Recent progress, developments, and issues in comparative fungal genomics

Tom Hsiang and David L. Baillie

Abstract: Biologists face an overwhelming richness of nucleotide and protein sequence data. As of the end of 2003, there were over 100 complete or almost complete nonviral genomes in publicly available databases. Most of these were bacterial, since prokaryotic genomes are generally much smaller in size than eukaryotic genomes. Among eukaryotes, fungi have some of the smallest genome sizes and, hence, represent the highest number of complete or almost complete genomes sequenced, with most of these released within the last 2 years. What are the genes that fungi have in common? Among these genes, which ones have homologs in plants, animals, or bacteria, and which ones are only found in fungi? Researchers are just beginning to be able to address these types of questions with data from high-throughput genomic sequencing. This paper examines some recent and possible future uses of fungal genomic data in comparative genome analyses, particularly as they relate to the study of fungal plant pathogens. Comparative genomics can facilitate research into the following areas: phylogenetics (via whole genome comparisons), targeted drugs (via unique target sites in pests), gene discovery (via conserved sequences), and gene function (via guilt by association). Each of these is discussed as well as the availability and ownership of the genomic data, and the concepts of homology and similarity.

Key words: bioinformatics, data mining, fungal genes, BLAST.

Résumé : Les biologistes font face à une impressionnante quantité de données sur les séquences de nucléotides et de protéines. À la fin de 2003, il y avait plus de 100 génomes non viraux complets ou pratiquement complets dans les bases de données librement accessibles. La plupart de ces génomes sont bactériens puisqu'ils sont habituellement beaucoup plus petits que les génomes eucaryotes. Parmi les eucaryotes, les champignons ont quelques-uns des plus petits génomes et, pour cette raison, ils constituent le groupe avec le plus grand nombre de génomes complètement ou presque complètement séquencés, la majorité de ces séquences l'ayant été au cours des deux dernières années. Quels sont les gènes partagés par les champignons? Parmi ces gènes, lesquels possèdent des homologues chez les plantes, les animaux ou les bactéries, et lesquels sont uniques aux champignons? Les chercheurs ne font que commencer à pouvoir répondre à ce genre de questions à l'aide des données du séquençage génomique à haut rendement. Le présent article se penche sur certaines utilisations récentes et potentielles des données génomiques sur les champignons dans des analyses comparatives de génomes, spécialement lorsqu'elles sont en lien avec l'étude des champignons phytopathogènes. La génomique comparative peut simplifier la recherche dans les domaines suivants : la phylogénétique (par comparaisons de génomes entiers), les médicaments à action élective (par l'intermédiaire de sites d'action spécifiques chez les ravageurs), la découverte de gènes (via les séquences conservées) et le rôle des gènes (par associations). Chacun de ces domaines est examiné, de même que la notion de propriété et de disponibilité des données génomiques ainsi que les concepts d'homologie et de similarité.

Mots clés : bioinformatique, exploitation des données, gènes fongiques, BLAST.

Introduction

As of the end of 2003, there were over 100 complete or almost complete nonviral genomes in publicly available databases (Thomson et al. 2003). Most of these were 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 have some of the smallest genome sizes (10–50 Mb) and, hence, represent the highest number of complete or almost complete genomes se-

Accepted 28 December 2003.

T. Hsiang.[1] Department of Environmental Biology, University of Guelph, Guelph, ON N1G 2W1, Canada.
D.L. Baillie. Department of Molecular Biology and Biochemistry, Simon Fraser University, Burnaby, BC V5A 1S6, Canada.

[1]Corresponding author (e-mail: thsiang@uoguelph.ca).

quenced. The first complete genomic sequence of a nonviral organism was that of *Haemophilus influenzae* Rd., a bacterium, in 1995. This was followed by the first eukaryote, *Saccharomyces cerevisiae* Hansen, in 1997. The first animal, *Caenorhabditis elegans* Maupas, followed in 1998, and the first plant, *Arabidopsis thaliana* L., in 2000. A few other animal species have been sequenced (*Drosophila melanogaster* Loew. in 2000, *Homo sapiens* L. in 2001, and *Fugu rubripes* Temminck & Schlegel, *Ciona intestinalis* L., *Mus musculus* L., and *Anopheles gambiae* Giles in 2002), as well as another plant species (*Oryza sativa* L. in 2002); but among eukaryotes, there has been a larger number of fungal species with complete or almost complete genomic sequences, mostly since 2001 (Table 1).

In addition to the fungal genomes listed in Table 1, there are privately held complete or almost complete fungal genomic data, including *Botrytis cinerea* Pers., *Cochliobolus heterostrophus* Drechs., and *Gibberella fujikuroi* (Sawada) Wollenw. at the Syngenta Torrey Mesa Research Institute, San Diego, Calif. (Turgeon et al. 2002). Genomic data of *Ashbya gossypii* (Ashby & Nowell) Guilliierm. are also held privately by Basel University and Syngenta AG, Basel, Switzerland, and those of *Aspergillus niger* Tiegh. and *Ustilago maydis* (DC.) Corda are held by Gene Alliance, Hilden and Konstanz, Germany, and Bayer AG, Leverkusen, Germany, respectively.

In 2000, the Fungal Genome Initiative, spearheaded by the Whitehead Institute for Biomedical Research, Cambridge, Mass., was formed to discuss and prioritize fungal genome sequencing. In February 2002, they released the first White paper on fungal species targeted for sequencing. Of the 15 fungi selected, the National Human Genome Research Institute (NHGRI), Bethesda, Md., agreed to fund the costs of sequencing 7, which have been completed or are almost completed. In June 2003, the Fungal Genome Initiative released the second White paper, which contains a list of 44 fungal sequencing targets, with an emphasis on 10 major genus clusters of related species (*Penicillium*, *Aspergillus*, *Histoplasma*, *Coccidioides*, *Fusarium*, *Neurospora*, *Candida*, *Schizosaccharomyces*, *Cryptococcus*, and *Puccinia*). Copies of the White papers, and more details on the status of these projects can be found at www.broad.mit.edu/annotation/fungi/fgi/history.html. 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), 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 purpose of the present paper is to provide an overview of developments in comparative genomics as well as some predictions for future directions in fungal comparative genomics. Comparative genomics can facilitate research into the following areas: phylogenetics (via whole genome comparisons), targeted drugs (via unique target sites in pests), gene discovery (via conserved sequences), and gene function (via guilt by association). 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.

Ownership of the genomic data

In 1991, the NHGRI and the U.S. Department of Energy developed a data-release policy whereby sequencing projects funded publicly should release their data within 6 months of generation. In 1996, the International Human Genome Sequencing Consortium adopted the "Bermuda Principles", which resulted in a policy of assembly data release within 24 h of generation. In early 2003, NHGRI issued a draft revision of release policies for sequencing data. In essence, sequencing projects funded publicly in the United States and the United Kingdom are required to release their data without imposing restrictions, while sequence users are reminded that they must provide proper citation of the data source and also 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 www.genome.gov/10506537. Situations have occurred where sequence generators felt that their prerogative to first publish on their own data has been preempted by other researchers who have analyzed the sequence data before full-genome release in a peer-reviewed publication (Marshall 2002). An editorial in the journal *Nature* reaffirms that journals will likely accept good research involving whole-genome analyses, whomever it comes from, since that is in the best interests of science (Anonymous 2003). A response to the editorial in *Nature* by several prominent bioinformatics researchers (Salzberg et al. 2003) asserts further that genome-sequence data must be available for all to use without restriction.

While the committees in the United States and the United Kingdom have provided these guiding principles, it is of course up to individual countries to decide on the accessibility of sequence data from publicly funded projects. For example, in Canada, some genomic data generated by publicly funded research institutions through government grants are not available publicly. The Canadian government currently does not have a policy on public access to data from government-funded gene-sequencing projects, and this situation should be addressed in light of the open policies established in the United States and the United Kingdom.

Homology

Comparative genomics involves comparisons of sequences to look for the presence or absence of homologs. Homology refers to similarity by descent and is qualitative rather than quantitative: two sequences are homologous or they are not (Doyle and Gaut 2000). In much of the molecular biology literature, homology has been used as a synonym for similarity, such as in statements 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 not usually the intended meaning (Doyle and Gaut 2000). For quantitative assessments of relationships, the terms identity and similarity

Table 1. Chronological listing of species genomes, showing source, release date, and size of plant, animal, and fungal genomes as well as of some phytopathogenic bacterial genomes and the first genomes sequenced from other major taxa.

			Genome	
Date published or released	Species	Taxon	Source and publication	Size
28/07/1995	*Haemophilus influenzae*	Bacterium	The Institute for Genomic Research, Rockville, Md. (TIGR) (Fleischmann et al. 1995)	2 Mb
23/08/1996	*Methanococcus jannaschii* Jones et al.	Archaea	TIGR (Bult et al. 1996)	2 Mb
12/06/1997	*Saccharomyces cerevisiae*	Fungus	*Saccharomyces* genome database (SGD) Stanford Genome Technology Center, Palo Alto, Calif. (Mewes et al. 1997b)	12 Mb
11/12/1998	*Caenorhabditis elegans*	Animal	(The *C. elegans* Sequencing Consortium 1998)	97 Mb
24/03/2000	*Drosophila melanogaster*	Animal	Berkeley Drosophila Genome Project (BDGP) (Adams et al. 2000)	138 Mb
03/02/2003	*Xylella fastidiosa* Wells et al.	Bacterium	Agronomical and Environmental Genomes (AEG), São Paulo, Brazil (Van Sluys et al. 2003)	3 Mb
14/12/2000	*Arabidopsis thaliana*	Plant	TIGR (Arabidopsis Genome Initiative 2000)	115 Mb
15/02/2001	*Homo sapiens*	Animal	(The International Human Genome Sequencing Consortium 2001)	2.9 Gb
May 2001	*Aspergillus fumigatus*	Fungus	Sanger Institute	29 Mb
14/12/2001	*Agrobacterium tumefaciens* (Smith & Townsend) Conn	Bacterium	University of Washington, Seattle, Wash. (Wood et al. 2001)	6 Mb
15/11/2001	*Encephalitozoon cuniculi* Levaditi et al.	Protist	Genoscope, Evry, France (Katinka et al. 2001)	3 Mb
11/12/2001	*Ralstonia solanacearum* (Smith) Yabuuchi	Bacterium	National Centre for Biotechnology Information, Bethesda, Md. (NCBI)	6 Mb
16/12/2002	*Phanerochaete chrysosporium* Burdsall	Fungus	Department of Energy's (DOE) Joint Genome Institute, Walnut Creek, Calif.	30 Mb
May 2002	*Candida albicans*	Fungus	Stanford Genome Technology Center (Tzung et al. 2001)	15 Mb
23/05/2002	*Xanthomonas campestris* pv. *campestris* Vauterin	Bacterium	Fundação de Amparo à Pesquisa do Estado de São Paulo, São Paulo, Brazil (da Silva et al. 2002)	5 Mb
June 2002	*Magnaporthe grisea* (Hebert) Barr	Fungus	Whitehead Institute for Biomedical Research, Cambridge, Mass.	40 Mb
25/07/2002	*Fugu rubripes*	Animal	DOE Joint Genome Institute (Aparicio et al. 2002)	330 Mb
04/10/2002	*Anopheles gambiae*	Animal	Sanger Institute, www.ensembl.org	278 Mb
05/04/2002	*Oryza sativa*	Plant	Beijing Genomics Institute, China (Yu et al. 2002) and International Rice Genome Sequencing Project (IRGSP) (Goff et al. 2002)	0.4 Gb
13/12/2002	*Ciona intestinalis*	Animal	DOE Joint Genome Institute (Dehal et al. 2002)	160 Mb
05/12/2002	*Mus musculus*	Animal	(Mouse Genome Sequencing Consortium 2002)	2.5 Gb
21/02/2002	*Schizosaccharomyces pombe*	Fungus	Sanger Institute (Wood et al. 2002)	14 Mb
Jan. 2003	*Aspergillus nidulans* (Eidam) Winter	Fungus	Whitehead Institute for Biomedical Research	31 Mb
Mar. 2003	*Fusarium graminearum*	Fungus	Whitehead Institute for Biomedical Research	40 Mb
28/03/2003	*Saccharomyces bayanus*	Fungus	Whitehead Institute for Biomedical Research (Kellis et al. 2003)	12 Mb
28/03/2003	*Saccharomyces mikatae*	Fungus	Whitehead Institute for Biomedical Research (Kellis et al. 2003)	12 Mb
28/03/2003	*Saccharomyces paradoxus*	Fungus	Whitehead Institute (Kellis et al. 2003)	12 Mb
31/03/2003	*Cryptococcus neoformans* subsp. D (Sanfelice) Vuillemin	Fungus	Stanford Genome Technology Center	24 Mb
07/04/2003	*Saccharomyces kluyveri* Phaff et al.	Fungus	Whitehead Institute for Biomedical Research (Cliften et al. 2003)	12 Mb
07/04/2003	*Saccharomyces castelli* Capriotti	Fungus	Whitehead Institute for Biomedical Research (Cliften et al. 2003)	12 Mb
07/04/2003	*Saccharomyces kadriavzevii* Naumov et al.	Fungus	Whitehead Institute for Biomedical Research (Cliften et al. 2003)	12 Mb
20/05/2003	*Cryptococcus neoformans* subsp. A	Fungus	Whitehead Institute for Biomedical Research	24 Mb
24/04/2003	*Neurospora crassa*	Fungus	Whitehead Institute for Biomedical Research (Galagan et al. 2003)	40 Mb

Table 1 (*concluded*).

Date published or released	Species	Taxon	Genome	
			Source and publication	Size
23/07/2003	*Ustilago maydis*	Fungus	Whitehead Institute for Biomedical Research	20 Mb
25/07/2003	*Coprinus cinereus* (Schaeff.) Gray	Fungus	Whitehead Institute for Biomedical Research	38 Mb
10/10/2003	*Phytophthora sojae* Hildebrand	Oomycete	DOE Joint Genome Institute	1.4 Gb
31/10/2003	*Phytophthora ramorum* Werres et al.	Oomycete	DOE Joint Genome Institute	1.4 Gb

Note: Information for this table was compiled from maine.ebi.ac.uk:8000/services/cogent, wit.integratedgenomics.com/GOLD, and www.broad.mit.edu/annotation/fungi/fgi/history. Since this paper was accepted, there have been several more fungal genomes publicly released: *Coprinopsis cinerea* (Schaeff.) Redhead et al., *Phakopsora pachyrhizi* Syd. & Syd., *Trichoderma reesei* Simmons, and *Ustilago maydis*.

are often employed, but the usage has been inconsistent. For nucleotides, both identity and similarity often refer to the same thing: the occurrence of the same nucleotide at the same (homologous) position. For protein sequences, identity has had the same usage as that for nucleotides, but similarity also includes matches with amino acids of similar triplet coding (2 out of 3) and similar chemical characteristics. Various programs such as FASTA (Pearson 1990) or BLAST (basic local alignment search tool; Altschul et al. 1990) serve to assess the matches between the query sequence and the subject sequence. Output from BLAST programs 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 provided. For example, a BLASTP analysis (protein query vs. protein database) of the *Saccharomyces cerevisiae* glucosidase protein YIL099W (549 amino acids) resulted in the following match with the *Neurospora crassa* Shear & Dodge glucosidase protein NCU01517: identities = 145/469 (30%), positives = 224/469 (47%). This indicates that out of the 549 amino acid sequence submitted as a query sequence, a 469 contiguous amino acid portion matched a sequence in the database; and in this match, 145 positions had the identical amino acid, while 224 positions had identical or similar amino acids.

What level of 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. A statistic that is often applied as a criterion for homology is the expect value (E), referring 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). To illustrate the relationship and differences between identity and E, we analyzed 15 BLASTP results of randomly selected yeast genes matched against the GenBank NR protein database (Fig. 1). From within each BLASTP analysis, we selected the match that was closest to $E = 10^{-5}$ and we recorded the identity level. Figure 1 shows that when the matching-sequence length is shorter, a greater number of matching positions is required to achieve the same E value; in essence, the E value accounts for both the percent identity and the length over which the matching occurs.

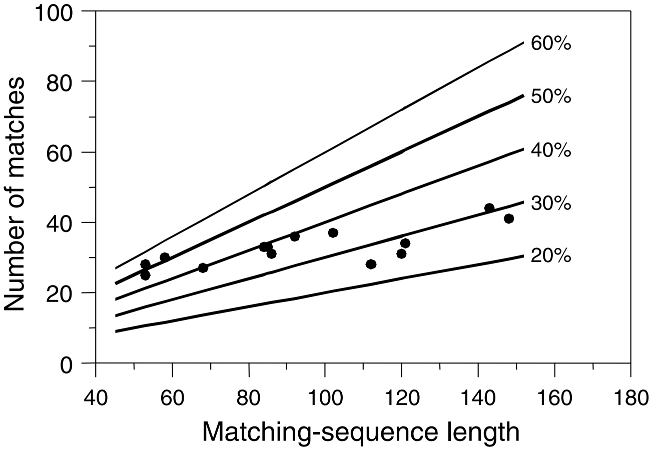
In many studies involving database searches with BLAST, a match with $E = 10^{-20}$ or less is considered a strong match, while matching below a threshold of $E = 10^{-5}$ is often considered as the criterion for homology (e.g., Keon et al. 2000; Kruger et al. 2002; Thomas et al. 2001, 2002). Some researchers consider $E < 10^{-1}$ to represent biological significance of the match, and have used the E value as a measure of statistical significance (Pertsemlidis and Fondon 2002). Pearson (1998) states that $E = 0.02$ could be used to infer homology with only a 2% chance of a false positive. 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.

A further complication is that there are at least three distinct types of homologs: orthologs, paralogs, and xenologs (Fitch 2000). Orthology describes the relationship between homologous genes found in different organisms, where the single ancestral gene was present in the most recent ancestor. Paralogy describes the relationship between homologous genes that arose by gene duplication, such as for 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 horizontal gene transfer into another organism. In a phylogenetic analysis, mixing paralogs, orthologs, or xenologs could result in a phylogeny that is correct for the genes, but not for the organisms (Fitch 2000). The difficulty is that it is sometimes not possible to establish a distinction among these different types of homologs with the data available.

Comparative genomics

A summary of some of the percentages of gene homologies between common model organisms is available at iubio.bio.indiana.edu:6780/all/hgsummary.html *Saccharomyces cerevisiae* shares, respectively, 25, 26, 20, 24, 26, 22, and 7% homology with genes of fruitfly (*D. melanogaster*), human (*Homo sapiens*), mouse (*Mus musculus*), mosquito (*Anopheles gambiae*), *Arabidopsis thaliana*, *Caenorhabditis elegans*, and *Escherichia coli* (Migula) Castellani & Chalmers. These percentages reflect the proportion of genes of *Saccharomyces cerevisiae* that have a match with protein sequences of the other organisms based on $E \leq 10^{-30}$.

Fig. 1. Relationship between number of positions matching and length of matching sequence, with lines of percent identity shown. All points were selected from BLASTP analyses to represent expect values (*E*) of 10^{-5}.



New insights into biology and evolution have been gained from studies of comparative genomics (Koonin et al. 2000) 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) such as phytoplankton (Fuhrman 2003), higher plants (Bennetzen 2002; Hall et al. 2002; Schmidt 2002; Shimamoto and Kyozuka 2002; Pertea and Salzberg 2002; Reiser et al. 2002), or animals (Ureta-Vidal et al. 2003). Only by making such comparisons, can many of the 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 although it was slightly smaller than the human genome, it was found that 99% of human genes have a homolog in the mouse genome (Mouse Genome Sequencing Consortium 2002). Among the genes exclusive to mouse, many are involved in the sense of smell. Also, during comparison of the two genomes, more predicted human genes were uncovered (Mouse Genome Sequencing Consortium 2002).

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* spp., 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).

Although full-genome comparisons would seem to allow questions in systematics to be settled definitively, there are several issues that need consideration and further investigation. 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 be as large as 11% for *Salmonella enterica* Le Minor & Popoff (McClelland et al. 2001) and 10% for *Pseudomonas aeruginosa* (Schroeter) Migula (Spencer et al. 2003). For *Pseudomonas 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* (Zopf) Lehmann & Neumann, 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 and Howlett 1993, 1995; Zolan 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 was not representative of the species.

Unique target sites in pests

One of the major purported benefits of 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 nonpathogenic, insights can be gained into target sites for antimicrobial activity, including novel fungicide target sites. Caution must be taken with this approach, since many agricultural pesticides, which turned out to have strong nontarget effects, often affected sites, in the host or other nontarget organisms, that 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 that 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 allow discovery of novel target sites present in pathogens and absent in the host (e.g., Kessler et al. 2002). A more comprehensive method for characterizing pharmacological targets may involve phylogenomics, where the evolutionary analyses of potential target sites are also considered (Searls 2003).

Gene prediction and gene function

While gene sequences are likely very accurate, with the level of estimatable error dependent on the sequencing procedure, annotation involves interpretation of the sequence and is often subject to error (Parkhill 2002). Gene prediction algorithms are based first on finding open reading frames larger than a given size (usually 100 amino acids), 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 that of known coding sequences in that organism (Parkhill 2002). In addition to similarity searches to assign function, there are nonsimilarity methods such as physical proximity and frequent cooccurrence (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* spp. (*S. cerevisiae*, *S. paradoxus* Bachinskaya, *S. mikatae* Naumov et al., and *S. bayanus* Saccardo) 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 *Saccharomyces 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 of guilt by association in their two-hybrid analysis of protein interactions in yeast. They 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.

The annotation of gene functions is likely to be a major bottleneck in genomics (Pallen 2002). 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. 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 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. There are attempts to classify genes from a variety of organisms into functional classes such as COG (cluster of orthologous genes; Rashidi and Buehler 2000; Tatusov et al. 2000) and MIPS (Munich Information Center for Protein Sequences, Max-Planck-Institut für Biochemie, Martinsreid, Germany; Mewes et al. 1997a). Rehm (2001) discusses some methods involved in sequence analyses, including functional assignment of genes.

For genes without known function, one method to determine functions is by gene knockout (Capecchi 1989). Prior to this breakthrough technique, researchers had already developed gene-transfer technology in mice in the early 1980s, but they could neither predict nor control where the transgene would be inserted into the genome of the target organism (Pray 2002). Using homologous recombination, Capecchi (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 polymerase chain reaction), cDNA sequencing, differential display, and more recently derived procedures such as microarray analyses (Lockhart and Winzeler 2000), serial analysis of gene expression (SAGE; Thomas et al. 2002), and analyses of expressed sequence tags (ESTs; Soanes et al. 2002). The EST database is the fastest growing segment in GenBank, and Jongeneel (2001) presents a good overview of searching for genes in various 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 is accumulating exponentially. The opportunities for data mining are concomitantly increasing exponentially.

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 that 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 reflecting the greater number of prokaryotic genomes available (Braun et al. 2000). For *Saccharomyces 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 *Caenorhabditis elegans* or *D. melanogaster* and concluded that functional genomics with human genes should involve higher fungi and not just the model organism *Saccharomyces cerevisiae*.

A massive comparative study of the genomes of *D. melanogaster*, *Caenorhabditis elegans*, and *Saccharomyces 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 that were similar in size and twice that of yeast, but that the multidomain proteins and signaling pathways in the animals were more complex than those of yeast. In another massive 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 did not code for any proteins. Noncoding 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 noncoding DNA actually contains hidden genes that work through RNA (Gibbs 2003). This significant discovery of hidden genes was facilitated by comparative genomics.

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. Among nine fungal-infected plant EST libraries, they found that an average of 5.6% of the sequences had the best match with a fungal genome, and 78% had the best match with a plant genome, while BLASTX of the GenBank NR database showed 1.8% and 70%, respectively. As the number of completely sequenced plant genomes increases, then the number of ESTs with no matches, when employing this method of analysis, should decrease.

Similarly, Xu et al. (2003) used 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 nonmatching sequences and attempted to amplify these sequences from infected and noninfected 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.

Fungal comparative genomics

As biologists who are interested in fungi, plant pathologists and mycologists are fortunate that the first sequenced eukaryote, *Saccharomyces cerevisiae*, still remains an important model for eukaryotic systems. Because of their small genomes, this yeast and other related yeasts have been sequenced and are receiving more attention (Salzberg 2003). Although a major emphasis is on how yeast genes relate to human genes, particularly in relation to human diseases, plant pathologists working with fungi and mycologists receive a side benefit of having a fungus with the most completely described eukaryotic genome.

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* (Ito & Kuribayashi) Drechsler ex Dastur, *Fusarium graminearum* Schwabe, and *B. cinerea* have more genes dedicated to secondary metabolism than do saprophytes such as *N. crassa*, *Ashbya gossypii*, and *Saccharomyces cerevisiae.* They found that the three phytopathogenic fungi were rich in peptide synthetases and polyketide synthases, whereas the saprophytes encoded few or none of these proteins. Yarden et al. (2003) contend that searches for differences between phytopathogenic fungi and nonphytopathogenic 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.

Papp et al. (2003) studied how gene dominance arises in the genome of *Saccharomyces cerevisiae.* According to the balance hypothesis of gene function, genes that are involved as part of protein–protein complexes must be in balance (optimal ratio of gene-copy number), and gene duplications or deletions would lead to lowered fitness (Papp et al. 2003). Papp et al. (2003) used genomic sequences of *Saccharomyces cerevisiae* to search for paralogs ($E \leq 10^{-2}$) to identify gene-family size. Then they compiled a list of interacting protein pairs that 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 they 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 *Candida albicans* (Robin) Berkhout with *Saccharomyces 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 *Candida 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 *Candida albicans* generates genetic variation (Tzung et al. 2001).

Kessler et al. (2002) used direct cDNA selection to increase the frequency of rare cDNA clones of *Aspergillus fumigatus* Fresenius and to reduce the frequency of abundant ones. They sequenced 3000 ESTs from normalized and nonnormalized libraries and found 2555 unigenes. The nonnormalized library contributed to 75% of all the redundant ESTs, demonstrating that normalization can greatly reduce redundancy. Kessler et al. (2002) compared these unique sequences with the genomes of *Saccharomyces cerevisiae, Schizosaccharomyces pombe* Lindner, and *Candida albicans* and found an average of 37% matches with each of these yeast genomes, and a total of 26% of sequences with matches in all three yeast genomes, using a criterion for homology of $E \leq 10^{-5}$ (Kessler et al. 2002). In addition, a match against GenBank NR showed 49% of the sequences without a database match. The authors concluded that these latter sequences could be *Aspergillus fumigatus* specific genes that could be used as potential candidates for novel antifungal targets specific to this fungus.

Wagner (2000) examined the ability of yeast 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 *Saccharomyces cerevisiae* that had 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). For *Saccharomyces cerevisiae*, they estimated that, among the gene deletions that 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.

Cliften et al. (2003) compared the genomes of six *Saccharomyces* spp. to find functional nonprotein-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.

Comparative genomics can be used to address many very fundamental questions in biology, and 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.

Essential fungal genes

Braun et al. (2000) conducted a whole-genome comparison between *Saccharomyces cerevisiae* and *N. crassa*. They found that *N. crassa*, with its larger genome, has more unique genes than *Saccharomyces 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 *Saccharomyces cerevisiae* was interpreted as gene loss from *Saccharomyces cerevisiae*. Braun et al. (2000) were also able to find genes in *N. crassa* that were not found in any nonfungal 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 posttranscriptional 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 *Schizosaccharomyces pombe* and found that 17.5% of these deletions were of essential genes. They then compared 450 proteins from two yeasts (*Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*) with those of Metazoa, plants, and prokaryotes in the GenBank nonredundant protein database and estimated that 80% of the essential genes of *Schizosaccharomyces 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 *Saccharomyces cerevisiae*, with the criterion for homology at $E \leq 10^{-5}$. With a greater number and taxonomic range of fungal genomes being sequenced every year, our ability to uncover genes that are conserved across many fungal taxa will be enhanced. We may then be able to determine which genes are exclusively fungal, which make fungi distinctive from other organisms.

“Cottage industry” versus large-scale fungal genomics

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 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 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 have been written with a multitude of authors, arising from laboratories that may have both high-powered molecular biology and computational tools; however, there is still a role for smaller research laboratories in comparative genomics. The fact that the massive computing power available to a supercomputing center may be able to process all the data and make the sequence comparisons in one day, a task that may take several months for a smaller 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. This paper has discussed just a few of the discoveries that are possible with 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 everyone to share and analyze to bring forth important discoveries of biological significance.

Acknowledgements

This article was written for the most part while T. Hsiang was on Research Leave at Simon Fraser University, Burnaby, B.C., working with D.L. Baillie. The authors are very grateful to Z. Punja for providing office space and access to the Simon Fraser University library and internet for T. Hsiang, and as well as for the fruitful discussions regarding fungal genomics. The authors are also very grateful to P. Goodwin for his helpful comments on the manuscript.

References

Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G., et al. 2000. The genome sequence of *Drosophila melanogaster.* Science (Washington, D.C.), 287: 2185–2195.

Alekshun, M.N. 2001. Beyond comparison — antibiotics from genome data? Nat. Biotechnol. 19: 1124–1125.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. 1990. Basic local alignment search tool. J. Mol. Biol. 215: 403–410.

Anonymous. 2003. Sacrifice for the greater good. Nature (London), 421: 875.

Aparicio, S., Chapman, J., Stupka, E., Putnam, N., Chia, J.M., Dehal, P., et al. 2002. Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. Science (Washington, D.C.), 297: 1301–1310.

Arabidopsis Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant, *Arabidopsis thaliana*. Nature (London), 408: 796–815.

Bennett, J.W., and Arnold, J. 2001. Genomics for fungi. *In* The Mycota: a comprehensive treatise on fungi as experimental systems for basic and applied research. Biology of the fungal cell. *Edited by* R.J. Howard and N.A.R. Gow. Springer-Verlag GmbH & Co., Berlin, Germany. pp. 267–297.

Bennetzen, J. 2002. Opening the door to comparative plant biology. Science (Washington, D.C.), 296: 60–63.

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

Bos, J.I.B., Armstrong, M., Whisson, S.C., Torto, T.A., Ochwo, M., Birch, P.R.J., and Kamoun, S. 2003. Intraspecific comparative genomics to identify avirulence genes from *Phytophthora*. New Phytol. 159: 63–72.

Braun, E.L., Halpern, A.L., Nelson, M.A., and Natvig, D.O. 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, P.D., Roberts, P.J., Spooner, B.M., and Panchal, G. 2003. On the reliability of published DNA sequences. New Phytol. 160: 43–48.

Buckingham, S. 2003. Programmed for success. Nature (London), 425: 209–215.

Bult, C.J., White, O., Olsen, G.J., Zhou, L., Fleischmann, R.D., Sutton, G.G., et al. 1996. Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. Science (Washington, D.C.), 273: 1058–1073.

Capecchi, M.R. 1989. Altering the genome by homologous recombination. Science (Washington, D.C.), 244: 1288–1292.

Cherry, M., Adler, C., Ball, C., Chervitz, S.A., Dwight, S.S., Hester, E.T., Jia, Y., Juvik, G., Roe, T., Schroeder, M., Weng, S., and Botstein, D. 1998. SGD: *Saccharomyces* genome database. Nucl. Acids Res. 26: 73–80.

Cliften, P.F., Hillier, L.W., Fulton, L., Graves, T., Miner, T., Gish, W.R., Waterston, R.H., and Johnston, M. 2001. Surveying *Saccharomyces* genomes to identify functional elements by comparative DNA sequence analysis. Genome Res. 11: 1175–1186.

Cliften, P., Sudarsanam, P., Desikan, A., Fulton, L., Fulton, B., Majors, J., Waterston, R., Cohen, B.A., and Johnston, M. 2003. Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. Science (Washington, D.C.), 301: 71–76.

Covert, S.F. 1998. Supernumerary chromosomes in filamentous fungi. Curr. Genet. 33: 311–319.

da Silva, A.C., Ferro, J.A., Reinach, F.C., Farah, C.S., Furlan, L.R., Quaggio, R.B., et al. 2002. Comparison of the genomes of two *Xanthomonas* pathogens with differing host specificities. Nature (London), 417: 459–463.

Date, S.V., and Marcotte, E.M. 2003. Discovery of uncharacterized cellular systems by genome-wide analyses of functional linkages. Nat. Biotechnol. 21: 1055–1062.

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

Dehal, P., Satou, Y., Campbell, R.K., Chapman, J., Degnan, B., De Tomeso, A., et al. 2002. The draft genome of *Ciona intestinalis*: insights into chordate and vertebrate origins. Science (Washington, D.C.), 298: 2157–2167.

Dennis, C. 2003. Draft guidelines ease restrictions on use of genome sequence data. Nature (London), 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.

Doyle, J.J., and Gaut, B.S. 2000. Evolution of genes and taxa: a primer. Plant Mol. Biol. 42: 1–23.

Edward, L., Braun, E.L., Halpern, A.L., Nelson, M.A., and Natvig, D.O. 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.

Fire, A., Xu, S., Montgomery, M.K., Kostas, S.A., Driver, S.E., and Mello, C.C. 1998. Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. Nature (London), 391: 806–811.

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

Fitch, W.M. 2000. Homology, a personal view on some of the problems. Trends Genet. 16: 227–231.

Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C.J., Tomb, J.F., Dougherty, B.A., and Merrick, J.M. 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae*. Science (Washington, D.C.), 269: 496–512.

Fraser, C.M., Eisen, J.A., and Salzberg, S.L. 2000. Microbial genome sequencing. Nature (London), 406: 799–803.

Fraser, C.M., Eisen, J.A., Nelson, K.E., Paulsen, I.T., and Salzberg, S.L. 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 (London), 424: 1001–1002.

Galagan, J.E., Calvo, S.E., Borkovich, K.A., Selker, E.U., Read, N.D., Jaffe, D., et al. 2003. The genome sequence of the filamentous fungus *Neurospora crassa*. Nature (London), 422: 859–868.

Gibbs, W.W. 2003. The unseen genome: gems among the junk. Sci. Am. 289(5): 46–53.

Goff, S.A., Ricke, D., Lan, T.H., Presting, G., Wang, R., Dunn, M., et al. 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). Science (Washington, D.C.), 296: 92–100.

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

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

Hedges, S.B., and Kumar, D. 2002. Vertebrate genomes compared. Science (Washington, D.C.), 297: 1283–1285.

Hofmann, G., McIntyre, M., and Nielsen, J. 2003. Fungal genomics beyond *Saccharomyces cerevisiae*. Curr. Opin. Biotechnol. 14: 226–231.

Hsiang, T., and Goodwin, P.H. 2003. Distinguishing plant and fungal sequences in ESTs from infected plant tissues. J. Microbiol. Methods, 54: 339–351.

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, M.D., Duprat, S., Cornillot, E., Metenier, G., Thomarat, F., Prensier, G., et al. 2001. Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi*. Nature (London), 414: 450–453.

Kato-Maeda, M., Rhee, J.T., Gingeras, T.R., Salamon, H., Drenkow, J., Smittipat, N., and Small, P.M. 2001. Comparing genomes within the species *Mycobacterium tuberculosis*. Genome Res. 11: 547–554.

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 (London), 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*). Fungal Genet. Biol. 29: 118–133.

Kessler, M.M., Willins, D.A., Zeng, Q., Del Mastro, R.G., Cook, R., Doucette-Stamm, L., Lee, H., Caron, A., McClanahan, T.K., Wang, L., Greene, J., Hare, R.S., Cottarel, G., and Shimer, G.H. 2002. The use of direct cDNA selection to rapidly and effectively identify genes in the fungus *Aspergillus fumigatus*. Fungal Genet. Biol. 36: 59–70.

Koonin, E.V., Aravind, L., and Kondrashov, A.S. 2000. The impact of comparative genomics on our understanding of evolution. Cell, 101: 573–576.

Kruger, W.M., Pritsch, C., Chao, S., and Muehlbauer, G.J. 2002. Functional and comparative bioinformatic analysis of expressed genes from wheat spikes infected with *Fusarium graminearum*. Mol. Plant–Microbe Interact. 15: 445–455.

Lockhart, D.J., and Winzeler, E.A. 2000. Genomics, gene expression and DNA arrays. Nature (London), 405: 827–836.

Lorenz, M.C. 2002. Genomic approaches to fungal pathogenicity. Curr. Opin. Microbiol. 5: 372–378.

Maher, B.A. 2003. The 0.1% portrait of human history. The Scientist, 17(13): 28–29.

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

McClelland, M., Sanderson, K.E., Spieth, J., Clifton, S.W., Latreille, P., Courtney, L., et al. 2001. The complete DNA sequence of *Salmonella enterica* serovar Typhimurium LT2. Nature (London), 413: 846–852.

Mellanby, K. 1992. The DDT Story. British Crop Protection Council, Farnham, Surrey, U.K.

Mewes, H.W., Albermann, K., Heumann, K., Liebl, S., and Pfeiffer, F. 1997a. MIPS: a database for protein sequences, homology data and yeast genome information. Nucl. Acids Res. 25: 28–30.

Mewes, H.W., Albertmann, K., Bahr, M., Frishman, D., Gkeissner, A., Hani, J., Heumann, K., Kleine, K., Maierl, A., Oliver, S.G., Pfeiffer, F., and Zollner, A. 1997b. Overview of the yeast genome, Nature (London), 387: 7–65.

Mira, A., Klasson, L., and Andersson, S.G.E. 2002. Microbial genome evolution: sources of variability. Curr. Opin. Microbiol. 5: 506–512.

Mitchell, T.K., Thon, M.R., Jeong, J.-S., Brown, D., Deng, J., and Dean, R.A. 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 (London), 420: 520–562.

Pallen, M. 2002. From sequence to consequence: in silico hypothesis generation and testing. Methods Microbiol. 33: 27–48.

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

Parkhill, J. 2002. Annotation of microbial genomes. Methods Microbiol. 33: 3–26.

Parkhill, J., Sebaihia, M., Preston, A., Murphy, L.D., Thomson, N., Harris, D.E., et al. 2003. Comparative analysis of the genome sequences of *Bordetella pertussis*, *Bordetella parapertussis*, and *Bordetella bronchiseptica*. Nat. Genet. 35: 32–40.

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

Pearson, W.R. 1990. Rapid and sensitive sequence comparison with FASTP and FASTA. Methods Enzymol. 183: 63–98.

Pearson, W.R. 1998. Empirical statistical estimates for sequence similarity searches. J. Mol. Evol. 276: 71–84.

Pertea, M., and Salzberg, S.L. 2002. Computational gene finding in plants. Plant Mol. Biol. 48: 39–48.

Pertsemlidis, A., and Fondon, J.W. 2002. Having a BLAST with bioinformatics (and avoiding BLASTphemy). Genome Biol. 2(10): 1–10.

Plummer, K.M., and Howlett, B.J. 1993. Major chomosomal length polymorphisms are evident after meiosis in the phytopathogenic fungus *Leptosphaeria maculans*. Curr. Genet. 24: 107–113.

Plummer, K.M., and Howlett, B.J. 1995. Inheritance of chromosomal length polymorphisms in the ascomycete *Leptosphaeria maculans*. Mol. Gen. Genet. 247: 416–422.

Pontecorvo, G. 1956. The parasexual cycle in fungi. Annu. Rev. Microbiol. 10: 393–400.

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

Rashidi, H.H., and Buehler, L.K. 2000. Bioinformatics basics. CRC Press, Boca Raton, Calif.

Rehm, B.H.A. 2001. Bioinformatic tools for DNA/protein sequence analysis, functional assignment of genes and protein classification. Appl. Microbiol. Biotechnol. 57: 579–592.

Reiser, L., Mueller, L.A., and Rhee, S.Y. 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., Williams, B.L, King, N., and Carroll, S.B. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature (London), 425: 798–804.

Rubin, G.M., Yandell, M.K., Wortman, J.R., Miklos, G., Nelson, C., Hariharan, I.K., et al. 2000. Comparative genomics of Eukaryotes. Science (Washington, D.C.), 287: 2204–2215.

Salzberg, S.L. 2003. Yeast rises again. Nature (London), 423: 233–234.

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

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

Searls, D.B. 2003. Pharmacophylogenomics: genes, evolution and drug targets. Nat. Rev. 2: 613–623.

Shimamoto, K., and Kyozuka, J. 2002. Rice as a model for comparative genomics of plants. Annu. Rev. Plant Biol. 53: 399–419.

Soanes, D.M., Skinner, W., Keon, J. Hargreaves, J., and Talbot, N.J. 2002. Genomes of phytopathogenic fungi and the development of bioinformatic resources. Mol. Plant–Microbe Interact. 15: 421–427.

Spencer, D.H., Kas, A., Smith, E.E., Raymond, C.K., Sims, E.H., Hastings, M., Burns, J.L., Kaul, R., and Olson, M.V. 2003. Whole-genome sequence variation among multiple isolates of *Pseudomonas aeruginosa*. J. Bacteriol. 185: 1316–1325.

Tatusov, R.L., Galperin, M.Y., Natale, D.A., and Koonin, E.V. 2000. The COG data-base: a tool for genome-scale analysis of protein functions and evolution. Nucl. Acids Res. 28: 33–36.

Thacker, P.D. 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 (Washington, D.C.), 282: 2012–2018.

The International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. Nature (London), 409: 860–921.

Thomas, J.W., Touchman, J.W., Blakesley, R.W., Bouffard, G.G., Beckstrom-Sternberg, S.M., Margulies, E.H., et al. 2003. Comparative analyses of multi-species sequences from targeted genomic regions. Nature (London), 424: 788–793.

Thomas, S.W., Rasmussen, S.W., Glaring, M.A., Rouster, J.A., Christiansen, S.K., and Oliver, R.P. 2001. Gene identification in the obligate fungal pathogen *Blumeria graminis* by expressed sequence tag analysis. Fungal Genet. Biol. 33: 195–211.

Thomas, S.W., Glaring, M.A., Rasmussen, S.W., Kinane, J.T., and Oliver, R.P. 2002. Transcript profiling in the barley mildew pathogen *Blumeria graminis* by serial analysis of gene expression (SAGE). Mol. Plant–Microbe Interact. 15: 847–856.

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

Till, B.J., Reynolds, S.H., Greene, E.G., Codomo, C.A., Enss, L.C., Johnson, J.E., et al. 2003. Large-scale discovery of point mutations with high-throughput TILLING. Genome Res. 13: 524–530.

Tisdall, J. 2003. Mastering PERL for bioinformatics. O'Reilly & Associates, Cambridge, Mass.

Tunlid, A., and Talbot, N.J. 2002. Genomics of parasitic and symbiotic fungi. Curr. Opin. Microbiol. 5: 513–519.

Turgeon, B.G., Kroken, S., Lee, B.-N., Baker, S.E., Amedeo, P., Catlett, N., Gunawardena, U., Wagner, E., Robbertse, B., Wu, J., Yoder, O.C., Glass, N.L., and Taylor, J.W. 2002. Comparative genomic analysis of fungal plant pathogens: secondary metabolites and mechanisms of pathogenesis. *In* Symposium: Functional Genomics of Plant Pathogen Interactions [online]. American Phytopathological Society Annual Meeting, 27–31 July 2002, Milwaukee, Wisc. Available from www.apsnet/org /online/feature/microbe/abstracts.html. [Abstr.]

Tzung, K.W., Williams, R.M., Scherer, S., Federspiel, N., Jones, T., Hansen, N., Bivolarevic, V., Huizar, L., Komp, C., Surzycki, R., Tamse, R., Davis, R.W., and Agabian, N. 2001. Genomic evidence for a complete sexual cycle in *Candida albicans*. Proc. Natl. Acad. Sci. U.S.A., 98: 3249–3253.

Uetz, P., Giot, L., Cagney, G., Mansfield, T.A., Judson, R.S., Knight, J.R., 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, J.M. 2000. A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. Nature (London), 403: 601–603.

Van Sluys, M.A., de Oliveira, M.C., Monteiro-Vitorello, C.B., Miykai, C.Y., Furlan, L.R., Camargo, L.E., et al. 2003. Comparative analyses of the complete genome sequences of Pierce's disease and citrus variegated chlorosis strains of *Xylella fastidiosa*. J. Bacteriol. 185: 1018–1026.

Wagner, A. 2000. Robustness against mutations in genetic networks of yeast. Nature Genet. 24: 355–361.

Wolf, Y.I., Rogozin, I.B., Grishin, N.V., Tatusov, R.L., and Koonin, E.V. 2001. Genome trees constructed using five different approaches suggest new major bacterial clades. BMC Evol. Biol. [serial online] 1: 8. Available from www.biomedcentral.com/1471-2148/1/8.

Wood, D.W., Setubal, J.C., Kaul, R., Monks, D.E., Kitajima, J.P., Okura, V.K., et al. 2001. The genome of the natural genetic engineer *Agrobacterium tumefaciens* C58. Science (Washington, D.C.), 294: 2317–2323.

Wood, V., Gwilliam, R., Rajadream, M.-A., Lyne, M., Lyne, R., Stewart, A., et al. 2002. The genome sequence of *Schizosaccharomyces pombe*. Nature (London), 415: 871–880.

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, R.G., Foley, K., Beheshti, J., Harris, N.L., 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, D.J., Freeman, S., Rodriquez, R.J., and Dickman, M.B. 2003. Fungal biology and agriculture: revisiting the field. Mol. Plant–Microbe Interact. 16: 859–866.

Yoder, O.C., and Turgeon, B.G. 2001. Fungal genomics and pathogenicity. Curr. Opin. Plant Biol. 4: 315–321.

Yu, J., Hu, S., Wang, J., Wang, G.K.S., Li, S.G., Liu, B., et al. 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). Science (Washington), 296: 79–92.

Zeng, Q., Morales, A.J., and Cottarel, G. 2001. Fungi and humans: closer than you think. Trends Genet. 17: 682–684.

Zolan, M.E. 1995. Chromosome-length polymorphism in fungi. Microbiol. Rev. 59: 686–698.