Minimum Phylogenetic Coverage: An Additional Criterion to Guide the Selection of Microbial Pathogens for Initial Genomic Sequencing Efforts

Stephen B. Goodwin

Crop Production and Pest Control Research, U.S. Department of Agriculture–Agricultural Research Service, Department of Botany and Plant Pathology, Purdue University, West Lafayette, IN 47907.

Accepted for publication 2 April 2004.

ABSTRACT

Goodwin, S. B. 2004. Minimum phylogenetic coverage: An additional criterion to guide the selection of microbial pathogens for initial genomic sequencing efforts. *Phytopathology* 94:800-804.

The scientific community has made a persuasive case to increase funding for genomic sequencing of microbial plant pathogens, and a number of objective criteria were developed to guide the selection of sequencing targets. However, this task still is complicated for fungi and Oomycetes, which are extremely diverse evolutionarily. Lack of a definite target number of organisms to be sequenced and inclusion of only a small segment of fungal diversity weaken the effort, which may reduce support and hinder efforts to increase the level of funding. These problems can be minimized by inclusion of phylogenetic relationship as an additional criterion to prioritize organisms for genomic sequencing. A phylogenetic analysis of 18S ribosomal DNA sequences revealed that many of the species proposed for genomic sequencing are very closely related, while other evolutionarily important groups, such as the Taphrinales, are under-

represented or ignored. By choosing one representative from each of the major clusters on the 18S tree, a minimum phylogenetic coverage (MPC) of the fungi and Oomycetes that have been proposed for genomic sequencing can be achieved by including seven Ascomycetes, four Basidiomycetes, and one Oomycete. A second round of 14 species could be sequenced to cover the major sub-branches within each of the 12 major clusters. This approach defines a tangible, achievable goal for genomic sequencing that could be used to lobby for additional funding. MPC also would assure support and participation by the largest number of plant pathologists because most would have access to a sequenced genome from a close relative of their organism of interest. Allocating funds for this effort might be achieved best through a competitive bidding process among sequencing centers rather than the traditional grants programs. MPC provides a logical criterion to help prioritize fungi, Oomycetes, and other microbial pathogens for genomic sequencing that generates a defined, tractable number of sequencing targets and could serve as a rallying point to unify plant pathologists toward achieving a common goal.

Prioritizing a list of plant pathogens for initial genomic sequencing efforts is a difficult challenge, particularly for fungi. The relatively large genome sizes of fungal plant pathogens limits the number that can be sequenced for a given amount of funding compared with that of bacteria or viruses. Furthermore, the fungi, and fungal-like organisms such as the Stramenopiles (class Oomycetes), are extremely divergent evolutionarily. This restricts the utility of genomic sequence information; data from one species may be of limited value for fungi that are distantly related. For this reason, a minimum phylogenetic coverage (MPC) of the fungal kingdom should be one of the goals of initial microbial sequencing efforts.

A number of objective criteria were developed to aid the choice of candidates for initial genomic sequencing. These include (i) economic importance of the organism and relevance to agriculture or forestry; (ii) unique biological or environmental traits; (iii) broad interest to a significant community of scientists; (iv) genetic tractability; and (v) availability of tools or other biological resources such as genomic libraries and genetic maps (4). Genetic

tractability of the host was another important consideration so that both sides of the host–pathogen interaction could be analyzed.

These are excellent criteria. However, the phylogenetic relationships among the organisms proposed for sequencing so far have not been emphasized. These relationships must be considered to broaden the utility of information obtained from initial sequencing efforts. One of the benefits from having multiple genomes sequenced will be in the area of comparative genomics, and this is more likely to yield globally important insights when comparing unrelated organisms with potentially different mechanisms of pathogenicity than by analysis of very closely related species. For example, a recent analysis of polyketide synthase genes in sequenced genomes from diverse species revealed major events in the evolution of these genes, and showed that the major clusters of polyketide synthases were present in most fungi (3), indicating that they were derived from a common ancestor and were not acquired by horizontal gene transfer. This study was limited by the low representation of fungal diversity included among sequenced genomes, yet still provided an unprecedented understanding of fungal polyketide synthases that was not possible only by analysis of close relatives. Additional insight into the possible biology of these genes could have been obtained by analysis of a broader phylogenetic spectrum, which could guide future analyses in functional genomics. Due to the high diversity of pathogenic fungi, choosing the broadest phylogenetic representation is the most logical way to distribute the initial sequencing funds to facilitate additional groundbreaking research through comparative genomic analysis.

MPC also will ensure the greatest participation by scientists in follow-up studies of functional genomics and other areas, which

Corresponding author: S. B. Goodwin; E-mail address: sgoodwin@purdue.edu

Names are necessary to report factually on available data. However, the USDA neither guarantees nor warrants the standard of the product, and the use of the name implies no approval of the product to the exclusion of others that also may be suitable.

Publication no. P-2004-0524-010

This article is in the public domain and not copyrightable. It may be freely reprinted with customary crediting of the source. The American Phytopathological Society, 2004.

currently is limited. For example, a recent request for proposals for the U.S. Department of Agriculture (USDA) National Research Initiative Integrated Program (information is available online from the USDA CSREES [Cooperative State Research, Education, and Extension Services]) in a section on "Functional Genomics of Agriculturally Important Organisms" said "Research in this area should address the characterization of the molecular mechanisms responsible for microbial processes, *enabled by the availability of a sequenced microbial genome or genomes*. Research activities should characterize on a large scale the function of genes or networks of genes *in microbe(s) having a completely, or almost completely, sequenced genome...* Projects are expected to *utilize microarrays* to analyze the spatial and/or temporal expression of sets of genes or to identify genes expressed under different environmental conditions or as part of particular metabolic or regulatory pathways." (emphasis added). This grant program is limited to those working on organisms with a sequenced genome; everyone else is excluded. Microarrays made from one species are not likely to work with those from different phylogenetic groups. Such a restricted focus for grant programs seriously limits participation by plant pathologists. Either we all must start working on *Magnaporthe oryzae* or *Fusarium graminearum*, for which the sequence and microarrays are available, or more pathogens should be sequenced quickly. Achieving MPC is essential or large numbers of plant pathologists will be excluded from future grant programs.

Given the high diversity within the fungal kingdom, it is somewhat surprising when multiple members of the same genus or order are given serious consideration for initial sequencing. Sequencing of closely related species certainly has value and should be completed eventually, but not until MPC has been achieved. For example, the top 13 species in the initial prioritized list in the APS White Paper, "*Microbial Genomic Sequencing: Perspectives of the American Phytopathological Society*," included two members of the genus *Fusarium* (*F. graminearum* and *F. moniliforme*) and six Sordariomycetes (*M. oryzae*, *F. graminearum*, *F. moniliforme*, *Colletotrichum graminicola*, *Cryphonectria parasitica*, and *Epichloë* sp.). All of these were included for very good reasons. However, entire groups of equally important fungi from unrelated groups were excluded or given much lower priority. The current method of choosing priorities could be improved greatly by including phylogenetic relationships among the selection criteria.

The main reason for the lack of phylogenetic diversity among the organisms proposed for sequencing probably is that most molecular biologists and many plant pathologists are not well versed in fungal phylogeny. Furthermore, the research of many scientists is justifiably concentrated on a few genetically tractable systems. So far, no trees showing the evolutionary relationships among the organisms proposed for sequencing are available. To rectify this situation, I have constructed a tree showing the phylogenetic relationships among the Ascomycetes and Basidiomycetes that have been proposed for genomic sequencing (Fig. 1). The Oomycetes are not related to the true fungi but were included on the tree because they often are incorrectly lumped with the fungi. This analysis was not intended to be comprehensive; additional species were added only to fill out branches of the tree or to rectify what I perceived as serious omissions. Obviously, many others could be added, but the existing tree provides a beginning. The tree shows a neighbor-joining (8) analysis of 18S (small subunit) ribosomal DNA sequences downloaded from GenBank, obtained from colleagues, or from my own sequencing efforts. The sequences were aligned with ClustalX (9) and the alignment was modified manually when necessary; this was particularly important to align the sequences of *Cryphonectria parasitica*, *Epichloë typhina*, and *Ustilago maydis*, which contained unique introns. The tree was made with Kimura's two-parameter method for estimating evolutionary distances (2). Gaps were ignored when preparing the final tree, which was displayed

with NJplot (5). The choice of gene(s) for phylogenetic analysis is not important as long as an accurate representation of past evolutionary history is obtained. The 18S region was chosen for analysis because it is the only sequence available for most proposed sequencing targets and is sufficiently conserved to permit easy alignment. The final tree and alignment were deposited in TreeBASE under accession no. SN1836. A more complete analysis of Ascomycete 18S sequences was published recently by Berbee (1).

This analysis of species proposed for genomic sequencing reveals that some groups are vastly over-represented, while others that do not have strong proponents in the scientific community should be included to ensure MPC. For example, two species of *Fusarium* have been proposed for sequencing in addition to *F. graminearum*, which was sequenced recently. Within the order Pleosporales, *Cochliobolus heterostrophus*, *Pyrenophora tritici-repentis*, *Alternaria alternata*, and *Alternaria brassicicola* all have been suggested for sequencing, even though they are extremely closely related. More distantly related species within the same order, such as *Leptosphaeria maculans* or *Phaeosphaeria* (anamorph *Stagonospora*) *nodorum*, are much more appropriate to achieve MPC but so far have been ignored. The Taphrinales is basal to all Ascomycetes and as such could yield important insight into early evolutionary events, yet also has no strong proponents within the scientific community and has not been strongly supported for sequencing. Obviously, the annotation of sequences from species chosen solely for phylogenetic completeness will not be aided by a large user community. However, the potential benefits of sequence comparisons over a greater phylogenetic range justify the inclusion of a few less-studied species among those chosen for initial genomic sequencing.

Scientists working on plant genomes have long recognized the advantages of a broad phylogenetic representation and have recommended that phylogenetic position, genome size, and experimental tractability be the only criteria used to choose species for sequencing, with phylogenetic position being the most important (6). These criteria lead to many poorly studied green plants being recommended for sequencing, even though they lack large user communities. The potential gains from including phylogenetically diverse, nonmodel species among those chosen for sequencing have encouraged plant scientists to expand their viewpoint. Scientists working on fungi and Oomycetes should develop a similarly broad vision.

By analysis of the phylogenetic tree for fungi and Oomycetes, it is possible to generate a short list for initial genomic sequencing efforts that would assure thorough representation for MPC. The phylogenetic tree was used to place all of the species proposed for sequencing into primary bins and sub-bins for the first and second rounds of sequencing, respectively (Table 1). Organisms for the first round of genomic sequencing should be chosen to represent the 12 primary bins. The sub-bins could be used to choose organisms that would broaden coverage after one representative from each of the primary bins has been sequenced. The choice of specific representative from each bin should be left to scientists working on each group based on the usual criteria; possible choices are indicated in Table 1. MPC could be achieved in the first round by sequencing seven Ascomycetes, four Basidiomycetes, and one Oomycete. So far, draft sequences of *M. oryzae*, *F. graminearum*, and *Ustilago maydis* are available, which represent bins 4, 5, and 10. Initial coverage of each of the remaining bins could be accomplished by sequencing *Aspergillus flavus*, *Mycosphaerella graminicola*, *Botryotinia fuckeliana*, *Cochliobolus heterostrophus*, and *Taphrina deformans* for the Ascomycetes, *Tilletia caries*, *Puccinia graminis*, and *Heterobasidion annosum* to represent the Basidiomycetes, and *Phytophthora infestans* or *P. sojae* from the Oomycetes. Fourteen additional species representing the remaining sub-bins should be added to ensure MPC in the second round of sequencing.

Fig. 1. Neighbor-joining analysis of 18S ribosomal DNA sequences from plant-pathogenic fungi and members of the class Oomycetes proposed for genomic sequencing. Sequences of additional species were added to fill out the tree. Species proposed for sequencing are indicated in larger, bold font; those that have been sequenced already are boxed. Most sequences were downloaded from GenBank. Species that are not represented in the 18S database are indicated to the right of their most likely positions. Orders are indicated by brackets to the right with three- or four-letter abbreviations as follows: Eur, Eurotiales; Doth, Dothideales; Ery, Erysiphales; Helo, Helotiales; Phyl, Phyllachorales; Hypo, Hypocreales; Sord, Sordariales; Dia, Diaporthales; Ophi, Ophiostomatales; Pleo, Pleosporales; Sacc, Saccharomycetales; Taph, Taphrinales; Schi, Schizosaccharomycetales; Till, Tilletiales; Ured, Uredinales; Usti, Ustilaginales; Bol, Boletales; Agar, Agaricales; Lyco, Lycoperdales; Cera, Ceratobasidiales; Sap, Saprolegniales; and Pyth, Pythiales. The correct orders for the genera *Cladosporium*, *Magnaporthe*, and *Venturia* are uncertain. Primary bins and sub-bins to achieve minimum phylogenetic coverage are indicated by the second set of brackets on the right; bracket numbers correspond to the bins in Table 1. All bootstrap values above 70 (percent of 1,000 replications) are indicated at the appropriate nodes. Branch lengths are proportional to genetic distance, which is indicated by a bar at the upper left. The primary branches leading to the true fungi and Oomycetes and that for the Uredinales clade were shortened artificially for a more compact presentation; these shortened branches are indicated by slashes. The Oomycetes deserve a separate analysis but were included here for continuity with previous lists.

An advantage of this system is that it is open-ended and can be updated easily by future input from the scientific community. Species not on the current tree could be added later and given a priority based on their phylogenetic distance from those chosen previously. It also may be necessary to make expanded trees, or to base them on other sequences, to ensure MPC of particular groups. For example, the current tree does not do justice to the Oomycetes, because important groups such as the Albuginaceae and Peronosporaceae were excluded due to a lack of suitable 18S sequences in GenBank. A recent analysis of 28S (large subunit) rDNA sequences (7) provided much better resolution and indicated that *Albugo candida* and *Bremia lactucae* probably should be included to ensure MPC of the Oomycetes. Trees for each major phylogenetic lineage should be constructed and updated

through a consensus of scientists working on each group. Once MPC has been achieved, sequencing of additional isolates of the same or closely related species could be decided on a case-by-case basis through the normal competitive grants process.

If MPC becomes a major consideration when prioritizing organisms for sequencing, then alternative means of funding the effort also should be explored. The current method, whereby scientists in academia and government write proposals through competitive grants programs, may not be the most efficient. A better approach might be to make a prioritized list by including MPC among the usual criteria for choosing sequencing targets, and then to allocate the available funding for sequencing through a competitive bidding process. Market forces among the competing institutions should lead to the most efficient sequence gen-

TABLE 1. Binning of species proposed for genomic sequencing efforts and suggestions for the best representative from each

Bin	Sub-bin	Order	Species ^a	Best bin representative ^b	Best sub-bin representative ^c
1	— ^d	Eurotiales	<i>Aspergillus flavus</i> <i>Penicillium</i> sp.	<i>Aspergillus flavus</i>	<i>Aspergillus flavus</i>
2	2A	Dothideales	<i>Cladosporium</i> sp.	—	<i>Cladosporium</i> sp.
	2B	Dothideales	<i>Cercospora kikuchii</i> <i>Mycosphaerella graminicola</i>	<i>Mycosphaerella graminicola</i>	<i>Mycosphaerella graminicola</i>
	2C	Dothideales	<i>Botryosphaeria</i> sp. <i>Dothidea</i> sp.	—	<i>Botryosphaeria</i> sp.
3	3A	Erysiphales	<i>Blumeria graminis</i>	—	<i>Blumeria graminis</i>
	3B	Helotiales	<i>Botryotinia fuckeliana</i> <i>Monilinia laxa</i> <i>Sclerotinia sclerotiorum</i>	<i>Botryotinia fuckeliana</i>	<i>Botryotinia fuckeliana</i>
4	4A	Phyllachorales	<i>Colletotrichum graminicola</i>	—	<i>Colletotrichum graminicola</i>
	4B	Hypocreales	<i>Verticillium dahliae</i>	—	<i>Verticillium dahliae</i>
	4C	Hypocreales	<i>Epichloë</i> sp. <i>Fusarium graminearum</i> <i>Fusarium moniliforme</i> <i>Fusarium oxysporum</i> <i>Trichoderma harzianum</i>	<i>Fusarium graminearum</i>	<i>Fusarium graminearum</i>
5	5A	Diaporthales	<i>Cryphonectria parasitica</i>	—	<i>Cryphonectria parasitica</i>
	5B	—	<i>Magnaporthe oryzae</i>	<i>Magnaporthe oryzae</i>	<i>Magnaporthe oryzae</i>
	5C	Ophiostomatales	<i>Ophiostoma novo-ulmi</i>	—	<i>Ophiostoma novo-ulmi</i>
6	6A	—	<i>Venturia inaequalis</i>	—	<i>Venturia inaequalis</i>
	6B	Pleosporales	<i>Leptosphaeria maculans</i> <i>Phaeosphaeria nodorum</i>	—	<i>Phaeosphaeria nodorum</i>
	6C	Pleosporales	<i>Alternaria alternata</i> <i>Alternaria brassicicola</i> <i>Cochliobolus heterostrophus</i> <i>Pyrenophora tritici-repentis</i>	<i>Cochliobolus heterostrophus</i>	<i>Cochliobolus heterostrophus</i>
7	—	Taphrinales	<i>Taphrina deformans</i>	<i>Taphrina deformans</i>	<i>Taphrina deformans</i>
8	—	Tilletiales	<i>Tilletia caries</i>	<i>Tilletia caries</i>	<i>Tilletia caries</i>
9	9A	Uredinales	<i>Cronartium ribicola</i> <i>Peridermium harknessii</i>	—	<i>Cronartium ribicola</i>
	9B	Uredinales	<i>Gymnosporangium juniperi-virginianae</i>	—	<i>Gymnosporangium juniperi-virginianae</i>
	9C	Uredinales	<i>Phakopsora meibomia</i> <i>Phakopsora pachyrhizi</i>	—	<i>Phakopsora pachyrhizi</i>
	9D	Uredinales	<i>Puccinia graminis</i> <i>Uromyces appendiculatus</i>	<i>Puccinia graminis</i>	<i>Puccinia graminis</i>
10	—	Ustilaginales	<i>Ustilago hordei</i> <i>Ustilago maydis</i>	<i>Ustilago maydis</i>	<i>Ustilago maydis</i>
11	11A	Agaricales	<i>Armillaria mellea</i> <i>Armillaria ostoyae</i> <i>Heterobasidion annosum</i>	<i>Heterobasidion annosum</i>	<i>Heterobasidion annosum</i>
	11B	Ceratobasidiales	<i>Rhizoctonia solani</i>	—	<i>Rhizoctonia solani</i>
12	12A	Saprolegniales	<i>Aphanomyces euteiches</i>	—	<i>Aphanomyces euteiches</i>
	12B	Pythiales	<i>Peronosclerospora sorghi</i> <i>Phytophthora infestans</i> <i>Phytophthora sojae</i> <i>Pythium ultimum</i>	<i>Phytophthora infestans</i>	<i>Phytophthora infestans</i>

^a All species proposed for sequencing from each bin are indicated, plus those that were apparent from the phylogenetic analysis.

^b These species represent the 12 primary bins for the initial round of sequencing. Representatives of bins 4 (*Fusarium graminearum*), 5 (*Magnaporthe oryzae*), and 10 (*Ustilago maydis*) have been sequenced already, while close relatives of the species in bins 1 (*Aspergillus nidulans*) and 7 (*Schizosaccharomyces pombe*) will be completed soon.

^c These 26 species cover all of the bins and sub-bins for primary and secondary rounds of sequencing.

^d Not needed or not applicable.

eration at the lowest cost. The scientific community should decide on the strain to be sequenced, which could be provided for free by a government or academic collaborator. The fungal genomics initiative at the National Institute of Health and the Community Sequencing Program at the Joint Genome Institute both allocate funding through a process similar to that described above: organisms are selected for sequencing based on scientific merit, and the sequencing is performed in house or by direct contracts with sequencing centers. This type of approach could serve as a more efficient model for the USDA.

Establishing MPC could be a major selling point for USDA program officers to lobby for additional funding from Congress. Previous initiatives to increase funding for microbial genomics did not specify how many species should be sequenced. Defining a target number of species by consideration of MPC would provide a clear, achievable goal, which might be easier to promote in Congress than requests for increased funding based solely on the potential benefits to science. Accomplishing MPC of fungal plant pathogens also would place the USDA as a contender with the National Science Foundation for achieving major genomics milestones, which would increase the scientific prestige of the USDA, The American Phytopathological Society (APS), and plant pathologists in general. Large sequencing centers such as the Whitehead Institute should be encouraged to consider MPC when prioritizing their lists of plant-pathogenic microorganisms for sequencing. If sequencing efforts are changed to a contract basis, then the competitive grants programs could focus on using the sequence information to do science rather than just sequencing.

Including MPC as an important criterion when choosing organisms for initial genomic sequencing efforts can guide and simplify the process of making a prioritized list. Sequencing 12 species in the first round and 14 in the second would achieve MPC for fungal and Oomycete plant pathogens. However, the Oomycetes are distinct from the true fungi and probably should be analyzed separately; they were included here only for continuity with previous lists. For a broader understanding of pathogenicity in general, the MPC concept should be expanded to encompass a larger phylogenetic spectrum of plant pathogens such as nematodes, parasitic plants, or even insects for future comparative genomics analyses. This approach could be adapted easily for these or other organisms, such as bacteria or even human pathogens. It is a logical, objective, and fair criterion to help allocate funding and would facilitate participation by the largest number of scientists in future endeavors. Allocating se-

quencing dollars on a contract basis would provide the greatest return for a given amount of funding and would allow the competitive grants programs to concentrate on what they do best—fund investigators to do science rather than sequencing. Achieving MPC could be a major rallying point to unify plant pathologists around a common goal. Sequencing of at least the first 12 species is a tangible, realistic milestone for APS and the USDA to strive for together.

ACKNOWLEDGMENTS

This work was supported by the USDA CRIS project 3602-22000-011-00D. I thank S. Kamoun for discussions about mechanisms of funding sequencing efforts and J. Cavaletto for generating the 18S sequences of *Cercospora zae-maydis*, *Cladosporium iridis*, *Mycosphaerella fijiensis*, *M. graminicola*, and *M. musicola*. L. Vaillancourt supplied the sequence of *Colletotrichum acutatum*, and L. Dunkle, S. Kamoun, M. Palm, and J.-R. Xu provided helpful comments on a previous draft of the manuscript. Published as paper 17375, Purdue University Agricultural Experiment Station.

LITERATURE CITED

1. Berbee, M. L. 2001. The phylogeny of plant and animal pathogens in the Ascomycota. *Physiol. Mol. Plant Pathol.* 59:165-187.
2. Kimura, M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* 16:111-120.
3. Kroken, S., Glass, N. L., Taylor, J. W., Yoder, O. C., and Turgeon, B. G. 2003. Phylogenomic analysis of type I polyketide synthase genes in pathogenic and saprobic ascomycetes. *Proc. Natl. Acad. Sci. USA* 100:15670-15675.
4. Leach, J. E., Gold, S., Tolin, S., and Eversole, K. 2003. A plant-associated microbe genome initiative. *Phytopathology* 93:524-527.
5. Perrière, G., and Gouy, M. 1996. WWW-Query: An on-line retrieval system for biological sequence banks. *Biochimie* 78:364-369.
6. Pryer, K. M., Schneider, H., Zimmer, E. A., and Banks, J. A. 2002. Deciding among green plants for whole genome studies. *Trends Plant Sci.* 7:550-554.
7. Riethmüller, A., Voglmayr, H., Göker, M., Weiß, M., and Oberwinkler, F. 2002. Phylogenetic relationships of the downy mildews (Peronosporales) and related groups based on nuclear large subunit ribosomal DNA sequences. *Mycologia* 94:834-849.
8. Saitou, N., and Nei, M. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4:406-425.
9. Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F., and Higgins, D. G. 1997. The CLUSTAL_X windows interface: Flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* 25:4876-4882.