue for phylogeographic investigations, and because these papers should not have an inherent taxonomic bias. Furthermore, we focus on DNA sequence data, broadly defined as being generated by assays that determine the identity of each nucleotide in a targeted genomic region. Compared to other classes of molecular markers, DNA sequence haplotypes and single nucleotide polymorphisms (SNPs) should be more informative about historical events and processes (Sunnucks 2000; Brumfield *et al.* 2003) operating over timescales most relevant to the discipline.

Approach

Literature survey

A database of potentially relevant papers was established via search functions on the publisher's website (www.onlinelibrary.wiley.com), using the term *'phylogeograph*'* occurring anywhere in an article's title, keywords, abstract or main text. This sampling was performed at 3-year intervals (i.e. 1992, 1995, 1998, ... 2013) and returned >1200 hits, which were then divided into sets and distributed among members of our laboratory groups. Papers under consideration were limited to those clearly identifiable as empirical phylogeographic research. As a working definition, we considered this to be any study concerned with the characterization of genetic diversity as a function of a species' geographic range and landscape context, motivated by a desire to make inferences regarding the historical demography of the focal taxon. Excluded papers were typically literature reviews, toolset and method development papers, strictly phylogenetically oriented papers and those focused on contemporary processes only.

The following information, partitioned by species (and by organellar vs. autosomal chromosomes when appropriate), was extracted from qualifying papers: (i) the total number of independent loci assayed, with the entire mtDNA (or cpDNA) genome treated as a single haploid locus; (ii) the total number of alleles sampled, where identical alleles contribute towards the count; (iii) the total length (base pairs) of DNA sequence generated, as measured by author-reported alignment lengths summed across loci; and (iv) the total number of SNPs identified (i.e. including parsimony-uninformative sites), summed across loci. For each data set, we also recorded whether it came from a paper that presented data from a single species vs. multiple codistributed species. In addition, we recorded the taxonomic group to which the focal species belongs, broadly classified as vertebrate, invertebrate, plant or 'other' (i.e. fungi, protists, algae and bacteria, grouped together due to low sample sizes). Unfortunately, we were not able to extract information on the geographical scale of sampling employed in surveyed studies because spatial coordinates were reported in an array of different (often unspecified) units and map projections, and with marked differences in resolution. Similarly, the types of research questions being addressed and approaches used were not readily amenable to meta-analysis; there was considerable variability in explicit reporting of which historical events or microevolutionary processes were under investigation, and it was not always clear which analyses contributed towards phylogeographic inferences. In retrieving the targeted data, we relied on the veracity of the information published in the main text and did not download Supporting Information (e.g. online appendices, or sequences archived in public databases) to verify counts of sequence length or SNPs presented in the manuscript.

When assembly of the final database was complete, 508 single-species data sets, drawn from 370 papers, were

included (Tables 1 and S1, Supporting information). The full list of data sources is given in Supporting Information, and the database is available on DRYAD (http://datadryad.org). Despite the need to exclude from downstream analyses those database entries that were recorded as 'not reported', our sample sizes were sufficient to discern trends regarding changes in the composition of phylogeographic data sets over time. Ultimately, our survey data were used to identify the magnitude and nature of improvements in empirical phylogeographic data sets over the past 20 years, the time point(s) associated with the most dramatic changes and which (if any) taxonomic group had the strongest role in leading the field forward. Our survey data also served as a basis for performing a coarse forward projection relating to the anticipated number of SNPs per data set for studies published in the near future (i.e. through to the end of the year 2016).

Analyses

Recorded variables were classified as either providing a measure of *data set size* (i.e. number of loci assayed, total number of alleles sampled), or being indicative of *potential information content* embedded within a given data set (i.e. number of SNPs). Together, our measures of data set size should reflect the extent of genomic and geographic sampling. In the case of information content, each SNP provides the basis on which coalescent (or other) genealogies can be constructed, and in turn, these generate estimates of key demographic parameters. For this category, we also considered a weighted metric, in which an increase in the overall value of a species' genetic data set was scaled by the number of independent loci from which alleles had been sampled (calculated as number of loci × total number alleles sampled). Although there can be several ways to achieve the same weighted value under this scheme (e.g. 1 locus × 1000 total alleles would be equivalent to 10 loci × 100 total alleles), this nonetheless provides a basic framework for considering the precision of allele frequency estimates, and the extent to which inherent among-locus variance can be appreciated. In attempting to understand how the role of organellar DNA markers in phylogeography has changed over time, we calculated the per cent reduction in the weighted metric that resulted from removing mtDNA or cpDNA loci from any given data set. This was examined in vertebrates, invertebrates and plants only.

Initially, we examined scatter plots of each recorded variable vs. year and used quantile–quantile probability plots to assess normality. As our goal was to fit a linear model in order to determine how well year predicts changes in data set size and information content, a $\log_{10}$ transformation was applied to all variables that describe genetic aspects of surveyed data sets. When examining changes in the number of loci (total, or autosomal loci only), a $1 + \log_{10}$ transformation was applied due to the presence of zeros in some cases. Subsequently, we examined all temporal trends using simple linear regression, owing to ease of interpretability.

To generate predictions about the information content of phylogeographic data sets in the near future, we used an automatic forecasting algorithm implemented in the 'FORECAST' package (Hyndman & Khandakar 2008) in R (R Development Core Team 2014). We focused on the total number of SNPs in a data set because this variable is commonly used by authors to report the success of applications of next-generation sequencing in phylogeography (e.g. Emerson *et al.* 2010; McCormack *et al.* 2012; Zellmer *et al.* 2012). Each data set from our survey that reported the total number of SNPs was ordered chronologically and then in ascending order within year. To generate predictions about future data sets, we used autoregressive integrated moving average (ARIMA) modelling. This includes three main parameters: autoregression, differencing and moving average. Following Hyndman & Khandakar (2008), we first determined the differencing parameter using the Kwiatkowski–Phillips–Schmidt test of stationarity (Kwiatkowski *et al.* 1992), and then used stepwise model selection based on the Akaike information criterion to determine the other two parameters. For the best model, parameter values were as follows: autoregression = 9, differencing = 2 and moving average = 0 (i.e. the latter could be effectively ignored). Accordingly, the number of SNPs reported for each data set was transformed using two previous data sets to compute differences $(y''_t = (y_t - y_{t-1}) - (y_{t-1} - y_{t-2})$, where y_t = number of SNPs at time t), in order to make the time series stationary. Using autoregression, the number of SNPs in each phylogeographic data set was predicted based on the number of SNPs in the nine previous data sets. After we fit this ARIMA model to the observed data, we extrapolated through to the end of the year 2016, with confidence intervals obtained following Hyndman *et al.* (2005).

Table 1 Summary of the number of papers examined and the number of papers and data sets retained for inclusion the literature survey, partitioned by year

Year	No. of papers with search term match	No. of relevant papers retained*	No. of single-species data sets
1992	18	3	3
1995	20	6	6
1998	48	17	23
2001	130	29	41
2004	195	60	69
2007	284	108	147
2010	266	71	121
2013	269	76	98
Total	1230	370	508

*Papers retained after quality control are listed in Supporting Information.

Trends over time

All taxa and study types combined

Over the past two decades, phylogeographic data sets have become progressively larger in size. This is evident in the

number of loci assayed (slope = 0.036, R^2 = 0.117, d.f. = 1,506, F = 66.793, P < 0.001) and the total number of alleles sampled (slope = 0.043, R^2 = 0.078, d.f. = 1,495, F = 41.781, P < 0.001). The magnitude of increase is substantial, with the median number of loci per data set tripling over the past 6 years (i.e. from 1 to 3 loci in 2007–2013). The median number of alleles sampled showed a similar increase over the same time frame (i.e. from 119 to 318 alleles). For the latter variable, the positive trend is attributable to the addition of autosomal markers, as the number of alleles sampled from organellar loci showed no significant change over time (P = 0.223). Even if phylogeographic studies based on data from only a single organellar locus are omitted from the set, the trend of increased sampling of alleles over time is still apparent (slope = 0.027, R^2 = 0.031, d.f. = 1,146, F = 4.972, P = 0.027). Taken together, increases in the number of loci and alleles sampled indicate that researchers have been allocating greater effort to both genomic and geographic sampling. While the development of analytical methods tailored towards multilocus DNA sequence datasets (e.g. Heled & Drummond 2008, 2010) would have influenced the genomic sampling strategy employed by empiricists, it appears that previous (Templeton 1998) and renewed (Fitzpatrick 2009) emphases on adequate geographic sampling have not been overlooked as a consequence.

Indicators of data set information content also showed notable improvements over time. For example, the total number of SNPs has increased (slope = 0.025, R^2 = 0.051, d.f. = 1,271, F = 14.667, P < 0.001), at a magnitude of ~1.3× over the past 6 years (i.e. median of 70 vs. 90 SNPs in 2007 vs. 2013). We considered the possibility that this trend may have been influenced by new methods for sequencing complete organellar genomes, as these have recently been applied in phylogeographic investigations (e.g. vertebrates, Morin *et al.* 2010; Keis *et al.* 2013; invertebrates, Winkelmann *et al.* 2013; plants, Mariac *et al.* 2014). However, gains in the number of SNPs per data set were not simply due to sequencing larger regions of effectively the same locus, as the number of organellar SNPs showed no significant change over time (P = 0.603). Our weighted metric, the product of loci × alleles, also showed a positive trend (slope = 0.078, R^2 = 0.103, d.f. = 1,495, F = 56.525, P < 0.001; Fig. 1), with a particularly marked increase in 2013. This probably reflects the influence of next-generation sequencing, which has been predicted to revolutionize the field of phylogeography (Carstens *et al.* 2012; Andrew *et al.* 2013).

The use of animal mtDNA and plant cpDNA sequence data as cornerstones of phylogeographic inference has decreased considerably over the years (slope = −0.025, R^2 = 0.115, d.f. = 1,479, F = 62.091, P < 0.001). This is not surprising, given the burgeoning sources of polymorphic nuclear genetic data (see Introduction). However, few data sets were comprised only of autosomal DNA sequences. This may reflect the importance of well-established molecular evolutionary rate estimates that exist for organellar protein-coding genes (Hewitt 2001), as these facilitate

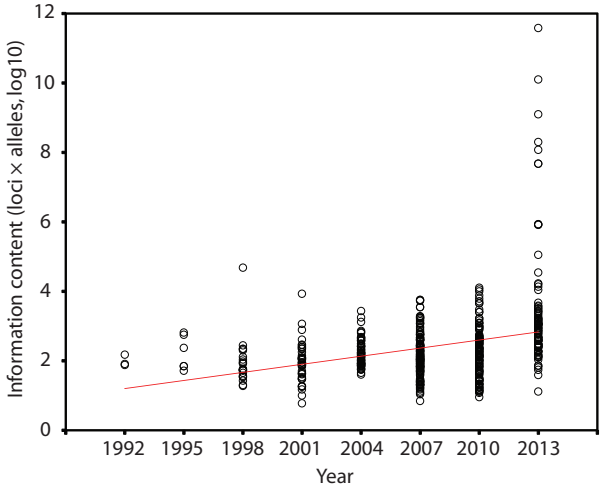

Fig. 1 Linear regression of a weighted metric (i.e. number of loci × total number of alleles sampled, log-transformed) that reflects *potential information content* of phylogeographic data sets, as function of time. The focal temporal period spans the past two decades and was sampled at 3-year increments (N = 497 data sets).

divergence time estimation. Additionally, when genetic markers with different modes of inheritance are used in combination, they yield unique insights into dispersal biology (e.g. animals, Prugnolle & de Meeus 2002; plants, Ennos 1994). Indeed, this information can be critical when interpreting landscape-level patterns of genetic diversity in the context of historical reconstruction (e.g. Pavlova *et al.* 2013). Another potential application of direct contrasts between markers with different modes of inheritance is the ability to distinguish between uni- and bidirectional admixture among distinct lineages at primary or secondary contact zones (e.g. Garrick *et al.* 2014). For these reasons, animal mtDNA and plant cpDNA sequence data are unlikely to become obsolete, but rather will continue to represent an important part of the phylogeography toolbox.

Single-species vs. comparative studies

Single-species and comparative studies both showed significant (P < 0.001) increases in the total number of loci assayed over time, and their trends also had similar trajectories (slope = 0.038 and 0.033, respectively). However, when considering the total number of alleles sampled, while both study types showed significant (P < 0.001) positive trends, single-species studies displayed a more pronounced increase over time (slope = 0.052) than comparative studies (slope = 0.039). Interestingly, the difference in allelic sampling between study types has recently begun to narrow. For example, in the year 2007, the median number of alleles sampled in single-species studies was almost quadruple that of comparative studies, whereas in 2013, the difference had been halved. If we tentatively consider the total number of alleles as a proxy for

density of sampling across a species' range, our survey indicates that sparse geographic sampling was associated with earlier comparative phylogeographic studies, but this is now in the process of being redressed.

Only single-species studies exhibited a significant increase over time in the total number of SNPs (slope = 0.028, $R^2 = 0.071$, d.f. = 1,207, $F = 15.830$, $P < 0.001$). The median number of SNPs in data sets from these studies was consistently larger than that of comparative studies, but the degree of difference between study types has diminished recently (e.g. a 2.5× difference in the year 2007, compared to a 1.3× difference in 2013). When our weighted metric was used to measure potential information content of data sets, single-species and comparative studies both showed significant ($P < 0.001$) increases over time, with similar trajectories (slope = 0.090 and 0.070, respectively).

The optimal balance between genomic vs. geographic sampling will usually depend on the timescales over which historical inferences are focused and/or the age of lineages under investigation (Maddison & Knowles 2006). Therefore, trends in resource allocation by researchers may be too complex to decipher with the present survey data. However, the generally larger size and greater information content of data sets generated by single-species studies represent an interesting trade-off. On one hand, such studies should achieve greater precision and confidence in estimated historical reconstructions. On the other hand, they lack the ability of comparative studies to statistically distinguish idiosyncratic from shared, community-wide responses to past environmental change (Arbogast & Kenagy 2001). Notably, of those phylogeographic studies that have made extensive use of next-generation sequencing, single-species studies clearly dominate (e.g. Emerson *et al.* 2010; McCormack *et al.* 2012; Zellmer *et al.* 2012; Catchen *et al.* 2013; Reitzel *et al.* 2013). One contributing factor may be that the field currently lacks analytical tools that can handle large, multilocus and multitaxon data sets.

Comparisons across taxonomic groups

The increase in size of phylogeographic data sets over time is observable across most major taxonomic groups. Vertebrates show the strongest trajectory of increase in the total number of loci assayed (slope = 0.040, cf. invertebrates, slope = 0.030). Vertebrates also showed the most marked increase for the total number of alleles sampled (slope = 0.051, cf. invertebrates, slope = 0.023). Based on our weighted measure of potential information content of data sets, only vertebrate and invertebrate studies exhibited a clear improvement (both $P < 0.001$; Fig. 2), and of the different taxonomic groups, only vertebrate studies showed a significant increase in the number of SNPs (slope = 0.045, $R^2 = 0.199$, d.f. = 1,147, $F = 36.512$, $P < 0.001$). Data sets assigned to the 'other' taxonomic category showed no change in data set size and information content, but for this group, sample sizes were limited (Table S1, Supporting information). Overall, our survey data showed that vertebrate studies continue to set the standard for phylogeography, as they have ever since the inception of the field (e.g. Avise 2000). Vertebrate studies have most directly benefited from the development of protocols for discovering and screening large subsets of DNA sequence polymorphisms, as members of this group are often used as exemplars for demonstrating proof-of-principal of the new approaches (e.g. Baird *et al.* 2008; Kerstens *et al.* 2009;

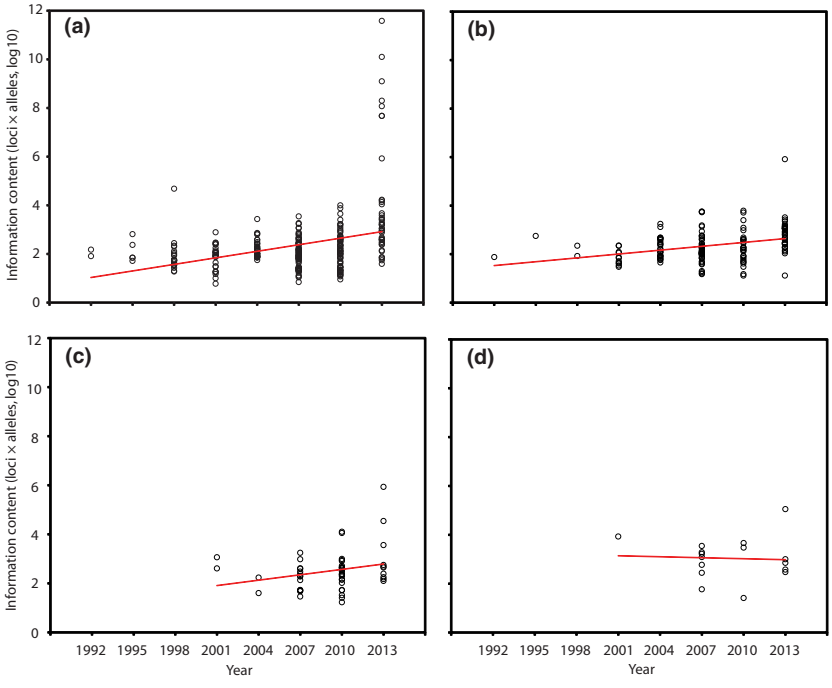

Fig. 2 Linear regression of a weighted metric (number of loci × total number of alleles sampled, log-transformed) as a function of time, partitioned by major taxonomic group. (a) vertebrates ($N = 272$ data sets). (b) invertebrates ($N = 153$). (c) plants ($N = 52$). (d) fungi, protists, algae and bacteria combined (i.e. 'other,' $N = 16$).

Faircloth *et al.* 2012; Peterson *et al.* 2012), and are overrepresented in terms of model organisms and sequenced genomes. Nonetheless, improvements also seen in invertebrate (and to some extent, plant) studies are encouraging; with the increasing application of next-generation sequencing, we expect this will continue for some time.

Data reporting deficiencies

An unexpected result was the realization that many of the papers included in our survey contained incomplete basic descriptions of data set characteristics. For example, 49% of the 508 total data sets had at least one 'not reported' entry for features such as total number of alleles sampled, total number of base pairs sequenced or total number of SNPs detected. Furthermore, 18% of data sets had multiple missing entries. One of the main deficiencies was caused by the increasingly common practice of reporting DNA sequence polymorphism summary statistics at the population level only (with geographic units often subjectively defined), rather than for the entire data set. This introduces redundancy with respect to number of SNPs (or unique haplotypes), making it difficult to evaluate the overall level of polymorphism in the data set. Also, for multilocus data sets in particular, it was common for sample sizes to be reported as a range (in unspecified units), rather than precise values, clearly identified as relating to the number of individuals (or alleles) sampled. We recognize that the deposition in public repositories of raw data that underpin published studies (Whitlock *et al.* 2010) could resolve some of these issues, provided it is associated with complete metadata. However, following Gurevitch *et al.* (2001), we advocate that key data set characteristics should appear in the main text of manuscripts describing empirical research. This will maximize the potential contribution of these studies towards addressing questions outside of the original research focus, via inclusion in meta-analyses.

Future phylogeographic data sets

In the year 2013, there were a handful of studies that stand out as having particularly large data sets, especially in the vertebrates (e.g. Fig. 2a). While this suggests that a linear model may be a poor fit for this taxonomic group, if we take our data at face value, it implies that in the near future, most phylogeographic data sets published in *Molecular Ecology* will consist of thousands of SNPs. Indeed, our forecast modelling suggested the number of SNPs per data set for studies published by the end of the year 2016 to reach ~20 000 (95% CI: 16 590–23 133), and this value represents more than a doubling over the immediately preceding 3-year period (Fig. 3). Given the ongoing development of high-throughput genotyping protocols, our forecast modelling should be considered exploratory. Nonetheless, it is thought-provoking—with such huge quantities of genetic data becoming attainable for almost any organism, what will become the next major constraint on accurately reconstructing evolutionary history? It is clear from our survey that the field of phylogeography is no longer data limited.

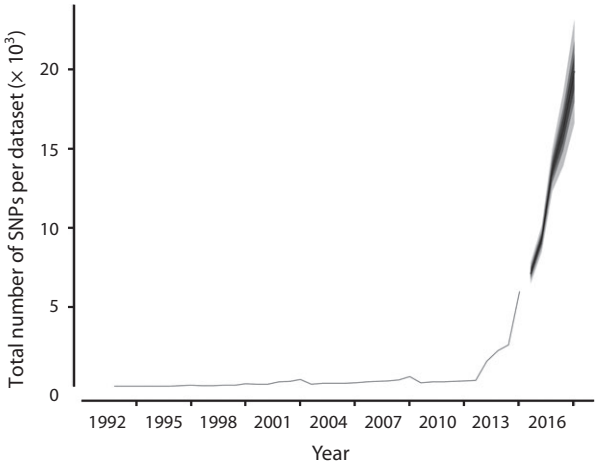

Fig. 3 Forward-time projection of the total number of single nucleotide polymorphisms (SNPs) per published phylogeographic data set, through to the end of the year 2016. Forecasts were generated using autoregressive integrated moving average (median values in black, 95% confidence intervals in pale grey), conditioned on survey data spanning 1992–2013, sampled at 3-year intervals. For each year, only the five highest values for the total number of SNPs are shown.

Outlook and conclusions

Dramatic increases in the size and information content of empirical phylogeographic data sets are opening the door to previously intractable questions. It is now becoming possible to reconstruct complex demographic histories that include multiple overlying events such as lineage splitting, population growth and decline, together with recurrent processes such as postdivergence gene flow (e.g. Carstens *et al.* 2013; Robinson *et al.* 2014a). In addition, dense genomic sampling is allowing a stronger focus on the relative importance of neutral vs. selective forces in driving microevolutionary change over time (Hickerson *et al.* 2010). This is being facilitated by the ability to categorize loci as outliers (i.e. those that are under selection; Antao *et al.* 2008), and through identifying environmental correlates of their alleles (Joost *et al.* 2007; Coop *et al.* 2010). The changing nature of phylogeographic data sets also provides new opportunities to integrate the field with related subdisciplines. For example, in the era of next-generation sequencing, the perceived distinction between landscape genetics and phylogeography (e.g. Wang 2010) increasingly represents a false dichotomy, as the resulting large DNA sequence data sets should be informative over a broad temporal spectrum. Indeed, the timescales on which inferences can be made are likely to depend more on geographic sampling of individuals than on choices relating to genetic data (Robinson *et al.* 2014a). Although analytical methods for handling large genomic data sets from nonmodel organisms are still in their infancy, early progress

is encouraging (Andrews & Luikart 2014; Robinson *et al.* 2014b).

Phylogeography is clearly at a turning point. Data set size and information content are improving dramatically, due to the large numbers of independent autosomal loci being assayed. Furthermore, the increase in genomic sampling seems not to have come at the expense of geographic sampling. While single-species phylogeographic studies of vertebrates—particularly fish—are leading the charge, concurrent improvements across a suite of taxonomic groups are evident. Addressing constraints related to analyses of these data sets, in terms of both upstream filtering and downstream inference of population genetic phenomena, will be critical for facilitating further growth of the field.

Acknowledgements

RCG was supported by a College of Liberal Arts Summer Research grant from the University of Mississippi. We thank Nolan Kane and two anonymous reviewers for constructive comments on the manuscript.

References

Andrew RL, Bernatchez L, Bonin A *et al.* (2013) A road map for molecular ecology. *Molecular Ecology*, **22**, 2605–2626.

Andrews KR, Luikart G (2014) Recent novel approaches for population genomics data analysis. *Molecular Ecology*, **23**, 1661–1667.

Antao T, Lopes A, Lopes RJ, Beja-Pereira A, Luikart G (2008) LOSITAN: a workbench to detect molecular adaptation based on a F_{st}-outlier method. *BMC Bioinformatics*, **9**, 323.

Arbogast B, Kenagy GJ (2001) Comparative phylogeography as an integrative approach to historical biogeography. *Journal of Biogeography*, **28**, 819–825.

Arbogast BS, Edwards SV, Wakeley J, Beerli P, Slowinski JB (2002) Estimating divergence times from molecular data on phylogenetic and population genetic timescales. *Annual Review of Ecology and Systematics*, **33**, 707–740.

Avise JC (2000) *Phylogeography: The History and Formation of Species*. Harvard University Press, Cambridge, Massachusetts, USA.

Backström N, Fagerberg S, Ellegren H (2008) Genomics of natural bird populations: a gene-based set of reference markers evenly spread across the avian genome. *Molecular Ecology*, **17**, 964–980.

Baird N, Etter P, Atwood T *et al.* (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE*, **3**, e3376.

Bazin E, Glémin S, Galtier N (2006) Response [5]. *Science*, **314**, 1390.

Beerli P, Felsenstein J (2001) Maximum likelihood estimation of a migration matrix and effective population sizes in *n* subpopulations by using a coalescent approach. *Proceedings of the National Academy of Sciences, USA*, **98**, 4563–4568.

Beheregaray LB (2008) Twenty years of phylogeography: the state of the field and the challenges for the Southern Hemisphere. *Molecular Ecology*, **17**, 3754–3774.

Bentley DR, Balasubramanian S, Swerdlow HP *et al.* (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, **456**, 53–59.

Bock DG, Andrew RL, Rieseberg LH (2014) On the adaptive value of cytoplasmic genomes in plants. *Molecular Ecology*, **23**, 4899–4911.

Brito PH, Edwards SV (2009) Multilocus phylogeography and phylogenetics using sequence-based markers. *Genetica*, **135**, 439–455.

Brumfield RT, Beerli P, Nickerson DA, Edwards SV (2003) The utility of single nucleotide polymorphisms in inferences of population history. *Trends in Ecology and Evolution*, **18**, 249–256.

Carling MD, Brumfield RT (2007) Gene sampling strategies for multilocus population estimates of genetic diversity (θ). *PLoS ONE*, **1**, e160.

Carstens BC, Lemmon AR, Lemmon AM (2012) The promises and pitfalls of next-generation sequencing data in phylogeography. *Systematic Biology*, **61**, 713–715.

Carstens BC, Brennan RS, Chua V *et al.* (2013) Model selection as a tool for phylogeographic inference: an example from the willow *Salix melanopsis*. *Molecular Ecology*, **22**, 4014–4028.

Catchen J, Bassham S, Wilson T *et al.* (2013) The population structure and recent colonization history of Oregon threespine stickleback determined using restriction-site associated DNA-sequencing. *Molecular Ecology*, **22**, 2864–2883.

Coop G, Witonsky D, Di Rienzo A, Pritchard JK (2010) Using environmental correlations to identify loci underlying local adaptation. *Genetics*, **185**, 1411–1423.

Edwards SV, Beerli P (2000) Perspective: gene divergence, population divergence, and the variance in coalescence time in phylogeographic studies. *Evolution*, **54**, 1839–1854.

Emerson KJ, Merz CR, Catchen JM *et al.* (2010) Resolving postglacial phylogeography using high-throughput sequencing. *Proceedings of the National Academy of Sciences, USA*, **107**, 16196–16200.

Ennos RA (1994) Estimating the relative rates of pollen and seed migration among plant populations. *Heredity*, **72**, 250–259.

Faircloth BC, McCormack JE, Crawford NG *et al.* (2012) Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Systematic Biology*, **61**, 717–726.

Felsenstein J (2006) Do we need more sites, more sequences, or more loci? *Molecular Biology and Evolution*, **23**, 691–700.

Fitzpatrick BM (2009) Power and sample size for nested analysis of molecular variance. *Molecular Ecology*, **18**, 3961–3966.

Garrick RC, Benavides E, Russello MA *et al.* (2014) Lineage fusion in Galápagos giant tortoises. *Molecular Ecology*, **23**, 5276–5290.

Gompert Z, Lucas LK, Buerkle CA *et al.* (2014) Admixture and the organization of genetic diversity in a butterfly species complex revealed through common and rare genetic variants. *Molecular Ecology*, **23**, 4555–4573.

Gurevitch J, Curtis PS, Jones MH (2001) Meta-analysis in ecology. *Advances in Ecological Research*, **32**, 199–247.

Hare MP (2001) Prospects for nuclear gene phylogeography. *Trends in Ecology and Evolution*, **16**, 700–706.

Heled J, Drummond AJ (2008) Bayesian inference of population size history from multiple loci. *BMC Evolutionary Biology*, **8**, 289.

Heled J, Drummond AJ (2010) Bayesian inference of species trees from multilocus data. *Molecular Biology and Evolution*, **27**, 570–580.

Hewitt GM (2001) Speciation, hybrid zones and phylogeography—or seeing genes in space and time. *Molecular Ecology*, **10**, 537–549.

Hey J, Nielsen R (2004) Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. *Genetics*, **167**, 747–760.

Hickerson MJ, Carstens BC, Cavender-Bares J *et al.* (2010) Phylogeography's past, present, and future: 10 years after Avise, 2000. *Molecular Phylogenetics and Evolution*, **54**, 291–301.

Hyndman RJ, Khandakar Y (2008) Automatic time series forecasting: the forecast package for R. *Journal of Statistical Software*, **26**, 1–22.

Hyndman RJ, Koehler AB, Ord JK, Snyder RD (2005) Prediction intervals for exponential smoothing using two new classes of state space models. *Journal of Forecasting*, **24**, 17–37.

Jarman SN, Ward RD, Elliott NG (2002) Oligonucleotide primers for PCR amplification of coelomate introns. *Marine Biotechnology*, **4**, 347–355.

Joost S, Bonin A, Bruford MW *et al.* (2007) A spatial analysis method (SAM) to detect candidate loci for selection: towards a landscape genomics approach to adaptation. *Molecular Ecology*, **16**, 3955–3969.

Keis M, Remm J, Ho SYW *et al.* (2013) Complete mitochondrial genomes and a novel spatial genetic method reveal cryptic phylogeographical structure and migration patterns among brown bears in north-western Eurasia. *Journal of Biogeography*, **40**, 915–927.

Kerstens HHD, Crooijmans RPMA, Veenendaal A *et al.* (2009) Large scale single nucleotide polymorphism discovery in unsequenced genomes using second generation high throughput sequencing technology: applied to turkey. *BMC Genomics*, **10**, 479.

Knowles LL (2004) The burgeoning field of statistical phylogeography. *Journal of Evolutionary Biology*, **17**, 1–10.

Kwiatkowski D, Phillips PCB, Schmidt P, Shin Y (1992) Testing the null hypothesis of stationarity against the alternative of a unit root. *Journal of Econometrics*, **54**, 159–178.

Maddison WP, Knowles LL (2006) Inferring phylogeny despite incomplete lineage sorting. *Systematic Biology*, **55**, 21–30.

Margulies M, Egholm M, Altman WE *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376–380.

Mariac C, Scarcelli N, Pouzadou J *et al.* (2014) Cost-effective enrichment hybridization capture of chloroplast genomes at deep multiplexing levels for population genetics and phylogeography studies. *Molecular Ecology Resources*, **14**, 1103–1113.

McCormack JE, Maley JM, Hird SM *et al.* (2012) Next-generation sequencing reveals phylogeographic structure and a species tree for recent bird divergences. *Molecular Phylogenetics and Evolution*, **62**, 397–406.

McCormack JE, Hird SM, Zellmer AJ, Carstens BC, Brumfield RT (2013) Applications of next-generation sequencing to phylogeography and phylogenetics. *Molecular Phylogenetics and Evolution*, **66**, 526–538.

Morin PA, Archer FI, Foote AD *et al.* (2010) Complete mitochondrial genome phylogeographic analysis of killer whales (*Orcinus orca*) indicates multiple species. *Genome Research*, **20**, 908–916.

Pavlova A, Amos JN, Joseph L *et al.* (2013) Perched at the mitonuclear crossroads: divergent mitochondrial lineages correlate with environment in the face of ongoing nuclear gene flow in an Australian bird. *Evolution*, **67**, 3412–3428.

Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE (2012) Double digest RADseq: an inexpensive method for *de novo* SNP discovery and genotyping in model and non-model species. *PLoS ONE*, **7**, e37135.

Pool JE, Nielsen R (2009) Inference of historical changes in migration rate from the lengths of migrant tracts. *Genetics*, **181**, 711–719.

Prugnolle F, de Meeus T (2002) Inferring sex-biased dispersal from population genetic tools: a review. *Heredity*, **88**, 161–165.

R Development Core Team (2014) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Reitzel AM, Herrera S, Layden MJ, Martindale MQ, Shank TM (2013) Going where traditional markers have not gone before: utility of and promise for RAD sequencing in marine invertebrate phylogeography and population genomics. *Molecular Ecology*, **22**, 2953–2970.

Robinson JD, Coffman AJ, Hickerson MJ, Gutenkunst RN (2014a) Sampling strategies for frequency spectrum-based population genomic inference. *BMC Evolutionary Biology*, **14**, 254.

Robinson JD, Bunnefeld L, Hearn J, Stone GN, Hickerson MJ (2014b) ABC inference of multi-population divergence with admixture from unphased population genomic data. *Molecular Ecology*, **23**, 4458–4471.

Rothberg JM, Hinz W, Rearick TM *et al.* (2011) An integrated semiconductor device enabling non-optical genome sequencing. *Nature*, **475**, 348–352.

Spinks PQ, Thomson RC, Barley AJ, Newman CE, Shaffer HB (2010) Testing avian, squamate, and mammalian nuclear markers for cross amplification in turtles. *Conservation Genetics Resources*, **2**, 127–129.

Sunnucks P (2000) Efficient genetic markers for population biology. *Trends in Ecology and Evolution*, **15**, 199–203.

Templeton AR (1998) Nested clade analyses of phylogeographic data: testing hypotheses about gene flow and population history. *Molecular Ecology*, **7**, 381–397.

Turchetto-Zolet AC, Pinheiro F, Salgueiro F, Palma-Silva C (2012) Phylogeographical patterns shed light on evolutionary process in South America. *Molecular Ecology*, **22**, 1193–1213.

Wang IJ (2010) Recognizing the temporal distinctions between landscape genetics and phylogeography. *Molecular Ecology*, **19**, 2605–2608.

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

Winkelmann I, Campos PF, Strugnell J *et al.* (2013) Mitochondrial genome diversity and population structure of the giant squid *Architeuthis*: genetics sheds new light on one of the most enigmatic marine species. *Proceedings of the Royal Society B: Biological Sciences*, **280**, 20130273.

Zellmer AJ, Hanes MM, Hird SM, Carstens BC (2012) Deep phylogeographic structure and environmental differentiation in the carnivorous plant *Sarracenia alata*. *Systematic Biology*, **61**, 763–777.

B.C.C. and R.C.G. conceived the study. All authors conducted the literature survey. B.C.C., C.H., R.C.G. and T.A.P. analysed the data. B.C.C., R.C.G. and T.A.P. drafted the manuscript. All authors read and approved the final version.

---

doi: 10.1111/mec.13108

Data accessibility

The complete literature survey database is available for download at http://datadryad.org under DRYAD Repository entry doi:10.5061/dryad.10sg7.

Supporting information

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

Table S1 Summary of the datasets retained for inclusion the literature survey, partitioned by broad taxonomic group, and type of phylogeographic study.