ELSEVIER

Applied Mycology
and Biotechnology
An International Series
Volume 6. Bioinformatics
2006 Elsevier Science B. V. All rights reserved

4

Issues in Comparative Fungal Genomics

Tom Hsiang[1] and David L. Baillie[2]
[1] Department of Environmental Biology, University of Guelph, Guelph, Ontario, N1G 2W1, Canada (thsiang@uoguelph.ca); [2] Department of Molecular Biology and Biochemistry, Simon Fraser University, Burnaby, B.C., V5A 1S6, Canada (baillie@sfu.ca).

Biologists face an overwhelming richness of nucleotide and protein sequence data. By the middle of 2005, there were almost 300 complete genomes that were publicly accessible. Most of these were archeal or bacterial since prokaryotic genomes are much smaller than eukaryotic genomes. Among eukaryotes, fungi, particularly yeasts, have some of the smallest genome sizes and hence represent the highest number of complete or almost complete genomes sequenced. By mid-2005, there were over 43 fungal genomes that were completely or almost completely sequenced and publicly accessible. What are the relationships among fungi and between fungi and other organisms? What type of genes and pathways are required for pathogenicity and other fungal lifestyles? Researchers are addressing these types of questions with data from high-throughput genomic sequencing. This review examines some recent uses of fungal genomic data in comparative genome analyses. Comparative genomics can facilitate research into the following areas: evolution, phylogenetics, targeted drugs, gene discovery, and gene function. Each of these is discussed as well as the availability and ownership of the genomic data, and the concepts of homology (homologs, orthologs, paralogs) and similarity.

1. INTRODUCTION

By the middle of 2005, there were almost 300 complete genomes that were publicly accessible (http://www.genomesonline.org). Most of these (87%) were archeal or bacterial since prokaryotic genomes range in size from 1 to 5 Mb (Fraser et al. 2000), and are much smaller than eukaryotic genomes, which range in size from 10 Mb to over 3 Gb. Among eukaryotes, fungi, particularly yeasts, have some of the smallest genome sizes (10 to 50 Mb) and hence represent the highest number of complete or almost complete genomes sequenced. By mid-2005, there were over 43 fungal genomes that were completely or almost completely sequenced and publicly accessible (Table 1). Most of these were released since 2003 (84%), but many of them (56%) are considered "posted" but not "published" (Hyman 2001).

Corresponding author: T. Hsiang

In addition to publicly accessible genomes, there are privately-held complete or almost complete fungal genomic data, including *Cochliobolus heterostrophus* and *Gibberella fujikuroi* by Syngenta Biotechnology at the Research Triangle Park, NC (Turgeon et al. 2002), and *Aspergillus niger* sequenced by Gene Alliance (an alliance of five German Companies) for DSM Food Specialties (Heerlen, The Netherlands).

In 2000, the Fungal Genome Initiative (FGI) was formed to discuss and prioritize fungal genome sequencing. The FGI is a partnership between the fungal research community and the Broad Institute (which evolved from the Whitehead Institute/MIT Center for Genome Research in 2004). In February 2002, the FGI released the First White Paper on fungal species targeted for sequencing. Of the 15 fungi selected, the National Human Genome Research Institute in the U.S.A. agreed to fund the costs of sequencing seven, which have been completed or are almost completed. In June 2003, the FGI released the Second White Paper which contains a list of 44 fungal sequencing targets, with an emphasis on 10 major clusters of related species (*Penicillium, Aspergillus*, *Histoplasma*, *Coccidioides*, *Fusarium*, *Neurospora*, *Candida*, *Schizosaccharomyces*, *Cryptococcus,* and *Puccinia*). In July 2004, the FGI released the Third White paper which contains a list of four more target fungal species: *Schizosaccharomyces octosporus*, *Schizosaccharomyces japonicus*, *Trichophyton rubrum* and *Batrachochytrium dendrobatidis*. Copies of the White Papers, and more details on the status of these projects can be found at http://www.broad.mit.edu/annotation/fungi/fgi/history.html.

Other sequencing centers which have been responsible for release of fungal genomes include the U.S. Department of Energy Joint Genome Institute (http://www.jgi.doe.gov), The Wellcome Trust Sanger Institute (http:// www.sanger.ac.uk), The Institute for Genomic Research (http://www.tigr.org), The Stanford Genome Technology Center (http://www-sequence.stanford.edu), The Génolevures Consortium (http://cbi.labri.fr/Genolevures), Genoscope (http:// www.genoscope.cns.fr), The University of Paris (http://www.igmors.u-psud.fr), and Washington University (http://genome.wustl.edu). Funding for these projects has usually been obtained from government sources.

Recent reviews on fungal genomics have concentrated on food industry applications (Hofmann et al. 2003), pathogenicity (Yoder and Turgeon 2001; Lorenz 2002; Mitchell et al. 2003; Tunlid and Talbot 2002; Bos et al. 2003), antifungal drug discovery (Firon and d'Enfert 2002; Jiang et al. 2002; Parkinson 2002), uncovering human genes with fungal homologs (Zeng et al. 2001), yeast comparative genomics (Piskur and Langkjaer 2004, Liti and Louis 2005), and fungal genomics from an agricultural perspective (Yarden et al. 2003). Bennett and Arnold (2001) published an excellent broad overview of fungal genomics. There is also a recent review of fungal genomics targeted toward a general audience (Thacker 2003). The current review has evolved from a previous one (Hsiang and Baillie 2004), and the purpose is to provide an update on developments in comparative fungal genomics. Comparative genomics can facilitate research into phylogenetics, targeted drugs, gene discovery, and gene function. Each of these aspects is discussed in the following sections, beginning with the availability and ownership of the genomic data, as well as the concepts of homology and similarity.

Table 1. Alphabetical listing of fungal genomes, showing year of first release, source, size, and current version. Information for this table was compiled from web searches, http://www.genomesonline.org (Bernal et al., 2001) and www-genome.wi.mit.edu/annotation/fungi/fgi/status.html.

First release	Species and strain	Genome source and publication[1]	Size[2]	File date and version[3]
2004	*Ashbya gossypii* ATCC 10895	Syngenta AG & Basel University (Dietrich et al. 2004) GenBank NC_005782 to 88	9 Mb	2004.3.4
2001	*Aspergillus fumigatus* AF293	TIGR (unpublished) GenBank NC_007194 to 201	29 Mb	2004.3.17
2003	*Aspergillus nidulans* FGSC-A4	Broad Institute (unpublished) GenBank AACD01000000	31 Mb	2003.6.20 Release 3
2005	*Botrytis cinerea* B05.10	Syngenta and Broad Institute (unpublished) http://www.broad.mit.edu/annotation/fgi/	30 Mb	2005.4.26
2002	*Candida albicans* SC5314	Stanford Genome Tech. Center (Tzung et al. 2001) http://www-sequence.stanford.edu/group/candida	16 Mb	2002.5.24 Assembly 19
2004	*Candida glabrata* CBS 138	Génolevures (Dujon et al. 2004). GenBank NC_005967 to NC_006036	12 Mb	2004.7.1
2004	*Candida guilliermondii* ATCC6260	Broad Institute (unpublished) GenBank AAFM01000000	12 Mb	2004.12.28 Assembly 1
2004	*Candida lusitaniae* ATCC 42720	Broad Institute (unpublished). GenBank AAFT01000000	16 Mb	2004.9.30 Assembly 1
2004	*Candida tropicalis* MYA-3404	Broad Institute (unpublished) GenBank AAFN01000000	30 Mb	2004.9.30 Assembly 1
2004	*Chaetomium globosum* CBS 148.51	Broad Institute (unpublished) GenBank: AAFU01000000	36 Mb	2004.12.10 Assembly 1
2004	*Coccidioides immitis* RS	Broad Institute (unpublished) GenBank AAEC01000000	29 Mb	2004.3.11 Assembly 1
2003	*Coprinus cinereus* Okayama 7	Broad Institute (unpublished) Gen Bank AACS01000000	38 Mb	2003.6.1 Assembly 1
2005	*Cryptococcus neoformans* JEC 21	TIGR (Loftus et al. 2005). GenBank NC_006670 to 94	21 Mb	2005-01-13
2003	*Cryptococcus neoformans* serotype A, strain H99	Broad Institute (unpublished) GenBank AACO01000000	20 Mb	2003.5.2 Assembly 1
2004	*Cryptococcus neoformans* Serotype B, strain R265	Broad Institute (unpublished) GenBank AAFP01000000	20 Mb	2004.8.18 Assembly 1
2003	*Cryptococcus neoformans* serotype D, strain B3501A	Stanford Genome Tech. Center (Loftus et al. 2005) www-sequence.stanford.edu/group/C.neoformans	18.5 Mb	2004.06.23 Assembly 040623
2004	*Debaryomyces hansenii* CBS 767	Génolevures (Dujon et al. 2004) GenBank NC_006043 to 49	12 Mb	2004.7.1
2001	*Encephalitozoon cuniculi* GB-M1	Genoscope (Katinka et al. 2001) GenBank NC_003229-42	3 Mb	2001.11.15
2003	*Fusarium graminearum* PH-1	Broad Institute (unpublished) GenBank AACM01000000	40 Mb	2003.10.03 Release 2
2003	*Fusarium verticillioides* 7600	Broad Institute (unpublished) http://www.broad.mit.edu/annotation/fgi/	36 Mb	2003.6.1 Assembly 2
2004	*Kluyveromyces lactis* NRRL Y-1140	Génolevures (Dujon et al. 2004). GenBank NC_006038 to 42	11 Mb	2004.7.1
2002	*Magnaporthe grisea* 70-15	Broad Institute (Dean et al. 2005). GenBank AACU01000000	40 Mb	2002.09.17 Release 2
2003	*Neurospora crassa* OR74A	Broad Institute (Galagan et al. 2003) GenBank AABX01000000	40 Mb	2005.2.17 release 7
2002	*Phanerochaete chrysosporium* RP-78	US DOE Joine Genome Inst. (Martinez et al. 2004) GenBank AADS00000000	36 Mb	2005.2.15 Release 2
2003	*Phytopthora infestans* T30-4	Broad Institute (unpublished). NCBI Trace Repository (http://www.ncbi.nlm.nih.gov/Traces)	237 Mb	2003.12.8
2003	*Phytophthora ramorum* UCD Pr4	US DOE Joint Genome Inst. (unpublished). http://genome.jgi-psf.org/ramorum1	65 Mb	2004.5.27 Release 1
2003	*Phytophthora sojae* P6497	US DOE Joint Genome Inst. (unpublished). http://genome.jgi-psf.org/sojae1	95 Mb	2004.05.27 Release 1
2004	*Podospora anserina*	University of Paris (unpublished).	34 Mb	2004.1.23

	S Mat+	http://podospora.igmors.u-psud.fr		Assembly 1
2004	*Rhizopus oryzae* RA 99-880	Broad Institute (unpublished) GenBank AACW01000000	40 Mb	2004.12.28 Release 1
2003	*Saccharomyces bayanus* MCYC 623	Washington University (Cliften et al. 2003) http://genome.wustl.edu/	12 Mb	2003.03.28
2003	*Saccharomyces castellii* NRRL Y-12630	Washington University (Cliften et al. 2003) http://genome.wustl.edu/	12 Mb	2003.04.07
1997	*Saccharomyces cerevisiae* S288C	SGD, Stanford (Mewes et al. 1997a). GenBank NC_001133 to 48	12 Mb	2005.8.1 Version 5
2004	*Saccharomyces cerevisiae* RM11-1a	Broad Institute (unpublished) GenBank AAEG01000000	12 Mb	2004.9.10 Assembly 1
2003	*Saccharomyces* *kudriavzevii* IFO 1802	Washington University (Cliften et al. 2003) http://www.genetics.wustl.edu	12 Mb	2003.04.07
2003	*Saccharomyces kluyveri* NRRL Y-12651	Washington University (Cliften et al. 2003) http://genome.wustl.edu/	12 Mb	2003.04.07
2003	*Saccharomyces mikatae* IFO 1815	Broad Institute (Kellis et al. 2003). http://www.broad.mit.edu/annotation/fgi/	12 Mb	2003.03.28
2003	*Saccharomyces paradoxus* NRRL Y-17217	Broad Institute (Kellis et al. 2003). http://www.broad.mit.edu/annotation/fgi/	12 Mb	2003.03.28
2002	*Schizosaccharomyces* *pombe* 972h	Sanger Institute (Wood et al. 2002) GenBank NC_003421 to 24	14 Mb	2005.6.20 Version 2
2005	*Sclerotinia sclerotiorium* 1980	Broad Institute (unpublished) http://www.broad.mit.edu/annotation/fgi/	38 Mb	2005.4.13 Assembly 1
2005	*Stagonospora nodorum* SN15	Broad Institute (unpublished) GenBank AAGI00000000	37 Mb	2005.1.17 Release 1
2003	*Trichoderma reesei* QM9414	US DOE Joint Genome Inst. (unpublished) GenBank AAIL01000000	35 Mb	2003.7.18 Release 1
2003	*Ustilago maydis* 521	Broad Institute (unpublished) GenBank AACP01000000	20 Mb	2004.4.1 Release 2
2004	*Yarrowia lipolytica* CLIB99	Génolevures (Dujon et al. 2004) GenBank NC_006067 to 72	21 Mb	2004.7.1

Genome source: in addition to the GenBank accession numbers listed, sequence data from the Broad Institute can also be obtained directly from the FTP site (ftp://ftp.broad.mit.edu/pub/annotation/ fungi/). [2] Size: Estimated size of the genome provided by the source; if no estimate is given, then the data file size is listed. [3] File date and version: the date of the most recent release (year.month.day) is provided as well as the current version. In general, "Release" or "Version" refer to a version of the released sequence data, and "Assembly" refers to the process of joining sequence reads into continguous consensus sequences with the final goal of complete chromosomal sequences.

2. OWNERSHIP OF THE GENOMIC DATA

In 1991, the US National Human Genome Research Institute (NHGRI) and the US Department of Energy developed a data release policy whereby publicly funded sequencing projects should release their data within 6 months. In 1996, the International Human Genome Research Consortium adopted the "Bermuda Principles" with a policy of release of assembly data within 24 hr of generation. In early 2003, NHGRI issued a revision of release policies, reaffirming the 1996 Principles, as well as adding that sequence traces should be in a public trace archive within one week of production, and that whole genome assemblies should be deposited as soon as possible in public databases after the data has passed set quality evaluation criteria.

In essence, the current policies state that publicly funded sequencing projects should release their data without restrictions, while sequence users should provide proper citation of the data source and keep in mind that the sequence generators would like to publish their own analyses of the sequence data (Dennis 2003). The full NHGRI report can be found at http://www.genome.gov/10506537.

Users of publicly available draft sequence data should consider that sequence generators require time from release of the first draft until the full sequence is sufficiently accurate for a full genome publication. For example, for the human genome, the first draft released in 2001 was considered 90% accurate while the completed version from 2003 was considered 99% accurate; however, this last 9% required as much time, effort and expense as the first 90% (International Human Genome Consortium 2004).

Situations have occurred where sequence generators felt that their prerogative to first publish using their data has been pre-empted by other researchers who have analyzed and published on the sequence data before full genome release in a peer-reviewed publication (Bell 2000, Hyman 2001, Marshall 2002). An Editorial in the journal Nature reaffirmed that journals will likely accept good research involving whole-genome analyses without restrictions on authorship, since that is in the best interests of science (Anon 2003). A response to the Editorial in Nature by several prominent bioinformatics researchers (Salzberg et al. 2003) asserts further that publicly funded genome sequence data should be available for use without restrictions.

3. HOMOLOGY

Comparative genomics involves comparisons of sequences to search for homologs. Homology is defined as similarity by descent. It is a qualitative measure rather than quantitative, since sequences are either homologous or not homologous (Doyle and Gaut 2000, Fitch 2000). In much of the molecular biology literature, homology is commonly used as a synonym for similarity, such as in a statement where two genes are said to be 75% homologous. It might be true that 75% of a gene shares common descent with another gene, while the remaining 25% does not, but this is usually not the intended meaning (Doyle and Gaut 2000). Instead of saying, "the two genes are 75% homologous", the statement should read, "the two genes are homologous with 75% similarity".

For quantitative assessments of relationships, the terms identity and similarity are often used, but the usage has been inconsistent. For nucleotides, both identity and similarity are used to refer to the occurrence of the same nucleotide at the same (homologous) position. For protein sequences, identity has the same usage as that for nucleotides, but similarity also includes matches with amino acids of similar triplet coding and similar chemical characteristics. For example, in the commonly used program, CLUSTALX (Jeanmougin et al. 1998), three characters are used in the multiple alignment to show conservation at each site: 'star' indicates positions which have a single, fully conserved residue; 'colon' indicates that one of the following strong groups is fully conserved (STA, NEQK, NHQK, NDEQ, QHRK, MILV, MILF, HY, and FYW); and 'period' indicates that one of the following weaker groups is fully conserved (CSA, ATV, SAG, STNK, STPA, SGND, SNDEQK, NDEQHK, NEQHRK, FVLIM, and HFY).

Various computer programs such as FASTA (Fast Alignment from Pearson 1990) or BLAST (Basic Local Alignment Search Tool from Altschul et al. 1990) have been used to assess the matches between a query sequence and a subject sequence. The output contains identity values for nucleotide or protein comparisons to indicate the percent matches between the query sequence and the matching database sequence. For protein searches, similarity values are also given in the output. For example, a BLASTP analysis (protein query vs. protein database) of the *S. cerevisiae* glucosidase protein YIL099W (549 amino acids) results in the following match with the *N. crassa* glucosidase protein NCU01517: Identities = 145/469 (30%), Positives = 224/469 (47%). This means that the 549 amino acid query sequence has a 469 amino acid portion which matched a sequence in the database, and in 469 amino acid portion, 145 positions were identical, and a further 79 (= 224 – 145) amino acids were similar. In this example, the 30% identical residues and an additional 17% similar residues resulted in 47% sequence similarity as indicated by 'Positives'.

Sequence similarity does not necessarily denote functional similarity. However, the more similar two sequences are, and by implication, the more recent the shared common ancestor, the more likely the retention of similar function (Webber and Ponting 2004). Structural similarity combined with sequence similarity increases the probability of homology (Webber and Ponting 2004) and of functional similarity.

What level of sequence identity or similarity is required to establish homology? For protein sequences, it is often said that 25% to 30% identity across a large segment is enough to call homologous. However, protein sequences may be homologous, yet not share statistically significant similarity (Pearson 1997), and conversely, protein sequences may share significant similarity in particular domains, yet not be truly homologous. A statistic often used as a criterion for homology is the expect value (e-value), which refers to: "the number of hits one can expect to see just by chance when searching a database of a particular size" (www.ncbi.nlm.nih.gov/BLAST/ blast_FAQs.shtml). E-value accounts for both the percent similarity and the length over which the matching occurs, such that very high similarity over only a very short stretch of sequence does not result in a strong e-value. Just as with probability values, lower e-values indicate more significant matching than higher e-values.

In many studies, e-values of 10^{-20} or less have been considered a strong match, while e-values less than 10^{-5} have often been used as the criterion for homology (e.g. Keon et al. 2000; Kruger et al. 2002; Thomas et al. 2001; Thomas et al. 2002). Pearson (1998) states that an e-value of 0.02 could be used for inferring homology with only a 2% chance of a false positive. Some researchers consider e-values less than 10^{-1} to represent biological significance of the match, and have used the e-value as a measure of statistical significance (Pertsemlidis and Fondon 2002). By increasing the e-value in a BLAST analysis, the chances are increased of detecting evolutionarily distant homologs, and some strategies for homologous gene detection involve increasing e-values above 1. However, by increasing the e-value, the chances are also increased of finding false positives. Another consideration is that e-value is directly proportional to the size of the database, such that a match against a local database, which is probably much smaller than the full GenBank database (www.ncbi.nlm.nih.gov/Genbank), will necessarily give a much higher e-value than for the exact same match as found in the GenBank database.

A further complication is that there are several distinct types of homologs: orthologs, paralogs and xenologs (Fitch 2000). Orthology is the relationship between homologous genes found in different organisms where the single ancestral gene was present in the most recent ancestor of the different organisms. Paralogy is the relationship between homologous genes which arose by gene duplication, such as members of a gene family found within the same organism. Xenology describes the relationship between two homologous genes found in different organisms where one gene was derived by lateral gene transfer into another organism. In phylogenetic analyses, if paralogs or xenologs are used in the place of orthologs, a phylogeny could result that is correct for the genes, but not for the organisms (Fitch 2000). The difficulty is that it is sometimes not possible to distinguish between these different types of homologs with the data available (Blattner et al. 1997).

4. COMPARATIVE GENOMICS

Insights into biology and evolution have been gained from studies of comparative genomics (Koonin et al. 2000, Hardison 2003) among bacteria (Fraser et al. 2000; Alekshun 2001; Fraser et al. 2002; Mira et al. 2002; Parkhill et al. 2003; Thomson et al. 2003) or eukaryotes (Rubin et al. 2000, Philip et al. 2005, Philippe et al. 2005) such as phytoplankton (Fuhrman 2003), higher plants (Bennetzen 2002; Hall et al. 2002; Schmidt 2002; Shimamoto and Kyozuka 2002; Pertea and Salzberg 2002; Resier et al. 2002; Kirst et al. 2003; Yu et al. 2005), protozoa (El-Sayed et al. 2005) or animals (Ureta-Vidal et al. 2003; Bofelli et al. 2004; Enard and Paabo 2004; Ptak et al. 2005). Through such comparisons, many secrets of a genome be revealed. For example, the tiger pufferfish (*Fugu rubripes*) was the second vertebrate genome sequenced after humans (Aparicio et al. 2002), and researchers were able to calculate the number of predicted genes conserved in both species or unique to either vertebrate. Genes conserved in these two divergent species after over 400 million years of evolution may have important functions. Although only one-ninth of the size of the human genome, the pufferfish genome has the same number of predicted genes, but with less repetitive DNA and shorter introns (Hedges and Kumar 2002).

The mouse genome was released shortly after that, and while slightly smaller that the human genome, 99% of human genes were found to have a homolog in mouse (Mouse Genome Sequencing Consortium 2002). During comparison of the two genomes, more predicted human genes were uncovered (Mouse Genome Sequencing Consortium 2002), and among the genes exclusive to mouse, many are involved in the sense of smell. Interestingly, among 33 pseudogenes uncovered in the completed sequence of the human genome, 10 may have been involved in olfactory reception (International Human Genome Consortium 2004). These pseudogenes are thought to have recently acquired one or more mutations that caused them to be nonfunctional, and among the 33, five were found to be still functional in chimpanzees (International Human Genome Consortium 2004). Increasing the number of genomes compared also increases the likelihood of detecting conserved sequences which are functional (Bofelli et al. 2004).

Chimpanzees (*Pan troglodytes*) are the closest relative to humans having diverged 5 to 7 million years ago, and the comparative genome analysis was released in late 2005 (Chimpanzee Sequencing and Analysis Consortium 2005). The sequences that can be directly compared between the two genomes are almost 99% identical, but when insertions and deletions are also considered, the similarity is closer to 96%. Compared to other mammals, certain classes of genes were found to be evolving more quickly in humans and chimpanzees including ones related to sound perception, nerve signal transmission, sperm production, and ion transport. More than 50 genes found in the human genome were not found in the chimpanzee genome.

5. PHYLOGENETICS

Complete-genome comparative analyses may also provide more definitive answers on phylogenetic assignments of organisms. Wolf et al. (2001) used different methods of tree construction based on complete genome data from diverse taxa of bacteria, and concluded that there were two primary prokaryotic domains. Datasets from the genomes of seven *Saccharomyces* species consisting of a few or a small number of genes often gave rise to conflicting topologies, whereas combined analysis of 8 or more genes yielded a tree with moderate bootstrap support (all branches over 70%), and a combined analysis of 20 or more genes yielded a single fully resolved tree with over 95% bootstrap support at all branches (Rokas et al. 2003). The implication of this research is that a larger number of genes is required in phylogenetic analyses to give more resolution.

Although full genome comparisons should seem to be able to settle questions in systematics, there are several issues that need consideration and further investigation. Soltis et al. (2004) demonstrated that even when whole genomes are used, if the number of taxa used is low, incorrect phylogenetic reconstructions can be obtained. A major controversy in metazoan systematics is the relationship among vertebrates, arthropods and nematodes. The Coelomata hypothesis argues that arthropods and vertebrates are more closely related because they have a true body cavity, while the Ecdysozoa hypothesis places arthropods as a sister group to nematodes. Using available genomic data, two research groups came to different conclusions, with Philip et al. (2005) supporting the Coelomata hypothesis while Philippe et al. (2005) supported the Ecdysozoa hypothesis. This demonstrates that even when starting with the same or similarly large sets of sequence data, different conclusions can be obtained depending on the analyses.

For species where multiple genomes have been sequenced or studied, researchers have found significant intraspecific variability (Bergthorsson and Ochman 1995). For bacterial species, these differences can as large as 11% for *Salmonella enterica* (McClelland et al. 2001) and 10% for *Pseudomonas aeruginosa* (Spencer et al. 2003). For *P. aeruginosa*, Spencer et al. (2003) concluded that loss, gain or rearrangements of large blocks of DNA were responsible for the significant intraspecific variability. The normal nucleotide substitution rate of 0.5% leads to some divergence between genomes (Spencer et al. 2003), and between any two humans, there is an average of 0.1% difference (Maher 2003). However, humans are different from most other species in having such a narrow genetic range, approaching that of asexually-reproducing species such as *Mycobacterium tuberculosis*, where variation is expected to be low (Kato-Maeda et al. 2001). For fungi, there may also be variable chromosome numbers (Covert 1998) and chromosome lengths (Plummer et al. 1993; Zolan 1995; Plummer et al. 1995; Dewar et al. 1997), in addition to variations in gene sequences between genomes of the same species. These factors could give rise to tremendous differences in genomic sequences, and the use of a particular genome in a phylogenetic assay could lead to biased results if the genome were not representative of the species.

A further consideration is that although genomes are said to be completely sequenced, they still contains gaps and usually exclude multiple copies of ribosomal genes and highly repetitive sequences. For example, the completed version of the human genome still contains 341 gaps that require new technology to complete (International Human Genome Consortium 2004). For the fungal genomes presented in Table 1, the statistics range from 90% to 100% complete. If gene absence or presence is used as an indicator of evolutionary relatedness (Huson and Steel 2004), then the occurrence of gaps and missing information in genomes could have a large effect on the results.

6. UNIQUE TARGET SITES IN PESTS

One of the major purported uses of microbial comparative genomics has been the discovery of antimicrobial target sites. By comparing the genomes of the host and of the pathogen, or of the pathogen and a species similar to the pathogen but non-pathogenic, insights can be gained into target sites for antimicrobial activity including novel fungicide target sites. Hsiang and Baillie (2005) found 17 uniquely fungal genes in their analyses of 14 fungal genomes compared with 2 genomes each of plants, animals and bacteria. They pointed out that seven of these 14 genes were already listed in U.S. patents dealing with antifungal drug discovery. Kessler et al (2002) compared 3000 cDNA sequences from *A. fumigatus* against genomes of three yeasts: *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, and *Candida albicans*. They found that 49% of the clones did not have a match at e-value $\leq 10^{-5}$, and concluded that these could be *A. fumigatus*-specific genes that could be used as potential candidates for novel antifungal targets specific to this fungus. Caution must be taken with this approach to antimicrobial research, since many agricultural pesticides which turned out to have strong non-target effects often affected sites in the host or other non-target organisms which were not homologous to the target site in the pest. For example, the insecticide DDT which affects the nervous system in insects turned out to also cause egg-shell thinning in birds, but the mechanism of action is not the same (Mellanby 1992). Similarly, many human therapeutic drugs turned out to have side-effects which are not related to their target sites. Despite these limitations, a major direction in the use of microbial sequences is to identify specific targets for inhibitor-based drug design (Wu et al. 2003). By searching for gene families that may be important in parasitic or pathogenic activities, and by comparing the presence of these genes in other organisms, specific targets for chemical inhibition may be identified. Many researchers have mentioned this issue as a strength of comparative genomics, and claim that it may be able to pinpoint novel target sites in pathogens which are absent in the host (e.g. Kessler et al. 2002). A more comprehensive method of characterizing pharmacological targets may involve phylogenomics, where the evolutionary analyses of potential target sites are also considered (Searls 2003).

7. GENE PREDICTION AND GENE FUNCTION

While gene sequences are likely to be very accurate, with the level of error estimatable based on the sequencing procedure used, annotation involves interpretation of the sequence and is often subject to error (Parkhill 2002), particularly if the annotation is automated (Nierman et al. 2005). Gene prediction algorithms are based first on finding open reading frames (ORF) larger than a given size (usually 100 aa), which have a start and stop codon in the same reading frame, and then determining whether the coding sequence has properties such as G+C content similar to known coding sequences in that organism (Parkhill 2002). In addition to similarity searches to assign function, there are non-similarity methods such as physically proximity and frequent co-occurrence (Parkhill 2002). Cliften et al. (2001) used comparative sequence analysis to identify conserved functional elements in several *Saccharomyces* genomes to predict genes. Kellis et al. (2003) compared the genomes of four *Saccharomyces* species (*S. cerevisiae, S. paradoxus, S. mikatae* and *S. bayanus*), and found a high degree of synteny across the genomes. By examining regulatory motifs and analyzing conservation of predicted gene sequences, they concluded that the proteome of *S. cerevisiae* could be reduced by approximately 500 predicted genes.

Once gene sequences are identified, how is function determined? Lockhart and Winzeler (2000) claim that "guilt by association" can allow for many groups of sequences to be simultaneously classified, since strong correlations between expression profiles may indicate similar functional assignments. Uetz et al. (2000) applied this concept in their two-hybrid analysis of protein interactions in yeast, and were able to identify interactions between proteins of known and unknown function, and shed light both on the existence of the interactions and on the possible roles of the proteins with undescribed function. Date and Marcotte (2003) extended this by using phylogenetic profiles to analyze pairwise coinheritance of genes within genomes to predict thousands of functional linkages and identify large-scale cellular systems. Nardone et al. (2004) describe how the use of conserved non-coding regulatory regions in cross-species comparisons can give insights into homologous transcriptional regulation.

The annotation of gene functions is a major bottleneck in genomics (Pallen 2002), and is one reason for the delay between genome release and publication (Table 1). Most genes have not yet been characterized. For example, although ~4000 of ~6000 predicted genes in yeast have been annotated (Cherry et al. 1998), it is not known how many of these annotations are accurate. In 2005, 8 years after this first eukaryotic genome was released in 1997, still only 66% of the 6591 open reading frames in *S. cerevisiae* were considered verified and characterized (http:// www.yeastgenome.org/cache/genomeSnapshot.html), while 22% were uncharacterized and 12% were considered dubious.

When analysis of the first draft of the human genome was published (International Human Genome Consortium 2001), they estimated 30,000 to 40,000 protein-coding genes. Before the draft was released, estimates ranged up to 120,000 (Liang et al. 2000), and other estimates based on the draft gave 65,000 to 75,000 transcriptional units (Wright et al. 2001). In 2004, the International Human Genome Consortium published on the completed human genome, and revised the estimate down to 22,287 gene loci with a total of 34,214 transcripts. Imanishi et al. (2004) investigated the function of 19,574 protein-coding human genes that were derived from experimental evidence, and were able to assign 50.1% of them to a functional group.

Predicted genes are often given a functional annotation that is derived from the BLAST hit with the lowest e-value, but this assignment of function makes the assumption that sequence similarity is equivalent to functional similarity, and, as discussed above, this is not always the case. Once an erroneous annotation is provided, it may become propagated throughout different databases and the original evidence may become difficult to track down (Pallen 2002). For example, Bridge et al. (2003) examined over 200 fungal ribosomal RNA sequences from publicly available databases, and concluded that 20% appeared to be misidentified, dubious or chimeric with 38% not linked to traceable material.

Comparative genomics provides a major route for the study of functional genomics. We may discover what is occurring in one organism because the same thing happens in another organism. Since model organisms such as *Saccharomyces cerevisiae* for fungi, *Arabidopsis thaliana* for plants, and *Caenorhabditis elegans* for nematodes, are among the best studied organisms in their respective taxa and have been completely sequenced, determination of gene function in one of these more easily manipulated organisms often gives insight into homologous functions in higher or larger organisms. Rehm (2001) discusses some methods involved in sequence analyses including functional assignment of genes. There are attempts to classify genes from a variety of organisms into functional classes such as GO (Gene Ontology)(Gene Ontology Consortium 2000), COG (Cluster of Orthologous Genes) (Rashidi and Buehler 2000; Tatusov et al. 2000), MIPS (Martinsreid Institute for Protein Sequences) (Mewes et al. 1997b), and InterPro (InterPro Consortium 2001).

For genes without known function, one method to determine function is by gene knockout (Capecchi 1989). Prior to this breakthrough technique, researchers had already developed gene transfer technology in mice in the early 1980's, but they could neither control nor predict where the transgene would be inserted into the genome of the target organism (Pray 2002). Using homologous recombination, Cappechi (1989) demonstrated that the transgene could be precisely aimed at a target site in the genome and the replacement of a specific gene with an inactive or mutated allele would knock out the function of this gene (Pray 2002). Other more recent methods for assessing gene function include RNA interference (RNAi) (Fire et al. 1998) and Targeted Induced Local Lesions in Genomes (TILLING) (Till et al. 2003).

Gene expression technologies are developing rapidly, and RNA detection includes standard procedures such as northern blots, RT-PCR (reverse transcription of RNA followed by PCR), cDNA sequencing, differential display, and more recently derived procedures such as microarray analyses (Lockhart and Winzeler 2000), serial analysis of gene expression

(SAGE, Velculescu et al. 1995) and analyses of expressed sequence tags (ESTs) (Soanes et al. 2002). ESTs are the fastest growing segment in GenBank, and Jongeneel (2001) presents a good overview of searching for genes in EST databases. These technologies for establishing gene function and expression are still developing, but the technologies for genomic sequencing have advanced at a far greater rate, and unexplored or lightly explored sequence data are accumulating exponentially.

8. COMPARATIVE GENOMICS BETWEEN FUNGI AND OTHER ORGANISMS

A genome represents the complete set of genes of an organism. This set includes all the instructions for maintenance, defense, growth and reproduction of the organism, and while a smaller genome is less expensive to maintain, it lacks the genetic flexibility of larger genomes (Fuhrman 2003). With greater complexity and larger genome sizes, the proportion of genes in a genome which can be found in other genomes in publicly available databases decreases. For prokaryotes, ~70% of the genes in any genome may be identified in other organisms, perhaps also reflecting the greater number of prokaryotic genomes available (Braun et al. 2000). For *S. cerevisiae*, which has one of the smallest eukaryotic genomes, more than 60% of the genes have a match in at least one other organism (Braun et al. 2000). However, for more complex eukaryotes such as *Caenorhabditis elegans* or *Arabidopsis thaliana*, the proportion of genes that have a match in other organisms is much smaller (Braun et al. 2000). Zeng et al. (2001) found almost 1000 human proteins with higher similarity to homologs in fungal genomes than in other animals, such as *C. elegans* or *Drosophila melanogaster*, and concluded that functional genomics with human genes should involve yeasts and higher fungi.

A massive comparative study of the genomes of *D. melanogaster, C. elegans,* and *S. cerevisiae* was conducted by over 50 researchers (Rubin et al. 2000) representing a wide array of agencies. They found that the two animal genomes had nonredundant protein sets which were similar in size and twice that of yeast, and that the multidomain proteins and signaling pathways in the animals were more complex than those of yeast. Another massive comparative genomics study (Thomas et al. 2003) compared a large genomic region in 13 vertebrate species including human, other primates, cat, dog, cow, pig, chicken, rodents, and fishes. Their analysis supported the closer phylogenetic relationship of primates to rodents than to the other mammals listed. They identified DNA segments that were conserved across a wide range of species but apparently not coding for any proteins. Non-coding DNA can represent a large part of the genome of an organism, such as 98% of the DNA in *Homo sapiens*, but some of this non-coding DNA actually contains hidden genes that work through RNA (Gibbs 2003). Roy and Gilbert (2005) examined the pattern of intron conservation in eukaryotes using seven fully sequenced genomes. They found that modern introns generally are very old and that 40% of the introns found in animals, plants and fungi date to their common ancestor.

There are also attempts using comparative genomics to distinguish between genes of the pathogen and that of the most in mixed libraries. Hsiang and Goodwin (2003) used the complete genomes of a plant and a fungal pathogen to assess the origin of ESTs from fungal-infected plant tissues. In trials with pure fungal or pure plant sequences, they showed that their method was better able to place the taxonomic origin of the sequences than a comparison with the GenBank NR database, and explained that since so many more plant genes have been investigated than fungal genes, a best match to a plant sequence from GenBank did not necessary ensure that the query sequence was of plant origin. Xu et al. (2003) used a similar method involving computational subtraction with human genome sequences to remove the human component from a cDNA library of virus-infected human tissue (27,840 sequences). They then designed primers for the remaining 32 non-matching sequences, and attempted to amplify these sequences from infected and non-infected tissues. Twenty-two were found to amplify from uninfected tissues, leaving 10 sequences, and all 10 of these sequences were found to match viral sequences (Xu et al. 2003). A major advantage of studying a human disease is that complete genomic data may be available for both the host and the pathogen, while for plant diseases, it is rare to have complete genomic sequences for both the host and pathogen. Furthermore, for fungal plant diseases, both the host and pathogen are eukaryotes and hence their sequences may be more difficult to distinguish, unlike human diseases where the important pathogens are mostly bacterial or viral.

9. FUNGAL COMPARATIVE GENOMICS

Fungal comparative genomics can be used to address many very fundamental questions in biology and evolution. As noted by Goswami and Kistler (2004), comparative genomics can give insights into evolution of gene clusters (Ward et al. 2002) and gene family expansions and extinctions (Kroken et al. 2003), and gene prediction using a reading frame conservation test (Kellis et al. 2003). Comparative analyses will also provide information on gene dispersion and loss, genome rearrangements, the acquisition of species-specific genes, and other mechanisms which should be applicable to eukaryotes in general (Goffeau 2004). Because of the greater number of fungal genomes currently available and soon to become available, comparative genomics with fungi should continue to be at the leading edge of the field of eukaryotic comparative genomics.

The complete genome sequences of particular fungal species also allows a full inventory of genes that might be related to sexual reproduction, particularly in species that are considered to be asexual. For example, the presence of certain types of mating genes in the genomic sequence of *Aspergillus fumigatus* suggested that it is able to mate and undergo meiosis (Paoletti et al. 2005). Similarly, Wong et al. (2003) found through datamining, the presence of genes involved mating and meiosis in the presumed asexual yeast, *Candida glabrata*. Although mating type genes can be found by using degenerate primers (e.g. Hsiang et al. 2003), such attempts have not always proven successful. The availability of complete genomic sequences provides the opportunity to datamine genomes for the presence of genes that might be involved in reproduction.

Yeast comparative genomics continues to be a highly active area of research (Grunfelder and Winzeler 2002, Dujon et al. 2004, Kellis et al. 2004, Piskur and Langkjaer 2004, Rokas and Carroll 2005, Fabre et al. 2005). Among the 43 genomes listed in Table 1, 18 are species of yeasts. Yeasts generally have smaller genome sizes than filamentous fungi, and were among the earliest genomes sequenced. For genera such as *Candida* and *Saccharomyces,* multiple species have been sequenced which allows for evolutionary comparisons within genera and between these two genera which diverged over 100 million years ago (Berbee and Taylor 2001, Heckman et al. 2001).

Other recent studies in fungal comparative genomics include a survey of *Aspergillus* species (Archer and Dyer 2004), since the genomes of four *Aspergillus* species have been sequenced (but not all are publicly available). Tekaia and Latge (2005) compared *A. fumigatus* to other fungal genomes and concluded that based on the presence of certain types of genes and enzymatic machinery, that *A. fumigatus* is a saprophyte and opportunistic invader of humans. Nierman et al. (2005) reviewed the progress on comparative genomics among *Aspergillus* species, and stated that the species are distantly related (compared to congeneric taxa among plants or animals), and that only 50% of each genome can be aligned with the corresponding region of the other genomes.

10. FUNGAL COMPARATIVE GENOMICS – EVOLUTIONARY BIOLOGY

Cliften et al. (2003) compared the genomes of six *Saccharomyces* species to find functional non-protein-coding sequences, such as gene regulatory elements. These are generally difficult to recognize because they are often short, degenerate and can be distant from the genes they control. By finding these "phylogenetic footprints", the authors were able to revise the catalog of yeast predicted genes, and to identify motifs that may be targets of transcriptional regulatory proteins. Schoch et al. (2003) inventoried the kinesin gene families in three filamentous fungi, *Botryotinia fuckeliana*, *Cochliobolus heterostrophus*, and *Gibberella moniliformis*, and compared these to two yeasts, *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*. They found that the filamentous species contained a constant set of 10 kinesins in nine subfamilies while the yeasts had much fewer kinesins. Kellis et al. (2004) compared the genomes of *S. cerevisiae* and *Kluyveromyces waltii* and concluded that *S. cerevisiae* arose from an ancient whole-genome duplication.

Zelter et al. (2004) looked for homologs of yeast calcium signalling machinery in *Neurospora crassa* and *Magnaporthe grisea* in a comparative genomics study. They found a greater number of homologs for various calcium signalling genes in the filamentous fungi than in yeast, and speculated that there was greater complexity in the filamentous forms because of their more complex cellular organization and possibly greater range of external signals in their natural habitats.

Dietrich et al. (2004) compared *S. cerevisiae* to the genome of *Ashbya gossypii*, a filamentous, ascomycetous, plant pathogen with a very small genome size (9.2 Mb). They found, using BLAST and FASTA, that 95% of the *A. gossypii* genes showed homology with *S. cerevisiae* genes, with percent identity values from 19% to 100%. Among *A. gossypii* genes, 90% showed homology and synteny with *S. cerevisiae* genes, 5% showed homology but not synteny, and 5% did not show homology, but were considered to be real genes because of the presence of homologues in other species. Through these comparisons, they found evidence that *S cerevisiae* resulted from a whole genome duplication or fusion of two related species.

Nielsen et al. (2004) examined intron loss and gain in four ascomycete species (*Magnaporthe grisea, Neurospora crassa, Fusarium graminearum,* and *Aspergillus nidulans*). Since the time of their divergence from the most recent comment ancestor over 300 million years ago, there have been up to 250 intron gains and 350 intron losses in each lineage, and the authors suggest that intron gain has been a major driving force in the evolution of fungi.

Fungi are good model organisms for the study of evolutionary biology using comparative genomics. First, the number of fungal genomes that have been sequenced is greater than that for other major eukaryotic taxa. Second, the relatively small and compact fungal genomes facilitate computational analyses. Third, ascomycetous yeast species alone cover the evolutionary range comparable to the entire phylum of chordates (Hedges and Kumar 2003).

11. FUNGAL COMPARATIVE GENOMICS - FUNGAL BIOLOGY

Papp et al. (2003) used genomic sequences of *S. cerevisiae* to search for paralogs (e-value $\leq 10^{-2}$) to identify gene family size. Then they compiled a list of interacting protein pairs which did not belong to the same gene family, and found that out of almost 7000 pairs, over 4300 had the two members with the same-sized gene families. They also found that members of large gene families were rarely involved in complexes, and supported the assertion that dominance is a by-product of physiology and metabolism rather than the result of selection to mask the effects of deleterious mutations (Papp et al. 2003).

Tzung et al. (2001) compared *C. albicans* with *S. cerevisiae* to assess whether genes important for sexual reproduction and meiosis might be present in *C. albicans*. The complete repertoire of genes related to sexual reproduction was not found, leading to the suggestion that *C. albicans* has alternative mechanisms of genetic exchange. Fungi are known to undergo asexual recombination under the parasexual cycle (Pontecorvo 1956), and the presence of homologs to genes involved in vegetative incompatibility suggests that this may be a method by which *C. albicans* generates genetic variation (Tzung et al. 2001).

Wagner (2000) examined the ability of *S. cerevisiae* to compensate for mutations and concluded that interactions among unrelated genes are the major cause of robustness against mutations. Gu et al. (2003) continued this line of research by studying a near complete set of single-gene-deletion mutants of *S. cerevisiae* with functional annotations. They found that for genes with paralogs, there was a greater probability of functional compensation than for singleton genes (Gu et al. 2003). They estimated for *S. cerevisiae*, that of the gene deletions which resulted in no phenotypic change, 25% were because of compensation by duplicate genes, and at least some of the remaining were because of alternative pathways.

Yoder and Turgeon (2001) compared the occurrence of selected protein families in genomes of selected pathogenic and saprophytic fungi, and concluded that the plant pathogens *Cochliobolus sativus, Fusarium graminearum,* and *Botrytis cinerea* have more genes dedicated to secondary metabolism than do saprophytes such as *Neurospora crassa, Ashbya gossypii*, and *S. cerevisiae*. They found that the three plant pathogenic fungi were rich in peptide synthetases and polyketide synthases, some of which are known to be virulence factors (Kroken et al. 2003) , whereas the saprophytes encoded few or none of these proteins. Yarden et al. (2003) contend that searches for differences between plant pathogenic fungi and nonpathogenic ones can be confounded when orthologous genes are present in both types of organisms, but the orthologous pathways may not be; hence, direct comparisons of presence or absence may be an oversimplification.

Gardiner and Howlett (2005) used previously characterized genes involved in sirodesmin biosynthesis in *Leptosphaeria maculans* to uncover a cluster of 12 genes putatively involved in gliotoxin production in *Aspergillus fumigatus*. The gliotoxin-related genes were identified by comparative genomics, since both gliotoxin and sirodesmin are epipolythiodioxopiperazine toxins. Further experimental work quantified gene expression using quantitative RT-PCR, and identified genes that were co-regulated and showed expression of timing correlated with gliotoxin production as measured by HPLC.

12. FUNGAL COMPARATIVE GENOMICS - ESSENTIAL FUNGAL GENES

Braun et al. (2000) conducted a whole genome comparison between *Saccharomyces cerevisiae* and *Neurospora crassa*. They found that *N. crassa*, with its larger genome, has more unique genes than *S. cerevisiae* by making comparisons with the GenBank protein database. The presence of a gene in *N. crassa* that could also be found in other organisms but not in *S. cerevisiae* was interpreted as gene loss from *S. cerevisiae*. They were also able to find genes in *N. crassa* that were not found in any non-fungal species in GenBank, and postulated that these were fungal-specific proteins (Braun et al. 2000).

Firon and d'Enfert (2002) reviewed some of the methods for identifying essential genes in fungal pathogens of humans, including transposon mutagenesis and post-transcriptional gene silencing. They contend that the characterization of genes essential for growth in fungal pathogens is an important step in development of novel antifungal drugs, as well as providing insights into biological diversity of fungi.

Decottignies et al. (2003) used a PCR-based gene deletion procedure on 100 genes of *S. pombe* and found that 17.5% of these deletions were of essential genes. They then compared 450 proteins from two yeasts (*S. cerevisiae* and *S. pombe*) with those of Metazoa, plants and prokaryotes in the GenBank nonredundant protein database, and estimated that 80% of the essential genes of *S. pombe* were shared with other eukaryotes, with half of these genes also found in prokaryotes, while only 10% of essential genes were fungal specific. Similar numbers were found for *S. cerevisiae*, with the criterion for homology at e-value $\leq 10^{-5}$. With a greater number and taxonomic range of fungal genomes being sequenced every year, our ability to uncover genes which are conserved across many fungal taxa will be enhanced. We may then be able to determine which genes are exclusively fungal that help make fungi distinctive from other organisms.

Strobel and Arnold (2004) compared cDNAs from the AIDs-related fungal pathogen *Pneumocystis carinii* to the saprophytes *Schizosaccharomyces pombe* and *Saccharomyces cerevisiae.* They identified 200 sequences shared with these other fungi and considered these to be essential genes. Because the cDNA library was thought to include half of all *P. carnii* genes, they then estimated the essential eukaryotic core to be approximately 400 genes.

Hsiang and Baillie (2005) searched for homologs of *Saccharmoyces cerevisiae* genes among 13 other fungal species. They found that out of the 6355 putative *Saccharomyces cerevisiae* genes, 3340 were present in at least 12 other fungal genomes (at e-value ≤ 10^{-5}). Of these 3340 genes, 938 had homologs in plants, animals and bacteria, while 17 were found to lack homologs in non-fungal species. These 17 core fungal genes did not seem to share peculiarities in GC content, codon usage patterns, or putative functional characteristics, and only one of these was considered to be essential from gene deletion studies.

13. FUNGAL COMPARATIVE GENOMICS - SMALL SCALE STUDIES

Bioinformatic tools are necessary to process the enormous amounts of genomic data that are generated. These tools include gene-matching algorithms, such as BLAST, and processing of output from such programs with computer scripts specifically written for these activities in languages, such as PERL (practical extraction and report language) (Tisdall 2003). As biologists, our goal in genomic studies is to enhance our understanding of the biology of the organisms and generally not just to catalogue the component parts (Lockhart and Winzeler 2000). Analytical tools are available to handle the masses of genetic data to generate results, but making biological interpretations from the results is a daunting task (Lockhart and Winzeler 2000). Most biologists do not consider themselves to be bioinformatics-enabled, but new computer programs should reduce the complexity of bioinformatic tools (Buckingham 2003). These tools are being directed toward the exponentially increasing amounts of genetic data, as well as toward categorizing the ever growing number of publications related to analysis and interpretation of such data (Buckingham 2003). These tools are generally freely available and can be downloaded from many websites on the Internet.

Many articles on comparative genomics studies have been written with a multitude of authors, arising from labs that may have both high-powered molecular biology and computational tools; however there is still a role for smaller research labs in comparative genomics. The fact that the massive computing power available to a super-computing center may be able to process all the data and make the sequence comparisons in one day, a task which may take several months for a smaller research program to conduct, doesn't outweigh the fact that the smaller research programs may come up with important novel ideas for an analysis which haven't been considered by the larger research programs. Although the learning curve can be quite steep for biologists, comparative genomic analyses can be conducted on common desktop computers using Windows, Mac, or Linux operating systems, and the results of these types of analysis can be very rewarding. Furthermore, genomes databases have been set up which allow users to search for homologs, and find current information on the annotation and physical location of loci in particular genomes. The January 2005 supplemental issue (Volume 33 Database Issue) of Nucleic Acids Research was devoted to descriptions of available genomic database resources.

14. CONCLUSION

This article has discussed just a few of the discoveries that are possible using comparative genomics, and certainly many more are possible. We encourage mycologists and plant pathologists to explore the use of the new tools of bioinformatics. After all, biologists do not usually hand over their data to statisticians for analysis and interpretation, but undertake the data analysis with the help of statisticians, since extensive training in biology is required to make many of the important biological interpretations from the results of statistical analyses of biological data. Similarly, with the ever-burgeoning amounts of sequence data, there is plenty for researchers to analyze to bring forth important discoveries of biological significance.

Acknowledgements: We gratefully acknowledge the research support provided by the Natural Sciences and Engineering Research Council of Canada.

REFERENCES

Alekshun MN (2001). Beyond comparison - antibiotics from genome data? Nature Biotech 19:1124-1125.

Altschul SF, Gish W, Miller W, Myers EW and Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403-10.

Anon (2003). Sacrifice for the greater good. Nature 421:875.

Aparicio S, Chapman J, Stupka E, Putnam N, Chia JM, Dehal P, Christoffels A, Rash S, Hoon S, Smit A etc. (2002). Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. Science 297:1301-1310.

Archer DB and Dyer PS (2004). From genomics to post-genomics in Aspergillus. Curr Op Microbiol 7: 499-504.

Bell E (2000) Publication rights for sequence data producers. Science 290:1696-1698.

Bennett JW and Arnold J (2001) Genomics for fungi. In: RJ Howard and NAR Gow, ed.The Mycota VIII: Biology of the fungal cell. Berlin: Springer-Verlag GmbH & Co, pp. 267-297.

Bennetzen J (2002) Opening the door to comparative plant biology. Science 296:60-63.

Berbee ML and JW Taylor (2001) Fungal molecular evolution: gene trees and geologic time. In: DJ McLaughlin, EG McLaughlin and PA Lemke, ed. The Mycota VII: Systematics and Evolution. Berlin: Springer-Verglab GmbH & Co, pp. 229-245.

Bergthorsson U and Ochman H (1995) Heterogeneity of genome sizes among natural isolates of *Escherichia colia*. J Bacteriol 10:5784-5789.

Bernal A, Ear U, Kyrpides N (2001) Genomes OnLine Database (GOLD): a monitor of genome projects world-wide. Nucl Acid Res 29:126-127.

Blattner FR, Plunkett G, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD, Rode CK, Mayhew GF, Gregor J, Davis NW, Kirkpatrick HA, Goeden MA, Rose DJ, Mau B and Shao Y (1997) The complete genome sequence of *Escherichia coli* K-12. Science 277:1453-74.

Bofelli D, Nobrega MA and Rubin EM (2004) Comparative genomics ate the vertebrate extremes. Nat Rev Genet 5:456-465.

Bos JIB, Armstrong M, Whisson SC, Torto TA, Ochwo M, Birch PRJ, and Kamoun S (2003) Intraspecific comparative genomics to identify avirulence genes from *Phytophthora.* New Phytol 159:63-72.

Braun EL, Halpern AL, Nelson MA and Natvig DO (2000) Large-scale comparison of fungal sequence information: mechanisms of innovation in *Neurospora crassa* and gene loss in *Saccharomyces cerevisiae*. Genome Res. 10:416-430.

Bridge PD, Roberts PJ, Spooner BM and Panchal G (2003) on the reliability of published DNA sequences. New Phytol 160:43-48.

Buckingham S (2003) Programmed for success. Nature 425:209-215.

Capecchi MR (1989) Altering the genome by homologous recombination. Science 244:1288-1292.

Cherry M, Adler C, Ball C, Chervitz SA, Dwight SS, Hester ET, Jia Y, Juvik G, Roe T, Schroeder M, Weng S and Botstein D (1998) SGD: *Saccharomyces* Genome Database. Nucl Acid Res. 26:73-80.

Chimpanzee Sequencing and Analysis Consortium (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437:69–87.

Cliften PF, Hillier LW, Fulton L, Graves T, Miner T, Gish WR, Waterston RH Johnston M (2001) Surveying *Saccharomyces* genomes to identify functional elements by comparative DNA sequence analysis. Genome Res 11:1175-1186.

Cliften PF, 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.

Covert SF (1998) Supernumerary chromosomes in filamentous fungi. Curr Genet 33:311-319.

Date SV and Marcotte EM (2003) Discovery of uncharacterized cellular systems by genome-wide analyses of functional linkages. Nat Biotech 21:1055-1062.

Dean RA, Talbot NJ, Ebbole DJ, Farman ML, Mitchell TK, Orbach MJ, Thon M, Kulkarni R, Xu JR, Pan H, Read ND, Lee YH, Carbone I, Brown D, Oh YY, Donofrio N, Jeong JS, Soanes DM, Djonovic S, Kolomiets E, Rehmeyer C, Li W, Harding M, Kim S, Lebrun MH, Bohnert H, Coughlan S, Butler J, Calvo S, Ma LJ, Nicol R, Purcell S, Nusbaum C, Galagan JE and Birren BW. 2005. The genome sequence of the rice blast fungus *Magnaporthe grisea*. Nature 21:980-986.

Decottignies A, Sanchez-Perez I and Nurse P (2003) *Schizosaccharomyces pombe* essential genes: a pilot study. Genome Res 13:399-406.

Dennis C (2003) Draft guidelines ease restrictions on use of genome sequence data. Nature 421:877-878.

Dewar K, Bousquet J, Dufour J and Bernier L 1997. A meiotically reproducible chromosome length polymorphism in the ascomycete fungus *Ophiostoma ulmi* (sensu lato). Mol Gen Genet 255:38-44.

Dietrich FS, Voegeli S, Brachat S, Lerch A, Gates K, Steiner S, Mohr C, Pohlmann R, Luedi P, Choi S, Wing RA, Flavier A, Gaffney TD and Philippsen P (2004) The *Ashbya gossypii* genome as a tool for mapping the ancient *Saccharomyces cerevisiae* genome. Science 304:304-307.

Doyle JJ and Gaut BS (2000) Evolution of genes and taxa: a primer. Plant Mol Biol 42:1-23.

Dujon B, Sherman D, Fischer G, Durrens P, Casaregola S, Lafontaine I, De Montigny J, Marck C, Neuveglise C, Talla E, etc. (2004) Genome evolution in yeasts. Nature 430:35-44.

El-Sayed NM, Myler PJ, Blandin G, Berriman M, Crabtree J, Aggarwal G, Caler E, Renauld H, Worthey EA, Hertz-Fowler C, etc. (2005) Comparative genomics of trypanosomatid parasitic protozoa. Science 309:404-9

Enard W and Pääbo S (2004) Comparative primate genomics. Ann Rev Genomics Hum Genet 5:351-78.

Fabre E, Muller H, Therizols P, Lafontaine I, Dujon B and Fairhead C (2005) Comparative genomics in hemiascomycete yeasts: evolution of sex, silencing and subtelomeres. Mol Biol Evol 22:856-73.

Fire A, Xu S, Montgomery MK, Kostas SA, Driver SE and Mello CC (1998) Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. Nature 391:806-811.

Firon A and d'Enfert C (2002) Identifying essential genes in fungal pathogens of humans. Trends Microbiol 10:456-462.

Fitch WM (2000) Homology, a personal view on some of the problems. Trends Genet 16:227-231.

Fraser CM, Eisen JA and Salzberg SL (2000) Microbial genome sequencing. Nature 406:799-803.

Fraser CM, Eisen JA, Nelson KE, Paulsen IT and Salzberg SL (2002) The value of complete microbial genome sequencing (You get what you pay for). J Bacteriol 184:6403-6405.

Fuhrman J (2003) Genome sequences from the sea. Nature 424:1001-1002.

Galagan JE, Calvo SE, Borkovich KA, Selker EU, Read ND, Jaffe D, FitzHugh W, Ma LJ, Smirnov S, Purcell S, etc. (2003) The genome sequence of the filamentous fungus *Neurospora crassa*. Nature 422:859-868.

Gardiner DM and Howlett BJ (2005) Bioinformatic and expression analysis of the putative gliotoxin biosynthetic gene cluster of As pergillus fumigatus. FEMS Microbiology Letters 248:241-248.

Gene Ontology Consortium (2000) Gene Ontology: tool for the unification of biology. Nature Genet 25: 25-29.

Gibbs WW (2003) The unseen genome: gems among the junk. Sci Amer 289(5):46-53.

Goffeau A (2004) Evolutionary genomics: seeing double. Nature 430:25-26.

Goswami RS and Kistler C (2004) Heading for disaster: *Fusarium graminearum* on cereal crops. Molecular Plant Pathology 5:515-525.

Grunenfelder B and Winzeler EA. (2002) Treasures and traps in genome-wide data sets: case examples from yeast. Nat Rev Genet 3:653-661.

Gu Z, Steinmetz L.M, Gu X, Scharfe C, Davis RW and Li WH (2003) Role of duplicate genes in genetic robustness against null mutations. Nature 421:63-66.

Hall AE, Fiebig A, and Preuss D (2002) Beyond Arabidopsis genome: opportunities for comparative genomics. Plant Physiol 129:1439-1447.

Hardison RC (2003) Comparative Genomics. PLoS Biol 1(2):e58

Heckman DS, Geiser DM, Eidell BR, Stauffer RL, Kardos NL and Hedges SB (2001) Molecular evidence for the early colonization of land by fungi and plants. Science 293:1129-1133.

Hedges SB and Kumar S (2002) Vertebrate genomes compared. Science 297:1283-1285.

Hedges SB and Kumar S (2003) Genomic clocks and evolutionary timescales. Trends Genet 19:200-206.

Hofman G, McIntyre M and Nielsen J (2003) Fungal genomics beyond *Saccharomyces cerevisiae*. Curr Opin Biotech 14:226-231.

Hsiang T and Baillie DL (2004) Recent progress, developments and issues in comparative fungal genomics. Can J Plant Pathol 26:19-30.

Hsiang T and Baillie DL (2005) Comparison of the yeast proteome to other fungal genomes to find core fungal genes. J Mol Evol 60:475-483.

Hsiang T and Goodwin PH (2003) Distinguishing plant and fungal sequences in ESTs from infected plant tissues. J Microbiol Meth 54:339-351

Hsiang T, Chen F and Goodwin PH (2003) Detection and phylogenetic analysis of mating type genes of *Ophiosphaerella korrae*. Can J Bot 81:307-315.

Huson DH and Steel M (2004) Phylogenetic trees based on gene content. Bioinformatics 20:2044-2049.

Hyman RW (2001) Sequence data: posted vs. published. Science 291:827.

Imanishi T, Itoh T, Suzuki Y, O'Donovan C, Fukuchi S, Koyanagi KO, Barrero RA, Tamura T, Yamaguchi-Kabata Y, Tanino M, etc. (2004) Integrative annotation of 21,037 human genes validated by full-length cDNA clones. PLoS Biology 2: 856-875.

International Human Genome Consortium (2001) Initial sequencing and analysis of the human genome. Nature 409:860-921.

International Human Genome Consortium (2004) Finishing the euchromatic sequence of the human genome. Nature 431:931-945.

InterPro Consortium (2001) The InterPro database, an integrated documentation resource for protein families, domains and functional sites. Nucl Acids Res 29:37-40.

Jeannmougin F, Thompson JD, Gouy M, Higgins DG and Gibson TJ. (1998) Multiple sequence alignment with Clustal X. Trends Biochem Sci 23:403-5.

Jiang B, Bussey H and Roemer T (2002) Novel strategies in antifungal lead discovery. Curr Opin Microbiol 5:466-471.

Jongeneel V (2001) Searching the expressed sequence tag (EST) databases: panning for genes. Brief Bioinform 1:76-92.

Katinka MD, Duprat S, Cornillot E, Metenier G, Thomarat F, Prensier G, Barbe V, Peyretaillade E, Brottier P, Wincker P, Delbac F, El Alaoui H, Peyret P, Saurin W, Gouy M, Weissenbach J, Vivares CP (2001) Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi*. Nature 414:401-402.

Kato-Maeda M, Rhee JT, Gingeras TR, Salamon H, Drenkow J, Smittipat N and Small PM (2001) Comparing genomes within the species *Mycobacterium tuberculosis*. Genome Res 11:547-554.

Kellis M, Birren BW, Lander ES (2004) Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. Nature 428:617-624.

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

Keon J, Bailey A and Hargreaves J (2000) A group of expressed cDNA sequences from the wheat fungal leaf blotch pathogen, *Mycosphaerella graminicola* (*Septoria tritici*). Fung Genet Biol 29:118-133.

Kessler MM, Willins DA, Zeng Q, Del Mastro RG, Cook R, Doucette-Stamm L, Lee H, Caron A, McClanahan TK, Wang L, Greene J, Hare RS, Cottarel G and Shimer GH (2002) The use of direct cDNA selection to rapidly and effectively identify genes in the fungus *Aspergillus fumigatus*. Fung Genet Biol 36:59-70.

Kirst M, Johnson AF, Baucom C, Ulrich E, Hubbard K, Staggs R, Paule C, Retzel E, Whetten R and Sederoff R (2003) Apparent homology of expressed genes from wood-forming tissues of loblolly pine (*Pinus taeda* L.) with *Arabidopsis thaliana*. PNAS USA 100:7383-7388.

Koonin EV, Aravind L and Kondrashov AS (2000) The impact of comparative genomics on our understanding of evolution. Cell 101:573-576.

Kroken S, Glass NL, Taylor JW, Yoder OC, Turgeon BG (2003) Phylogenomic analysis of type I polyketide synthase genes in pathogenic and saprobic ascomycetes. PNAS USA 100:15670-15675

Kruger WM, Pritsch C, Chao S and Muehlbauer GJ (2002) Functional and comparative bioinformatic analysis of expressed genes from wheat spikes infected with *Fusarium graminearum*. Mol Plant-Microbe Interact 15: 445-455.

Liang F, Holt I, Pertea G, Karamycheva S, Salzberg SL, Quackenbush J (2000) Gene Index analysis of the human genome estimates approximately 120,000 genes. Nat Genet 25:239-240.

Liti G and Louis EJ (2005) Yeast genome evolution and comparative genomics. Ann Rev Microbiol 59:135-153.

Lockhart DJ and Winzeler EA (2000) Genomics, gene expression and DNA arrays. Nature 405:827-836.

Loftus BJ, Fung E, Roncaglia P, Rowley D, Amedeo P, Bruno D, Vamathevan J, Miranda M, Anderson IJ, Fraser JA, etc. (2005) The genome and transcriptome of *Cryptococcus neoformans*, a basidiomycete fungal pathogen of humans. Science 307:1321-1324.

Lorenz MC (2002) Genomic approaches to fungal pathogenicity. Curr Opin Microbiol 5:372-378.

Maher BA (2003) The 0.1% portrait of human history. The Scientist, June 30, 2003.

Marshall, E (2002) DNA sequencer protests being scooped with his own data. Nature, 295:1206-1207.

McClelland M, Sanderson KE, Spieth J, Clifton SW, Latreille P, Courtney L, Porwollik S, Ali J, Dante M, Du F, etc. (2001) Complete genome sequence of *Salmonella enterica* serovar Typhimurium LT2. Nature 413:852-846.

Mellanby K (1992) The DDT Story. British Crop Protection Council, Farnham, Surrey, UK.

Mewes HW, Albertmann K, Bahr M, Frishman D, Gkeissner A, Hani J, Heumann K, Kleine K, Maierl A, Oliver SG, Pfeiffer F and Zollner A. (1997a) Overview of the yeast genome. Nature 387(suppl):7-65.

Mewes HW, Albermann K, Heumann K, Liebl S and Pfeiffer F (1997b) MIPS: a database for protein sequences, homology data and yeast genome information. Nucl Acids Res 25:28-30.

Mira A, Klasson L and Andersson SGE (2002) Microbial genome evolution: sources of variability. Curr Opin Microbiol 5:506-512.

Mitchell TK, Thon MR, Jeong JS, Brown D, Deng J and Dean RA (2003) The rice blast pathosystem as a case study for the development of new tools and raw materials for genome analysis of fungal plant pathogens. New Phytol 159:53-61.

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

Nardone J, Lee DU, Ansel KM. and Rao A (2004) Bioinformatics for the 'bench biologist': how to find regulatory regions in genomic DNA. Nat Immunol 5:768-774.

Nielsen CB, Friedman B, Birren B, Burge CB and Galagan JE (2004) Patterns of intron gain and loss in fungi. PLoS Biol 2:2234-2242.

Nierman WC, May G, Kim HS, Anderson MJ, Chen D and Denning DW (2005) What the *Aspergillus* genomes have told us. Medical Mycol 43:S3 - S5.

Pallen M (2002) From sequence to consequence: in silico hypothesis generation and testing. Meth Microbiol 33:27-48.

Paoletti M, Rydholm C, Schwier EU, Anderson MJ, Szakacs G, Lutzoni F, Debeaupuis JP, Latge JP, Denning DW and Dyer PS (2005) Evidence for sexuality in the opportunistic fungal pathogen *Aspergillus fumigatus*. Curr Biol 15:1242-1248.

Papp B, Pal C and Hurst LD (2003) Dosage sensitivity and the evolution of gene families in yeast. Nature 424:194-197.

Parkhill J (2002) Annotation of microbial genomes. Meth Microbiol 33:3-26.

Parkhill J, Sebaihia M, Preston A, Murphy LD, Thomson N, Harris DE, Holden MT, Churcher CM, Bentley SD, Mungall KL, etc. (200) Comparative analysis of the genome sequences of *Bordatella pertussis, Bordatella parapetussis,* and *Bordatella bronchiseptica*. Nat Genet 45:32-40.

Parkinson T (2002) The impact of genomics on anti-infectives drug discovery and development. Trends Microbiol 10:S22-S26.

Pearson WR (1990) Rapid and sensitive sequence comparison with FASTP and FASTA. Meth Enzymol 183:63-98.

Pearson WR (1997) Identifying distantly related protein sequences. CABIOS 13:324-332.

Pearson WR (1998) Empirical statistical estimates for sequence similarity searches. J Mol Evol 276:71-84.

Pertea M and Salzberg SL (2002) Computational gene finding in plants. Plant Mol Biol 48:39-48.

Pertsemlidis A and Fondon JW (2002) Having a BLAST with bioinformatics (and avoiding BLASTphemy). Genome Biol 2(10):1-10.

Philip GK, Creevey CJ and McInerney JO (2005) The Opisthokonta and the Ecdysozoa may not be clades: stronger support for the grouping of plant and animal than for animal and fungi and stronger support for the Coelomata than Ecdysozoa. Mol Biol Evol 22:1175-1184.

Philippe H, Lartillot N and Brinkmann H (2005) Multigene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa and Protostomia. Mol Biol Evol 22:1246-1253.

Piskur J and Langkjær RB (2004) Yeast genome sequencing: the power of comparative genomics. Mol Microbiol 53:381-389.

Plummer KM and Howlett BJ (1993) Major chomosomal length polymorphisms are evident after meiosis in the phytopathogenic fungus *Leptosphaeria maculans*. Curr Genet 24:107-113.

Plummer KM and Howlett BJ (1995) Inheritance of chromosomal length polymorphisms in the ascomycete *Leptosphaeria maculans*. Mol Gen Genet 247:416-22.

Pontecorvo G (1956) The parasexual cycle in fungi. Ann Rev Microbiol 10:393-400.

Pray L (2002) Refining transgenic mice. The Scientist 16(13):34.

Ptak SE, Hinds DA, Koehler K, Nickel B, Patil N, Ballinger DG, Przeworski M, Frazer KA and Pääbo S (2005) Fine-scale recombination patterns differ between chimpanzees and humans. Nat Genet 37:429-434

Rashidi HH, and Buehler LK (2000) Bioinformatics Basics. Boca Raton: CRC Press.

Rehm BHA (2001) Bioinformatic tools for DNA/protein sequence analysis, functional assignment of genes and protein classification. Appl Microbiol Biotechnol 57:579-592.

Reiser L, Mueller LA and Rhee SY (2002) Surviving in a sea of data: a survey of plant genome data resources and issues in building data management systems. Plant Mol Biol 48:59-74.

Rokas A and Carroll SB (2005) More genes or more taxa? The relative contribution of gene number and taxon number to phylogenetic accuracy. Mol Biol Evol 22:1337-1344.

Rokas A, Willaims BL, King N and Carroll SB (2003) Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425:798-804.

Roy SW and Gilbert W (2005) Complex early genes. PNAS USA 102:1086-1991.

Rubin GM, Yandell MD, Wortman JR, Gabor Miklos GL, Nelson CR, Hariharan IK, Fortini ME, Li PW, Apweiler R, etc. (2000) Comparative genomics of Eukaryotes. Science 287:2204-2215.

Salzberg S, Birney E, Eddy S and White O (2003) Unrestricted free access works and must continue. Nature 422:801.

Schmidt R (2002) Plant genome evolution: lessons from comparative genomics at the DNA level. Plant Mol Biol 48:21-37.

Schoch CL, Aist JR, Yoder OC and Turgeon BG (2003) A complete inventory of fungal kinesins in representative filamentous ascomycetes. Fungal Genet Biol 39:1-15.

Searls DB (2003) Pharmacophylogenomics: genes, evolution and drug targets. Nature Rev 2:613-623.

Shimamoto K and Kyozuka J (2002) Rice as a model for comparative genomics of plants. Ann Rev Plant Biol 53:399-419.

Soanes DM, Skinner W, Keon J, Hargreaves J and Talbot NJ (2002) Genomes of phytopathogenic fungi and the development of bioinformatic resources. Mol. Plant-Microbe Interact 15:421-427.

Soltis DE, Albert VA, Savolainen V, Hilu H, Qiu YL, Chase MW, Farris JS, Stefanovic S, Rice DW, Palmer JD and Soltis PS (2004) Genome-scale data, angiosperm relationships, and 'ending incongruence': a cautionary tale in phylogenetics. Trends Plant Sci 19:477-483

Spencer DH, Kas A, Smith EE, Raymond CK, Sims EH, Hastings M, Burns JL, Kaul R and Olson MV (2003) Whole-genome sequence variation among multiple isolates of *Pseudomonas aeruginosa*. J Bacteriol 185:1316-1325.

Strobel G and Arnold J (2004) Essential eukaryotic core. Evolution 58:441-446.

Tatusov RL, Galperin MY, Natale DA and Koonin EV (2000) The COG data-base: a tool for genome-scale analysis of protein functions and evolution. Nucl Acids Res 28:33-36.

Tekaia F, Latge JP (2005) *Aspergillus fumigatus*: saprophyte or pathogen? Curr Opin Microbiol 8:385-92.

Thacker PD (2003) Understanding fungi through their genomes. BioScience 53:10-15.

The *C. elegans* Sequencing Consortium (1998) Genome sequence of the nematode *C. elegans*: a platform for investigating biology. Science 282:2012-2018.

Thomas JW, Touchman JW, Blakesley RW, Bouffard GG, Beckstrom-Sternberg SM, Margulies EH, Blanchette M, Siepel AC, Thomas PJ, McDowell JC, Maskeri B, Hansen NF, Schwartz MS, Weber RJ, etc. (2003) Comparative analyses of multi-species sequences from targeted genomic regions. Nature 424:788-793.

Thomas SW, Glaring MA, Rasmussen SW, Kinane JT and Oliver RP (2002) Transcript profiling in the barley mildew pathogen *Blumeria graminis* by serial analysis of gene expression (SAGE). Mol Plant-Microbe Interact 15:847-856.

Thomas SW, Rasmussen SW, Glaring MA, Rouster JA, Christiansen SK and Oliver RP (2001) Gene identification in the obligate fungal pathogen *Blumeria graminis* by expressed sequence tag analysis. Fung Genet Biol 33:195-211.

Thomson N, Sebaihia M, Cerdeno-Tarraga A, Bentley S, Crossman L and Parkhill J (2003) The value of comparison. Nature Rev Microbiol 1:11-12.

Till BJ, Reynolds SH, Greene EA, Codomo CA, Enns LC, Johnson JE, Burtner C, Odden AR, Young K, Taylor NE, Henikoff JG, Comai L and Henikoff S (2003) Large-scale discovery of induced point mutations with high-throughput TILLING. Genome Res 13:524-530.

Tisdall J (2003) Mastering PERL for Bioinformatics. Cambridge, Massachusetts: O'Reilly & Associates.

Tunlid A, and Talbot NJ (2002) Genomics of parasitic and symbiotic fungi. Curr Opin Microbiol 5:513-519.

Turgeon BG, Kroken S, Lee BN, Bsaker SE, Amedeo P, Catlett N, Gunawardena U, Wagner E, Robbertse B, Wu J, Yoder OC, Glass NL and Taylor JW (2002) Comparative genomic analysis of fungal plant pathogens: secondary metabolites and mechanisms of pathogenesis. APS Symposium on Functional Genomics of Plant Pathogen Interactions, Milwaukee, Wisconsin, July 27-31, 2002.

Tzung KW, Williams RM, Scherer S, Federspiel N, Jones T, Hansen N, Bivolarevic V, Huizar L, Komp C, Surzycki R, Tamse R, Davis RW and Agabian N (2001) Genomic evidence for a complete sexual cycle in *Candida albicans*. PNAS 98:3249-3253.

Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, Qureshi-Emili A, Li Y, Godwin B, Conover D, Kalbfleisch T, Vijayadamodar G, Yang M, Johnston M, Fields S and Rothberg JM (2000) A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. Nature 403:601-603.

Ureta-Vidal A, Ettwiller L and Birney E (2003) Comparative genomics: genome-wide analysis in metazoan eukaryotes. Nat Rev Genet 4:251-262

Wagner A (2000) Robustness against mutations in genetic networks of yeast. Nat Genet 24:355-361.

Ward TJ, Bielawski JP, Kistler HC, Sullivan E and O'Donnell K (2002) Ancestral polymorphism and adaptive evolution in the trichothecene mycotoxin gene cluster of phytopathogenic Fusarium. PNAS USA 99:9278–9283.

Webber C and Ponting CP (2004) Genes and homology. Curr Biol 14:R332-R333.

Wolf YI, Rogozin IB, Grishin NV, Tatusov RL and Koonin EV (2001) Genome trees constructed using five different approaches suggest new major bacterial clades. BMC Evol Biol 1:8.

Wong S, Fares MA, Zimmermann W, Butler G and Wolfe KH (2003) Evidence from comparative genomics for a complete sexual cycle in the 'asexual' pathogenic yeast *Candida glabrata.* Genome Biol 4:R10.

Wood V, Gwilliam R, Rajandream MA, Lyne M, Lyne R, Stewart A, Sgouros J, Peat N, Hayles J, Baker S, Basham D, Bowman S, Brooks K, Brown D, Brown S, Chillingworth T, Churcher C, etc. (2002) The genome sequence of *Schizosaccharomyces pombe*. Nature 415: 871-880.

Wright FA, Lemon WJ, Zhao WD, Sears R, Zhuo D, Wang J-P, Yang HY, Baer T, Stredney D, Spitzner J, Stutz A, Krahe R and Yuan B (2001) A draft annotation and overview of the human genome. Genome Biol 2(2): research0025.1-0025.18.

Wu Y, Wang X, Liu X and Wang Y (2003) Data-mining approaches reveal hidden families of proteases in the genome of malaria parasite. Genome Res 13:601-616.

Xu Y, Stange-Thomann N, Weber G, Bo R, Dodge S, David RG, Foley K, Beheshti J, Harris NL, Birren B, Lander E and Meyerson M (2003) Pathogen discovery from human tissue by sequence-based computational subtraction. Genomics 81:329-335.

Yarden O, Ebbole DJ, Freeman S, Rodriquez RJ and Dickman MB (2003) Fungal biology and agriculture: revisiting the field. Molec Plant-Microbe Interact 16:859-866.

Yoder OC and Turgeon BG (2001) Fungal genomics and pathogenicity. Curr Opin Pl Biol 4:315-321.

Yu J, Wang J, Lin W, Li S, Li H, Zhou J, Ni P, Dong W, Hu S, Zeng C, etc. (2005) The genomes of *Oryza sativa*: A history of duplications. PLoS Biol 3(2):38.

Zelter A, Bencina M, Bowman BW, Yarden O and Read ND (2004) A comparative genomic analysis of the calcium signaling machinery in *Neurospora crassa*, *Magnaporthe grisea*, *Aspergillus fumigatus* and *Saccharomyces cerevisiae*. Fung Genet Biol 41:827-841.

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

Zolan ME (1995) Chromosome-length polymorphism in fungi. Microbiol Rev 59:686-698.