A practical guide to fungal genome projects: strategy, technology, cost and completion

Li-Jun Ma and Natalie D. Fedorova*

a The Broad Institute, Cambridge, MA, USA; b Infectious Disease Program, The J. Craig Venter Institute, Rockville, MD 20850, USA

(Received 26 January 2010; final version received 5 February 2010)

Since the publication of the first fungal genome in 1996, over 400 fungal genome projects have been launched, which makes fungi the most sequenced eukaryotic taxon. In this review, we sought to address the two major challenges associated with such projects, such as obtaining genomic sequence with high quality and attaching relevant biological information to the genomic sequence and features. Using the available literature and our own experiences in fungal genomics, we outline here the critical steps and quality control procedures for each stage of the genome project from library construction, sequencing to assembly and annotation, which will provide practical guidelines to optimize genome sequencing and annotation strategies. In addition, a brief overview of available comparative genomics and visualization tools is provided to assist investigators along the course of biological discovery. We also highlight how the success of a genomic project depends on close collaboration among experts from biological and computational sciences.

Keywords: fungal genomics; genome sequencing; genome assembly; annotation; gene finder; next generation sequencing

Introduction

Since the late 1970s, technological advancements in DNA sequencing have continued to fuel the exponential growth of sequence databases. The Next Generation Sequencing (NGS) technologies developed in the recent years have further accelerated the growth and decentralization of genome sequencing from specialized centers to research institutions and universities poses numerous challenges for the biologists. Since fungi became the most sequenced eukaryotic taxon (genomesonline.org/gold_statistics.htm; www.ncbi.nlm.nih.gov/genomeprj), fungal biologists are among the very first scientists that have to deal with these challenges, as well as the new discoveries made through various fungal genomics studies.

The first and possibly the most critical step in any genome project is the selection of organism to be sequenced based on medical, agricultural, or industrial significance, ecological diversity and phylogenetic breadth. These considerations are comprehensively covered in the companion article by Zhiqiang An and others entitled “China’s fungal genomics initiative: a whitepaper.” For each selected organism, the ultimate objective is to generate a complete genomic sequence and categorize its functional features, such as genes, promoters, etc., as well as to discover genetic determinants of important biological properties. In this review, we will focus on the practical aspects of the genome project and present a brief overview of current sequencing technologies as well as genome assembly, annotation and analysis tools. The main objective is to provide guidelines to help researchers find the most cost effective sequencing and annotation strategies and to find the right tools for downstream analyses of the sequence data.

Genome resources

Since the publication of the first fungal genome, the model yeast Saccharomyces cerevisiae in 1996 (Goffeau et al. 1996), over 400 fungal genome projects have been launched and over 100 been brought to at least the draft assembly phase (www.ncbi.nlm.nih.gov/genomeprj). The rapid growth of available sequence data, analysis tools and databases, makes it a daunting task for fungal researchers to sort out the best sequencing, annotation and analysis strategies. Fortunately a number of public genomic resources have been developed to help the researcher to overcome some of the challenges.

To determine the taxonomic position of the organism, one can search the NCBI Taxonomy database (ncbi.nlm.nih.gov/Taxonomy) with the organism name (common or scientific). This will generate links to taxon summary pages containing the Entrez records table with the numbers of records in the Nucleotide, Nucleotide EST, Nucleotide GSS, and Protein databases. The Taxonomy database contains information on 54,000 fungal species out of the estimated 1.5 million fungal species found in nature (Hawksworth and Rossman 1997). Phylogenetic information on the target species can be obtained using other web resources such as The Fungal Tree of Life Project web site (tolweb.org/tree/phylogeny.html) and the Fungal 18S Ribosomal RNA RefSeq Targeted
The availability of a so-called reference genome from the same choice of sequencing platform can be affected by the enhancing the impact of the project. For instance, the determining the sequencing and annotation strategies, and genomic sequence data will help in estimating the cost, unigene). To estimate the number of protein or EST from the NCBI UniGene database (www.ncbi.nlm.nih.gov/ (Appendix Table 1).

In addition, it contains information on many mitochondrial and plasmid fungal genome projects (Appendix Table 1).

A comprehensive understanding of the available genomic sequence data will help in estimating the cost, determining the sequencing and annotation strategies, and enhancing the impact of the project. For instance, the choice of sequencing platform can be affected by the availability of a so-called reference genome from the same or a very closely species. In addition, genomic data from close relatives can facilitate genome annotation and improve gene models. Of course, the comparative studies of both genomic and transcriptomic data are among the most powerful tools in understanding the biological significance of the sequences and are frequently used in post-genomic studies.

Whole transcriptome sequences are very useful, but at the moment only five complete transcriptome sets are available from the NCBI UniGene database (www.ncbi.nlm.nih.gov/unigene). To estimate the number of protein or EST sequences available for a specific organism, one can use the NCBI Entrez search engine and the type the species name (e.g. Aspergillus parasiticus [organism]) and mRNA (biomol_mrna [properties]) in the query. EST sequences can be also accessed via the Database of Expressed Sequence Tags (dbEST). It is a division of GenBank that hosts short single-pass reads of cDNA (transcript) sequences. For resequencing projects, the genomic sequence can be also used as a query in searches against the Nucleotide EST Database.

### DNA extraction and library construction

The quality of sequencing data is directly affected by the integrity of template. While specific requirements may vary depending on the nature of the initial DNA or RNA material, such as gDNA, cDNA, BACs, mRNA-Seq, small RNA-Seq and other nucleic acid (NA) material, quality control remains the most critical step for good sequencing library preparation. The presence of polysaccharides or any other particulate matter has been shown to interfere with library preparation and negatively affect the yield.

Having a sufficient quantity of the starting material is also critical to ensure additional QC steps at various stages of library preparation. Nucleic acid (NA) concentration measurement can be a surprisingly challenging step. Nanodrop readings are often misleading by 100–200%, due to the presence of polysaccharides or other NAs in the sample. It is preferable to estimate DNA concentration by gel densitometry and using a fluorometric assay, such as SYBR Green or the Quant-IT PicoGreen dsDNA assay kit.

In addition to quality and quantity, the NA size distribution also needs to be checked in the initial sample prep and at various stages of library preparation. Sample requirements can vary depending on the type of library (fragment or paired end) and the sequencing method being employed. For example, paired end plasmid and fosmid libraries require high quality intact, high molecular weight (hmw) DNA. Smaller fragmented DNA is only recommended for the small insert libraries with an insert size less than 3 kb. Large quantities of hmw gDNA are required for large insert libraries to ensure optimization of shearing protocols. For instance, up to 60 μg of hmw gDNA may be necessary for a 20-kb paired-end library sequenced using the 454 technology, even though a fragment library can be constructed with as little as 3–5 μg gDNA.

Extraction of fungal gDNA remains a challenge for many species, several reliable methods have been developed including some published at Protocols Online (www.protocol-online.org). To avoid shearing of gDNA during pipetting, many protocols recommend not using columns and cutting off the ends of the pipette tips before handling gDNA. Many sequencing centers also prefer NA samples to be shipped as a pellet in 70% ethanol (plus salt)

---

**Table 1. Taxonomic distribution of sequenced fungal species.**

| Taxonomic clade | Sequenced speciesa | Total no. speciesb | Total no. articlesc |
|----------------|-------------------|--------------------|---------------------|
| Ascomycetes    | 97                | 31,287             | 69,070              |
| Basidiomycetes | 20                | 17,062             | 9,696               |
| Microsporidian | 6                 | 520                | 800                 |
| Chytridiomycota| 2                 | 441                | 201                 |
| Blastocladiomycetes | 1 | 36          | 147                 |
| Unclassified fungi | 2   | 3,805        | 41                  |
| Total          | 128               | 53,151             | 79,955              |

Notes: aThe number of fungal species from each clade being sequenced according to the NCBI Genome Projects database. 
bThe total number of species from each clade according the NCBI Taxonomy database. 
cThe number of records of annual publications that contain information about each taxonomic clade.
overnight in dry ice. As a potent inhibitor of some sequencing reaction enzymes, the use of EDTA should be minimized.

Another complication that can significantly affect a project’s outcome is the availability of the homokaryon phase. Ideally, a genomic DNA should be extracted from the homokaryon phase with haploid nuclei. Even though genomic DNA isolated from the heterokaryotic and vegetative diploid (or dikaryon) phases can be used, it is usually at a much higher price. Depending on the assembly stringency, the span of scaffolds for a heterokaryon genome can be almost twice the size of the homokaryon genome from the same species, demonstrated in recently sequenced wood decay Basidiomycete fungus Postia placent (Martinez et al. 2009) and the potato fungus Rhizoctonia solani (Marc Cubeta, 2009, unpublished data). Equally important is that a diploid genome may create complications at both the assembly and annotation steps.

Contamination is another potential hurdle for a genome project; therefore, it is extremely important to ensure that the DNA sample comes solely from the target organism by validation steps, such as PCR assays. We also strongly recommend users to generate small number of data as a test run before committing to full-scale sequencing efforts. The testing reads should be analyzed using Blast to detect any potential contamination. The read quality can be accessed by a few metrics, including the total number of bases, read length distribution, GC content distribution, and BLAST n/p hits. For instance, a marked difference in the GC content between the new reads and the reference data for the target organism could be indicative of a possible contamination issue.

To minimize the chance of a stain mix-up, it is desirable to obtain the strains directly from a public repository such as the Fungal Genetics Stock Center (FGCS) or Fungal Biodiversity Center (www.cbs.knaw.nl/). It is also recommended or required by many funding agencies that the sequenced strains should be deposited in one of the publicly accessible strain centers and the sequence data should be released to public databases to ensure the best usage of the resources generated through the genomic project.

Sequencing instruments

In the late 1970s, two pioneering sequencing technologies, the Maxam–Gilbert (Maxam and Gilbert 1977) and the Sanger (Sanger and Coulson 1975) methods, were developed almost simultaneously. Even though the Maxam–Gilbert sequencing method was widely used initially, improvements in the Sanger method, especially the development of automated high-throughput DNA sequence analyzers, made the latter the sequencing method of choice and dominated the sequencing landscape in the early period of the genomic era. For some time, incremental improvements were made to the Sanger technology through four-color sequencing instrument optimization, resulting in improved base quality and read length. The use of automation and ever improving assembly software delivered continuous improvements to sequenced genome quality at continually declining costs. Since the early 2000s, several new technologies, called the Next Generation Sequencing (NGS) technologies, have been developed and those continue to revolutionize the field of genomics. The most successful ones produce large-scale sequence data in a much shorter time and with a dramatic cost reduction.

Until very recently, reading sequence bases from a single DNA molecule had not been possible. Therefore, amplifying individual DNA molecules through in vitro cloning remained a necessary step for increasing the signal intensity. Such amplification steps are implemented in both conventional Sanger methods and most widely applied NGS technologies, including 454 (www.454.com), Illumina (www.illumina.com) and SOLiD (www.solid.appliedbiosystems.com). Although, it is accomplished differently in each NGS platform, this step fundamentally separates the NGS technologies from the Sanger method. Briefly, the Sanger method captures, amplifies and sequences each DNA fragment one at a time. In contrast, all the NGS methods generate millions of high-density clonal DNA fragments in parallel and produce sequences for the millions of fragments simultaneously. This lowers sequencing cost and reduces the sequencing generation time. New alternative sequencing technologies, such as Single Molecule Sequencing by Helicos BioSciences (www.helicosbio.com) and Pacific Bioscience (www.pacificbiosciences.com), are now being developed to detect sequence from a single molecule, and eliminate the need for the amplification step altogether. Significantly, such method will overcome the problems associated with amplification bias; therefore reveal the true diversity within a population. Further development of single molecule sequencing by Pacific Biosciences, focusing on generating long kb-sized reads, is expected to widely impact sequencing and assembly of new genomes or repetitive regions.

Table 2 summarizes the main features of the four most commonly used sequencing platforms. In general, the Sanger method produces long sequence reads of high quality, which contributes to its wide application over a quite long period of time. The pairing information from whole genome sequencing (WGS) libraries in combination with directed sequencing efforts to finish the genome has delivered many relatively complete fungal genomes sequenced by this method. The major downside of Sanger method is its high cost and low speed relative to the NGS platforms. By increasing read length and improving base quality, the 454 technology is now becoming the platform of choice for most de novo fungal sequencing projects around the world. The 454 sequencing of a set of 3, 8 and 20 kb
libraries is often used as a more cost effective alternative to Sanger plasmid and fosmid libraries. Other NGS technologies, such as the Illumina and SOLiD platforms, have a greater reduction in cost and increased total yield of nucleotide bases, albeit at the expense of reads lengths and accuracy. This enables the generation of high sequence coverage, but also requires an increased computational expense in handling the data. Uniquely mapping short reads have been used successfully in genome re-sequencing projects and in transcriptome sequencing studies, further improvements in read length and quality are expanding their applications towards use in de novo assembly (Butler et al. 2008; Li et al. 2010).

Selection of sequencing strategy

Two main parameters are frequently used to plan a sequencing project: (i) desired sequencing depth and (ii) the desired linkage or connectedness. The depth is calculated by multiplying the instrument per run sequence yield by the number of runs and divided by the genome size. Linkage, on the other hand, is determined by the read length and by the linking power of paired-end libraries. It is often estimated as scaffold N50, a length-weighted average of scaffold size, meaning that the average nucleotide in an assembly will appear in a scaffold of N50 size or greater. For example, for a collection of scaffolds of sizes 7, 4, 3, 2, 2, 1, and 1 kb, the N50 scaffold length equals 4, since 50% of the total length (10 kb) is located in scaffolds greater or equal to 4 kb. Having paired-end libraries in addition to fragment libraries is critical for large eukaryotic projects, because it can facilitate assembly of repeat regions.

The choice of the sequencing strategy for a genomic project depends on the considerations of the strengths and weaknesses of different technologies and their ability to meet the depth and linkage requirements (Table 2); and it also depends on the project objective, as a genome project can be accomplished at different levels, such as finished genome, draft genome, or gene-only (i.e. exon-only) sequencing. Depth is chosen to balance the cost of additional sequence with the benefits to assembly connectivity and accuracy.

| Instruments | Sanger (ABI 3730) | NGS 454 FLX | NGS Illumina GA2 | NGS AB SOLiD |
|-------------|------------------|-------------|-----------------|--------------|
| Cost/kb ($)       | 1                | 0.1         | 0.0025          | <0.002       |
| Read length (bp)  | 750              | 500         | 36–125          | 35–75        |
| Days/Gb           | 650              | 1           | 0.3             | 0.3          |

Note: aThe sequencing capacity of each platform listed here as several factors can influence the actual output. In addition, based on past experience, improvements in sequencing chemistry and/or sequencing platform will likely increase their capacity in the near future.

Genome finishing

A genome is considered finished if its sequence quality conforms to the Bermuda standard, i.e. an accuracy of 99.99% (or less than one sequencing error in 10,000 nt) with no gaps (www.genome.gov/10001812). Any eukaryotic genome contains certain regions that are difficult to sequence (often regions with highly repetitive DNA). Thus, “completed” genome sequences are rarely ever complete, and terms such as “working draft” or “essentially complete” have been used to more accurately describe the status for many genome projects. Even when every base pair of the genomic sequence is determined, it may still contain errors in the consensus sequence.

Draft genome assembly

A working draft assembly can be produced more quickly with a much smaller budget than a finished draft assembly. Although there is no absolute standard to define a “draft assembly”, the standardized Sanger sequencing approach had delivered predictable High Quality Draft (HQD) assemblies that represent up to 99% of the genomic sequences with high base quality (at 7–8 × sequence coverage). The N50 scaffold size of a draft assembly range from over 5 Mb (e.g. Fusarium graminearum) to 10 kb, depending on the chromosome size and the repetitiveness of the genome. For creating a high quality assembly, high quality reads and paired-end reads from large insert size libraries are always desirable. According to the NCBI Genome Project database, most sequenced genomes are in the draft assembly state without their chromosomal structure defined.

Genes only approach

Along with the development of high yield, low cost NGS technologies, RNA-seq come to be “a new revolutionary tool” for transcriptomics (Wang et al. 2009). It can be been used to catalogue all actively transcribed elements from small RNAs to protein-coding genes by creating de-novo assembly of the short reads to define the transcriptome. The RNA-seq data can also be used to measure gene expression with the unprecedented dynamic range and precision. Paired-end reads are also useful in transcriptome sequencing projects by resolving differences between paralogous genes and by identifying alternatively spliced transcripts (Yassour et al. 2009). If the main objective is to obtain the complete set of genes, then whole transcriptome sequencing could be a cost-effective approach. The main challenge associated with this approach is to assess the completeness of the transcriptome. However, it will not bring information regarding repetitive sequences, the genome structure, chromosomal organization or physical coordinates of the genes.
RNA-seq library construction typically starts with polyA RNA, which has been purified from total RNAs, followed by fragmentation and conversion to double stranded cDNA. A minimal 10 μg total RNA, measured after DNase treatment and prior to the reverse transcriptase (RT) step, is required to ensure sequencing data integrity. One way to further increase the number of distinct transcripts in a given number of sequenced reads is to use normalized cDNA libraries, where highly abundant transcripts are reduced in relative abundance. Reannealing and double stranded nuclease are typically used to normalize transcripts, since the most abundant transcripts anneal first. This is an important step especially for prokaryotic samples that cannot be enriched using oligo-dT beads. Normalized cDNA can be amplified using the same adapters used during the RT step if necessary. Construction of normalized libraries can be outsourced to biotech companies, such as Evrogene (Moscow, Russia) or Express Genomics (http://www.express-genomics.com/).

**Genome assembly**

Until technology advances to routinely generate long DNA sequence fragments with high accuracy, sequence reads will continue to be generated from relatively short DNA fragments. As a result, complete genomic sequences spanning entire chromosomes have to be reconstructed by piecing these fragments together using the Whole Genome Assembly (WGA) process. WGA is an extremely computationally expensive process, which has been only made possible through major advances in the software development (International Human Genome Sequencing Consortium, 2001). The most commonly used software packages include Arachne (Batzoglou et al. 2002; Jaffe et al. 2003), SEQAID (Peltola et al. 1984), CAP (Huang and Madan 1999), PHRAP (Green 1994), TIGR assembler (Sutton et al. 1995), AMASS (Kim and Segre 1999) and many others. The advent of the NGS technologies sparked the development of a new generation of assembly programs such as Newbler from 454 Life Sciences (www.454.com), Velvet (Zerbino and Birney 2008), ALLPATHS (Butler et al. 2008) and SOAPdenovo (soap.genomics.org.cn/).

Briefly, the reads that originated from the same genomic location are identified by sequence-overlap and pieced together into larger, contiguous fragments or contigs (Figure 1). In turn, contigs are ordered and oriented into scaffolds (i.e. supercontigs) based on the paired-end reads information obtained from paired-end libraries. The success of the assembly process largely depends on the accurate placement of each read in the correct genomic location and of the each contig into the right scaffold. This process can be complicated by misplacement of sequence reads caused by repetitive sequences and by sequence errors introduced through the sequencing processes. For example, if a read is shorter than the associated repeat, it can be mapped to multiple locations and its placement can only be arbitrarily selected from one of these locations. A correct placement of the partner read from a read pair could correct this problem and find a correct location as long as the other paired-end read originates from an adjacent non-repeat unique region. The frequency of correct placement is increased by having paired-end insert sizes consequentially larger than the repeat classes contained in the genome being sequenced. Since the repeat classes are often unknown, it is important to have sequencing libraries with a range of insert sizes (e.g. 3, 8 and 20 kb).

The final step of the assembly process may involve mapping of the obtained scaffolds to the fungal chromosomes. In the past, such a process depended on the presence of a physical map. Today, a WGA assembly can be linked to the chromosomes using a genetic linkage map, if available. However, because the development of genomic sequence often surpasses the mapping process, many sequenced genomes have neither genetic nor physical mapping information associated with them. The development of the optical mapping technique, which involves capturing individual hmw DNA molecules elongated and fixed onto derivatized glass surfaces, followed by digestion in situ by selected restriction endonucleases (Schwartz et al. 1993), fills that gap. A detailed optical map, sometimes called optical restriction map, can be used to order and orientate the contigs, and to link all of the sequence in the assembly into whole chromosomes in a very cost efficient manner (Nierman et al. 2005; Ma et al. 2009).

Potential red flags in the assembly may include a lower than expected total scaffold span based on the estimated genome size (e.g. Platypus (Warren et al. 2008)), low coverage (<25 × sequence coverage in NGS reads), and low linkage (<12 × in paired-end clone coverage). The low linkage problem can be caused by the low amount of large-insert data (e.g. fosmid, mate-pair libraries), the presence of repetitive elements, and by mutations in the E. coli shotgun library prepared for Sanger sequencing (Sanger et al. 1977). Low coverage can be caused by random or non-random under-sampling, which can be fixed by more sequencing on the same platform or on a different platform and/or with paired-end libraries. Contaminating DNA from bacteria or parasites and endosymbionts and the presence of heterokaryons or diploid genomes can also complicate the assembly process.

**Genome annotation**

Genome annotation is a process that integrates lines of evidence to predict structure and function of certain genome features, such as coding genes, non-coding genes (ncRNAs), 5′ and 3′ untranslated regions of genes (UTRs), non-coding genes (ncRNAs), repeats, etc. For practical purposes, annotation is typically focused on the prediction of gene models for protein coding genes, i.e. a “parts list”
for the entire proteome. The results of genome annotation and associated assemblies are ultimately submitted to their final destination: the public archival databases at NCBI, DDBJ, and EMBL.

Surprisingly, there are virtually no universally agreed upon standards in eukaryotic genome annotation except for the input file format requirements from these public databases. This fact accounts for a great deal of variation in annotation quality among sequenced eukaryotic genomes, in association with the project objectives, funding resources and the pipelines used to generate the data. An accurate gene annotation is the foundation of a genome project, as inaccurate gene models in the public database could hinder many downstream analyses, including comparative and functional genomics, phylogenetic analysis, evolutionary biology and crystallography, to name but a few. Nevertheless, the goal to produce an accurate annotation is still an over-arching objective. Despite significant improvements in eukaryotic gene prediction algorithms, correct exon–intron structures can be only found in a handful of major genome organisms. Even these genomes may contain thousands of completely unannotated and misannotated genes (Brent and Guigo 2004; Wei et al. 2005).

Efforts using comparative genomics to improve the fungal annotation in model system yeast (Kellis et al. 2003) and Aspergillus genomes (Fedorova et al. 2008) have provided systematic methods to improve annotation using conservation signatures. However, particularly for genomes without comparative data, the predicted gene models had multiple problems, such as missing exons, fused or split gene models and even missing genes. Only a fraction (20%) of the orthologous core Aspergillus proteome had consistent gene structures across five species and discrepancies in gene structure and content were observed in up to 30% of orthologous gene models (N.D. Fedorova, 2009, unpublished data). Even though some of these errors can be corrected by extensive manual curation, this is very expensive and labor-intensive, and could be decreased by improvements to high-throughput methods.

To predict a gene structure, sequence- and evidence-based approaches are available.

**Sequenced-based annotation**
Sequence-based annotation determines reading frame, translation start and stop sites, splice boundaries, and the validity of short ORFs based on the sequence composition and signatures of these sites. The most used genefinders including ab initio gene prediction, (e.g. Augustus (Stanke et al. 2004), Fgenesh, Geneid), sequence similarity based (e.g. GeneWise (Birney et al. 2004)), and consensus-based (EvidenceModeler (Haas et al. 2008)) software. Atypical splice signals that evade gene finders can complicate predictions of splice sites, protein coding exons and gene model construction. To further improve gene structures, an ab initio gene finder can be trained with a training set generated from cDNA and proteogenomics data.

**Evidence-based annotation:- cDNA sequencing and proteogenomics**
Recently, cDNA sequencing has emerged as a cost-effective approach to improving annotation quality of eukaryotic genomes. The sequence reads can be either aligned to the reference directly or first assembled and then aligned to the genome using alignment tools, such as GMAP (Wu and Watanabe 2005). The application of such approach in annotating Cryptococcus neoformans (Tenney et al. 2004) resulted in 60% accuracy rate. In other words, 60% of the gene models were predicted correctly at every coding base and splice site, confirmed by RT-PCR and direct cDNA sequencing when compared to purely computationally predicted gene models. With the advent of the NGS technologies, the cDNA sequencing step can be taken even further to provide deep sequencing of the fungal transcriptome, even though methods to incorporate NGS reads into genome annotation are more complicated (Denoeud et al. 2008; Yassour et al. 2009).

Similar to cDNA sequencing, proteogenomics that complements nucleotide-based annotation has consistently and significantly improved the quality of genome annotation in humans, Arabidopsis, Toxoplasma, Drosophila, C. elegans, and many others (Tanner et al. 2007; Castellana et al. 2008; Merrihew et al. 2008; Xia et al. 2008; Findlay et al. 2009). It is routinely used at the J. Craig Venter Institute to validate existing annotations (Castellana et al. 2008). Briefly, mass spectrometry data is searched against an expanded protein catalog including both a six-frame translation of the entire fungal genome, and numerous potentially novel splice sites. Identified Expressed Peptide Tags (EPTs) are then mapped back on to the genome and compared to the current gene models. There are a variety of simple gene model updates, such as novel splice sites, novel exons or 5’ in-frame extension, that can be performed with an automated gene updating mechanism. For more complex peptide mapping, a gene predictor (e.g. Augustus) can be utilized to combine the extrinsic peptide evidences with current gene models to produce the best new model.

**Annotation pipelines**
Ideally, a genome annotation pipeline will combine and evaluate all the evidence collected for each gene locus, including empirical data such as cDNA sequences and proteogenomics data, and computed results, such as gene predictions from various genefinders, blast results, conserved domain mapping (Figure 1). A weighted matrix can be assigned to different evidence based on the different confidence levels.
For instance, empirical evidence, such as a full-length cDNA provides much higher confidence for the locus comparing to a single ab initio prediction. All lines of evidence are then utilized to generate a consensus gene-set using a program such as EvidenceModeler (Haas et al. 2008). Since many loci lack experimental evidence, more than five independent methods (such as SNAP, GlimmerHMM, FGENESH, and GeneMarkHMM, in addition to Augustus) are recommended for constructing good consensus gene models. After automated gene structure updates are finalized, manual quality control is suggested, aided by automated comparison to ESTs of genes in other species. Visualization tools such as Argo (Engels et al. 2006), or the Apollo genome browser (Ed et al. 2009) can be used to display all alignment evidence.

Well-known genome browsers include the Ensembl Genome Browser (Stalker et al. 2004), the NCBI Entrez Map Viewer (Wolfsberg 2007), UCSC’s Golden Path Genome Browser (Kent et al. 2002), Apollo (Ed et al. 2009), and Argo (www.broad.mit.edu/annotation/argo/). In addition to genome browser functionalities, the Argo browser provides an interface for annotation of all publicly released fungal genomes conducted through the Fungal Genome Initiative at the Broad. As with many other genome browsers, Argo software enables the user to upload his own annotation for comparison to reference genomes stored in the database. Argo also enables users to inspect introns, Poly(A)+ signals, and to efficiently perform local alignments. Fungal-specific genome browsers include SNUGB (Jung et al. 2008), the Saccharomyces Genome Database (SGD) (Issel-Tarver et al. 2002), the Aspergillus Genome Database (AspGD; www.aspgd.org), the Central Aspergillus Data Repository (CADRE) (Mabey et al. 2004), the Ashbya Genome Database (AGD) (Gattiker et al. 2007), the Yeast Gene Order Browser (YGOB) (Byrne and Wolfe 2005), and many others.

**Functional annotation and the research community**

Once gene structure improvements are completed, functional annotation can be performed using tools to assign conserved protein domains (such as HMMER, Interproscan), Gene Ontology (GO) terms (such as Blast2Go), Enzyme Classification (EC) numbers, metabolic/signaling/regulatory pathways, motifs, and cell localization signals to the gene set. Supervision of skilled curators is highly recommended to evaluate the quality of the output.

With collective knowledge, manual curation for a limited number of genes, which are of high interest to the community or selected for potential incorrect annotation, can be very beneficial to the project. It is often labor intensive, but the effort could dramatically improve the genome annotation quality, as demonstrated by the Aspergillus nidulans genome re-annotation project (Wortman et al. 2009) and Fusarium graminearum manual annotation efforts (Guldener et al. 2006). Ideally, such process should involve the research communities directly through online genome databases, such as the Aspergillus Genome Database (AspGD) (Arnaud et al. 2010).

Another consequence of community involvement is increased visibility of the project, as well as a unifying impact on the research community. As a result of such efforts, some fungal research communities, such as the Aspergillus and Fusarium communities, have became very well organized. Each community holds annual satellite meetings in conjunction with larger conferences to exchange scientific discoveries and to access priority listing of resources to be generated and shared among the community. Since the publication of the first A. fumigatus genome sequence (Nierman et al. 2005), NCBI PubMed records of annual publications on this fungus have increased 10-fold in five years.

**Visualization and understanding of the sequencing data**

Genomics emphasizes the complexity of biological systems and reveals the challenges we have to overcome to correlate complex biological traits with genomic features. Visualization, one of the critical components in organizing these complex datasets, is critical for the success of such process. The most commonly used visualization tools include genome browsers, dotplots, and syntenic maps, which biologists to interrogate genomic features and view aligned orthologous regions between genomes.

A dotplot is another useful tool that can display an alignment of two sequences as a set of local aligned segments. It has been widely used in whole genome comparative studies to identify large-scale evolutionary changes and to define conserved syntenic regions, which can help to delineate the evolutionary relationships between species. Multiple software tools, such as Dotter and gff2aplot, have been developed to facilitate the dotplot analysis (Sonnhammer and Durbin 1995). The comparative genome browser Combo, developed as part of the Argo software package (www.broad.mit.edu/annotation/argo/), also provides a dynamic view of whole genome alignments. Users can load both genomic sequences and sequence annotation data in FASTA or GFF format. The software tool can also import alignment data from BLAST (Altschul et al. 1990) and PatternHunter (Li et al. 2003; 2004), as well as any file in tab-delimited format.

The success of a genomic project depends on the thorough understanding of the biological system, which requires close collaboration between experimental and computational scientists. Typically, biologists are most comfortable with a hypothesis-driven research, but can sometimes get lost in this mostly data-driven environment because they may lack the tools and/or skills required to handle large datasets. At the other extreme, most computational scientists have a difficult time translating the sequence bases into
biological traits. As two indispensable parts of a whole, an active, trusting and mutual-beneficial collaboration between these two groups of project participants becomes the pillar of each genomic project.

As a result, one of the most critical steps in a genome project is to assemble an interdisciplinary team of contributors with specific roles, including experts in computer science, bioinformatics and fungal biology. To better meet the needs of the research community, interested parties should participate in the annotation effort and help to map out the research field and to identify the knowledge gaps.

Although the actual size of the global fungal research community is not precisely known, biannual meetings, such as the Fungal Genetics and the European Community Fungal Genetics meetings, are attended by thousands of researchers from around the world. Local meetings such as the 2009 China Fungal Genome Initiative Symposium can be an additional breeding ground for collaborations.

Applying comparative genomics among a group of carefully selected organisms that maximize the phenotypic variation within shortest divergence time could be another key to the success. A good design could enable scientists to formulate and test hypothesis of increasing strength that may guide the project development. During this process, it is important to keep focused on the few key findings that most likely will have the most impact on the field. An old Chinese proverb states that “an overly focused vision on a tree may block the global view of the forest”. It is important to identify and constantly revisit the main biological questions through the course of every genome project.

Summary

This article is designed to overview the critical steps in fungal genome sequencing, assembly and annotation, and to provide a brief introduction to existing bioinformatics tools and databases. The guidelines can help optimize strategies and shorten the discovery path in finding the genetic determinants of the most interesting phenotypic traits. We also provide a brief overview of comparative genomic analysis and visualization tools available to investigators. Ultimately, the success of each project depends on an inter-dependent, collaborative, and trusting research team including experts in bioinformatics and fungal biology, as well as on allocating sufficient computational resources as early in the project as possible.

Acknowledgements

This project has been funded in part with federal funds from the US National Science Foundation, the United States Department of Agriculture, the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services under contract numbers, 2005-35600-16405, 2008-35600-04691, N01-AI-30071 and HHSN272200900007C. We would like to thank the faculty and staff of the J. Craig Venter Institute (Rockville, MD, USA) and the Broad Institute of MIT and Harvard (Cambridge, MA, USA) who helped us to acquire the latest updates on the technological innovations in sequencing and annotation for this review. We are especially grateful to William C. Nierman and Elisabeth Caler from the J. Craig Venter Institute and Christina Cuomo, Chad Nusbaum and Brian Haas from the Broad Institute. In addition, we would like to thank the 2009 China Fungal Genome Initiative Symposium participants for providing the inspiration for this article.

References

Altschul, SF, Gish, W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol. 215(3): 403–410. Arnaud MB, Chibucos MC, Costanzo MC, Crabtree J, Inglis DO, Lotia A, Orvis J, Shah P, Skrzypek MS, Binkley G, Miyasato SR, Wortman JR, Sherlock G. 2010. The Aspergillus Genome Database, a curated comparative genomics resource for gene, protein and sequence information for the Aspergillus research community. Nucleic Acids Res. 38(Database issue): D420–427. Batzoglou S, Jaffe DB, Stanley K, Butler J, Gnerre S, Mauceli E, Berger B, Mesirov JP, Lander ES.2002. ARACHNE: a whole-genome shotgun assembler. Genome Res. 12(1): 177–189.
Huang X, Madan A. 1999. CAP3: A DNA sequence assembly program. Genome Res. 9: 868–877.

Issel-Tarver L, Christie KR, Dolinski K, Andrade R, Balakrishnan R, Ball CA, Binkley G, Dong S, Dwight SS, Fisk DG, Harris M, Schroeder M, Sethuraman A, Tse K, Weng S, Botstein D, Cherry. 2002. Saccharomyces Genome Database. Methods Enzymol. 350: 329–346.

Jaffe DB, Butler j, Gnerre S, Mauceli E, Lindblad-Toh K, Mesirov JP, Zody MC, Lander ES. 2003. Whole-genome sequence assembly for mammalian genomes: Arachne 2. Genome Res. 13: 91–96.

Jung K, Park J, Choi J, Park B, Kim S, Ahn K, Choi J, Choi D, Kang S, Lee YH. 2008. SNPUGB: a versatile genome browser supporting comparative and functional fungal genomics. BMC Genomics 9: 586.

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.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. 2002. The human genome browser at UCSC. Genome Res. 12: 996–1006.

Kim S, Segre AM. 1999. AMASS: a structured pattern matching approach to shotgun sequence assembly. J Comput Biol. 6: 163–186.

Li M, Ma B, Kisman D, Tromp J. 2003. PatternHunter II: highly sensitive and fast homology search. Genome Inform. 14: 164–175.

Li M, Ma B, Kisman D, Tromp J. 2004. PatternHunter II: highly sensitive and fast homology search. J Bioinform Comput Biol. 2: 417–439.

Li R, Fan W, Tian G, Zhu H, He L, Cai J, Huang Q, Cai Q, Li B, Bai Y, Zhang Z, Zhang Y, Wang W, Li J, Wei F, Li H, Jian M, Li J, Zhang Z, Nielsen R, Li D, Gu W, Yang Z, Xuan Z, Ryder OA, Chi-Ching Leung F, Zhou Y, Cao J, Sun X, Fu Y, Fang X, Guo X, Wang B, Hou R, Shen F, Mu B, Ni P, Lin R, Qian W, Wang G, Yu C, Nie W, Wang J, Wu Z, Liang H, Min J, Wu Q, Cheng S, Ruan J, Wang M, Shi Z, Wen M, Liu B, Ren X, Zheng H, Dong D, Cook K, Shan G, Zhang H, Kosiol C, Xie X, Liu Z, Zheng H, Li Y, Steiner CC, Tsan-Yuk Lam T, Lin S, Zhang Q, Li G, Tian J, Gong T, Liu H, Zhang D, Fang L, Ye C, Zhang J, Hu W, Xu A, Ren Y, Zhang G, Bruford MW, Li Q, Ma L, Guo Y, An N, Hu Y, Zheng Y, Shi Y, Li Z, Liu Q, Chen Y, Zhao J, Qu N, Zhao S, Tian F, Wang X, Wang H, Xu L, Liu X, Vinar T, Wang Y, Lam TW, Yiu SM, Liu S, Zhang H, Li D, Huang Y, Wang X, Yang G, Jiang Z, Wang J, Qin N, Li L, Li J, Bolund L, Kristiansen K, Wong GKS, Olson M, Zhang X, Li S, Yang H, Wang J, Wang J. 2010. The sequence and de novo assembly of the giant panda genome. Nature 463: 311–317.

Ma, L. J., A. S. Ibrahim, Skory C, Grabherr MG, Burger G, Butler M, Elias M, Idnurm A, Lang BF, Sone T, Abe A, Calvo SE, Corrochano LM, Engels R, Fu J, Hansberg W, Kim JM, Kodira1 CD, Koehrsen MJ, Liu B, Miranda-Saavedra D, O’Leary S, Ortiz-Castellanos L, Poulter R, Rodriguez-Romero J, Ruiz-Herrera J, Shen YQ, Zeng Q, Galagan J, Birren BW, Cuomo CA., Wicks BL. 2009. Genomic analysis of the basal lineage fungus Rhizopus oryzae reveals a whole-genome duplication. PLoS Genet. 5: e1000549.

Mabey JE, Anderson MJ, Giles PF, Miller CJ, Attwood TK, Paton NW, Bornberg-Bauer E, Robson GD, Oliver SG, Denning DW. 2004. CADRE: the Central Aspergillus Data REpository. Nucleic Acids Res. 32(Database issue): D401–5.

Martinez D, Challacombe J, Morgenstern I, Hibbett D, Schmoll M, Kubicek CP, Ferreira P, Ruiz-Duenas FJ, Martinez AT, Kersten P, Hammel KE, Vandern Wymelenberg A, Gaskell J,
Lindquist E, Sabat G, Splinter BonDurant S, Larondo LF, Canessa P, Vicuna R, Yadav J, Doddapaneni H, Subramanian V, Pisabarro AG, Levin J, Oguiza JA, Master E, Herrissat B, Coutinho PM, Harris P, Magnuson JK, Baker SE, Bruno K, Kenealy W, Hoegerg PJ, Kiles U, Ramayia P, Lucas S, Salamov A, Shapiro H, Tu H, Chee CL, Misra M, Xie G, Teter S, Yaver D, James T, Mokrejs M, Propisek M, Grigoriev IV, Breetin T, Rokhsar D, Berka R, Cullen D. 2009. Genome, transcriptome, and secretome analysis of wood decay fungus Postia placenta supports unique mechanisms of lignocellulose conversion. Proc Natl Acad Sci USA 106: 1954–1959.

Maxam AM, Gilbert W. 1977. A new method for sequencing DNA. Proc Natl Acad Sci USA 74: 560–564.

Merrihew GE, Davis C, Ewing B, Williams G, Käll L, Fureen BE, Noble WS, Green P, Thomas JH, MacCoss MJ. 2008. Use of shotgun proteomics for the identification, confirmation, and correction of C. elegans gene annotations. Genome Res. 18: 1660–1669.

Nieman WC, Pain A, Anderson MJ, Wortman JR, Kim HS, Arroyo J, Bertram R, Abe K, Archer DB, Bermejo C, Bennett J, Bowyer P, Chen D, Collins M, Couless R, Davies R, Dyer PS, Farman M, Fedorova N, Feldbyum TV, Fischer R, Fosker N, Fraser A, Garcia JL, Garcia MJ, Goble A, Goldman GH, Gomi K, Griffith-Jones S, William RG, Haas B, Haas H, Harris D, Horiiuchi H, Huang J, Humphray S, Jiménez J, Keller N, Khouri H, Kitamoto K, Kobayashi T, Konzack S, Kulkarni R, Kumagai T, Lafton A, Latge JP, Li W, Lord A, Lu C, Majoros WH, May GS, Miller BL, Mohamoud Y, Molina M, Monod M, Mouyna I, Mulligan S, Murphy L, O’Neil S, Paulsen I, Penalva11 MA, Pertz A, Price C, Pritchard BL, Quail MA, Rabbinowitz E, Rawlins N, Rajandream MA, Reichard U, Renault H, Robson GD, Rodriguez de Cordoba S, Rodriguez-Pena JM, Ronning CM, Rutter S, Salzberg SL, Sanchez M, Sanchez-Ferrero JC, Saunders D, Seeger K, Squares R, Squares S, Takeuchi M, Tekaia F, Turner G, Vazquez de Aldana CR, Weidman J, White O, Woodward J, Yu JH, Fraser C, Galagan JE, Asai K, Machida M, Hall N, Barrell B, Denning DW. 2005. Genomic sequence of the pathogenic and allergenic filamentous fungus Aspergillus fumigatus. Nature 438: 1151–1156.

Peltola H, Soderlund h, Ukkonen E. 1984. SEQAIQ: A DNA sequence assembling program based on a mathematical model. Nucleic Acids Res. 12 (1 Pt 1): 307–321.

Sanger F, Coulson AR. 1975. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J Mol Biol. 94: 441–448.

Sanger F, Nicklen S, Coulson AR. 1977. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA 74: 5463–5467.

Schwartz DC, Li X, Hernandez LI, Ramnarain SP, Huff EJ, Wang YK. 1993. Ordered restriction maps of Saccharomyces cerevisiae chromosomes constructed by optical mapping. Science 262: 110–114.

Sonntag EM, Durbin R. 1995. A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. Gene 167: GC1–10.

Stalker J, Gibbins B, Meidl P, Smith J, Spooner W, Hotz HR, Cox AV. 2004. The Ensembl Web site: mechanics of a genome browser. Genome Res. 14: 951–955.

Stanke M, Steinkamp R, Waack S, Morgenstern B. 2004. AUGUSTUS: a web server for gene finding in eukaryotes. Nucleic Acids Res 32(Web Server issue): W309–312.

Sutton G, White O, Adams MD, Kerlavage AR. 1995. TIGR assembler: A new tool for assembling large shotgun sequencing projects. Genome Sci Technol. 1: 9–19.

Tanner S, Shen Z, Ng J, Florea L, Guigó H, Briggs SP, Bafna V. 2007. Improving gene annotation using peptide mass spectrometry. Genome Res. 17: 231–239.

Tenney AE, Brown RH, Vaske C, Lodge JC, Doering TL, Brent MR. 2004. Gene prediction and verification in a compact genome with numerous small introns. Genome Res. 14: 2330–2335.

Wang Z, Gerstein M, Snyder M. 2009. RNA-Seq: a revolutionary tool for transcriptomics. Nature Rev Genet. 10: 57–63.

Warren WC, Hillier LW, Graves JAM, Birney E, Ponting CP, Grutzner F, Belov K, Miller W, Clarke L, Chinwalla AT. 2008. Genome analysis of the platypus reveals unique signatures of evolution. Nature 453: 175–183.

Wei C, Lamesch P, Anumagum M, Rosenberg J, Hu P, Vital M, Brent MR. 2005. Closing in on the C. elegans ORFeome by cloning TWINSCAN predictions. Genome Res. 15: 577–582.

Wolfsberg TG. 2007. Using the NCBI Map Viewer to browse genomic sequence data. Current Protocols in Bioinformatics [internet]. Available from: http://www.ncbi.nlm.nih.gov/pubmed/18428781.

Wortman JR, Gilsenan JM, Joardar V, Deegan J, Clutterbuck J, Andersen MR, Archer D, Bencina M, Braus G, Coutinho P, von Döhrn H, Doenan J, Driessen AJ, Durék P, Espeso E, Fekete E, Filippi M, Estrada CG, Geyens S, Goldman D, Groot PW, Hansen K, Harris SD, Heinemekk T, Helmaeckert K, Herrissat B, Hofmann G, Homan T, Horio T, Horuiuchi H, James S, Jones M, Karaffa L, Karanyi Z, Kato M, Keller N, Kelly DE, Kiel JAKW, Kim JM, van der Klei JJ, Kliis FM, Kovalchuk A, Kraševac N, Kubicek CP, Liu B, MacCabe A, Meyer V, Mirabito P, Miskei M, Mos M, Mullins J, Nelson DR, Nielsen J, Oakley BR, Osmani SA, Pakula T, Paszewski A, Paulsen I, Pilsyk S, Pócsi I, Punt PJ, Ram AFJ, Ren Q, Robellet X, Robson G, Seiboth B, van Solingen P, Specht T, Sun J, Taheri-Talesh N, Takeshita N, Ussery D, van Kuyk PA, Visser H, van de Vondervoort PJI, de Vries RP, Walton J, Xiang Y, Xiong Y, Zeng AP, Brandt BW, Cornell MJ, van den Honde CAMJJ, Visser H, Oliver SJ, Turner G. 2009. The 2008 update of the Aspergillus nidulans genome annotation: a community effort. Fungal Genet Biol. 46(Suppl 1): S2–13.

Wu TD, Watanabe CK. 2005). GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics 21: 1859–1875.

Xia D, Sanderson SJ, Jones AR, Prieto JH, Yates JR, Bromley E, Tomley FM, Lal K, Sinden SE, Brunk BP, Roos DS, Wastling JM. 2008. The proteome of Toxoplasma gondii: integration with the genome provides novel insights into gene expression and annotation. Genome Biol. 9: R116.

Yassour M, Kaplan T, Fraser HB, Levin JZ, Pfiffner J, Adiconis X, Schrot G, Luo S, Khrebtukova I, Ginirke A, Nusbaum C, Thompson DA, Friedman N, Regev A. 2009. Ab initio construction of a eukaryotic transcriptome by massively parallel mRNA sequencing. Proc Natl Acad Sci USA 106: 3264–3269.

Zerbino DR, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 18: 821–829.
## Appendix

### Appendix Table 1. Sequenced fungal species.

| Species                        | Description                                                                 | Division       | Size Mb | No. haploid chromosomes | Sequencing status  |
|-------------------------------|-----------------------------------------------------------------------------|----------------|---------|-------------------------|--------------------|
| Acremonium alcalophilum       | Fungus that produces cellulose-degrading Low-Temperature-Active Enzyme       | Ascomycota     | 40      | In progress             |                    |
| Ajellomyces capsulatus        | Dimorphic fungus (anamorph Histoplasma capsulatum) that causes histoplasmosis | Ascomycota     | 28      | 7                       | Draft assembly     |
| Ajellomyces dermatitidis      | Dimorphic fungus (aka Blastomyces dermatitidis) that causes blastomycosis | Ascomycota     | 75      | In progress             | Draft assembly     |
| Alternaria brassicicola       | Plant pathogenic fungus                                                    | Ascomycota     | 29      | 9                       | Draft assembly     |
| Arthroderma gypseum           | Dermatophyte fungus (anamorph: Microsporum gypseum) that causes skin infections in humans | Ascomycota     | 23      | 5                       | Draft assembly     |
| Ascosphaera apis              | Pathogenic fungus that causes the honey bee larval disease chalkbrood      | Ascomycota     | 24      | In progress             | Draft assembly     |
| Aspergillus carbonarius       | Fungus that produces ochratoxin A (OTA)                                     | Ascomycota     | 35      | 8                       | Complete           |
| Aspergillus clavatus          | Rarely pathogenic fungus that is a close relative of two major pathogenic species | Ascomycota     | 36      | 8                       | Complete           |
| Aspergillus flavus            | Filamentous fungus that produces aflatoxin and causes aspergillosis       | Ascomycota     | 30      | 8                       | Complete           |
| Aspergillus fumigatus         | Filamentous fungus that causes more infections worldwide than any other mold | Ascomycota     | 37      | 8                       | Complete           |
| Aspergillus niger             | Filamentous fungus used in biotechnology                                    | Ascomycota     | 37      | 8                       | Complete           |
| Aspergillus oryzae            | Filamentous fungus used in the production of fermented foods and beverages | Ascomycota     | 37      | 8                       | Complete           |
| Aspergillus parasiticus       | Filamentous fungus important for its production of aflatoxins             | Ascomycota     | 35      | 8                       | Complete           |
| Aspergillus terreus           | Filamentous fungus that produces statins, which are used as cholesterol-lowering drugs | Ascomycota     | 34      | 8                       | In progress         |
| Beauveria bassiana            | Insect fungus (teleomorph: Cordyceps bassiana) that is important for biological control of parasitic insects | Ascomycota     | 100     | In progress             | Draft assembly     |
| Blumeria graminis             | Pathogenic fungus that causes powdery mildew on grasses                    | Ascomycota     | 38      | In progress             | Draft assembly     |
| Botryotinia fuckeliana        | Pathogenic fungus (aka Botrytis cinerea) that causes gray mold rot in plants | Ascomycota     | 16      | 8                       | Complete           |
| Candida albicans              | Yeast that is the most common human fungal pathogen                        | Ascomycota     | 16      | 8                       | Complete           |
| Candida dubliniensis          | Clinically important yeast-like fungus of the Candida genus               | Ascomycota     | 16      | 8                       | In progress         |
| Candida glabrata              | Pathogenic ascomycete yeast that is the second most frequent causative agent of candidiasis in humans. | Ascomycota     | 12      | 13                      | Complete           |
| Candida neerlandica           | A yeast isolated from pressed yeast cake and from the digestive tracts of beetles. | Ascomycota     | 24      | 14                      | Draft assembly     |
| Candida parapsilosis          | Clinically important yeast-like fungus of the genus Candida               | Ascomycota     | 15      | 6                       | Draft assembly     |
| Candida tropicalis            | Pathogenic hemiascomycete yeast used in industry and sequenced for a comparative genomics study | Ascomycota     | 36      | In progress             | Draft assembly     |
| Chaetomium globosum           | Filamentous fungus that can infect humans and produce mycotoxins in buildings | Ascomycota     | 16      | In progress             | Draft assembly     |
| Clavispora lusitaniae         | Pathogenic yeast (aka Candida lusitaniae) distantly related to Candida albicans and used in the laboratory | Ascomycota     | 29      | In progress             | Draft assembly     |
| Coccidioides immitis          | Pathogenic fungus that is morphologically indistinguishable from Coccidioides posadasii; both cause the disease coccidioidomyicosis | Ascomycota     | 29      | 4                       | Draft assembly     |
| Coccidioides posadasii        | A dimorphic ascomycete causing coccidioidomyicosis and morphologically indistinguishable from Coccidioides immitis. | Ascomycota     | 24      | 9                       | Draft assembly     |

(Continued)
Appendix Table 1. (Continued).

| Species                        | Description                                                                                                                                                                                                 | Division    | Size Mb | No. haploid chromosomes | Sequencing status       |
|--------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------|---------|-------------------------|-------------------------|
| Debaryomyces hansenii          | Halotolerant and the dominant yeast species found in the production of surface-ripened cheese.                                                                                                                    | Ascomycota  | 12      | 7                       | Complete                |
| Emericella nidulans            | Filamentous fungus (aka Aspergillus nidulans) that is one of the most extensively studied organisms in genetics and biochemistry.                                                                                | Ascomycota  | 31      | 8                       | Complete                |
| Eremothecium gossypii          | A filamentous plant pathogen (aka Ashbya gossypii) closely related to Saccharomyces cerevisiae and used in industry for the production of Vitamin B2.                                                            | Ascomycota  | 9       | 7                       | Complete                |
| Fusarium oxysporum             | Pathogenic fungus that infects humans and plants                                                                                                                                                             | Ascomycota  | 15      | In progress             | EST only                |
| Fusarium sporotrichioides      | Fungus that produces toxic trichotheccenes, such as T-2 toxin.                                                                                                                                               | Ascomycota  | 46      | 12                      | Complete                |
| Gibberella circinata           | The causal agent of pitch canker in pine trees.                                                                                                                                                              | Ascomycota  | 57      | 13                      | Draft assembly          |
| Gibberella moniliformis        | Pathogenic fungus (aka Fusarium verticillioides) that produces fumonisin mycotoxins and causes kernel and ear rot of maize.                                                                                   | Ascomycota  | 32      | In progress             | Draft assembly          |
| Gibberella zeae                | Pathogenic fungus (anamorph Fusarium graminearum) that causes head blight (scab) in wheat and barley.                                                                                                        | Ascomycota  | 33      | 7                       | Draft assembly          |
| Glomerella graminicola         | Pathogenic fungus (aka Colletotrichum graminicola) that causes anthracnose disease of maize.                                                                                                                   | Ascomycota  | 18      | 16                      | In progress             |
| Glomerella clavigera           | Pine pathogenic fungus (aka Ophiostoma clavigerum) associated with Mountain Pine Beetle.                                                                                                                      | Ascomycota  | 10      | 6                       | Complete                |
| Hypocrea jecorina              | Filamentous fungus (anamorph: Trichoderma reesei; teleomorph: Hypocrea jecorina) important to industry for its cellulase production.                                                                      | Ascomycota  | 14      | 10                      | In progress             |
| Hypocrea virens                | Filamentous fungus (anamorph: Trichoderma virens; teleomorph: Hypocrea virens) that is a mycoparasite of Fusarium graminearum.                                                                                  | Ascomycota  | 12      | 8                       | Draft assembly          |
| Kazachstania exigua            | Saccharomyces sensu lato yeast (known as Saccharomyces exigus or Kazachstania exigua) sequenced for a comparative genomics study.                                                                                | Ascomycota  | 10      | 6                       | Complete                |
| Kluyveromyces lactis           | Petite-negative hemiascomycete yeast used for genetic studies and industrial applications.                                                                                                                   | Ascomycota  | 10      | 6                       | Complete                |
| Kluyveromyces marxianus        | Polymorphic hemiascomycete yeast sequenced for a comparative genomics study.                                                                                                                                  | Ascomycota  | 12      | 8                       | Draft assembly          |
| Lacazia lboi                   | Pathogenic fungus (aka Paracoccidioides lboi, Blastomyces lboi, and Loboa lboi) that causes chronic infection of the skin in humans and dolphins.                                                              | Ascomycota  | 10      | 8                       | In progress             |
| Lachancea kluveri              | Petite-negative budding yeast (known as Saccharomyces kluveri or Lachancea kluveri) sequenced for comparative genomics studies.                                                                                | Ascomycota  | 10      | 8                       | In progress             |
| Lachancea thermotolerans       | Hemiascomycete yeast (anamorph: Kluyveromyces thermotolerans) sequenced for a comparative genomics study.                                                                                                     | Ascomycota  | 11      | 8                       | Draft assembly          |
| Lachancea waltii               | Yeast that is related to Saccharomyces cerevisiae but is relatively uncharacterized.                                                                                                                         | Ascomycota  | 10      | 8                       | In progress             |
| Species                              | Description                                                                 | Division       | Size Mb | No. haploid chromosomes | Sequencing status   |
|--------------------------------------|-----------------------------------------------------------------------------|----------------|---------|-------------------------|---------------------|
| **Lodderomyces elongisporus**        | Yeast that forms ascospores and is the closest sexual relative of *Candida albicans* | Ascomycota     | 16      |                         | Draft assembly      |
| **Magnaporthe grisea**               | Fungus that causes rice blast disease                                         | Ascomycota     | 40      | 7                       | Complete            |
| **Magnaporthe poae**                 | A turfgrass pathogen that causes summer patch of Poa species, creeping bentgrass, and fine-leaved fescues. | Ascomycota     | 23      |                         | In progress         |
| **Microsorum canis**                 | Dermatophyte fungus that causes tinea capitis                                 | Ascomycota     |          |                         | Draft assembly      |
| **Myeliophthora thermophila**        | Fungus that produces thermostable enzymes                                     | Ascomycota     |          |                         | In progress         |
| **Mycosphaerella fijiensis**         | Pathogenic fungus that causes black leaf streak disease in bananas           | Ascomycota     | 68      |                         | In progress         |
| **Mycosphaerella graminicola**       | Pathogenic fungus that causes *Septoria tritici* blotch in wheat             | Ascomycota     | 39      |                         | In progress         |
| **Naumovia castellii**               | Saccharomyces sensu lato budding yeast (aka *Naumovia castellii*) sequenced for a comparative genomics study of *Saccharomyces* species | Ascomycota     |          |                         | Draft assembly      |
| **Nectria haematococca**             | Filamentous fungus (aka *Fusarium solani*) that is a plant pathogen           | Ascomycota     | 40      | 17                      | Draft assembly      |
| **Neosartorya fischeri**             | Pathogenic fungus (aka *Aspergillus fischeri*, anamorph *Aspergillus fischerianus*) that can cause aspergillosis and keratitis | Ascomycota     | 35      | 8                       | Complete            |
| **Neurospora crassa**                | Orange bread mold; Filamentous fungus that has been a central model organism in genetics and biochemistry | Ascomycota     | 43      | 7                       | Draft assembly      |
| **Paracoccidioides brasiliensis**    | Pathogenic fungus that causes paracoccidioidomycosis                          | Ascomycota     | 30      | 5                       | Draft assembly      |
| **Penicillium chrysogenum**          | Filamentous fungus (aka *Penicillium notatum*) used in industrial penicillin production | Ascomycota     | 34      | 4                       | Complete            |
| **Penicillium marneffei**            | Pathogenic dimorphic fungus that causes penicilliosis                         | Ascomycota     | 26      |                         | Complete            |
| **Phaeosphaeria nodorum**            | Fungus (anamorph *Stagonospora nodorum*) that causes leaf and glume blotch disease in wheat | Ascomycota     | 37      |                         | Draft assembly      |
| **Pichia angusta**                   | Methylotrrophic hemiascomycete yeast used in research and industry, and sequenced for a comparative genomics study | Ascomycota     | 9       | 6                       | In progress         |
| **Pichia farinosa**                  | A halotolerant and osmotolerant ascomycete yeast.                             | Ascomycota     | 13      | 7                       | In progress         |
| **Pichia guilliermondii**            | Nonpathogenic yeast (anamorph *Candida guilliermondii*) closely related to pathogenic *Candida* yeasts | Ascomycota     | 12      |                         | Draft assembly      |
| **Pichia pastoris**                  | Yeast that widely used in production of recombinant proteins                  | Ascomycota     | 9       | 4                       | Complete            |
| **Pichia stipitis**                  | Yeast that has a high capacity for xylose fermentation                        | Ascomycota     | 15      | 8                       | Complete            |
| **Pneumocystis carinii**             | Pathogenic fungus that causes pneumonia in rats                               | Ascomycota     | 8       | 17                      | In progress         |
| **Pneumocystis jirovecii**           | Pathogenic fungus (aka *Pneumocystis carinii* f. sp. *hominis*) that causes pneumonia in immunocompromised people | Ascomycota     | 7       |                         | In progress         |
| **Pneumocystis marina**              | Pathogenic fungus (aka *Pneumocystis carinii* f. sp. *muris*) that causes pneumonia in mice | Ascomycota     | 6       |                         | In progress         |
| **Podospora anserina**               | Filamentous fungus used as a model organism for studying several biological processes | Ascomycota     | 34      | 7                       | In progress         |
Appendix Table 1. (Continued).

| Species                        | Description                                                                                       | Division     | Size Mb | No. haploid chromosomes | Sequencing status |
|--------------------------------|--------------------------------------------------------------------------------------------------|--------------|---------|-------------------------|-------------------|
| Pyrenophora tritici-repentis    | Pathogenic fungus that causes tan spot of wheat                                                  | Ascomycota   | 37      | 11                      | Draft assembly    |
| Saccharomyces bayanus           | Saccharomyces sensu stricto budding yeast used in winemaking and sequenced for comparative genomics studies | Ascomycota   | 11      | 16                      | Draft assembly    |
| Saccharomyces cerevisiae        | Baker’s yeast, brewer’s yeast; Budding yeast that is a major model organism and of great industrial importance | Ascomycota   | 12      | 16                      | Complete          |
| Saccharomyces kudriavzevii      | Saccharomyces sensu stricto budding yeast sequenced for a comparative genomics study of Saccharomyces species | Ascomycota   | 16      | 16                      | Draft assembly    |
| Saccharomyces mikatae           | Saccharomyces sensu stricto budding yeast sequenced for comparative genomics studies of Saccharomyces species | Ascomycota   | 12      | 16                      | Draft assembly    |
| Saccharomyces paradoxus         | Saccharomyces sensu stricto budding yeast used in winemaking and sequenced for a comparative genomics study of Saccharomyces species | Ascomycota   | 11      | 16                      | Draft assembly    |
| Saccharomyces pastorianus       | Lager brewing yeast                                                                             | Ascomycota   |         |                         | Draft assembly    |
| Saccharomyces servazzii         | Saccharomyces sensu latu yeast sequenced for a comparative genomics study                        | Ascomycota   | 12      | 12                      | In progress       |
| Saccharomycescodes              | Budding yeast that produces secondary fermentation products                                      | Ascomycota   |         |                         | In progress       |
| Schizosaccharomyces lundigii    | Dimorphic fission yeast being sequenced for a comparative genomics study                         | Ascomycota   |         |                         | Draft assembly    |
| Schizosaccharomyces japonicus   | Fission yeast being sequenced for a comparative genomics study                                  | Ascomycota   | 14      | 11                      | Draft assembly    |
| Schizosaccharomyces octosporus  | Fission yeast that is an excellent model organism                                                  | Ascomycota   | 14      | 3                       | Complete          |
| Schizosaccharomyces pombe       | Fungal plant pathogen that has the broadest known host range                                      | Ascomycota   | 38      |                         | Draft assembly    |
| Sordaria macrospora             | Filamentous ascomycete fungus that is used as a model organism for fungal cell and developmental biology | Ascomycota   | 39      | 7                       | In progress       |
| Talaromyces stipitatus          | Non-pathogenic filamentous fungus closely related to the pathogenic fungus Penicillium marneffei | Ascomycota   |         |                         | Complete          |
| Thielavia terrestris            | Soil-borne thermophilic fungus                                                                    | Ascomycota   |         |                         | In progress       |
| Trichoderma atroviride          | Filamentous fungus that is a mycoparasite                                                       | Ascomycota   | 40      |                         | Draft assembly    |
| Trichophyton equinum            | Dermatophyte fungus that causes ringworm in horses                                               | Ascomycota   | 24      |                         | Draft assembly    |
| Trichophyton rubrum             | Dermatophyte fungus that is the most frequent cause of fungal skin infections in humans         | Ascomycota   | 22      | 5                       | Draft assembly    |
| Trichophyton tonsurans          | Dermatophyte fungus that causes scalp infections                                                | Ascomycota   | 22      |                         | Draft assembly    |
| Trichophyton verrucosum         | Dermatophyte fungus causing ringworm in cattle                                                   | Ascomycota   |         |                         | In progress       |
| Uncinocarpus reesii             | Non-pathogenic fungus closely related to the pathogenic Coccidioides species                    | Ascomycota   | 30      |                         | Draft assembly    |
| Vandervaltizyma polyporital     | Yeast (aka Kluyveromycetes polysporus) that represents the post-WGD lineage most divergent from Saccharomyces cerevisiae | Ascomycota   |         |                         | Draft assembly    |
| Verticillium albo-atrum          | A soil-borne plant pathogen                                                                       | Ascomycota   | 30      |                         | Draft assembly    |
| Verticillium dahliae             | Pathogenic fungus that causes Verticillium wilt disease in plants                                | Ascomycota   | 28      | 7                       | Draft assembly    |
| Yarrowia lipolytica              | Alkane-using yeast that shares several properties with filamentous fungi                         | Ascomycota   | 20      | 6                       | Complete          |
| Species                        | Description                                                                 | Division         | Size Mb | No. haploid chromosomes | Sequencing status |
|-------------------------------|------------------------------------------------------------------------------|------------------|---------|--------------------------|-------------------|
| Zygosaccharomyces rouxii      | Halotolerant and osmotolerant yeast that causes food spoilage; sequenced for a comparative genomics study | Ascomycota       | 12      | 7                        | In progress       |
| Mortierella verticillata      | Zygomycete fungus that causes zygomycosis in animals                        | Zygomycota       |         |                          | In progress       |
| Rhizopus oryzae               | Pathogenic filamentous fungus that is the most common cause of mucormycosis | Zygomycota       | 40      |                          | Draft assembly    |
| Coprinopsis cinerea           | Multicellular basidiomycete fungus (aka Coprinus cinereus) that is an excellent model for studying development and regulation | Basidiomycetes   | 37      | 13                       | Draft assembly    |
| Filobasidiella bacillispora   | Fungus (teleomorph: Filobasidiella bacillispora, anamorph: Cryptococcus bacillispora) which is presented as pathogenic and environmental varieties | Basidiomycota    | 20      | 14                       | Draft assembly    |
| Filobasidiella neoformans     | Pathogenic fungus (teleomorph: Filobasidiella neoformans, anamorph: Cryptococcus neoformans) that causes cryptococcosis | Basidiomycota    | 19      | 14                       | Complete          |
| Laccaria bicolor              | Mycorrhizal fungus (common name: bicolored deceiver) that symbiotically associates with Poplar trees | Basidiomycota    | 65      |                          | Complete          |
| Lentinula edodes              | Fungus commonly known as the shiitake mushroom, it is one of the most consumed mushrooms | Basidiomycota    | 33      | 8                        | In progress       |
| Malassezia globosa            | Fungus associated with some skin disorders, including pityriasis versicolor | Basidiomycota    | 9       |                          | Draft assembly    |
| Malassezia restricta          | Fungus associated with some skin disorders, including seborrheic dermatitis and dandruff | Basidiomycota    |         |                          | Draft assembly    |
| Moniliophthora perniciosa     | Pathogenic fungus (aka Crinipellis perniciosais and Marasmius perniciosus) that causes Witches’ broom disease (WBD) of cacao | Basidiomycota    | 39      | 7                        | Draft assembly    |
| Phakopsora meibomiae          | Fungus that causes soybean rust                                              | Basidiomycota    | 50      |                          | In progress       |
| Phakopsora pachyphiza         | Virulent fungus that causes soybean rust                                     | Basidiomycota    | 30      | 10                       | Draft assembly    |
| Phanerochaete chrysosporium   | Most intensively studied white-rot fungus                                     | Basidiomycota    | 10      |                          | Draft assembly    |
| Postia placenta               | Fungus that causes brown rot of wood                                          | Basidiomycota    | 33      |                          | Complete          |
| Puccinia graminis             | Pathogenic fungus that causes stem rust                                       | Basidiomycota    | 81      | 18                       | Draft assembly    |
| Puccinia triticina            | Fungus that causes wheat leaf rust                                            | Basidiomycota    | 100     |                          | Draft assembly    |
| Rhizoctonia solani            | Soil fungus that causes disease on many agricultural crops worldwide including eggplant, pepper, potato, and tomato. | Basidiomycota    | 80      |                          | In progress       |
| Rhodosporidium babjevae       | Carotenoid-producing yeast                                                   | Basidiomycota    | 20      | 14                       | In progress       |
| Schizopyllum commune          | Widespread, wood-decaying fungus, also an opportunistic human pathogen       | Basidiomycota    | 38      | 14                       | Draft assembly    |
| Serpula lacrymans var. lacrymans | Fungus that causes dry rot in wood building constructions                      | Basidiomycota    | 25      |                          | In progress       |
| Tremella mesenterica          | Jelly fungus (aka Witch’s butter) that parasitic on wood decay fungi         | Basidiomycota    | 20      |                          | In progress       |
| Ustilago maydis               | Pathogenic fungus that causes corn smut disease                               | Basidiomycota    | 20      | 23                       | Complete          |
| Allomyces macrognus           | Chytrid fungus that is well-characterized                                    | Blastocladiomycota| 30     |                          | In progress       |
| Batrachochytrium dendrobatidis | Chytrid fungus that caused the recent declines of global amphibian populations | Chytridiomycota  | 23      | 20                       | Draft assembly    |
| Spizellomyces punctatus       | Chytrid fungus whose mitochondrial tRNAs are edited                          | Chytridiomycota  | 24      |                          | Draft assembly    |
| Antonospora locustae          | Parasitic fungus (aka Nosema locustae) that is used to prepare a grasshopper suppression bait | Microsporidia    | 2       |                          | In progress       |
| Species                              | Description                                                                 | Division  | Size Mb | No. haploid chromosomes | Sequencing status |
|--------------------------------------|-----------------------------------------------------------------------------|-----------|---------|--------------------------|-------------------|
| *Encephalitozoon cuniculi*           | An obligate intracellular parasite at the base of Kingdom Fungi             | Microsporidia | 2       | 11                        | Complete          |
| *Enterocytozoon bieneusi*            | Microsporidium that is clinically significant                              | Microsporidia | 6       | 3                         | Draft assembly    |
| *Nosema bombycis*                    | The microsporidia (fungus) that causes silkworm pebrine disease            | Microsporidia | 15      | 18                        | In progress       |
| *Nosema ceranae*                     | An obligate intracellular parasitic fungus at the base of the fungal kingdom that infect honey bees. | Microsporidia |          |                           | Draft assembly    |
| *Octosporea bayeri*                  | Large microsporidian fungus that is a parasite of Daphnia magna            | Microsporidia | 24      |                           | Draft assembly    |
| **Average**                          |                                                                             |           | 28      | 10                        |                   |

Note: *The table was generated based on the data obtained from the NCBI Genome Projects database.*