Fungal Genome Initiative

A White Paper for Fungal Genomics

July 10, 2004

Submitted on behalf of: The Fungal Genome Initiative Steering Committee
Corresponding author: Bruce Birren
The Broad Institute of Harvard and MIT
320 Charles Street, Cambridge, MA 02141 USA.
Phone: 617-258-0900. E-mail: bwb@broad.mit.edu.

Summary

Since 2001 fungal scientists and members of the Broad Institute have been pursuing a comprehensive plan to apply comparative sequence analysis to study the evolution of eukaryotic cellular processes and the molecular basis for fungal pathogenesis. In this white paper we seek approval to sequence four fungi to continue these efforts:

- ***Schizosaccharomyces octosporus*** and ***Schizosaccharomyces japonicus*** — to improve the annotation of the *S. pombe* genome sequence.
- ***Trichophyton rubrum*** — the cause of athlete's foot and the first organism to be sequenced in the class of fungi that adhere to human skin.
- ***Batrachochytrium dendrobatidis*** — the causal organism in the recent dramatic decline of global amphibian populations.

The sequences of these four genomes, totaling 88 Mb, will both directly contribute to the study of key fungal pathogens and advance the utility of *S. pombe*, a widely used model for molecular and cellular research.

Background and rationale

Although *S. cerevisiae, S. pombe* and several filamentous fungi are well-established research organisms, the vast majority of fungi remain very poorly understood. The fact that over 800 million years (myr) has elapsed since the divergence from the common ancestor of the fungal kingdom (Figure1) means that these few well-studied organisms do not adequately represent the diversity found within the many fungi that play critical roles in human health and industry. This diversity can be appreciated by noting that just within the small portion of the fungal tree containing the Aspergilli (Figure1), there are numerous species spanning an evolutionary distance greater than human and fish. Hence, providing genome sequences for additional fungi is crucial to jumpstart genomic approaches to fighting fungal pathogens and boosting the utility of fungal models for eukaryotic biology and evolution.

In 2001 fungal researchers and scientists at the Broad Institute (then the Whitehead Institute Center for Genome Research) established a formal relationship through the creation of a Steering Committee for the Fungal Genome Initiative (FGI). The goal was to identify and sequence organisms that were: 1) important for defining fungal diversity; 2) individually critical for medical, industrial or research purposes; and 3) maximally valuable for comparative purposes. With guidance from the Genome Resources and Sequencing Priority Panel (GRASPP), the FGI Steering Committee has elected to focus on fungi that are tied to specific questions in human health and biology and that serve as important models of eukaryotic genome evolution.

Even as mammalian sequencing accelerates, fungal genome sequencing and comparisons remain vital to human health and our quest to completely understand the human genome. First, they spur development and validation of tools needed for whole genome analysis in advance of many complete mammalian genomes. The comparison of several yeast

genomes have been instrumental in framing the approach for comparative discovery of human genes and conserved non-coding sequences (Cliften et al. 2003; Kellis et al. 2003). Similarly, the availability of many complete fungal genomes is leading us to invent automated methods to align these sequences, recognize synteny and extract the genomic information years before we will be able to do this with mammals. The fact that the filamentous fungi have larger genomes and more complex gene structures than yeast make these new comparisons even more useful as models for the mammalian comparisons ahead. Second, the ancient origins of the fungal tree mean that comparisons of these fungi with the genomes of higher eukaryotes identify features that reflect the origins of our genes and genomic elements. Third, as our catalogue of sequences from fungal pathogens grows, so does our ability to recognize fungal-specific genes as potential targets for diagnostics and therapy. A critical barrier to developing effective antifungals with acceptable side effects is the extent of the biology and cellular components they share with the eukaryotic host. Developing a fungal-specific and pathogen-specific gene catalogue will focus future research efforts on these proteins.

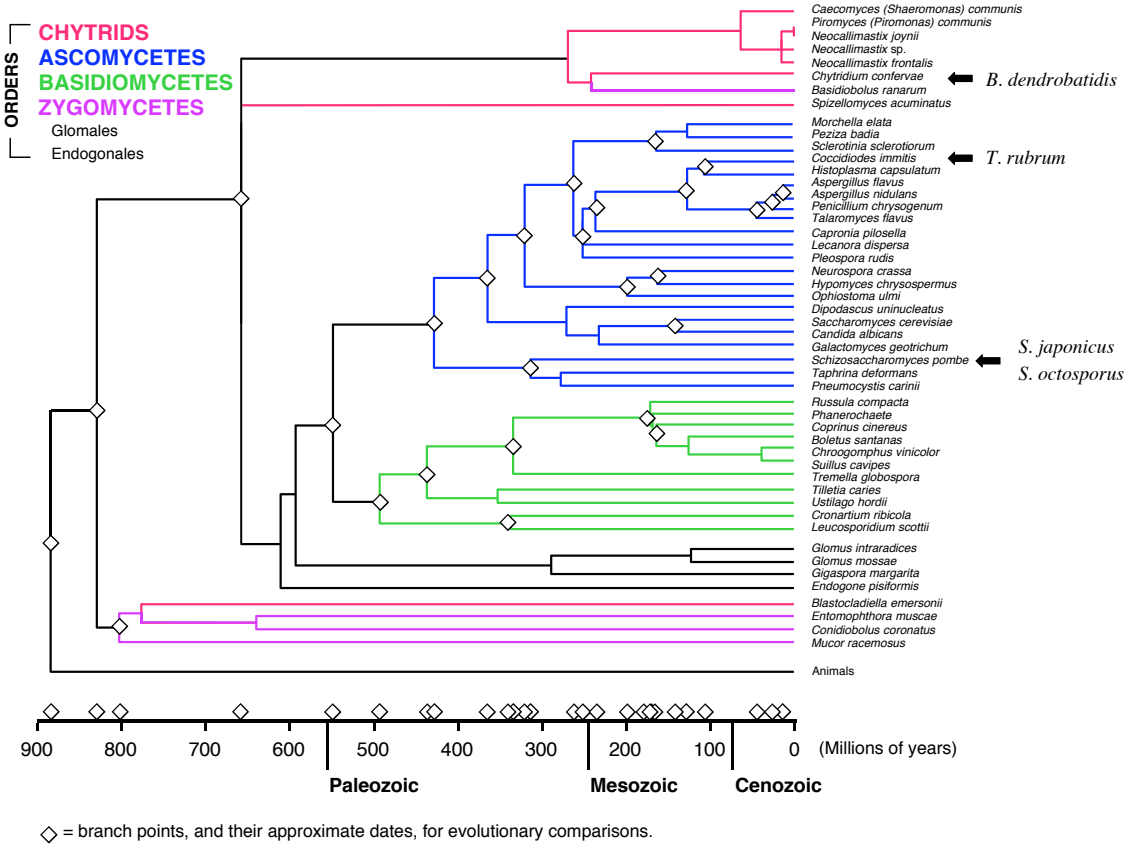

Figure 1: Fungal phylogenetic tree. The four organisms proposed for sequencing in the current white paper are indicated on the far right of the figure. The arrows point to the nearest relative in the fungal phylogenetic tree.

Over the past 3 years, the FGI Steering Committee has evaluated over 100 candidate organisms formally nominated for sequencing by members of the community. The four fungi in this white paper represent the Steering Committee's recommendation as the most important organisms among current candidates.

Biological importance of the *Schizosaccharomycetes*

S. pombe has been a leading model system for eukaryotic cell biology for over 20 years. Its simple life cycle, powerful genetics and biochemical accessibility make it an ideal experimental organism. It has been particularly important in the areas of cell cycle regulation, chromosome dynamics, DNA damage checkpoints, cytokinesis and chromatin structure. Its strengths largely reflect the fact that these are areas in which *S. pombe* is mechanistically more similar to metazoans than *S. cerevisiae*. Like metazoans, *S. pombe* has relatively large chromosomes with large repetitive centromeres, large, ill-defined, inefficient origins, and intron-rich genes. In addition to having a similar genome structure, *S. pombe* and metazoans have similar genome-packaging mechanisms; *S. pombe* telomeres and heterochromatin are composed of proteins similar to those found in metazoans.

Additional features and processes that are similar between *S. pombe* and metazoans but absent or highly diverged in *S. cerevisiae* include: large repetitive centromeres, complex, redundant and inefficient origins of replication; distinct metaphase chromosomes; gene regulation by histone methylation; chromodomain heterochromatin proteins; gene and transposon regulation by the RNAi pathway; telomere binding proteins; G2/M cell cycle control; cytokinesis; mitochondrial translation code; Signalosome pathway and Spliceosome components.

The growing use of *S. pombe* as a model for eukaryotic biology is reflected in an increase in the number of labs working with the organism. In 1999 approximately 100 labs were working on *S. pombe*, and in most cases *S. pombe* was the major focus of the lab. This year that estimate has doubled, with most of the growth coming from labs that have now begun to use *S. pombe*. The vitality of *S. pombe* research can be seen in the growth of publishing and Internet usage. A PubMed search for "pombe" returns 6,500 publications, with over 800 in the last year. The GeneDB website, which provides curated genome databases, generates over 5,000 hits a day for *S. pombe*. Likewise, pombe.net, a community resource maintained by Susan Forsburg at USC, gets over 1,000 hits a day.

The advantages of comparative genomics in the *Schizosaccharomycetes*

The *S. pombe* genome sequence has been a tremendous resource for the fission yeast community and for researchers looking for a simple model of eukaryotic cellular processes. However, the impact of the *S. pombe* sequence is limited by the state of its annotation. To accelerate the annotation process and increase the utility of the existing genome sequence we propose to use the strategy of sequence comparison that has been so successful for the *Hemiascomycetes*. There are two areas in particular that would benefit

extraordinarily from comparative genomics: annotation of coding sequences and identification of conserved non-coding regions as candidate regulatory elements.

Gene annotation

Although the *S. pombe* genome is actively being annotated using manual methods, the completeness in the current *S. pombe* gene catalogue lags well behind that of *S. cerevisiae* due to the more complex structure of the *S. pombe* genes and the smaller amount of supporting evidence. The greater intron density in *S. pombe* has made the initial automated annotations less accurate than *S. cerevisiae*, increasing the reliance on the slower process of manual annotation. Analysis of the *Saccharomyces sensu stricto* species has demonstrated that comparative genomics can efficiently identify coding regions and produce high-quality annotations of compared genomes (Cliften et al. 2003; Kellis et al. 2003). Comparative annotation of the *Schizosaccharomycetes* will also be of great interest because it will identify conserved non-protein-coding genes, which are currently impossible to identify from sequence alone. For instance, many polyadenylated non-coding RNAs have been identified in *S. pombe* cDNA libraries (Watanabe et al. 2002). Their function is mysterious, although it is speculated that they may be involved in RNAi-dependent gene regulation. A comparative annotation of such genes would be very useful in understanding their function. An accurately annotated genome would greatly increase the power of *S. pombe* as a model organism, making it easier to identify mammalian homology.

Motif discovery

In addition to the annotation of expressed genes, comparative genomics within the *Schizosaccharomycetes* offers a powerful method for identifying regulatory and structural sequence motifs. Such an analysis will be particularly interesting because of the relatively complicated nature of *S. pombe* chromosomes, which has thus far hampered attempts to identify functional elements within centromeres and replication origins. Thus, in addition to the transcriptional regulatory elements that will be identified, motif discovery promises to provide a powerful way to identify and study the organization, function and evolution of conserved sequences important in genome structure and metabolism. In particular, such an analysis should identify well-understood features such as origins and centromeres, as well as less well-defined features such as cohesion sites, recombinational enhancers, MARs/SARs, heterochromatin boundary elements, and perhaps other unexpected features. For example, *S. pombe* appears to use transposon long terminal repeats (LTRs) to regulate genes in a RNAi-dependent mechanism; comparisons will help to identify regulatory transposons (Schramke and Allshire 2003).

Species selection

Of the three *Schizosaccharomyces* species described here, the two most suitable for the goals of comparative genome annotation and motif discovery are *S. octosporus* and *S. japonicus* (Figure 2). Pairwise nucleotide identity between these species and *S. pombe* for well-conserved genes, such as gamma tubulin and U6 RNA, ranges from 78% to 89%. These three species therefore form a cluster that resembles *S. cerevisiae*, *S. mikatae* and *S. bayanus* in evolutionary distance (Kellis et al. 2003). This range of sequence similarity has proven to be extremely powerful for comparative purposes, both for coding region

annotation and regulatory and structural motif discovery. Conversely, another candidate, *S. kambucha,* appears to be too closely related to *S. pombe* to be informative for the comparisons described; the two are 95% identical over the 15-kb mating type locus (Singh and Klar 2002).

The choice of the *S. japonicus* genome has the added advantage that this species, unlike *S. pombe*, is highly invasive and forms true mycelia (Sipiczki et al. 1998). *S. japonicus* is already attracting interest as a experimental organism in its own right, with over two dozen papers published, and has the potential to be a useful model for invasive fungal growth. Furthermore, the comparison between the *S. pombe* and *S. japonicus* genomes will provide insight into the genetic determinants of invasive growth.

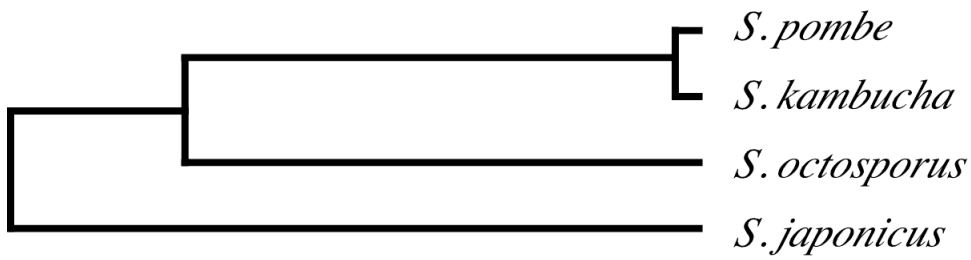

Figure 2: A phylogenetic tree of the *Schizosaccharomycetes*. The topology of the tree is based on mitochondrial protein sequence and is supported at each branch at over 95% confidence by both maximum likelihood and distance methods (Bullerwell et al. 2003). The relative branch lengths are based on nuclear protein and RNA sequences.

First sequence of a dermatophyte: *Trichophyton rubrum* Strain D12

T. rubrum and other dermatophytes are by far the most common fungal infections of humans. Thirty to seventy percent of adults are asymptomatic carriers of these fungi that cause skin and nail infections, including athlete's foot. Transmission of these fungal infections is common in public facilities, including gyms, camps, prisons, and military bases. Dermatophytes also cause acute and severe problems, for example, the crippling 1968 epidemic in the Mekong valley during the Vietnam War. While not fatal, dermatophyte infections cause tremendous pain and account for significant costs to society. Infections occur on the feet, nails, torso, groin, head, and arms. Dermatophyte infections of the head (tinea capitis) are a significant and ignored problem in Afro-American populations in urban settings in the United States. Over-the-counter drugs are available but do not eliminate the fungus and infections regularly reoccur. Further, these drugs are not effective against more severe infections.

Thirty-one of the 42 known dermatophyte species are pathogenic to humans. The species are distinguished by morphology and mating criteria. The species are divided into geophilic (soil dwelling), zoophilic (animal-specific), and arthropophilic (human-specific) species. *T. rubrum* is a cosmopolitan arthropophilic dermatophyte found throughout the world. Once an infection with *T. rubrum* has been controlled by therapy,

the patient is a life-long carrier of the organism. Similar to other arthropophilic species, *T. rubrum* appears to have lost its mating ability.

Fifty-eight percent of the dermatophyte species isolated at clinical laboratories are *T. rubrum*, 27% are *T. mentagrophytes*, 7% are *T. verrucosum*, and 3% are *T. tonsurans*. Based on these frequencies, *T. rubrum* is the best choice for the genome sequencing of a dermatophyte. The strain selected for sequencing is a particularly virulent strain, isolated during the above-mentioned Vietnam War Mekong Valley outbreak. The strain was also the main *T. rubrum* strain used in the development of the antifungal drug terbinafine.

T. rubrum researchers focus on the epidemiology, clinical case reports, strain relatedness, and drug susceptibilities of the organism. Clinical microbiology laboratories report on the isolation and identification of various dermatophyte species. Susceptibility levels in response to many current drugs are known for numerous strains and species. However, there is surprisingly little basic investigative biological research for a group of organisms that cause the most common fungal infection in humans. The availability of the genome sequence would augment the work in many of these areas, especially basic research including drug susceptibility and resistance. A complete genome sequence for *T. rubrum* would also allow current technologies and techniques to be applied to the dermatophytes for the development of new diagnostics, therapies, and vaccines. A group of researchers has been assembled who work on dermatophytes, who support the sequencing effort and who agree on the strain selected by the FGI Steering Committee for sequencing. The group includes clinical researchers, epidemiologists, basic researchers, and those interested in drug development.

The genome of *T. rubrum* is estimated to be 35 Mb and can be separated into four chromosomes using pulsed-field gel electrophoresis. The repeat content has been estimated to be typical for a filamentous fungus, at 10–15%, and the genes sequenced to date show an average AT content of 50%. *T. rubrum* strain D12 is haploid, avoiding complications in sequence assembly due to polymorphism.

First sequence from the chytrid order highlights an emerging pathogen

Chytrids, a deeply rooted fungal branch

Phylogenetic studies based on rDNA and on whole mitochondrial genome sequences indicate that the Chytridiomycota (chytrids) are basal in the fungal clade. Chytrids are unique among the true fungi in possessing zoospores, which move using posterior flagella. In this aspect, chytrids most closely resemble the protist ancestors of all fungi and animals. It is thought that the Chytridiomycota diverged from the ancestors of the later arising ascomyctes and basidiomyctes more than 600 myr ago (Figure 1). Thus, the many sequenced representatives of the distantly related ascomycetes and basidiomycetes have limitations in explaining the poorly understood biology and evolution of the chytrids. The genome sequence of the first chytrid will have great value for comparative genomics — both within the fungal clade and also with the sister animal clade.

Chytrid species are commonly found in aquatic and damp terrestial environments. As members of both aquatic and terrestrial microbial communities, chytrids are parasites and saprobes of many microscopic organisms (e.g., pollen, algae, and invertebrates) and play an important ecological role in the degradation of recalcitrant materials, such as chitin, keratin, and cellulose. A few chytrid species are parasites of higher plants, the most notable of which is *Synchytrium endobioticum*. The discovery of this potato pathogen in Nova Scotia in October 2000 raised concerns about spread of this pathogen to the United States, especially since this fungus can survive in the soil up to 40 years. Species of *Olpidium* also are of agricultural importance because they serve as a vector for viruses of crops. The lack of supporting resources greatly hampers studies of the many chytrid species that disrupt agriculture and ecological systems.

Batrachochytrium dendrobatidis

B. dendrobatidis is a non-filamentous (monocentric) chytrid firmly within the order Chytridiales, based on ultrastructural characteristics and molecular sequence data. Chytridiales is the largest (~80 genera and 500 species) of the five chytrid orders.

Batrachochytrium is a pathogen of amphibians implicated as a primary causative agent of amphibian declines. This recently emerging pathogen was identified in 1998 as the cause of amphibian deaths in Australian and Central America (Berger et al. 1998). More recently, *B. dendrobatidis* has been implicated in the dramatic and well-publicized die-offs of frog species in North America, South America, Europe and Africa (Berger et al. 1998; Daszak et al. 1999; Bosch et al. 2001; Bradley et al. 2002). This fungus invades the top layers of skin cells and causes thickening of the keratinized layer. Because amphibians drink and breathe through their skin, the fungus may kill them by disrupting these mechanisms. Alternatively, the fungus may be secreting a toxin.

Pulsed-field gel electrophoresis of the approximately 20 *B. dendrobatidis* chromosomes provides an estimate of the haploid genome size as roughly 20 Mb. The GC content of the genome is estimated to be 40%, based on the sequences of 34 randomly cloned DNA regions.

B. dendrobatidis is cultured as a diploid. From a sample of 8,500 bp across 14 loci of allelic sequence, the polymorphism rate is estimated to be 1 in 1000 bases on average. This rate did not significantly vary between 34 strains of *B. dendrobatidis* tested (Morehouse et al. 2003). Recent assemblies of the chimpanzee and dog genomes at the Broad Institute have demonstrated that we can perform high quality assemblies of polymorphic genomes with this rate of polymorphism. In particular, the chimpanzee genome assembly displays an average polymorphism rate of 1 in 1000 bases. This 4.5X assembly yielded a contig N50 of 15.7 kb and scaffold N50 of 8.6 Mb. For the dog, the approximately 10% of the 8X assembly displaying a polymorphism rate of 1 in 1000 or greater is entirely found in large contigs of at least 100 kb. The complete dog assembly has a contig N50 of 123 kb and scaffold N50 of 41.6Mb. From this data, we expect to assemble the vast majority of the *B. dendrobatidis* genome. We propose to sequence to 10X depth to aid in the assembly of regions with high polymorphism rates.

B. dendrobatidis as an experimental system

B. dendrobatidis is readily grown in axenic culture. No resting spore or sexual stages are known. Dozens of cultures isolated from North America, Central America, and Australia of *B. dendrobatidis* are available, including the type strain. The NSF-sponsored IRCEB project concerning the role of disease in amphibian decline is defining the population genetics of *B. dendrobatidis*. These scientists, as well as the new chytrid systemetists trained by the NSF PEET grant have an immediate interest in and use for the sequence. An additional value of the chytrid sequence is the tremendous benefit to other fungal genomic studies through use of the sequence of a basal fungus for comparative genomics. Annotation of genes in *B. dendrobatidis* will be aided by the existing 1,500 EST sequences and ongoing work to generate another 3,500 sequences.

Sequencing strategy and coverage

Whole-genome shotgun sequence will be obtained as paired reads from both plasmid and Fosmid clones obtained by random shearing of genomic DNA. For each genome, 90–95% of the total sequence coverage will be produced from plasmid subclones (using a combination of 4-kb and 10-kb inserts), and 5–10% of the coverage will be produced from the Fosmid clones. Sequence reads from each genome will be assembled prior to any alignment with related genome sequences. For each genome, we have targeted 7X coverage with the goal of producing high-quality, independent assemblies of each organism. Although Fosmid links could be omitted from the *Schizosaccharomyces* spp. without suffering a loss in base pair quality of the assembly, the information on genome rearrangement contributed by the greater long-range continuity resulting from inclusion of Fosmid links will provide additional valuable information about genome rearrangements between the species being compared.

The desired coverage and number of reads needed for each of these genomes is shown in Table 1. Current combined average read lengths (715 bp) and pass rates (87%) for all read types have been used for these calculations.

Table 1: Genome size, desired coverage and required reads

Organism	Est. genome size (Mb)	Total coverage	Approx. number of reads
S. octosporus	14	7X	162,000
S. japonicus	14	7X	162,000
T. rubrum	40	7X	461,000
B. dendrobatidis	20	10X	330,000
Total	**88**		**1,188,000**

Community and resources

Existing strong connections between the fungal research community and the Broad Institute provide access to the materials and information needed to perform these studies. Genomic DNA from *S. octosporus* and *S. japonicus* will be provided by Dr. Nick Rhind at the University of Massachusetts Medical School. DNA from *T. rubrum* will be supplied by Dr. Ted White at the Seattle Biomedical Research Institute. DNA from *B. dendrobatidis* will be provided by Joyce Longcore (University of Maine) and her collaborators. To ensure preparation of high quality DNA for successful library construction attempts, scientists at The Broad Institute provide detailed instructions regarding DNA preparation methods and will provide kits when needed.

References

Berger L, Speare R, Daszak P, Green DE, Cunningham AA, Goggin CL, Slocombe R, Ragan MA, Hyatt AD, McDonald KR, Hines HB, Lips KR, Marantelli G, Parkes H. 1998. Chytridiomycosis causes amphibian mortality associated with population declines in the rain forests of Australia and Central America. *Proc Natl Acad Sci USA* **95**(15): 9031–9036.

Bosch J, Martinez-Solano I, Garcia-Paris M. 2001. Evidence of a chytrid fungus infection involved in the decline of the common midwife toad (*Alytes obstetricans*) in protected areas of central Spain. *Biological Conservation* **97:** 331–337.

Bradley GA, Rosen PC, Sredl MJ, Jones TR, Longcore JE. 2002. Chytridiomycosis in native Arizona frogs. *J Wildl Dis* **38(1):** 206–212.

Bullerwell CE, Leigh J, Forget L, Lang BF. 2003. A comparison of three fission yeast mitochondrial genomes. *Nucleic Acids Res*. **31(2):** 759-68.

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* **301:** 71–76.

Daszak P, Berger L, Cunningham AA, Hyatt AD, Green DE, Speare R. 1999. Emerging infectious diseases and amphibian population declines. *Emerg Infect Dis* **5(6):** 735–748.

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.

Morehouse EA, James TY, Ganley AR, Vilgalys R, Berger L, Murphy PJ, Longcore JE. 2003. Multilocus sequence typing suggests the chytrid pathogen of amphibians is a recently emerged clone. *Mol Ecol* **12(2):** 395–403.

Schramke V and Allshire R. 2003. Hairpin RNAs and retrotransposon LTRs effect RNAi and chromatin-based gene silencing. *Science* **301:** 1069–1074.

Singh G and Klar AJ. 2002. The 2.1-kb inverted repeat DNA sequences flank the mat2,3 silent region in two species of Schizosaccharomyces and are involved in epigenetic silencing in *Schizosaccharomyce pombe*. *Genetics* **162:** 591–602.

Sipiczki M, Takeo K, Yamaguchi M, Yoshida S, Miklos I. 1998. Environmentally controlled dimorphic cycle in a fission yeast. *Microbiology* **144 (Pt 5):** 1319–1330.

Watanabe T, Miyashita K, Saito TT, Nabeshima K, Nojima H. 2002. Abundant poly(A)-bearing RNAs that lack open reading frames in *Schizosaccharomyce pombe*. *DNA Res* **9:** 209–215.