Fungal Genome Initiative

A White Paper for Fungal Comparative Genomics

June 10, 2003

Submitted by The Fungal Genome Initiative Steering Committee
Corresponding authors: Bruce Birren, Gerry Fink, and Eric Lander, Whitehead Institute
Center for Genome Research, 320 Charles Street, Cambridge, MA 02141 USA
Phone 617-258-0900; E-mail bwb@genome.wi.mit.edu

1. Overview

The goal of the Fungal Genome Initiative is to provide the sequence of key organisms across the fungal kingdom and thereby lay the foundation for work in medicine, agriculture, and industry. The fungal and genomics communities have worked together for over 2 years to choose the most informative organisms to sequence from the more than 1.5 million species that comprise this kingdom. The February 2002 white paper identified an initial group of 15 fungi. These fungi present serious threats to human health, serve as important models for biomedical research, and provide a wide range of evolutionary comparisons at key branch points in the 1 billion years spanned by the fungal evolutionary tree.

The Fungal Genome Initiative (FGI) has garnered attention from a broad group of scientists through presentations at meetings, publications, and the release of its first genome sequences. The biological community's interest in the project has grown steadily, resulting in nearly 100 nominations of organisms to be sequenced. Simultaneously, the methods and strategies for effective comparative studies have been clarified by recent whole-genome comparisons of yeasts. Recognizing the power of these comparative approaches, the FGI Steering Committee has identified a coherent set of 44 new fungi as immediate targets for sequencing with an emphasis on clusters of related species.

In this white paper, we propose to sequence additional fungi that include well-studied models important to human health and welfare from poorly understood regions of the fungal kingdom. The data obtained from these fungi will support comparative analyses of other fungi that are also critical to human health and that serve as research models. The FGI will thus propel research in medicine, industry, and agriculture, and will spur progress in computational studies of eukaryotic biology and evolution though the rapid release of sequence from clusters of genomes that are closely related to the most important fungi in research and medicine.

2. History and Promise of Fungal Genomics

2.1 Impact of the yeast sequence. The sequence of the genome of *Saccharomyces cerevisiae* was a landmark in genomics (Goffeau et al. 1996). It made possible the first global studies of eukaryotic gene expression and gene function (e.g., Giaever et al. 1999; Winzeler et al. 1999; Birrell et al. 2001; Ideker et al. 2001; Ooi et al. 2001; Lee et al. 2002) and provided scientists working on human genes access to a deep body of knowledge about how those genes functioned, and allowed use of the exquisite tools of yeast genetics to further interrogate protein functions and interactions (Dodt et al. 1996; Foury et al. 1997; Primig et al. 2000; Bennett et al. 2001; Jorgensen et al. 2002; Segal et al. 2003).

2.2 Other fungal sequencing. Since the sequencing of *S. cerevisiae*, progress on other fungal genomes has been limited. The sequence for the fission yeast *Schizosaccharomyces pombe* was published only in 2002 (Wood et al. 2002) and the sequence for the first filamentous fungus, *Neurospora crassa*, was published this year (Galagan et al. 2003). Although several other fungal projects are underway outside of the FGI, to date most have yielded only fragmentary sequence and most of these genomes are not freely available. In fact, the freely available genome sequences generated through the FGI represent the largest and most complete set of fungal genome sequences yet produced.

The paucity of fungal genome sequencing is remarkable considering the exceptional contributions that fungal genome sequences can provide to the study of eukaryotic biology and human medicine. Fungal genomes are relatively modest in size (7–40 Mb) and contain few repeats. They are thus ideal targets for whole-genome shotgun (WGS) sequencing. Further, the high gene density of fungi makes them extremely cost effective in terms of eukaryotic gene discovery. For example, *S. cerevisiae* contains a gene approximately every 2 kb (Goffeau et al. 1996), while the larger Neurospora genome averages a gene every 3.7 kb (Galagan et al. 2003).

Within fungal genomes lies the evolutionary history of the origins of many important biological processes found in higher eukaryotes, and their experimental tractability make fungi among the most useful model systems in cell biology. Fungal cellular physiology and genetics share key components with animal cells, including multicellularity, cytoskeletal structures, development and differentiation, sexual

reproduction, cell cycle, intercellular signaling, circadian rhythms, DNA methylation and regulation of gene expression through modifications to chromatin structure, and programmed cell death. The shared origins of the genes responsible for these fundamental biological functions between humans and fungi make understanding the history and function of fungal genes and genomes of vital interest to human biology.

2.3 Insights from the sequence of *N. crassa*. The draft sequence of *N. crassa* underscores the potential of this project (Galagan et al. 2003). The 10,000 predicted genes in this 40-Mb genome correspond to more than twice the number found in the fission yeast *S. pombe* and only about 25% fewer than found in Drosophila. More than 4100 of these predicted proteins lack significant matches to known proteins in the public databases and more than 5800 — a number equal to the entire *S. cerevisiae* genome — lack significant matches to genes in either *S. cerevisiae* or *S. pombe*. These data confirm the early stage of genomic characterization of the filamentous fungi. In addition, when compared with other sequenced eukaryotes, a full 1421 predicted *N. crassa* proteins show their best match to proteins in either plants or animals. Of these, 584 lack high-scoring matches to genes in either sequenced yeast. These genes point to aspects of biology that are shared between filamentous fungi and higher eukaryotes that in many cases are not found within the yeasts.

Despite the fact that Neurospora is the most extensively studied filamentous fungus, the genome sequence revealed some important surprises. Among those revealed by comparative genomics was the presence of genes putatively involved in secondary metabolism. Secondary metabolism, the synthesis of small bioactive molecules, is a key aspect of the biology of filamentous fungi, producing well-known antibiotics and toxins. With the exception of carotenoid and melanin pigment synthesis, Neurospora was not known to possess secondary metabolism. Nonetheless, a number of non-ribosomal polypeptide synthetases and polyketide synthases were identified. The production of secondary metabolites may prove to be an important aspect of the toxicity of fungal pathogens, and represent a potential new diagnostic opportunity.

Although not a pathogen, the *N. crassa* genome revealed a number of genes similar to genes required for pathogenesis in fungal plant pathogens. In several cases, the only known functions for many of these genes within the pathogens was their requirement for pathogenicity. Hence, the value of obtaining genome sequences from model fungi for understanding pathogenesis is extremely high, both because of the intrinsic value the model fungi have as tractable experimental systems and because of the additional power obtained from comparative genome analyses.

3. Rationale for a Fungal Sequencing Program

We propose that fungal genomics should be approached in a kingdom-wide manner — that is, by selecting a set of fungi (rather than choosing individual fungi in isolation) that maximizes the overall value through a comparative approach. In this way the sequence of well-chosen organisms will not only enable ongoing research on that organism but will enhance the value of other sequences through comparative studies of the evolution of genes, chromosomes, regulatory and biochemical pathways and pathogenesis.

In this white paper, we describe a collection of fungi that have been selected to further expand our understanding of the fungal kingdom and to provide much deeper insight into several extremely important fungal groups. The comparative genomic approach we describe will accelerate the pace of discovery for sorely needed diagnostic tools and therapeutic agents. We begin by outlining the general considerations that justify a program for fungal sequencing.

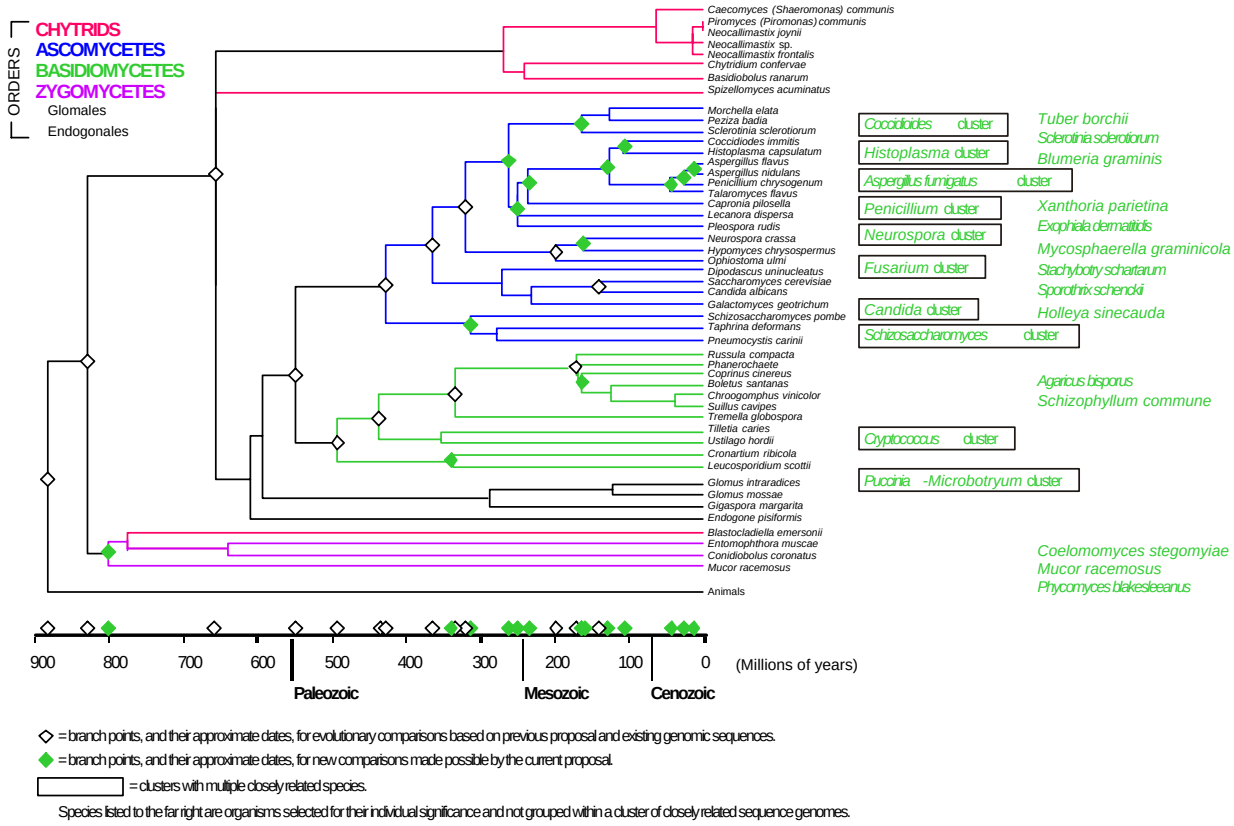

FIGURE 1. Phylogenetic tree.

3.1 Impact on human health. Fungal pathogens are devastating to human health. Fungal infections have lethal consequences for the growing population of patients immunocompromised with AIDS or therapeutically immunosuppressed after cancer chemotherapy or transplantation surgery. Fungal disease now represents as much as 15% of all hospital-acquired infections. Emerging fungal infections represent an equally serious threat to healthy human populations, including severe allergic reactions to fungal spores and molds. Identifying effective therapies against these eukaryotes has been more difficult than for bacteria, and, as a result, few effective antifungals are currently available. The worldwide market for antifungals is projected to reach $6.5 billion by 2008. Most of the existing drugs have serious side effects, and resistance to these compounds is an increasing problem. Genome sequence from pathogenic fungi will be the most efficient step in identifying potential targets for therapeutic intervention and vaccination among the largely unknown set of fungal proteins.

One of the greatest needs clinically is the availability of diagnostics that can provide facile and accurate identification of particular fungal species. Genome sequences provide the opportunity for unique DNA probes that could be used for identification. Our work on Neurospora suggests that fungi may also have pathways yielding secondary metabolites whose presence in the urine or blood would be indelible signatures of the organism causing the infection.

In addition to their role as pathogens, the importance of fungi to human health includes their production of a vast array of secondary metabolites, including toxins and carcinogens that destroy human and animal foodstuffs. The ability to produce and secrete these metabolites underlies their role in the development and production of critical pharmaceuticals with billions of dollars in annual sales, including

antibiotics such as penicillin and the cephalosporins, as well as the cholesterol-lowering statins and cyclosporin.

Fungi also exert a heavy influence on agriculture and our ability to feed the world's population. Fungal plant pathogens destroy vast amounts of crops in the field and after harvest each year. For example, in the United States, where over $600 million is spent annually on agricultural fungicides, crop losses due to fungi exceed $200 billion annually. Genome sequence will be paramount for understanding fungal infection as well as host/pathogen interactions. Sequence data will also provide crucial information on how these organisms reproduce, persist in the environment and interact with their hosts. In addition to their role as pathogens, fungi have additional vital but poorly understood positive roles in agriculture. For example, the mycorrhizal fungi that grow interdependently with plant roots are critical for nutrient uptake by plants.

3.2 Impact on human biology. Although *S. cerevisiae* provides key insights into the function of many human proteins, analysis of filamentous fungi reveal a much larger set of proteins shared with humans (Zeng et al. 2001; Galagan et al. 2003). A deeper sampling of fungal genes will rapidly increase the number of human proteins for which we can access homologues in model organisms. Genome sequence from many diverse fungi, coupled with comparative and functional genomics, will advance our understanding of the eukaryotic proteome. In doing so, we will learn not only how to manipulate fungi, but also how to manipulate human physiology for the treatment of metabolic and infectious diseases.

3.3 Impact on comparative genomics and evolutionary science. Comparative genomics and evolutionary genomic studies hold great promise, but these fields are still in their infancy. For example, most mammalian comparative analyses consist of pair-wise sequence alignment of regions. Only once has this involved complete genomes (Mouse Genome Sequencing Consortium 2002). Recently, alignment of a single region from multiple genomes has provoked great excitement about the ability to do this on a genome-wide basis (Boffelli et al. 2003). While of great value, these studies have been largely defined by the limited availability of mammalian genome sequence. A fungal sequencing program represents an ideal system for developing comparative methods for eukaryotic genomes because:

- The small genome sizes allow for the comparison of many complete eukaryotic genomes for small cost and effort.
- The fungal kingdom, with more than 1 million different species, displays extraordinary diversity.
- Genomes can be selected representing a wide variety of evolutionary distances, ranging from less than 5 million years to approximately 1 billion years.
- Genomes can be selected representing specific branch points in the phylogenetic tree to illuminate the molecular basis for key biological innovations.
- Fungi offer outstanding opportunities to study natural populations and evolution. For example, there are over 4000 well-characterized natural isolates of Neurospora deposited at the Fungal Genetics Stock Center, taken from widespread and ecologically diverse regions.

Analyses of representative genomes distributed across the fungal tree are expected to provide the molecular basis for understanding the extraordinary diversity that has arisen over the estimated 1 billion years since divergence from a common ancestor.

3.4 Lessons from yeast comparative studies. Recent work using budding yeast illustrates the power of whole-genome comparative analysis for studying genome evolution, gene identification and gene regulation (Cliften et al. 2003; Kellis et al. 2003). For example, the alignment of high-quality assemblies of four closely related yeasts — *S. cerevisiae*, *S. paradoxus*, *S. bayanus*, and *S. mikatae* — revealed all large-scale chromosomal rearrangements between the species. In addition, the conservation of gene sequences during evolution permits real genes to be distinguished from random open reading frames (ORFs). By these means, ~500 previously annotated yeast ORFs were suggested to be deleted from the gene catalogue, an additional 188 small genes were detected, and the boundaries of more than 300 genes

were revised. On the one hand, it is sobering that these many corrections were recognized for the heavily studied *S. cerevisiae*. One the other hand, these data reveal the power of evolutionary comparisons to recognize and correct genes, a technique applicable to other fungi.

Finally, the yeast comparisons provided the ability to detect the small signals of genetic regulatory elements above the noise of surrounding non-conserved sequence, using computational methods alone. Without prior knowledge of the function of individual genes or factors, virtually all known regulatory motifs were discovered. These methods will be of great value to the study of filamentous fungi, for which there is presently very little information about the components of transcriptional and post-transcriptional regulation.

These studies suggest a clear strategy for further fungal sequencing, namely to sequence close relatives of the most important fungi and use multiple alignments of whole genome assemblies to identify functionally conserved elements and investigate genome evolution. Because the genomes selected for comparative sequencing are also used in laboratory studies, research on these organisms to will be dramatically accelerated. The prime motivation is to use comparative genomics to illuminate to the greatest degree possible the biology and genetics of organisms of proven importance.

The species of yeast used in the recent comparative analyses of yeast genomes (Kellis et al. 2003; Cliften et al. 2003) were specifically chosen based on their evolutionary distance as being maximally informative for alignment and identification of functional elements. The FGI Steering Committee has used the extensive resources of the community to identify fungi for genome sequencing that are appropriately close to important fungi for which there is also genome sequence. This has been necessary because the existing sequence data do not support effective comparisons. In fact, at present, the only filamentous fungi for which multiple genome alignments can be attempted are for *Aspergillus fumigatus*, *A. nidulans*, and *A. oryzae*. However, these species are more distantly related than would be ideal for comparative analysis.

4. History of the FGI

4.1 Origins. In November 2000, Dr. Gerry Fink invited a small group of fungal geneticists and biologists to discuss ways to accelerate the slow pace of fungal genome sequencing. Participants included academic and industrial fungal scientists as well as those with experience in genome sequencing and analysis.

The group concluded that the dearth of publicly available fungal genome sequence was a major barrier to biomedical research. A broad initiative was conceived in which organisms would not be selected one at a time, but would be considered as part of a cohesive strategy. The primary selection criteria endorsed were:

- Importance of the organism in human health and commercial activities.
- Value of the organism as a tool for comparative genomics.
- Presence of genetic resources and an established research community.

These principles were laid out in a draft white paper that described a broad, collaborative Fungal Genome Initiative. This white paper was circulated during the summer of 2001 among federal agencies and served as the direct inspiration for NHGRI's process for prioritizing organisms for sequencing.

4.2 Steering Committee. To provide advice and oversight of a sustained effort, a Steering Committee was organized, consisting of:

Gerry Fink, Whitehead Institute for Biomedical Research, Steering Committee, Chair
Ralph Dean, North Carolina State University; Fungal Genetics Policy Committee, Chair
Peter Hecht, Microbia, Inc.
Joe Heitman, Duke University
Ron Morris, UMDNJ-Robert Wood Johnson Medical School
Matthew Sachs, Oregon Health and Science University
John Taylor, University of California, Berkeley
Mary Anne Nelson, University of New Mexico

Bruce Birren, Whitehead Institute for Biomedical Research

These fungal biologists represent a cross-section of interests in mycology and fungal genetics and are responsible for setting the direction of the project and reviewing progress. The Steering Committee meets annually, having met most recently in March 2003 at the International Fungal Genetics Conference at Asilomar, California. Throughout the year, the Steering Committee interacts through e-mail and telephone conference calls. In addition, regular meetings take place between the Committee Chair and the Whitehead Institute/MIT Center for Genome Research (WICGR) staff.

4.3 Ongoing community input. In November 2001, the Steering Committee convened an NHGRI and NSF sponsored workshop on Fungal Genomics in Alexandria, Virginia. Over 60 attendees representing academic, government and industrial interests in medical, agricultural, industrial, evolutionary, basic biological fungal research and informatics discussed the genome resources that were most needed to spur research and development in their areas of interest. The workshop produced strong endorsement of an initiative that would include rapid sequencing and public release of many fungal genomes chosen as part of a single, coherent plan.

One outcome of the workshop was formalization of the communication channels between the FGI Steering Committee and the broader research community. Presentations about the FGI at major fungal conferences keep the community apprised of progress and always include the solicitation of community input. Notice of new data releases are sent through our own mailing list and that of the Fungal Genetics Stock Center. New nominations of candidate fungi arrive weekly along with other e-mail correspondence (FGI_Info@genome.wi.mit.edu). To date, nearly 100 nominations of organisms for sequencing have been considered by the Steering Committee. Of particular importance in evaluating these nominations has been the FGI Steering Committee's connection with other groups interested in fungal phylogeny and pathogenesis. Specifically, the input from scientists associated with the NSF-funded Fungal Tree of Life project (Dr. J. Spatafora, PI) has been valuable in identifying fungi of the appropriate evolutionary distance for our purposes, and our contacts with the American Phytopathological Society have provided a great deal of useful information.

4.4 FGI website, data release and user community. Since the February 2001 release of the *N. crassa* genome sequence, WICGR has maintained a growing set of resources for fungal genomics on the web. The FGI website (www-genome.wi.mit.edu/annotation/fungi/fgi/) lists the status of all FGI projects and provides forms for submitting sequencing candidates. Currently, genome assemblies are available through the FGI website for five filamentous fungi and three yeasts. These fungal databases have received over 2.9 million hits since their launch and currently average 440,000 hits/month. This level of use demonstrates the wide appeal of these data, including scientists engaged in comparative genomics, evolutionary studies, fungal biology, infectious disease and computational biology.

4.5 WICGR fungal collaborations. WICGR has forged many strong collaborative relationships involving fungal genome sequencing and analysis. In fact, each fungal sequencing project represents a successful collaboration between WICGR and a research community of varying size. In the case of *N. crassa*, WICGR organized the community analysis project that engaged over 70 scientists in analysis of the genome sequence. In other cases, WICGR is helping to coordinate comparative analyses for multiple genome sequences, such as with *Aspergillus fumigatus*, *A. oryzae*, and *A. nidulans*. A similar collaboration exists with The Institute for Genomic Research (TIGR) and a consortium of university-based labs to jointly analyze genome sequence from two different Cryptococcus species. In several instances, WICGR has built alliances directly responsible for public release of data previously held in private hands, such as with Monsanto, Bayer, and Exelixis.

5. Sequencing Progress

The WICGR has fully released 8 fungal genome assemblies, including 5 from the 15 on our first white paper (Table 1). Funds to sequence 2 of the original 15 genomes, *Magnaporthe grisea* and *Fusarium graminearum*, were provided by awards outside of NHGRI. Two more assemblies are in "pre-release", currently available on our website to collaborators for quality review pending full public release later this month. Sequence data for all FGI genomes are made available in advance of assembly according to NHGRI policy on rapid data release by regular deposition of traces at the NCBI trace repository. We are on schedule to release seven high-priority fungi as per our plan provided to NHGRI. Difficulty obtaining DNA samples of sufficient quality for Fosmid cloning required us to revise the exact order of genomes to be sequenced, highlighting our need and ability to remain flexible when working from a list of targets.

Table 1. WICGR Fungal Genome Releases

Species	Status	Predicted genome size (Mb)	Assembled bases (Mb)	N50 scaffold size (Mb)	Accession #
Neurospora crassa	annotated assembly released	40	38,044,343	0.61	AABX01000000
Magnaporthe grisea	annotated assembly released	40	37,878,070	1.6	trace repository
Aspergillus nidulans	assembly released	31	30,068,514	2.44	AACD01000000
Fusarium graminearum	assembly released	40	36,093,143	5.36	AACM01000000
Cryptococcus neoformans, serotype A	assembly released	20	19,223,796	1.3	AACO01000000
Ustilago maydis	assembly in pre-release	20	19,762,689	0.82	trace repository
Coprinus cinereus	assembly in pre-release	38	36,259,524	2.06	trace repository
Coccidiodes immitis	in sequencing	29	—	—	—
Rhizopus arrhizus	in sequencing	35	—	—	trace repository
Saccharomyces paradoxus	assembly released	12	11,570,000	0.509	AABY00000000
Saccharomyces mikatae	annotated assembly released	12	11,220,000	0.334	AABZ00000001
Saccharomyces bayanus	annotated assembly released	12	11,320,000	0.234	AACA00000002

—, not applicable.

6. Sequencing Approach

6.1 Deep-shotgun sequencing. For each fungus, we propose to generate a high-quality draft sequence. Specifically, we will produce assemblies representing 8X whole-genome shotgun (WGS) sequence from paired-end reads obtained from 4-kb plasmids (80%), 10-kb plasmids (10%) and 40-kb Fosmids (10%). All libraries will be prepared from randomly sheared genomic DNA. Our experience with shotgun sequencing of fungal genomes indicates that 8X sequence coverage produces a high-quality draft assembly with the vast majority of each genome (well over 96%) present in the assembly. Further, these assemblies achieve long-range continuity as a result of the links provided by the Fosmid end sequences and the combined physical coverage of all libraries, which is approximately 50X. Although fungal genomes vary considerably in structure and nucleotide composition, a typical 8X assembly yields N50 contig sizes of 30–110 kb, and the N50 scaffold size of ~2 Mb. The high quality of this draft sequence is sufficient for most of the comparative studies this project will support. For example, our yeast comparative studies employed roughly this coverage (Kellis et al. 2003).

6.2 Polishing and finishing. For genomes that serve as particularly important references, it may be valuable to further improve the quality of the assembly. We propose that all clones required for automated polishing and/or finishing be retained and that the decision to carry out subsequent polishing and/or targeted finishing work be prioritized by NHGRI staff on the basis of evolving assessment of cost and capacity. One important feature of these high-quality draft assemblies is that almost all gaps and low-quality regions are small and are spanned by Fosmid subclones. This provides the opportunity for rapid and efficient improvement in the quality of the assembly. Fosmids that span gaps or low-quality regions can be automatically identified and used as templates for highly automated directed sequencing.

7. Detailed Description of Organisms

The organisms, the rationale for sequencing, and relevant genome information are summarized in Table 2 and are described in more detail on the following pages. Note: The original whitepaper named 44 fungi and described them in detail. Of these, we present descriptions here of only the four that were designated High Priority after review. The original white paper with the complete descriptions can be found at: http://www-genome.wi.mit.edu/annotation/fungi/fgi/candidates.html

Table 2. Organism Summaries

	Name	Significance	Est. size (Mb)
MEDICINE			
Candida cluster	*Candida albicans* (strain WO-1)	Most common human pathogen, related to laboratory strain being sequenced (SC5314)	16
	Candida tropicalis	The second most pathogenic of the *Candida* species; extremely close relative to *C. albicans* to be used for genomic comparison	16
	Lodderomyces elongisporus	Closest sexual relative to *C. albicans* and source of haploid genome for comparison to *C. albicans*	16
	Candida lusitaniae	Haploid relative of *C. albicans* with different codon usage and known complete sexual/meiotic cycles	16
	Candida krusei	Haploid relative of *C. albicans* with different codon usage and known complete sexual/meiotic cycles	16
	Candida guillermondii	Haploid relative of *C. albicans* with different codon usage and known complete sexual/meiotic cycles	16
Aspergillus fumigatus cluster	*Neosartorya fischeri*	Sexual *Aspergillus* species and closest relative to *A. fumigatus*, the #2 U.S. health problem. Also a causative agent of aspergillosis	25–30
	Aspergillus clavatus	Close relative to *A. fumigatus*, the second most common fungal pathogen in the U.S.	25–30
Fusarium cluster	*Fusarium verticillioides*	Pathogenic fungus that infects immunosuppressed patients	46
	Fusarium solani	Deadly threat to immunosuppressed, especially neutropenic and transplant patients	40
	Fusarium oxysporum	Pathogenic fungus that infects immunosuppressed patients	33
Histoplasma cluster	*Paracoccidioides brasiliensis*	Most prevalent systemic mycose in Latin America	25
	Blastomyces dermatitidis	Causative agent of blastomycosis, the principle systemic mycoses	28
Coccidioides cluster	*Unicinocarpus reesei*	Closest known relative of *Coccidioides* species, most severe of the U.S. systemic mycoses and select agents.	30
Penicillium cluster	*Penicillium marneffei*	Only dimorphic *Penicillium*; cause of grave pneumonia in AIDS patients	22–33
	Penicillium mineoleutium	Close relative *Penicillium* marneffei	30
	Stachybotrys chartarum	Black mold, major indoor environmental threat	40
	Sporothrix schenckii	Pathogenic dimorphic fungus with worldwide distribution	40
	Exophiala (Wangiella) dermatitidis	Causative agent of human dermatomycoses; excellent model for other dematiaceous fungal pathogens	19
Cryptococcus cluster	*Cryptococcus neoformans variety gattii* (Serotype B)	Encapsulated basidiomycete; leading cause of infectious meningitis	20
	Cryptococcus neoformans variety gattii (Serotype C)	Encapsulated basidiomycete; leading cause of infectious meningitis	20

Cryptococcus cluster	*Tremella fuciformis*	Close relative of *Cryptococcus*; well-developed sexual fruiting body to compare with *Cryptococcus*; obligate mycoparasite on wood rot fungi	20
COMMERCE			
	Penicillium chrysogenum	Primary penicillin source; reveals genomic effects of mutagenesis and selection	34
	Aspergillus niger	Widely used for the industrial production of enzymes and metabolites	36
	Agaricus bisporus	Most widely cultivated mushroom, annual worldwide production valued at $4.5 billion	40
	Tuber borchii	Edible truffle; ectomycorrhizal fungus, with experimental plantation	34
EVOLUTION / FUNGAL DIVERSITY			
	Saccharomyces cerevisiae RM11-1a	Natural isolate, now used in laboratory studies	12
Neurospora cluster	*Neurospora tetrasperma*	Pseudohomothallic species, diverged from *N. crassa* about 2.5 million years ago	40
	Podospora anserina	Model filamentous fungus, supports comparative studies of *Neurospora*	34
Schizosaccharomyces cluster	*Schizosaccharomyces japonicus*	Supports comparative analysis of model yeast, *S. pombe*	14
	Schizosaccharomyces octosporus	Supports comparative analysis of model yeast, *S. pombe*	14
	Schizosaccharomyces kambucha	Supports comparative analysis of model yeast, *S. pombe*	14
	Schizophyllum commune	Major model for mushroom-forming fungi	38
	Phycomyces blakesleeanus	Model filamentous zygomycete	30
	Mucor racemosus	Model for dimorphic growth among poorly understood zygomycetes	39
	Xanthoria parietina	Lichen; classic example of a mutualistic symbiotic relationship	30–40
	Coelomomyces stegomyiae	Haploid chytrids, the main eukaryotic pathogen of mosquito larvae	50
AGRICULTURAL			
	Mycosphaerella graminicola	Cause of septoria tritici leaf blotch, the second most important wheat disease in the U.S.	32–40
	Blumeria graminis	Cause of most important leaf disease of barley; most intensively studied powdery mildew species at the molecular level	35–45
	Sclerotinia sclerotiorum	Broadest known host-range of any plant pathogen; infects more than 408 species	26–44
	Holleya sinecauda	Pathogen on mustard seeds, close relative to *Saccharomyces* and *Candida*	8.5
Puccinia-Microbotryum cluster	*Microbotryum violaceum*	Best studied pathogen in natural plant populations; closely related with *Puccinia* species	25–30
	Puccinia triticina	Wheat leaf rust, the most widespread and common wheat pathogen in the U.S. and worldwide	90
	Puccinia striiformis	Basidiomycete fungal pathogen of crops and grasses	90

7.1 *Candida* cluster

SIGNIFICANCE: *Candida albicans* is the most common human fungal pathogen, likely because it is generally a benign commensal that resides on the mucosal surfaces of most if not all organisms. *C. albicans* is capable of causing superficial infections (vaginitis, thrush) in normal hosts, and severe systemic infections in immunocompromised hosts. *C. albicans* is diploid, but recent discoveries herald a new era in understanding the sexual cycle and its role in virulence. These advances stemmed directly from the genome project and the discovery of the MAT locus.

The genome of one particular strain, SC5314 is nearing completion at Stanford, however assembling the shotgun sequence has proven difficult due to polymorphisms in the diploid genome. We propose to sequence a second *C. albicans* isolate, clinical isolate WO-1, to provide important information about the pathogenic form of *C. albicans*. This isolate differs considerably from the SC5314 strain. It is the most carefully characterized MTL-homozygous strain for both white-opaque switching, a phenotypic change that correlates with host cell specificity and mating. We also propose to sequence two species closely related to *C. albicans*: *C. tropicalis* and *Lodderomyces elongisporus.* The asexual diploid yeast *C. tropicalis* is the second most pathogenic of the *Candida* species. Unlike *C. albicans*, which is a normal commensal on human mucous membranes, the detection of *C. tropicalis* is more often associated with the development of deep fungal infections. *L. elongisporus* is the only known ascosporogenous species in the *C. albicans* clade. It usually produces a single ascospore.

GENERAL DESCRIPTION: *Candida* species belong to Ascomycota, Saccharomycotina subphyla and all species can be cultured under laboratory conditions. The species we propose to sequence are descended from an ancestral strain in which the CTG codon was reconfigured ~150 million years ago.

GENOME FACTS: The genome size of *Candida* species is around 20 Mb. All the sequences can be readily aligned to the *C. albicans* strain SC5314 assembly (available at www-sequence.stanford.edu/), which will expedite assembly and annotation.

COMMUNITY: There is a very large research community, over 200 investigators, working on basic research using *C. albicans*. The genomic data of these closely related species will greatly improve the genome assembly and annotation of *C. albicans*, which will accelerate the understanding of both the pathogenesis of this fungal pathogen and the genome divergence among these species. The genomic DNA of *C. albicans* (strain WO-1) will be provided by Dr. David Soll at University of Iowa. *C. tropicalis* and *L. elongisporus* DNA will be provided by Dr. Clete Kurtzman at the USDA in Peoroa.

7.2 *Saccharomyces cerevisiae* — RM11-1a a natural isolate

SIGNIFICANCE: *Saccharomyces cerevisiae* is arguably the most important model organism for studies of genetics and eukaryotic biology. As the first eukaryote to have its genome sequenced, it has also become the model of choice for functional and comparative genomics. To date, the sequence of only a single laboratory strain of *S. cerevisiae*, S288C, is available, and a sequence of an independent natural isolate would greatly enhance biological, genomic, and evolutionary studies. High-quality draft sequence of three related *Saccharomyces* species was recently generated (Kellis et al. 2003). The sequence divergence between *S. paradoxus* (the closest of these species) and *S. cerevisiae* is 20%, considerably greater than that between human and rhesus macaque. In contrast, the sequence divergence between RM11 and S288C is estimated to be 0.5–1%, approaching that between human and chimp. RM11 sequence would thus fill an important gap in one of the key objectives of the Fungal Genome Initiative — to compare genomes that represent a variety of evolutionary distances and relationships. The comparison of *S. cerevisiae* with other *Saccharomyces* species found a number of differences in *S. cerevisiae*, and the authors pointed out that "sequencing of additional strains will be required to determine whether [these changes] represent differences in the S288C strain alone or in *S. cerevisiae* in general." Sequence of RM11, an independent natural isolate that shares no recent history with S288C, will provide this information. Indeed, we have found a number of polymorphisms between S288C and RM11 that represent new mutations in S288C relative to natural isolates of *S. cerevisiae* and other yeast species and that alter the function of the protein products in S288C, perhaps reflecting adaptation to the new selective regime in the laboratory environment. Thus some conclusions drawn from studies in laboratory strains may not accurately represent the biology of *S. cerevisiae*. Sequence of a natural isolate will allow identification of such mutations genome-wide and enable studies of the true wild-type alleles, as well as shedding light on natural genetic variation within *S. cerevisiae*. Sequence of RM11 will also *identify S. cerevisiae* genes deleted in S288C and quantify the rate of intraspecific telomeric "genome churning." RM11 has been used as a model for mapping loci that affect gene expression and other complex phenotypes, and the sequence will greatly facilitate positional cloning of the genes involved.

GENERAL DESCRIPTION: RM11-1a is a haploid derivative of Bb32(3), a natural isolate collected by Robert Mortimer from a California vineyard. It has high spore viability (80–90%) when crossed with different lab strains. Strains of both mating types and with a number of auxotrophic markers are available. RM11 has been subject of extensive phenotypic characterization, including growth under a wide range of conditions and gene expression profiling. Measurements of gene expression in RM11, S288C, and segregants from a cross between them show that 1/3–1/2 of the genome is differentially expressed, and that these differences are due to at least 500–1000 separate loci, some of which affect expression in *cis* and others in *trans*. Sequence of RM11 will greatly facilitate rapid identification of these loci and comprehensive characterization of regulatory variation. RM11 also has significantly longer life span than laboratory yeast strains and accumulates age-associated abnormalities at a lower rate. It is being used in studies of aging.

GENOME FACTS: The *S. cerevisiae* genome is 12.16 Mb with 16 chromosomes. Availability of finished S288C sequence will allow assembly of the RM11 sequence from a low-coverage (3–4X) shotgun, analogous to assembly of the chimp genome with the human genome as the reference. Thus the RM11 sequence can be generated with only ~10% of the sequencing capacity required for 10X coverage of a new 40-Mb fungal genome. Sequencing a haploid strain will ensure that assembly and analysis will not be complicated by polymorphic sites. Sequence divergence from S288C is estimated at 1 in 100–200 bp, and this sequence variation is distributed throughout the genome, confirming that RM11 shares no recent history with S288C.

COMMUNITY: In addition to several labs that have been working with RM11 specifically, the sequence will be of interest to the large wine yeast community, as well as to the broader *S. cerevisiae* community. More generally, the usefulness of the sequence will extend far beyond the yeast community to scientists with interests in natural genetic variation, comparative genomics, and evolution. Leonid Kruglyak (FHCRC) will supply haploid genomic DNA and has assembled a team to analyze differences between the S288C and RM11 genomes.

References

Birrell GW, Giaever G, Chu AM, Davis RW, Brown JM (2001) A genome-wide screen in *Saccharomyces cerevisiae* for genes affecting UV radiation sensitivity. *Proc. Natl. Acad. Sci. USA* 98: 12608–12613.

Bennett CB, Lewis LK, Karthikeyan G, Lobachev KS, Jin YH, Sterling JF, Snipe JR, Resnick MA (2001) Genes required for ionizing radiation resistance in yeast. *Nat. Genet.* 29: 426–434.

Boffelli D, McAuliffe J, Ovcharenko D, Pachter L, Rubin EM (2003) Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science* 299: 1391–1394.

Cliften P, Sudarsanam P, Desikan A, Fulton L, Fulton B, Majors J, Waterston R, Cohen BA, Johnston M (2003) Finding functional features in Saccharomyces genomes by phylogenetic footprinting. *Science* Published online May 29, 2003.

Dodt G, Braverman N, Valle D, Gould SJ (1996) From expressed sequence tags to peroxisome biogenesis disorder genes. *Ann. NY Acad. Sci.* 804: 516–523.

Foury F (1997) Human genetic diseases: A cross-talk between man and yeast. *Gene* 195: 1–10.

Galagan JE, Calvo SE, Borkovich KA, Selker EU, Read ND, Jaffe D, FitzHugh W, Ma LJ, Smirnov S, Purcell S, Rehman B, Elkins T, Engels R, Wang S, Nielsen CB, Butler J, Endrizzi M, Qui D, Ianakiev P, Bell-Pedersen D, Nelson MA, Werner-Washburne M, Selitrennikoff CP, Kinsey JA, Braun EL, Zelter A, Schulte U, Kothe GO, Jedd G, Mewes W, Staben C, Marcotte E, Greenberg D, Roy A, Foley K, Naylor J, Stange-Thomann N, Barrett R, Gnerre S, Kamal M, Kamvysselis M, Mauceli E, Bielke C, Rudd S, Frishman D, Krystofova S, Rasmussen C, Metzenberg RL, Perkins DD, Kroken S, Cogoni C, Macino G, Catcheside D, Li W, Pratt RJ, Osmani SA, DeSouza CP, Glass L, Orbach MJ, Berglund JA, Voelker R, Yarden O, Plamann M, Seiler S, Dunlap J, Radford A, Aramayo R, Natvig DO, Alex LA, Mannhaupt G, Ebbole DJ, Freitag M, Paulsen I, Sachs MS, Lander ES, Nusbaum C, Birren B (2003) The genome sequence of the filamentous fungus *Neurospora crassa*. *Nature* 422: 859–868.

Giaever G, Shoemaker DD, Jones TW, Liang H, Winzeler EA, Astromoff A, Davis RW (1999) Genomic profiling of drug sensitivities via induced haploinsufficiency. *Nat. Genet.* 21: 278–283.

Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, Louis EJ, Mewes HW, Murakami Y, Philippsen P, Tettelin H, Oliver SG (1996) Life with 6000 genes. *Science* 274: 563–567.

Ideker T, Thorsson V, Ranish JA, Christmas R, Buhler J, Eng JK, Bumgarner R, Goodlett DR, Aebersold R, Hood L (2001) Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science*. 292: 929–934.

Jorgensen P, Nishikawa JL, Breitkreutz BJ, Tyers M (2002) Systematic identification of pathways that couple cell growth and division in yeast. *Science* 297: 395-400.

Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* 423: 241–254.

Lee TL, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I, Zeitlinger J, Jennings EG, Murray HL, Gordon DB, Ren B, Wyrick JJ, Tagne JB, Volkert TL, Fraenkel E, Gifford DK, Young RA (2002) Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* 298: 799–804.

Mouse Genome Sequencing Consortium (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature* 420: 520–562.

Ooi SL, Shoemaker DD, Boeke JD (2001) A DNA microarray-based genetic screen for nonhomologous end-joining mutants in *Saccharomyces cerevisiae*. *Science* 294: 2552–2556.

Primig M, Williams RM, Winzeler EA, Tevzadze GG, Conway AR, Hwang SY, Davis RW, Esposito RE (2000) The core meiotic transcriptome in budding yeasts. *Nat. Genet.* 26: 415–423.

Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, Friedman N (2003) Module networks: Identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet.* 34: 166–176.

Winzeler EA, Shoemaker DD, Astromoff A, Liang H, Anderson K, Andre B, Bangham R, Benito R, Boeke JD, Bussey H, Chu AM, Connelly C, Davis K, Dietrich F, Dow SW, El Bakkoury M, Foury F, Friend SH, Gentalen E, Giaever G, Hegemann JH, Jones T, Laub M, Liao H, Davis RW, et al. (1999) Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science* 285: 901–906.

Wood V, Gwilliam R, Rajandream MA, Lyne M, Lyne R, Stewart A, Sgouros J, Peat N, Hayles J, Baker S, Basham D, Bowman S, Brooks K, Brown D, Brown S, Chillingworth T, Churcher C, Collins M, Connor R, Cronin A, Davis P, Feltwell T, Fraser A, Gentles S, Goble A, Hamlin N, Harris D, Hidalgo J, Hodgson G, Holroyd S, Hornsby T, Howarth S, Huckle EJ, Hunt S, Jagels K, James K, Jones L, Jones M, Leather S, McDonald S, McLean J, Mooney P, Moule S, Mungall K, Murphy L, Niblett D, Odell C, Oliver K, O'Neil S, Pearson D, Quail MA, Rabbinowitsch E, Rutherford K, Rutter S, Saunders D, Seeger K, Sharp S, Skelton J, Simmonds M, Squares R, Squares S, Stevens K, Taylor K, Taylor RG, Tivey A, Walsh S, Warren T, Whitehead S, Woodward J, Volckaert G, Aert R, Robben J, Grymonprez B, Weltjens I, Vanstreels E, Rieger M, Schafer M, Muller-Auer S, Gabel C, Fuchs M, Dusterhoft A, Fritzc C, Holzer E, Moestl D, Hilbert H, Borzym K, Langer I, Beck A, Lehrach H, Reinhardt R, Pohl TM, Eger P, Zimmermann W, Wedler H, Wambutt R, Purnelle B, Goffeau A, Cadieu E, Dreano S, Gloux S, Lelaure V, Mottier S, Galibert F, Aves SJ, Xiang Z, Hunt C, Moore K, Hurst SM, Lucas M, Rochet M, Gaillardin C, Tallada VA, Garzon A, Thode G, Daga RR, Cruzado L, Jimenez J, Sanchez M, del Rey F, Benito J, Dominguez A, Revuelta JL, Moreno S, Armstrong J, Forsburg SL, Cerutti L, Lowe T, McCombie WR, Paulsen I, Potashkin J, Shpakovski GV, Ussery D, Barrell BG, Nurse P, Cerrutti L (2002) The genome sequence of *Schizosaccharomyces pombe*. *Nature* 415: 871–880.

Zeng Q, Morales AJ, Cottarel G (2001) Fungi and humans: Closer than you think. *Trends. Genet.* 17: 682–684.

Appendix — Fungi nominated by community for sequencing (as of June 2003)

ASCOMYCOTA

Mycosphaerella graminicola
Cenococcum geophilum
Aspergillus niger
Aspergillus versicolor
Neosartorya fischeri
Aspergillus clavatus
Penicillium chrysogenum
Penicillium roquefortii
Penicillium marneffei
Penicillium mineoleutium
Xanthoria parietina
Ramalina menziesii
Fusarium oxysporum
Fusarium verticillioides
Fusarium solani
Fusarium proliferatum
Stachybotrys chartarum
Sporothrix schenckii
Ophiostoma umli
Cryphonectria parasitica
Blastomyces dermatitidis
Paracoccidioides brasiliensis
Uncinocarpus reesei
Blumeria graminis
Sclerotinia sclerotiorum
Neurospora tetrasperma
Podospora anserina
Septoria lycopersici
Tuber borchii
Tuber melanosporum
Exophiala (Wangiella) dermatitidis
Saccharomyces cerevisiae (RM11-1a)
Candida albicans (WO-1)
Candida tropicalis
Lodderomyces elongisporus
Candida lusitaniae
Candida krusei
Candida guillermondii
Holleya sinecauda
Eremothecium gossypii
Schizosaccharomyces japonicus
Schizosaccharomyces octosporus
Schizosaccharomyces kambucha
Tolypocladium inflatum
Cordyceps militaris
Epichloe typhina
Ceratocystis fimbriata
Colletotrichum
Corollospora maritima
Xylaria hypoxylon
Leotia lubrica
Botrytis cinerea
Pyrenophora trici-repentis
Morchella esculenta
Taphrina deformans

BASIDIOMYCOTA

Schizophyllum commune
Agaricus bisporus
Microbotryum violaceum
Trichodoma
Tremella fuciformis
Tremella mesenterica
*Cryptococcus neoformans variety gattii (*Serotype B*)*
*Cryptococcus neoformans variety gattii (*Serotype C*)*
Tsuchiyaea wingfieldii,
Filobasidiella depauperata
Filobasidiella flava
Filobasidiella xianghuijun.
Puccinia triticina
Puccinia striiformis
Amanita phalloides
Flammulina velutipes
Armillaria
Cantharellus cibarius
Phallus impudicus
Phellinus pinii
Leptosphaeria maculans
Stagonospora nodorum

ZYGOMYCOTA

Phycomyces blakesleeanus
Mucor racemosus

CHYTRIDIOMYCOTA

Coelomomyces stegomyiae
Coelomomyces utahensis
Allomyces macrogynus
Blastocladiella emersonii

GENUSES NOMINATED BY RESEARCH COMMUNITY

Caldina (compare to lichen)
Letharia (compare to lichen)
Trichoderma (industry, biocontrol)
Rhytisma
Cladonia
Arthonia
Orbilia
Septobasidium
Sporidium
Tilletia
Polyporus
Glomus
Smittium
Chytridium