# Emerging Technologies

*This section highlights new and emerging areas of technology and methodology. Topics may range from hardware and software, to statistical analyses and technologies that could be used in ecological research. Articles should be no longer than a few thousand words, and should be sent to the editors, David Inouye (e-mail: inouye@umd.edu) or Sam Scheiner (e-mail: sschein@nsf.gov).*

## Advancing Synthetic Ecology: A Database System to Facilitate Complex Ecological Meta-Analyses

### Introduction

Meta-analysis is the process whereby statistical techniques are used to analyze quantitatively the results from multiple studies. A common synthetic tool in the medical and social sciences since the 1980s (reviewed by Schulze 2004), meta-analysis is now increasingly being used by ecologists, and meta-analytic approaches are being developed to deal with the specific characteristics of ecological data (Gurevitch and Hedges 1999, 2001, Gurevitch et al. 2001, Gates 2002, Lajeunesse and Forbes 2003). A survey of the top 20 journals in ecology (ranked by ISI Impact Factor) using the Web of Science database reveals a striking increase in the number of published meta-analyses over time, from an average of fewer than 5 per year in the mid-1990s to more than 30 per year in 2006 (Fig. 1). This increase in the use of meta-analysis likely represents progress and maturation in the field of ecology, as more quantitative methods are now being used in attempts at research synthesis.

However, the same survey demonstrated that, although the number of ecological meta-analyses published per year is increasing, the analytical complexity of these studies is not (Fig. 1). Of 188 meta-analyses, only 15 of these (8%) conducted multifactor analyses, simultaneously examining more than one predictor variable. In ecology, some scientific questions may be addressed with meta-analysis by simply estimating an overall effect size, such as the average effect of a particular experimental treatment, or

the average correlation between two variables. However, most questions in ecology are more complex, asking how such effects change in response to one or more explanatory variables. For this reason, many manipulative experiments in ecology are multifactorial, exploring the main and interactive effects of multiple factors on a response of interest (e.g., Goldberg and Scheiner 2001). In our survey, almost 90% of the studies collected data on multiple predictor variables, but did not conduct simultaneous multifactor analyses, instead analyzing predictors one at a time. These findings indicate that although ecologists are increasingly recognizing the importance of synthesis, single-factor models dominate meta-analysis efforts in ecology.

The limited complexity of meta-analyses in part reflects limitations in available statistical methodologies and software (i.e., MetaWin [Rosenberg et al. 2000]). Methodologies for fitting multifactor mixed models in meta-analysis remain largely undeveloped and/or unavailable to most researchers in ecology, although advances are being made (e.g., Hughes et al. 2002, Borer et al. 2005, Hoeksema and Forde 2008). A second limiting factor, and the focus of this paper, is that for the results of complex meta-analysis to be meaningful, the number of studies examined should greatly outweigh the number of explanatory variables explored (Thompson and Higgins 2002), and such large data sets are not easily constructed with data management tools readily available to ecologists. Simple spreadsheets quickly become unwieldy and encourage the introduction of errors during data entry, particularly in collaborative efforts. Relational database software can address some of these issues, but it is logistically difficult for multiple collaborators to contribute data to a single database file. In this paper, we describe an online data collection and management system that was designed specifically to facilitate complex ecological meta-analyses. Our system is structured to calculate effect sizes for meta-analysis, allows for simultaneous use by multiple collaborators, includes mechanisms to minimize data entry errors, and automates the creation of custom output data sets. This data management technique is adaptable to many individual ecological problems, and its widespread adoption could facilitate more complex meta-analyses in ecology and hence progress on synthetic ecological questions.

## A data system for complex meta-analyses

### *Background*

The process of gathering, organizing, and handling the data for our own ecological meta-analysis efforts revealed the need for an advanced data management system. The main purpose of our meta-analysis was to determine the conditions under which the addition of mycorrhizal fungi can be beneficial to plants (Hoeksema et al., *unpublished manuscript*). Mycorrhizal fungi form one of the most prevalent symbioses in nature; they develop intimate associations with the roots of most plants, delivering nutrients and other benefits in exchange for photosynthates (Smith and Read 1997). Studies yield conflicting results regarding whether the addition of mycorrhizal fungi improves plant health, and it is likely that many factors (e.g., growth environment, life history traits of plant hosts, number and identity of fungal partners) influence the outcome of the symbiosis. Our goal was to use meta-analysis to identify the conditions, if any, under which mycorrhizal fungal addition is beneficial to plants.

Several characteristics of our meta-analysis efforts necessitated the creation and use of a sophisticated data management system. First, we planned to collect a large amount of data and therefore recognized

the need for a database that was flexible with respect to size and possessed the ability to scale up to fit our needs. Our initial meta-analysis literature search yielded 1853 articles, many of which reported the results of multiple independent experiments. Because the goal of the meta-analysis was to identify conditions in which the effect of mycorrhizal inoculation was beneficial to plants, we aimed to collect data on a large number of independent or explanatory variables, many different response variables, and other experimental treatment factors. A sophisticated management system was required to organize and track this large amount of data. Second, in order to calculate meta-analysis effect sizes, we needed a database system that could match control and experimental means, sample sizes, and standard deviations for all of the response variables of interest. Third, this project was the result of a working group that involved 17 international collaborators, and therefore required a system that could be used both concurrently and remotely.

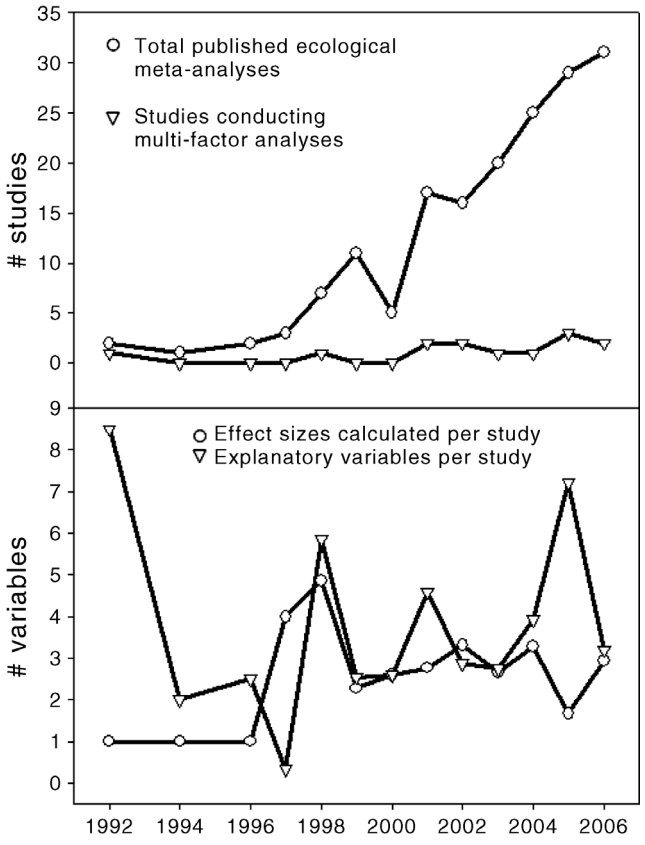
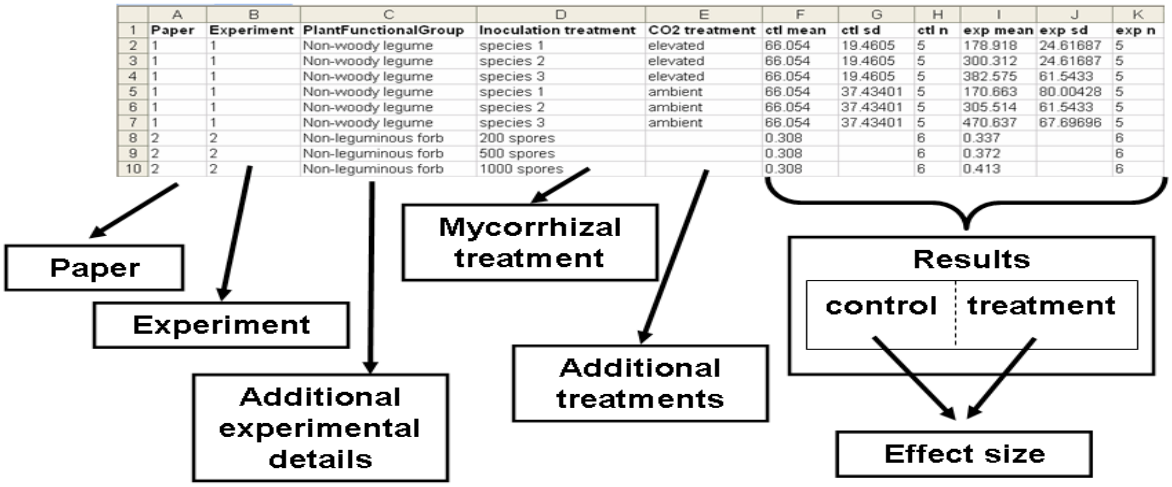



Fig. 1. The total number of meta-analyses published in the top 20 ecology journals (ranked by Thompson ISI impact factor) per year is increasing, but their analytical complexity remains low as indicated by the number of effect sizes calculated, the number of independent variables examined, and the number of studies conducting multi-factor analyses.

*MycoDB: database and Web-interface*

We created a relational database and custom Web interface, called MycoDB, to house and organize our meta-analysis data, to facilitate its simultaneous entry by multiple scientists, and to allow calculations and customized outputs. MycoDB is organized in a hierarchical manner. The data are first structured according to each research paper from which they were extracted. The papers table contains details about each unique article, such as author names and year published. These tables are then linked to another set of tables that contain details about each experiment in that paper. Because multiple experiments are often published in one paper, the papers table is linked to data from multiple experiments, the second level of organization. The experiments table contains data regarding many independent variables that are fixed for each individual experiment, such as length of experiment, species used, and other experimental conditions of interest. The third level in the hierarchy links experimental treatments with each experiment. In our meta-analysis, each experiment contained a mycorrhizal manipulation, but could also contain additional factors (e.g., nutrient addition, $CO_2$ manipulation). The treatments table is then linked to the fourth level in the hierarchy, response variables. In this step, multiple treatments—or "treatment sets" when there are many varying treatments in a single experiment—are related to one or




| | A | B | C | D | E | F | G | H | I | J | K |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Paper | Experiment | PlantFunctionalGroup | Inoculation treatment | CO2 treatment | ctl mean | ctl sd | ctl n | exp mean | exp sd | exp n |
| 2 | 1 | 1 | Non-woody legume | species 1 | elevated | 66.054 | 19.4605 | 5 | 178.918 | 24.61687 | 5 |
| 3 | 1 | 1 | Non-woody legume | species 2 | elevated | 66.054 | 19.4605 | 5 | 300.312 | 24.61687 | 5 |
| 4 | 1 | 1 | Non-woody legume | species 3 | elevated | 66.054 | 19.4605 | 5 | 382.575 | 61.5433 | 5 |
| 5 | 1 | 1 | Non-woody legume | species 1 | ambient | 66.054 | 37.43401 | 5 | 170.663 | 80.00428 | 5 |
| 6 | 1 | 1 | Non-woody legume | species 2 | ambient | 66.054 | 37.43401 | 5 | 305.514 | 61.5433 | 5 |
| 7 | 1 | 1 | Non-woody legume | species 3 | ambient | 66.054 | 37.43401 | 5 | 470.637 | 67.69696 | 5 |
| 8 | 2 | 2 | Non-leguminous forb | 200 spores | | 0.308 | | 6 | 0.337 | | 6 |
| 9 | 2 | 2 | Non-leguminous forb | 500 spores | | 0.308 | | 6 | 0.372 | | 6 |
| 10 | 2 | 2 | Non-leguminous forb | 1000 spores | | 0.308 | | 6 | 0.413 | | 6 |

Fig. 2. Example output of MycoDB in flat spreadsheet format showing data from two papers, each reporting the results of a single experiment. MycoDB itself is structured in a multi-dimensional, relational manner such that data are never repeated.

more response variables to create a corresponding "result set" that includes a value for each response variable. In other words, treatments and responses are combined in all factorial combinations to create results where mean, standard deviation, and replication values can be entered. Fig. 2 shows a sample spreadsheet output version of data from two papers that were entered into MycoDB, highlighting the levels of organization from papers to experiments to treatments to results, which are then used to calculate meta-analysis effect sizes. Paper 1 conducted a 3 × 2 factorial experiment with a single nonwoody leguminous plant species; the first factor was mycorrhizal inoculation (three different species or levels) and the second factor was $CO_2$ manipulation (ambient or elevated). Paper 2 tested the effect of a single factor, mycorrhizal inoculation at three application levels, on a nonleguminous forb species. Control (ctl) and experimental (exp) means were used to calculate meta-analysis effect sizes. It is important to note that Fig. 2 is not an illustration of MycoDB itself, which is instead structured in a multidimensional, relational manner, such that data are never repeated. We created MycoDB using Microsoft Access as the database platform, because it was available in our Web hosting environment, but other alternatives such as MySQL, Microsoft SQL Server, or Oracle may be preferred for very large data sets.

Although relational databases are not new to data management efforts in ecology, several features of MycoDB were designed specifically for unique features of meta-analyses. The most critical need of our system, and the innovation that was most useful to meta-analyses, was matching control treatment set data with noncontrol treatment set data to facilitate calculation of meta-analysis effect sizes across a wide array of experiments. Matching the control and noncontrol treatments began with a database design that would accept unlimited variations of full factorial experiments. Data for each experiment were linked to a list of treatment factors and a list of the levels within each factor. Treatment sets were created using all factorial combinations of these levels. For example, Experiment 1 in Fig. 2 contained two experimental factors: mycorrhizae and $CO_2$. The mycorrhizal treatment contained three levels (fungal species 1, fungal species 2, and fungal species 3), and the $CO_2$ treatment contained two levels (elevated and ambient), such that the experiment contained six unique treatment sets, each stored as a separate treatment set.

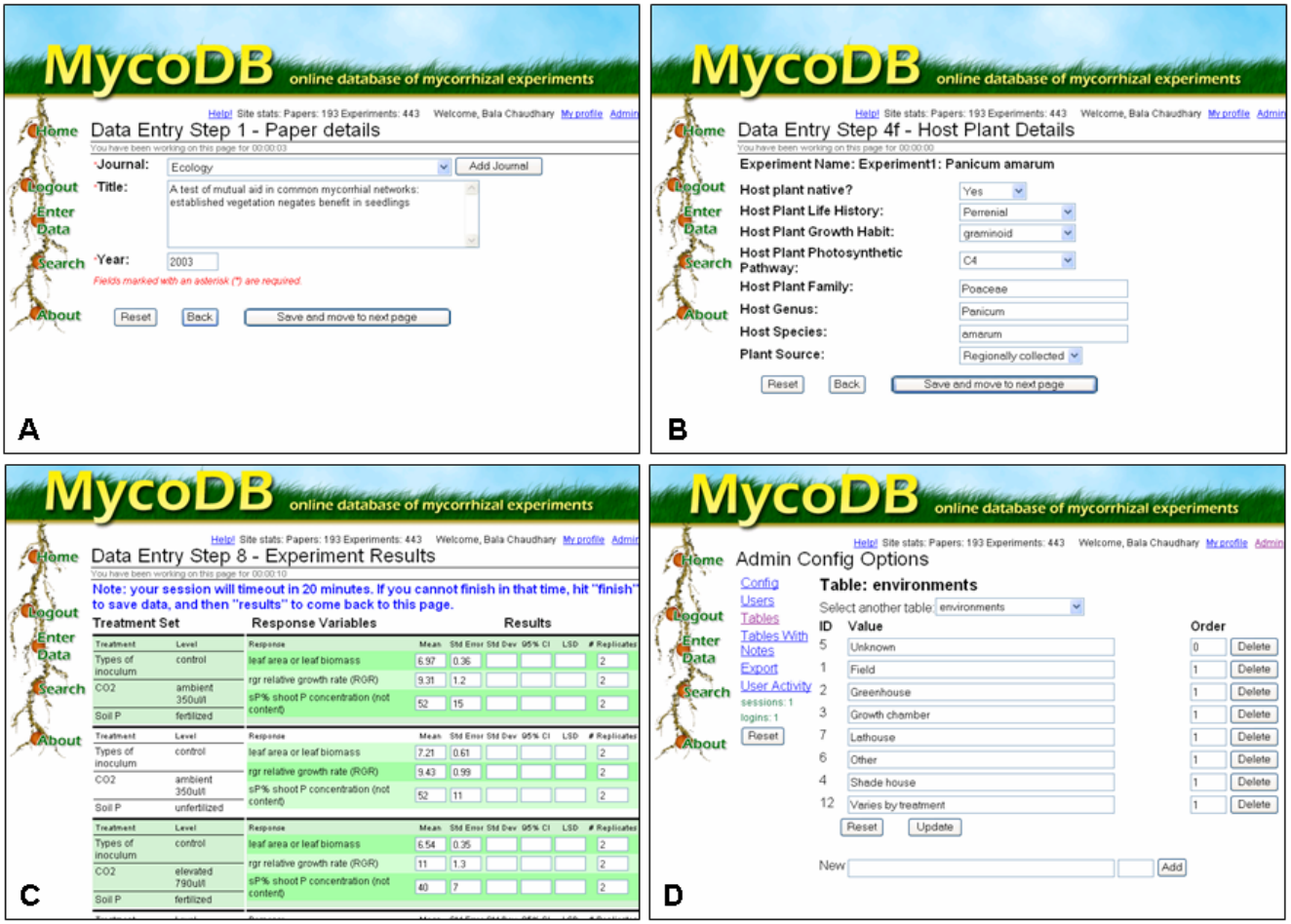



Fig. 3. Example screen shots of the data-entry web interface for MycoDB.

Each treatment set is assigned a control treatment set that matches it exactly, except that the mycorrhizal treatment level is not inoculated. For each response and treatment set combination, a new row is created in an experimental results table. After all of the results for an experiment are entered, it is possible to calculate effect sizes by comparing a response variable value from a noncontrol treatment set to the same response variable value from its control treatment set. Our database calculates log response ratios, but could easily be set up to calculate automatically other meta-analysis metrics such as Hedges' d.

MycoDB was programmed with an option to output random subsets of data from the complete data set. Nonindependence among data points can derive from multiple sources in meta-analysis data sets (Gurevitch and Hedges 1999), and one key type of nonindependence is generated when multiple effect sizes are calculated from a single experiment, using the same control for comparison with multiple treatments. These effect sizes are not statistically independent, and thus can bias the analyses by over-representing the number of independent replicates derived from a single experiment. This bias can be controlled through analyses with smaller data sets that contain only a single randomly chosen effect size from each experiment, providing useful comparison to analyses of the full data set. MycoDB can export

all data, as well as a random subset of the data that includes only a single randomly chosen effect size from each experiment.

We developed a Web site interface to facilitate data entry by members of our working group. The Web site comprises a series of forms that are linked to the database tables, all housed on a server at Northern Arizona University (Fig. 3). Basic details about each paper (A) as well as independent variables (B) and response variables (C) can be collected; certain users can be granted administrative privileges, which conveys the ability to add, remove, or edit drop-down menu choices (D). This structure enables multiple users to access and modify the same database concurrently. Each working group member can log onto the Web site using a unique username and password, making the data traceable to the user who entered it and simplifying error checking. By using the Web, the data storage is centralized and made constantly available, which made our large group collaboration possible. The commonly used alternative—e-mailing copies of a spreadsheet or database file—does not allow simultaneous access to the data and can become confusing when merging data from different versions of files. The Web interface combined with the database can also prevent mistakes by providing fixed data entry options in the form of drop-down menus. This approach prevents typical errors, standardizes the number of levels for each independent variable (easing analyses), and streamlines the process of collecting data in different units (e.g., kg/ha vs. ppm). We used "Classic" ASP to create the Web interface, because it was available in our hosting environment, but any other Web language would be suitable. A generic version of the entire MycoDB application, as well as the database, is provided as supplementary materials.

## Advantages of using a data management system

### *Facilitation of complex meta-analyses*

The use of a data management system has the potential to improve the depth of our synthetic understanding in ecology, as demonstrated by our mycorrhizal meta-analysis. Utilizing MycoDB and its associated Web interface, our working group initially collected data from nearly 2000 laboratory and field trials, which enabled the usage of a multifactor meta-regression approach and multimodel inference to determine the relative importance of 10 different explanatory variables to plant response to inoculation with mycorrhizal fungi. We found nitrogen fertilization and plant functional groups to be the most important explanatory variables, and that inoculation was more beneficial to plants when the soil community was more complex (J. D. Hoeksema et al., *unpublished manuscript*).

Conducting less complex analyses in our mycorrhizal meta-analysis might have led to spurious conclusions and potentially erroneous management recommendations. For example, across all studies in our analysis, the average response to inoculation was positive. Had we stopped at this analysis stage, recommendations for ecosystem management might have included the purchase and application of mycorrhizal fungal inoculum in all circumstances in which plant productivity is valued. Instead, in explorations of single-factor models as part of our overall analysis, we found that the size and occasionally direction of effects of particular factors often differed substantially when effects of other factors were not controlled in the same statistical model. By conducting multi-factor analyses with multiple explanatory variables, our results showed that the benefit conferred by the addition of mycorrhizal fungi depends more on the functional group of the plant, the fertility of the soil, and the biotic complexity of the soil,

compared to other factors. Given that the majority of ecological meta-analyses conducted single-factor analyses (Fig. 1), there is considerable potential for spurious interpretations leading to poor ecosystem management decisions.

*A large volume of quality data*

A main advantage of using a data management system was that, after relatively few hours of data entry per person, our working group was able to compile a tremendous amount of complex data. The large amount of data collected in our mycorrhizal meta-analysis efforts was effectively *the* reason we could examine multiple predictors simultaneously in analyses. In fact, our mycorrhizal meta-analysis contained 2.5× more papers, 12× more response variables, and 6× more explanatory variables than the average ecological meta-analysis in our survey. It is possible that ecological meta-analyses include fewer studies and response or explanatory variables because of a general paucity of published data regarding the particular topic being quantitatively summarized. However, we argue that data management limitations have also contributed to the small amount of data summarized in most ecological meta-analyses.

Utilization of a data collection and management system can also improve the quality of meta-analysis data in specific ways. First, a single copy of the database is maintained on a secure server, reducing errors associated with version tracking. Second, a Web interface for data collection with drop-down menus (Fig. 3) can reduce errors, particularly in the case of collaborative efforts where multiple researchers are entering data. Generally, spreadsheets are used to organize data, information is extracted from articles, and typed into spreadsheet fields (Gurevitch and Hedges 2001). In collaborations, the spreadsheet is e-mailed to multiple researchers so that they can enter data from other articles. Problems can arise not only with data entry typos, but also when all collaborators have the ability to add an unlimited number of categories to each field. For example, in our mycorrhizal meta-analysis we wanted to determine whether the environment in which an experiment is conducted (e.g. field, greenhouse, growth chamber) affects the plant response to mycorrhizal inoculation. Using a limited number of choices for drop-down menus, we could restrict the choices and thus the number of levels for this factor. This approach is important because statistically testing for differences among groups in ANOVA-type analyses requires more degrees of freedom as the number of levels increases (Sokal and Rohlf 1995).

*Promotion of a priori thought and collaboration*

The process of database development encouraged our working group to put a priori thought into the process of question development, analysis, and data generation, thereby improving the inferences we were able to make from the meta-analysis. It has been argued that science culture generally focuses more on data analysis than on question formulation, thus deterring front-end critical thinking and questions that make biological sense (Burnham and Anderson 2002). Accepted theory, expert background knowledge, and prior information should be carefully incorporated into the early hypothesis formulation stage of research (Chatfield 1995). We used our group's collective knowledge to generate a list of 44 candidate variables that could influence plant response to mycorrhizal inoculation. This process was initially necessary to determine the technical requirements for MycoDB prior to its construction, but in the end served to promote a priori thought and improved hypothesis development for our mycorrhizal meta-analysis.

## Conclusions

Meta-analyses are becoming an increasingly common mode of synthesis in the field of ecology, yet few conduct analyses with more than one response variable, independent variable, or factor. As a result, researchers may be preferentially addressing low-complexity syntheses, as demonstrated by a survey of recently published studies. We argue that limitations in data management have contributed to this phenomenon, and describe the creation of MycoDB, a relational database and associated Web interface that facilitated more complex meta-analyses in our own research efforts. Progress in synthetic research in ecology, including complex meta-analysis, should be more rapid as additional informatics tools become available. We have presented one example. As the amount of ecological data requiring synthesis increases, and collaborations become larger and more interdisciplinary, future meta-analyses will benefit from such data management techniques.

## Acknowledgments

This work was conducted as a part of the "Narrowing the Gap between Theory and Practice in Mycorrhizal Management" working group, supported by the National Center for Ecological Analysis and Synthesis, which is supported by the National Science Foundation (NSF grant DEB-0072909), the University of California at Santa Barbara, and the state of California. The project was also supported by funding from the Radcliffe Institute for Advanced Study at Harvard University. We thank Northern Arizona University for hosting the MycoDB and its associated Web site. Additional funding was provided by the Northern Arizona University e-Learning Center to L. L. Walters. and a National Science Foundation IGERT Fellowship (NSF grant DGE-0549505) to V. B. Chaudhary. Nancy C. Johnson and Catherine A. Gehring provided helpful comments on earlier drafts.

## Literature cited

Burnham, K. P., and D. R. Anderson. 2002. Model selection and multimodel inference: a practical information-theoretic approach. Second edition. Springer-Verlag, New York, New York, USA.

Chatfield, C. 1995. Model uncertainty, data mining and statistical-inference. Journal of the Royal Statistical Society Series A. Statistics in Society 158:419–466.

Gates, S. 2002. Review of methodology of quantitative reviews using meta-analysis in ecology. Journal of Animal Ecology 71:547–557.

Goldberg, D. E., and S. M. Scheiner. 2001. ANOVA and ANCOVA: field competition experiments. Pages 46–67 *in* S. M. Scheiner and J. Gurevitch, editors. Design and analysis of ecological experiments. Second edition. Chapman and Hall, New York, New York, USA.

Gurevitch, J., P. S. Curtis, and M. H. Jones. 2001. Meta-analysis in ecology. Advances in Ecological Research 32:199–247.

Gurevitch, J., and L. V. Hedges. 1999. Statistical issues in ecological meta-analyses. Ecology 80:1142–1149.

Gurevitch, J., and L. V. Hedges. 2001. Meta-analysis: combining the results of independent experiments. Pages 347–369 *in* S. M. Scheiner and J. Gurevitch, editors. Design and analysis of ecological experiments. Second edition. Chapman and Hall, New York, New York, USA.

Hoeksema, J. D., and S. E. Forde. 2008. A meta-analysis of factors affecting local adaptation between interacting species. American Naturalist 171:275–290.
Lajeunesse, M. J., and M. R. Forbes. 2003. Variable reporting and quantitative reviews: a comparison of three meta-analytical techniques. Ecology Letters 6:448–454.
Rosenberg, M. S., D. C. Adams, and J. Gurevitch. 2000. MetaWin: statistical software for meta-analysis. Version 2. Sinauer Associates, Sunderland, Massachusetts, USA.
Schulze, R. 2004. Meta-analysis: a comparison of approaches. Hogrefe and Huber, Cambridge, Massachusetts, USA.
Smith, S. E., and D. J. Read. 1997. Mycorrhizal symbiosis. Second edition. Academic Press, New York, New York, USA.
Sokal, R. R., and F. J. Rohlf. 1995. Biometry: the principles and practice of statistics in biological research. Third edition. W. H. Freeman, New York, New York, USA.
Thompson, S. G., and J. P. T. Higgins. 2002. How should meta-regression analyses be undertaken and interpreted? Statistics in Medicine 21:1559–1573.

[1]V. Bala Chaudhary, [2]Lawrence L. Walters, [3]James D. Bever, [4]Jason D. Hoeksema, [5]Gail W.T. Wilson

[1]Department of Biological Sciences, Northern Arizona University, Flagstaff, AZ 86011-5640
[2]e-Learning Center, Northern Arizona University, Flagstaff, AZ 86011-5682
[3]Department of Biology, Indiana University, Bloomington, IN 47405
[4]Department of Biology, University of Mississippi, University, MS 38677
[5]Department of Natural Resource Ecology and Management, Oklahoma State University, Stillwater, OK 74078

## SUPPLEMENT

Sample database and web application. (*Ecological Archives* B091-001-S1).