Evolution, 53(6), 1999, pp. 1961–1965

ENHANCING THE RETRIEVABILITY OF POPULATION GENETIC SURVEY DATA? AN ASSESSMENT OF ANIMAL MITOCHONDRIAL DNA STUDIES

Paul L. Leberg[1] and Joseph E. Neigel[2]
Department of Biology, University of Louisiana, Lafayette, Louisiana 70504
[1]*E-mail: pll6743@louisiana.edu*
[2]*E-mail: jneigel@louisiana.edu*

Abstract.—Surveys of genetic variation in natural populations represent a valuable and often irreplaceable resource. It may be desirable to reanalyze data as new methods are developed for comparisons with other populations or for comparisons with the same populations at different times. We evaluated existing mechanisms of data preservation in a survey of 627 published surveys of mitochondrial DNA variation in animal and found that over half of the datasets (56%) contained insufficient information for reanalysis. In many cases, publication of complete data would not have added excessively to the length of the publication. Because at present, publications represent the main archive of population genetic data, we offer recommendations for how the essential data from mtDNA surveys can be presented in a form that is complete and concise.

Key words.—Database, GenBank, geographic variation, mitochondrial DNA.

Received January 26, 1999. Accepted June 15, 1999.

The Value of Genetic Marker Survey Data

Population genetics, ecologists, and conservation biologists have conducted thousands of surveys of genetic markers (reviewed in Smith and Wayne 1996). Most of these studies had very specific aims, such as the reconstruction of population history, detection of population bottlenecks, or estimation of levels of gene flow (reviewed in Avise 1994). Surveys of molecular markers are generally undertaken to answer specific questions and, if successful, are described in a publication that addresses those questions.

With due recognition of the value of individual studies and publications, we submit that their full value could be much greater. A publication may be based on the best methods of analysis and interpretation that are available, but better methods are likely to follow. At some later time, it may even become necessary to reexamine the conclusions of earlier studies by reanalysis of their data. The estimation of Wright's (1951) index of population subdivision, F_{ST}, provides an example of how this need may arise. Weir and Cockerham (1984) surveyed published estimates of F_{ST} from allozyme data and found (among other problems) that many authors confounded the actual variance in allele frequencies among populations with the variance produced by sampling from populations. As a result, original estimates of F_{ST} were upwardly biased. More importantly, some authors did not describe how they estimated their reported values and thus made it impossible to critically evaluate them. The long-term value of these studies has been limited by the quality of their analysis rather than the quality of their data.

It may be desirable to reanalyze data with newer methods, even if the data were correctly analyzed when first published. For example, in some early estimates of F_{ST} from mtDNA data, each polymorphic site was generally treated as a separate locus. Although this approach is valid, it may not make the best use of the information provided by DNA sequences (Hudson et al. 1992). More recently, methods have been introduced that use genealogical models to estimate migration parameters from mtDNA sequence data (Slatkin and Maddison 1989; Neigel and Avise 1993). Finally, data may be reanalyzed to address entirely new questions. For example, data from earlier publications have been used to estimate gene flow among populations (Slatkin 1985) and to determine if a population has experienced a recent demographic bottleneck (Luikart and Cornuet 1998).

Comparative studies are another source of added value for surveys of molecular variation. Comparisons of multiple allozyme (Nevo et al. 1984) and mtDNA (Avise 1994) surveys have revealed interesting phylogenetic, ecological, and biogeographic patterns. However, such comparisons are only effective when there is a common way to present the findings of each survey. As shown above, accepted methods of data analysis are prone to change, and biases may be introduced when studies are compared on the basis of original parameter estimates. Reanalysis of data may be the best means to remove these biases.

Comparisons between surveys conducted at different times in the history of a population or species are especially valuable. There is arguably no better way to examine the effects of either natural or anthropgenic factors on the distribution of genetic markers. Temporal changes in the frequencies of genetic markers have been used to estimate effective population size (Waples 1989), assess the loss of variation from population bottlenecks (Tarr et al. 1998), and detect the breakdown of barriers to hybridization (Spaak 1996). Because many of these changes are widespread and irreversible consequences of human activities, our opportunities to establish baselines for "natural conditions" and thus to detect subsequent changes are steadily diminishing. Furthermore, because it is difficult to predict where the next oil spill, exotic introduction, or pathogen outbreak may occur, nearly every survey of molecular variation should be viewed as a potential "last chance" to bank such data for future use.

TABLE 1. Categorization of published datasets of mtDNA variation based on the type of additional data that would be needed for reanalysis.[1] Some datasets were incomplete in multiple ways.

	Incomplete data on					
Type of genetic survey	Number of datasets	Haplotype differences	Haplotype similarities	Haplotype frequencies	Sampling locations[2]	Total incomplete datasets
RFLP	257	113	11	77	15	152 (59.1%)
RS	191	63	2	61	16	119 (62.3%)
PCR RFLP	51	15	6	7	1	21 (41.1%)
SEQ	128	0	38	27	13	57 (44.5%)
Total	628	191	58	172	45	349 (55.6%)
Percentage of total		30.6%	9.2%	27.5%	7.2%	

[1] Datasets were taken from 460 publications, of which some publications contained multiple datasets. A bibliography of publications that contained these datasets can be found on our web site (http://seahorse.louisiana.edu/PGDB).

[2] Because precise collection locations are often unknown for pelagic organisms, this category excludes surveys of marine mammals and fishes.

MECHANISMS OF DATA PRESERVATION AND THEIR EFFECTIVENESS

If there are compelling reasons to preserve the data generated by surveys of molecular variation, there is also a need for practices that insure this preservation. We suggest the use of three criteria to evaluate the adequacy of data preservation. First, datasets should be reasonably complete, so that they can be subject to reanalysis. Second, the data should be accessible, so that the process of retrieval is not a barrier. Third, the data should be secure, so that it is not likely to be lost. At present there are several ad hoc mechanisms by which data from past surveys can be accessed. These mechanisms are: (1) extraction from scientific publications; (2) direct communication with providers of data; and (3) retrieval from organized databases. We used our three criteria for adequate data preservation to consider the suitability of each of these mechanisms.

We assessed current mechanisms for the preservation of molecular marker survey data and the provision of access to this data by attempting to retrieve data from published studies of geographic variation in animal mtDNA. We choose mtDNA because the two methods most commonly used to assess mtDNA variation, restriction fragment analysis and DNA sequencing, are relatively standardized and thus easily evaluated. Distinct mtDNA sequence variants are designated as haplotypes, and most studies include a comparison of the haplotypes found in a survey as well as a survey of the distribution of these haplotypes among sampled localities. There have also been a fairly large number of mtDNA studies published since 1979 and this number is certainly increasing.

Scientific Publications

When data from surveys of molecular variation are published in the scientific literature, they effectively become part of a data archive that comprises journal articles, symposium proceedings, and academic books. This archive is duplicated in universities around the world, and thus is both accessible and secure. Furthermore, the text of each publication generally provides details about how data were collected and analyzed. However, unlike a scientific database, the data deposited in the scientific literature is not well structured, standardized, or referenced and it would be a mistake to assume that this archive could fulfill the functions of a database. For surveys of molecular variation, there is little consistency in how data are presented and generally no requirements for publication of complete datasets. Although relatively complete data can usually be presented in as few as two tables (see below), such tables may become quite large and their publication may be viewed as prohibitively expensive by authors or journal editors.

Because the scientific literature represents the main archive for surveys of molecular variation, we estimated the proportion of published mtDNA studies that provided what we considered to be the minimum amount of data necessary for reanalysis. We attempted to conduct a thorough, but not exhaustive, survey of papers published from 1979 to the summer of 1997. We divided studies into categories based on the methods that were employed to assess mtDNA variation: unmapped restriction fragment length polymorphisms detected in whole mtDNA genomes (RFLP), polymorphisms of mapped restriction sites detected in whole mtDNA genomes (RS), restriction fragment length polymorphisms within amplified segments of mtDNA (PCR-RFLP), and sequences of segments of the mtDNA genome (SEQ). Individual publications often contained data for multiple species or data generated by multiple techniques for a single species. Therefore, we used the individual dataset as the unit of observation in our investigation. A dataset was defined as any comparison of conspecifics from at least three localities that was based on a single type of mtDNA data.

We defined a dataset as complete if it included descriptions of molecular characters (restriction fragments, restriction sites, or nucleotides in DNA sequences) used to differentiate haplotypes, the characters shared by haplotypes, adequate descriptions of the locations sampled, and the numbers of individuals with each haplotype that was sampled from each location. Detailed criteria can be found on our web site (http://seahorse.louisiana.edu/PGDB). Datasets in publications were not considered incomplete if the missing information could be obtained from other publications in the peer-reviewed literature.

A large proportion of datasets that were based on restriction fragments and sites failed to provide adequate information on differences between haplotypes (see Table 1). In many cases authors noted that polymorphisms were detected with a specific enzyme, but failed to indicate whether one haplotype differed from another by the loss or gain of a site or by an indel. Without such information, even basic forms of reanalysis are impossible, and such datasets would be of limited use for comparisons with data from similar surveys. Of datasets that provided descriptions of the nature of polymorphisms, several failed to describe how many restriction sites or fragments were shared among haplotypes (Table 1). A surprisingly large number of published datasets failed to adequately report the distributions of haplotypes among sampled locations (Table 1). In many cases, the haplotypes found at each location were listed, but the haplotype frequencies were not provided. Only a small proportion of the studies lacked information on the locations of the collection sites (Table 1).

In summary, for the average mtDNA survey, the data presented is more likely than not to be incomplete. Overall, 56% of the surveyed datasets contained incomplete data (Table 1). The alarmingly high proportion demonstrates the importance of other methods for data archival and retrieval.

Direct Communication

When the data provided in a scientific publication are incomplete, it is generally assumed, and often stated, that readers can contact the authors to obtain additional data. Such data cannot be considered either readily accessible or secure. Authors may fail to preserve records of their data in the face of career changes, retirements, or the simple passage of time, and it may not be obvious who should be contacted for data after an author's death. It may be difficult to retrieve data from old notebooks or obsolete computer media. Authors may also be unwilling to provide data if they feel it would take up too much of their time or if they would prefer not to have their data reanalyzed.

We evaluated the feasibility of an investigator obtaining data directly from scientists by sending written requests for data to the authors of 30 publications that contained incomplete datasets. We addressed these requests to the senior author of each publication, unless another author was designated as the corresponding author; the request was mailed to the most current address we could locate. In each letter, we made a specific request for information or clarification that would allow us to extract data from the publication. We explained that we wished to deposit their published data in a database of mtDNA surveys that was funded by the National Science Foundation. If we did not receive a response within three months, we considered authors to have not responded for the purposes of this survey.

Of the 30 authors we attempted to contact for data that was not included in their original publication, only five responded. Only one provided all of the information that we requested, and one other provided some of the requested data. Two authors directed us to publications that contained different, but more complete datasets. The final respondent wrote that the requested information could not be retrieved. Perhaps more authors may have responded to more persuasive or persistent requests; however, these results demonstrate that mtDNA survey data are often not readily available from authors.

Databases

Public databases have the potential to provide secure and accessible depositories for the data generated by surveys of molecular markers. The value of public databases and the willingness of the scientific community to support them is demonstrated by GenBank, a widely used repository for sequence data (Burks et al. 1991). This database is considered to be of so much value to the research community that submission of sequence data to GenBank is a stated requirement for publication in many journals. However, although GenBank is quite flexible, it was not designed as a database for surveys of variation. GenBank lacks both guidelines for the organization of such data and mechanisms to ensure that if survey data are submitted, they are sufficiently complete.

We found that GenBank has not been a particularly effective archive for mtDNA survey data. Among 128 datasets with sequence information, there were no apparent GenBank submissions for 51 (40%). Many of these datasets were published in journals with policies that state the submission of sequence data to GenBank is a requirement for publication. For 30% of the datasets with sequence information, it was not possible to retrieve usable sequence data from either GenBank or the text of the publication. Of the 77 datasets with at least one sequence available from GenBank, eight did not provide sufficient information to match the GenBank entries with the sequences described in the corresponding publication.

We have shown that mtDNA survey data has not been adequately preserved by any of the mechanisms that are presently in operation. However, we do not fault the authors of mtDNA studies, the agencies that have funded their work, or the journals that publish their findings. Population genetic surveys do not fit into a single mold, nor would we advocate that they should. It is because they have been so successful that we now recognize that it is important to preserve their data.

RECOMMENDATIONS

We could employ either of two approaches to improve the preservation and accessibility of population genetic survey data. Ideally, data from population genetic surveys would be archived in a public database similar to GenBank. Databases not only store data, but also greatly facilitate access. We are currently exploring the feasibility of such a database; however, its development and full implementation may take several years and its success will depend on support from the community of potential users. Thus, at present, the most effective way to preserve population genetic survey data is to publish complete datasets. We advocate that authors, reviewers, and editors consider this important function of publications and make reasonable efforts to verify that the data from a published study is either available in the publication itself or in a publicly accessible database such as GenBank. Although the type of data generated for different studies will certainly vary, we suggest that most share the common structure that we have described, and on this basis we offer several recommendations for what data should be considered essential and how it may be effectively presented.

TABLE 2. An example of a table that shows the restriction fragments that define each haplotype. The essential data in this table is a listing of every haplotype by number, followed by the presence or absence of every restriction fragment. We also recommend that the sizes of the fragments be listed in a footnote, as shown below. Essentially the same format can be used for restriction site data, with a listing of map positions rather than fragment sizes. In this example, the number of individuals with each haplotype and composite haplotype designations are provided as additional information.

Haplotype	*Apa*I	*Bam*HI	*Eco*RI	*Pst*I	*Sau*3A	*N*	Composite
1	111010	0011	101	0001	11111111111	35	AAAAA
2	011010	0011	011	1110	11111111111	39	BABBA
3	000110	1101	001	1110	11111111111	10	CBCBA
4	000001	1101	001	1110	11111111111	16	DBCBA

Sizes of fragments: *Apa*I: 0.71 kb, 0.85, 1.1, 2.9, 3.8, 5.2; *Bam*HI: 1.1, 1.7, 2.9, 3.7, 5.1; *Eco*RI: 1.2, 2.3, 2.5, 3.1, 5.4; *Pst*I: 1.3, 2.9, 3.7, 6.5; *Sau*3A: 0.25, 0.38, 0.41, 0.81, 1.2, 1.3, 1.7, 1.8, 1.9, 2.1, 2.3.

Restriction Fragment and Site Characters

For restriction fragments and sites, it is common for authors to present only their character states (present or absent), without additional information. Some analyses are based solely on the states of polymorphic or "informative" characters (e.g., maximum parsimony), which may explain why these are often the only data provided. However, there are two situations in which more information would be needed. The first is when data from two or more separate studies are combined or compared. It is then necessary to match equivalent characters from each study; this is not possible when characters are defined ambiguously. The second is when information on invariant characters is needed, either because they are not invariant in every dataset or because they are needed to estimate parameters such as sequence divergence.

Information on restriction fragments or sites is best presented in a table. The format we suggest (exemplified by Table 2) is often used in publications that provide complete datasets. In the first column, each haplotype is identified by a unique label; either a short descriptive term or number. The remainder of the table shows the presence/absence character states that define each haplotype. We recommend that authors follow the standard convention for denoting presence by a "1," absence by a "0," and, if necessary, "?" when the state is unknown. We strongly recommend against the use of blank spaces. Blanks make it more difficult to determine which column a symbol is in, for both humans and optical character recognition software. A specific restriction enzyme should be identified for each fragment or site. If character states are arrayed in columns, they can be grouped under the names of each restriction enzyme. We also suggest that information is provided on the approximate sizes of fragments or the locations of sites relative to some reference position. This information may be placed in a footnote immediately beneath the table (see Table 2). Although we did not use inclusion of this information as criteria for dataset completeness, it would be needed to match equivalent haplotypes that have been identified in separate studies. If site locations or fragment sizes are included in a table, additional figures that display gels or restriction site maps may no longer be needed, with a consequent reduction in publication length. To reduce the size of this table, another table or footnote can be used to list invariant characters.

DNA Sequence Data

Because there is a public database for sequence information, there may be little need to publish entire sequences. However, sequences in GenBank are only useful when their accession numbers are matched with the haplotype designations used in a publication (Table 3). A list of accession numbers in the publication also verifies that the sequences have, in fact, been submitted to GenBank. We recommend submission of every variant sequence to GenBank, both to avoid ambiguity and to facilitate automated retrieval. If sequences differ by indels, alignment information should also be included. The program Sequin, which is available from the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/Sequin), should be used to prepare sequences for submission. This program allows a group of aligned sequences to be submitted as a "population study," and each sequence is given a unique accession number.

Localities and Haplotype Distributions

A table should be used to provide locality information and haplotype distributions (exemplified by Table 4). The first few columns of the table should provide information on the sampled locations themselves. Locations may be designated by a concise label or abbreviation, followed by a brief description such as a nearby town or landmark. We strongly recommend that latitude and longitude coordinates also be included in the table. Coordinates are easy to obtain and, in most cases, are more precise then verbal descriptions. If the geographic scale of a study is very small (i.e., less than 1 km^2), it may be more effective to provide local Cartesian

TABLE 3. An example of a table that shows the DNA sequences that define each haplotype. The essential data in this table is a list of each haplotype by label and a GenBank accession number for its sequence. The table can also be used to show polymorphic positions. The first haplotype is used as a reference sequence; for the other haplotypes, periods indicate agreement with the reference, dashes indicate deletions, and letters indicate base substitutions. This format facilitates visual comparison of the sequences and provides landmarks that could be used to reconstruct the sequence alignment. In addition, the number of individuals with each haplotype is provided.

Haplotype	Accession number	Nucleotide Position																	N
		77	97	171	172	173	174	198	201	218	227	238	257	268	319	321	322	323	
1	Z25682	t	a	a	c	g	c	c	a	c	g	t	t	a	c	t	a	c	35
2	Z25683	.	.	.	.	.	.	.	t	a	.	a	.	t	.	.	–	.	39
3	Z25684	.	g	g	.	.	.	.	.	.	a	.	t	c	.	.	–	.	10
4	Z25685	c	.	.	.	–	–	.	.	.	c	.	a	t	c	.	–	.	16

TABLE 4. An example of a table that provides the location of each collection site and the number of individuals of each haplotype that was found at that site. Haplotype frequencies can be used in place of numbers, but only if the total number of individuals from each site is also provided.

			Haplotypes			
Site	Location	Geographic coordinates	1	2	3	4
1	8 miles N of Fairbanks, Alaska	64°55.9′N 147°41.5′W	20	0	0	0
2	7 miles NW of Fargo, North Dakota	46°58.4′N 96°58.1′W	15	5	0	0
3	10 miles SW of Columbia, Missouri	38°47.7′N 92°22.1′W	0	30	0	0
4	2 miles NE of Lafayette, Louisiana	30°18.1′N 91°59.8′W	0	0	10	10

coordinates, in appropriate units, for each location. If the compass orientation of the Cartesian axes and the geographic coordinates of the Cartesian origin are also provided, these local coordinates are more convenient than global coordinates and just as informative. The placement of precise location information in a table may eliminate the need for a map of the sampled locations, although maps that show locations may be useful as interpretive tools. After the locality information, the remainder of the columns should contain either the number or frequency of haplotypes at each location (see Table 4). If frequencies are presented, an additional column should be used to provide the total number of individuals sampled from each location. Histograms or pie diagrams are sometimes used to display haplotype frequency data. We agree with Tufte (1983) that such data are often better presented in tables; however, if diagrams are used, we suggest that numerical values be placed adjacent to each bar or pie section. In some cases, locations are pooled for purposes of analysis. This may be justified for statistical reasons, but it represents a loss of information that could be valuable in the future. We suggest that pooled data are not presented in the primary table, but in a separate figure or table.

If the above guidelines are followed, in most cases the information needed for analysis of a mtDNA survey can be presented in two tables. Placement of primary data in well-organized tables, rather than text or figures, not only facilitates the later extraction of this data from publications, but in many cases may be a more effective means of presentation (Tufte 1983). For many of the publications that we surveyed, this format would have reduced the numbers and space requirements of tables, figures, and text that were used to present the same information. This format also facilitates the transfer of data to digital media, especially if optical character recognition is used. We recognize that our suggested format may not be suitable for all publications that include data on mtDNA variation. For those cases, we suggest that authors, reviewers, and editors consider other means to make the full dataset available. Compared to the effort and expense required to acquire data on mtDNA variation, only a small investment is needed to preserve this data in a fairly complete form.

Most of the published studies we examined were funded by governmental agencies. Government agencies often have the expectation, if not the requirement, that investigators "share with other researchers, at no more than incremental cost and within a reasonable time, the data, samples, physical collections and other supporting materials created or gathered in the course of the work" (National Science Foundation Grant General Conditions [GC-1], October 1998). In our study, requests for data from individual authors were usually not successful. At present, it appears that the most effective way to meet the expectations of funding agencies is to ensure that data is accessible to potential users by publication in the peer-reviewed literature. This access will enhance research opportunities within the scientific the community and increase the effectiveness of public funding in extending the growth of scientific knowledge.

ACKNOWLEDGMENTS

We thank E. Hines and J. Waits for their assistance with literature searches. This research as funded the Grant DBI9630311 from the National Science Foundation.

LITERATURE CITED

Avise, J. C. 1994. Molecular markers, natural history, and evolution. Chapman and Hall, New York.

Burks, C., M. Cassidy, M. J. Cinkosky, K. E. Cumella, P. Gilna, J. E. D. Hayden, G. M. Keen, T. A. Kelley, M. Kelley, D. Kristofferson, and J. Ryals. 1991. GenBank. Nuleic Acids Res. 19 (Suppl.):2221–2225.

Hudson, R. R., M. Slatkin, and W. P. Madison. 1992. Estimation of levels of gene flow from DNA sequence data. Genetics 132: 583–589.

Luikart, G., and J. M. Cornuet. 1998. Empirical evaluation of a test for identifying recently bottlenecked populations from allele frequency data. Conserv. Biol. 12: 228–237.

Neigel, J. E., and J. C. Avise. 1993. Application of a random walk model to geographic distributions of animal mitochondrial DNA variation. Genetics 135:1209–1220.

Nevo, E., A. Beiles, and R. Ben-Shlomo. 1984. The evolutionary significance of genetic diversity: ecological, demographic and life history correlates. Lect. Notes Biomath 53:13–213.

Slatkin, M. 1985. Rare alleles as indicators of gene flow. Evolution 39:53–65.

Slatkin, M., and W. P. Maddison. 1989. A cladistic measure of gene flow inferred from the phylogenies of alleles. Genetics 123:603–613.

Smith, T. B., and R. K. Wayne, eds. 1996. Molecular genetic approaches in conservation. Oxford Univ. Press, New York.

Spaak, P. 1996. Temporal changes in the genetic structure of the *Daphnia* species complex in Tjeukeneer, with evidence for backcrossing. Heredity 76:539–548.

Tarr, C. L., S. Conant, and R. C. Fleischer. 1998. Founder events and variation at microsatellite loci in an insular passerine bird, the Laysan finch (*Telespiza cantans*). Mol. Ecol. 7:719–731.

Tufte, E. R. 1983. The visual display of quantitative information. Graphics Press, Cheshire, CT.

Waples, R. S. 1989. A generalized approach for estimating effective population size from temporal changes in allele frequency. Genetics 121:379–391.

Weir, B. S., and C. C. Cockerham. 1984. Estimating *F*-statistics for the analysis of population structure. Evolution 38:1358–1370.

Wright, S. 1951. The genetical structure of populations. Ann. Eugen. 15:323–354.

Corresponding Editor: R. Burton