MOLECULAR ECOLOGY

Molecular Ecology (2012) **21**, 4925–4930

NEWS AND VIEWS

OPINION

Recommendations for utilizing and reporting population genetic analyses: the reproducibility of genetic clustering using the program STRUCTURE

KIMBERLY J. GILBERT,*† ROSE L. ANDREW,*‡ DAN G. BOCK,*‡ MICHELLE T. FRANKLIN,*§ NOLAN C. KANE,*‡¶ JEAN-SÉBASTIEN MOORE,*† BROOK T. MOYERS,*‡ SÉBASTIEN RENAUT,*‡ DIANA J. RENNISON,*† THOR VEEN* and TIMOTHY H. VINES***

**Biodiversity Research Centre, University of British Columbia, 6270 University Blvd, Vancouver, BC, Canada, V6T 1Z4; †Department of Zoology, University of British Columbia, 6270 University Blvd, Vancouver, BC, Canada, V6T 1Z4; ‡Department of Botany, University of British Columbia, 6270 University Blvd, Vancouver, BC, Canada, V6T 1Z4; §Department of Biological Sciences, Simon Fraser University, 8888 University Drive, Burnaby, BC, Canada, V5A 1S6; ¶Department of Ecology and Evolutionary Biology, University of Colorado at Boulder, Boulder, CO, 80309, USA; **Molecular Ecology Editorial Office, 6270 University Blvd, Vancouver, BC, Canada, V6T 1Z4*

Abstract

Reproducibility is the benchmark for results and conclusions drawn from scientific studies, but systematic studies on the reproducibility of scientific results are surprisingly rare. Moreover, many modern statistical methods make use of 'random walk' model fitting procedures, and these are inherently stochastic in their output. Does the combination of these statistical procedures and current standards of data archiving and method reporting permit the reproduction of the authors' results? To test this, we reanalysed data sets gathered from papers using the software package STRUCTURE to identify genetically similar clusters of individuals. We find that reproducing STRUCTURE results can be difficult despite the straightforward requirements of the program. Our results indicate that 30% of analyses were unable to reproduce the same number of population clusters. To improve this, we make recommendations for future use of the software and for reporting STRUCTURE analyses and results in published works.

Correspondence: Kimberly J. Gilbert, Fax: (604) 822-2416; E-mail: kgilbert@zoology.ubc.ca

Keywords: population clustering, population genetics, reproducibility, STRUCTURE

Received 15 May 2012; revised 18 July 2012; accepted 24 July 2012

Introduction

The reproducibility of scientific research is fundamental to maintaining scientific rigor and advancing science (Price 2011). Full experimental replication provides the most thorough means of verifying published empirical results, but this approach can be impractical due to the difficulty in obtaining identical samples and large financial and time commitments (Peng 2011). Diminishing costs and advancing technology have resulted in a plethora of large genetic data sets, while at the same time, there has been an increase in the complexity of software applications. A previous investigation into the reproducibility of microarray studies found that few were fully repeatable, as many suffered from ambiguity in the methods, discrepancy in the results, and lack of available data or software (Ioannidis *et al.* 2009). Maintaining the rigor of today's scientific research may therefore prove a more difficult task than expected, as both the empirical results and the often complex analyses need to be reproducible. Efforts to encourage and implement data archiving and sharing are expanding, and these create the opportunity to test the validity and reproducibility of scientific results (Whitlock *et al.* 2010).

Reproducing results within the field of molecular ecology is especially difficult because biological samples are unique to their particular place and time, and subsequent samples may reflect different ecological or evolutionary forces (Wolkovich *et al.* in press). Researchers therefore tend to test the same overarching hypothesis with samples from different taxa and locations, in the hope of arriving at a more general and repeatable pattern. However, drawing broad conclusions from the results of many studies is ineffective when the results of the individual studies cannot be reproduced from their underlying data. It is thus essential to test the reproducibility of statistical analyses at the level of individual papers as well. To examine how well we could recreate the results from typical molecular ecology studies, we investigate, as an example, the reproducibility of studies that used genotype data to identify genetically similar clusters of individuals with STRUCTURE (Pritchard *et al.* 2000). Many studies use clustering results based on STRUCTURE to perform further analyses, making it an important foundation upon which inferences are built. We ask whether (i) archived data sets are sufficiently complete and well annotated that they can be reused, (ii) published articles specify all the methodological details necessary to

reproduce the analysis, and (iii) where possible, the same conclusions can be reached by reanalysing the archived data. Although reproducibility has many different aspects, we use it here to mean the agreement between results obtained through analysing identical data sets using the same analytical method but under different conditions (different observers, computers and starting points in computer algorithms). We reanalysed 23 articles from 2011 that used STRUCTURE to infer genetic clustering and also checked the level of data completeness and methodology reporting in an additional 37 articles.

Methods

Obtaining data sets

We gathered STRUCTURE data sets associated with 23 papers published in 2011: 21 from *Molecular Ecology*, and two from the journal *PLoS One*. Data were obtained from the online data repository Dryad (Dryad Digital Repository) in November 2011, NCBI GenBank, or from the supplementary material accompanying the paper. With one exception, we excluded papers where data were archived on GenBank due to the difficulty of compiling individual accessions into the correct format for STRUCTURE.

For a broader assessment of data set completeness and methods reporting, we also included 37 data sets obtained by contacting the authors of original research papers published in *PLoS One*, *PLoS Genetics* and *BMC Evolutionary Biology* and collected as part of a separate study (T.H. Vines *et al*., unpublished data).

The program STRUCTURE

The freely available Bayesian clustering program STRUCTURE (Pritchard *et al.* 2000) is the most commonly used application to infer population structure, with over 5000 citations in Web of Science as of June 2012. STRUCTURE uses multilocus genotype data to describe and visualize population structure based on allele frequencies of the data.

STRUCTURE is capable of analysing a variety of genotype data, including both codominant markers (microsatellite and single nucleotide polymorphism, SNP; Pritchard *et al.* 2000) and dominant markers (amplified fragment length polymorphism, AFLP; Falush *et al.* 2007). The model uses Markov chain Monte Carlo (MCMC) simulations to estimate the group membership of each individual, assuming Hardy–Weinberg and linkage equilibrium within groups, random mating within populations and free recombination between loci (Pritchard *et al.* 2000). Due to the random walk characteristic of the MCMC methods, STRUCTURE outputs are not expected to produce identical results, yet the approach should be robust enough to yield identical conclusions when reproduced. The program is initiated with a required text file containing individual genotype data and labels as well as optional information on population assignment, sampling sites or locus names. The user specifies several essential parameters regarding the ancestry model, the allele frequencies model, the length of the burn-in (initial runs of the simulation during which data are not retained to ensure results are not dependent on initial conditions), length of run time (number of MCMC repetitions during which data are retained), the number of independent replicates of each set of parameters and the range of number of clusters (K values) to be tested. These can be specified directly in the graphical user interface or in a separate text file when run in the command line. In addition, it is possible to specify extra parameters, mainly regarding the Markov chain process as well as a sampling location prior. The details of the model are described by Pritchard *et al.* (2000, 2007) and Falush *et al.* (2003, 2007).

STRUCTURE outputs are typically analysed to infer the optimal K by one or a combination of methods. In the method described in Pritchard *et al.* (2000), the optimal K is chosen by plotting the log probability of the data (referred to as ln $\Pr(X|K)$ in STRUCTURE's manual, Pritchard *et al.* 2007) against a range of K values and selecting the K with the highest ln $\Pr(X|K)$ or the one after which the trend plateaus, while also taking into account the consistency of the groupings across multiple runs with the same K. An alternative method, described by Evanno *et al.* (2005), formalizes an ad hoc approach based on plotting the second-order rate of change in ln $\Pr(X|K)$ for successive Ks (referred to as ΔK) against a range of K values, and selecting the true K based on where the maximal value of this distribution occurs. As emphasized in the STRUCTURE manual (Pritchard *et al.* 2007), selecting the optimal K can be quite a subjective procedure and is best inferred when the biology and history of the organism are taken into account. Replicate STRUCTURE runs can be combined using the software programs CLUMPP (Jakobsson & Rosenberg 2007) and STRUCTURE HARVESTER (Earl & von Holdt 2011). Bar plots depicting the ancestry proportions (or membership coefficients, Q) of individuals in each cluster can then be created, for example with the software DISTRUCT (Rosenberg 2004), to visualize the population clusters.

Analysing data sets

We followed procedures for analysis as described in the methods section of each publication and used default settings for parameters that were not specified. Several publications performed multiple STRUCTURE analyses, which we counted independently for a total of 34 analyses. We made use of the Bioportal computing resource (https://www.bioportal.uio.no/; Kumar *et al.* 2009) or local desktop computers. Output was compiled with STRUCTURE HARVESTER and processed following the authors' description, including CLUMPP analysis where appropriate. We first assessed whether we could reproduce the K values from the original study based on the methods used by the authors. Then, whenever possible, membership coefficient bar plots were visually compared by multiple authors of the present study to assess whether our results were a true reproduction of the original results. When we concluded the same value of K as the authors, we deemed the analysis as reproduced

unless the membership coefficient bar plots showed strikingly different results.

For the broader survey of data set completeness, we evaluated whether the sample size and number of loci described in the paper matched the obtained data set from both the data sets obtained from online material (23 studies) and email correspondence (37 studies). To check the overall standards for reporting parameter settings within either the methods section or in a supplemental file, we also recorded the number of 'essential' parameter settings (range of *K* values tested, length of burn-in, length of MCMC repetitions, number of independent replicates, the admixture model and allele frequencies model) given in the paper or supplemental material for all of the above analyses.

Results

Of the 23 papers, we attempted to reanalyse using data from supplementary materials or online repositories, two papers did not have archived data present at the time, making their reanalysis impossible. Three papers (13%) provided data where the number of individuals and/or loci specified in the publication did not match those present in the data set, or the authors performed their STRUCTURE analysis on an unspecified subset of the archived data. Of these 23 papers, three selected *K* using the Pritchard method (Pritchard *et al.* 2000), seven used the Evanno method (Evanno *et al.* 2005), eight used a combination, one used a nonparametric Wilcoxon test, two did not specify their method, one used no standard method and rather utilized $K = 2$ to identify hybrid individuals and one discussed a comparison of two *K* values obtained in a previous study. We therefore also did not assess the reproducibility of these final two papers, leaving 19 papers (containing 30 analyses) that we attempted to reproduce. See Table S1 (Supporting information) for full characteristics of all analyses.

We were able to reproduce the authors' inference of *K* for 70% (21 of 30) of the analyses (Fig. 1). All of the successfully reproduced data sets consisted of microsatellite genotypes. In general, microsatellite data sets were analysed using longer burn-in and MCMC run lengths as well as more independent replicates; however, there was no significant difference in an overall proxy for run length ([length of burn-in + length of MCMC repetitions] × number of independent replicates) between analyses that were reproduced and those that were not ($t = -0.0617$, d.f. = 13.564, *P*-value = 0.95). Comparing these parameters individually, we found a trend of longer burn-in ($t = -1.8706$, d.f. = 26.991, *P*-value = 0.072) but not of more MCMC repetitions ($t = -1.6537$, d.f. = 21.677, *P*-value = 0.11) or an increase in the number of independent replicates ($t = 1.1442$, d.f. = 7.511, *P*-value = 0.29) for reproduced studies. Comparison of the *K* values chosen by the original authors versus our reanalysed *K* results showed a significant correlation of 0.5934 ($t = 3.9703$, d.f. = 29, *P*-value = 0.0004; Fig. 2).

We also assessed the completeness and description of all 60 data sets that we obtained and found 35% to be either incorrectly or insufficiently described by the authors. We found that 17 data sets did not match the description given in the paper, most typically because the data contained a different number of loci or individuals than suggested by the paper. Lastly, four papers did not give any clear description of the number of individuals, loci or both used in their STRUCTURE analysis, making it impossible to judge how well the archived data matched the data analysed by the authors.

Authors' descriptions of the essential parameters used to run STRUCTURE varied markedly, ranging from 0 described parameters to a maximum of 6 (median = 6). We found a significant difference in number of essential parameters between two of the journals ($t = 3.31$, d.f. = 40.27, *P*-value = 0.015), with *Molecular Ecology* having a mean of 5.7 parameters specified and *PLoS One* 4.6 (*PLoS Genetics*, 4.7 and *BMC Evol. Biol.*, 4.8). Overall, length of burn-in ranged from 1000 to 50 000 000 (median = 50 000), while MCMC repetitions ranged from 10 000 to 500 000 000 (median = 450 000). Independent replicates ranged from 3 to 100 (median = 10).

Discussion and recommendations

Reproducibility is a foundation of scientific research. The widespread application of STRUCTURE makes it an ideal case study to test the ability to reproduce molecular ecology results that rely on large data sets and complex algorithms. As STRUCTURE results often serve as the underpinnings for

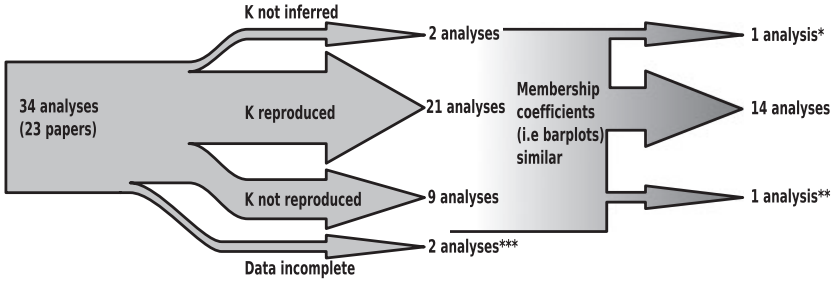

Fig. 1 Results of STRUCTURE reanalyses. Initial branching arrows show the numbers of analyses resulting in different outcomes at the point of selecting a *K* value. The subsequent arrows show the numbers of analyses successfully reaching the point of matching membership coefficients. Size of arrowheads is proportional to number of analyses present. *When *K* was not inferred, we attempted to match membership coefficients across all *K* values (still only counted as 1 analysis). **When *K* was not reproduced, we compared membership coefficients at the authors' chosen *K*. ***For incomplete data, analyses could not be run.

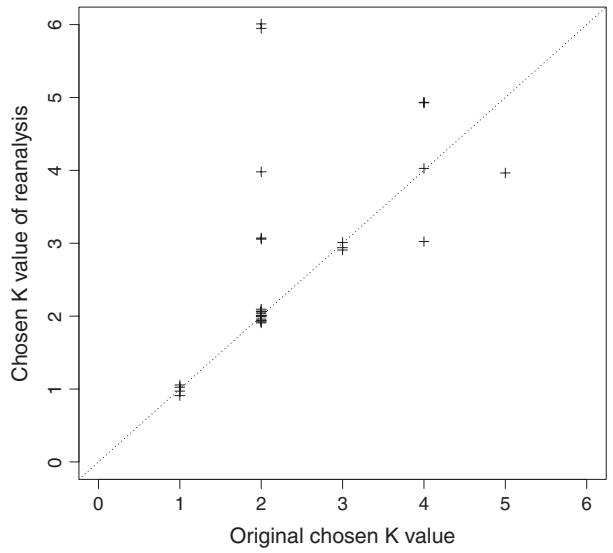

Fig. 2 Comparison of *K* values for original studies versus reproduced studies. Dotted line indicated the 1:1 line, points are jittered for better visualization.

further analyses and conclusions within a study, it is important to assess whether the implementation of the program, subsequent analysis and associated conclusions are properly reported and can be reproduced.

We find that reproduction of STRUCTURE results can be difficult to achieve, despite the straightforward input requirements of the program (a genotype file and two parameter files). Our results show that 30% of analyses are not reproducible. A large factor in the failure to reproduce these analyses was the availability of data in a form that could be readily understood by researchers not familiar with the study system. Had we included studies with no data available as the starting point for our reanalysis, our assessment of failure to reproduce would have been even higher, particularly for journals without a strongly enforced data archiving policy (see T.H. Vines *et al.*, unpublished data for further discussion of data accessibility).

We recognize that assessing reproducibility is inherently difficult. Our main evaluation criterion (same *K* value) is only a small part of full reproducibility of these studies, but the most objective one. Furthermore, it is difficult to disentangle the nonreproducibility caused by the stochastic nature of the program from that caused by both discrepancies in data sets available versus those used by authors and their reported methods. The trend of longer burn-in lengths in reproduced studies suggests that at least a portion of the poor reproducibility of some studies is due to the inherent stochasticity of the Monte Carlo approach itself. In at least one case, we can attribute our failure to reproduce the study to insufficiently described, complex analyses performed; however, there seemed to be no other outstanding characteristics of nonreproducible studies. It is important to note that although STRUCTURE is the most commonly used program, in some instances, other methods may be more appropriate for a given data set. For example, performing a PCA allows examination of variability within clusters, other Bayesian methods such as the program INSTRUCT (Gao *et al.* 2007) allows inbred genotypes to be used, TESS (Chen *et al.* 2007) utilizes spatial information, and BAPS (Corander *et al.* 2003, 2004, 2006; Corander & Marttinen 2006;) aids in detection of admixed individuals. Using the right program is not only essential to drawing correct conclusions, but may also improve reproducibility of results. Further discussion of additional approaches can be found in Latch *et al.* (2006) and François & Durand (2010).

In addition, we may have judged a study to be nonreproducible despite differences in the final results that may or may not have biological significance. The correlation between original and reproduced *K* values implicates this, yet there is still clearly room for improvement. With such widespread use within its field, it is important that users of STRUCTURE properly implement the software, regardless of whether or not they possess a full understanding of the algorithm underlying the analysis. To ensure that published results can be reproduced, we make the following recommendations for future users of the program. Although our study is specific to STRUCTURE, many of these recommendations are applicable to other types of analysis.

1 For archiving purposes, authors should be encouraged to provide the final version of both the genotype and parameter files. We propose that authors archive genotype data from all individuals. If only a subset was used in the analysis, these individuals should be clearly identified in the same file so that this information is retained. The parameter files include all the settings used in the analysis, hence archiving the entire file avoids any confusion regarding use of default settings when not explicitly stated by the authors. When using the graphical user interface version of the software, the parameters can be exported from the program in text format for archiving purposes.
2 Authors should ensure that burn-in and run lengths are sufficient. We found remarkable variation in parameters affecting the computational demands of the analysis (burn-in time, MCMC repetitions, and replicate runs). Though we found no significant difference between an overall proxy for run length and reproducibility and only a slight trend individually for burn-in time, given the advances in computing power, we feel that the proposed minimum requirements, dating back to the software's advent more than a decade ago, should be increased. It is difficult to set a standard, as variability across data sets in the number of loci, their levels of polymorphism, and the amount of population structure present all also contribute to the program's ability to successfully detect the appropriate *K* (Rosenberg *et al.* 2001; Latch *et al.* 2006; Gao & Starmer 2007). We would advise a minimum of at least 100 000 burn-in iterations and MCMC repetitions for each run, and much longer burn-in will be required for some data sets. Comparing a range of run durations may help to determine the appro-

priate run length, and it is always advisable to choose a longer burn-in and run length. To confirm that burn-in is adequate, it is also important check for convergence in values of summary statistics (particularly α, F, D, and the likelihood) that are estimated by the program, as recommended in the Structure manual (Pritchard *et al*. 2007). Additional independent replicate runs are of great importance as they limit the influence of stochasticity and increase the precision of the parameter estimates. That is especially true when using the Evanno method, which requires an estimate of variance. In at least one reanalysis we performed, only five replicate runs were used, which may explain the failure to reproduce results (the chosen K) in this particular study. We recommend 20 replicates as used by Evanno *et al*. (2005).

3 Proper reporting of the methods used to analyse STRUCTURE results is vital for inferring K. Whether the method outlined by Pritchard *et al*. (2000) or by Evanno *et al*. (2005) or both are used to select K should be clearly stated, as well as any biological factors that have influenced the choice of K. Special attention should be given to the comparison of $K = 1$ versus greater values, as the Evanno method is not capable of performing this comparison. We advise that results are reported in the form of the graph of the natural logarithm of the likelihood of the data given K (if the Pritchard method was used) and the ΔK graph (if the Evanno method was used) as well as the bar plot(s) showing individual assignments for the given K or comparison across plausible K values. Ideally, for full reproducibility of a study, membership coefficients for each individual should also be provided. These results should be examined within each replicate to determine how much stochasticity is present before runs are averaged, as well as after averaging all replicate runs.

Conclusion

A substantial proportion of STRUCTURE results were not reproducible, despite the relative simplicity of the procedure, requiring only a genotype file and associated parameter settings. Our recommendations on how to archive data sets analysed with STRUCTURE should reduce the component of nonreproducibility due to uncertainty of parameter choice or lack of clarity in the data analysed, but some discrepancies will no doubt still persist. We hope that scientists will increasingly acknowledge the concept of scientific reproducibility in the future and be aware of practices they can enact both for better data archiving and better implementation of other similar programs in their analyses.

Acknowledgements

We would like to thank Mike Hart, Elizabeth Kleynhans and Sam Yeaman for their help in analysing data and all of the many authors who contributed their data sets at our requests. We also thank Mike Whitlock, three anonymous reviewers, and subject editor Richard Abbott for providing helpful comments that considerably improved earlier versions of the manuscript.

References

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.

Corander J, Marttinen P (2006) Bayesian identification of admixture events using multilocus molecular markers. *Molecular Ecology*, **15**, 2833–2843.

Corander J, Walmann P, Sillanpaa MJ (2003) Bayesian analysis of genetic differentiation between populations. *Genetics*, **163**, 367–374.

Corander J, Walmann P, Marttinen P, Sillanpaa MJ (2004) BAPS2: enhanced possibilities for the analysis of genetic population structure. *Bioinformatics*, **20**, 2363–2369.

Corander J, Marttinen P, Mäntyniemi S (2006) Bayesian identification of stock mixtures from molecular marker data. *Fishery Bulletin*, **104**, 550–558.

Dryad. The Dryad Digital Repository: http://datadryad.org/

Earl DA, von Holdt BM (2011) STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources*, **4**, 359–361.

Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular Ecology*, **14**, 2611–2620.

Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, **164**, 1567–1587.

Falush D, Stephens M, Pritchard JK (2007) Inference of population structure using multilocus genotype data: dominant markers and null alleles. *Molecular Ecology Notes*, **7**, 574–578.

François O, Durand E (2010) Spatially explicit Bayesian clustering models in population genetics. *Molecular Ecology Resources*, **10**, 773–784.

Gao X, Starmer J (2007) Human population structure detection via multilocus genotype clustering. *BMC Genomics*, **8**, 34.

Gao H, Williamson S, Bustamante CD (2007) A Markov Chain Monte Carlo approach for joint inference of population structure and inbreeding rates from multilocus genotype data. *Genetics*, **176**, 1635–1651.

Ioannidis JPA, Allison DB, Ball CA *et al*. (2009) Repeatability of published microarray gene expression analyses. *Nature Genetics*, **41**, 149–155.

Jakobsson M, Rosenberg NA (2007) CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics*, **23**, 1801–1806.

Kumar S, Skjaeveland A, Orr R *et al*. (2009) AIR: a batch-oriented web program package for construction of supermatrices ready for phylogenomic analyses. *BMC Bioinformatics*, **10**, 357.

Latch EK, Dharmarajan G, Glaubitz JC, Rhodes Jr OE (2006) Relative performance of Bayesian clustering software for inferring population substructure and individual assignment at low levels of population differentiation. *Conservation Genetics*, **7**, 295–302.

Peng RD (2011) Reproducible research in computational science. *Science*, **334**, 1226–1227.

Price M (2011) To replicate or not to replicate. *Science Careers*, doi: 10.1126/science.caredit.a1100133.

Pritchard J, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.

Pritchard JK, Wen W, Falush D (2007) *Documentation for STRUCTURE software: Version 2.2*, Available from http://pritch.bsd.uchicago.edu/software.

Rosenberg NA (2004) Distruct: a program for the graphical display of population structure. *Molecular Ecology Notes*, **4**, 137–138.

Rosenberg NA, Burke T, Elo K *et al*. (2001) Empirical evaluation of genetic clustering methods using multilocus genotypes from 20 chicken breeds. *Genetics*, **159**, 699–713.

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

Wolkovich EM, Regetz J, O'Connor MI (2012) Advances in global change research require open science by individual researchers. *Global Change Biology*, **18**, 2102–2110.

The authors are interested in ecology, evolution, and genetics. They are part of a discussion group focusing on enhancing data availability and reproducibility.

Data accessibility

All data used for this study is listed in the supplemental information.

Supporting information

Additional Supporting Information may be found in the online version of this article.

Appendix S1 Papers and data used in our study.

Table S1 Characteristics of STRUCTURE runs performed. Reanalyses followed the same settings as the original paper, using defaults when nothing was specified.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

doi: 10.1111/j.1365-294X.2012.05754.x