Comparative genomics and gene finding in fungi

Marina Axelson-Fisk and Per Sunnerhagen

Abstract

In the spring of 2005, we had access to 18 fully sequenced fungal genomes, and more are coming rapidly. New approaches and methods are being developed to harvest this information source to derive functional predictions and understanding of genome anatomy. Comparative genomics also tells us stories about the evolution of yeasts and filamentous fungi, and the genome rearrangements that marked their history. For example, several genes encoding proteins required for heterochromatin formation and RNA interference have been lost uniformly throughout the *Hemiascomycetes*, although some genes remain in a few species in a scattered pattern.

Being the first eukaryote to have its genome fully sequenced, *Saccharomyces cerevisiae* was the forerunner for *in silico* methods of genome annotation in general, and gene finding in particular. Lessons learned from the comparatively simple genome of this budding yeast have paved the way for efficient genome analysis in other fungi as well as eukaryotes in general.

Several fungal species are of important applied interest for mankind, and so it is essential to utilise comparative genomics to derive functional information about them. The set of fungal genomes: simple, related in evolution, and with a high density of functional information, can serve as a highly efficient test bed for the further development of comparative genomics.

1 Comparative genomics

Comparative genomics is on the rise as a potent tool in molecular biology. Comparisons of single sequences, protein or nucleic acid, preceded comparisons of whole genomes by two decades. Classical similarity searches of amino acid sequences identified orthologues and paralogues of proteins from widely divergent species, and comparison of ribosomal RNA sequence was used to determine phylogenetic relationships. Since these are among the most highly conserved features that can be derived directly from genomes, comparisons over long evolutionary distances are possible and desirable. More recently, the availability of many fully sequenced genomes has made possible a broad collection of comparative exercises. For instance, study of closely related species allow identification of syntenic blocks in chromosomes, conservation of *cis*-regulatory sequences, spreading of repetitive sequence elements, development of pseudogenes etc.

Topics in Current Genetics, Vol. 15
P. Sunnerhagen, J. Piškur (Eds.): Comparative Genomics
DOI 10.1007/4735_111 / Published online: 11 November 2005

Comparative genomics attains its full power only when experimental genetic and molecular biology data are available from at least one of the species. Prominent cases are mammalian genomes (mouse, rat, and human), where functional data from mouse and human can be drawn upon, and the nematodes *Caenorhabditis elegans vs. C. briggsae.* Among plants, full genome sequences are available from *Arabidopsis thaliana* and rice (*Oryza sativa*), and more genome sequences are underway. Studies of genomes from higher plants face the obstacle of quite differing sizes, ranging from 1.2×10^8 bp (*A. thaliana*) to over 1.5×10^{10} bp (*e.g.* some *Allium* species).

1.1 Comparative genomics of fungi

The publicly available genomes from yeasts and filamentous fungi, 18 at the time of writing, represent a unique resource for comparative genomics, by two arguments. First, a wide range of evolutionary distances is represented, from separation times between 5 and 20 MYr (within the "*Saccharomyces sensu stricto*" group; Kellis et al. 2003) to 600 – 1200 MYr (between basidiomycetes and ascomycetes; Heckman et al. 2001; Douzery et al. 2004). Second, among the 18 species, many are genetically tractable experimental organisms. Thus, it is possible to directly verify inferences from genome comparisons using molecular genetics, opening up a multitude of interesting possibilities. Further, analysis of pathway conservation can reveal if whole signalling or metabolic pathways, or branches thereof, are missing or differently wired in some species (see Chapter 6 by Krantz and Hohmann in this volume).

Comparative genomics of yeasts has been reviewed with emphasis on the protein-coding complements of the different species (Herrero et al. 2003). The potential of comparative genomics of closely related *Saccharomyces* species for identification of regulatory elements has recently been highlighted (Kellis et al. 2004b), and the usefulness of genome sequencing for shedding light on phylogenetic relationships among yeasts has also been emphasised (Piškur and Langkjaer 2004). The purpose of the present volume is to draw attention to the considerable potential of a combination of bioinformatics and experimental approaches utilising information from the many fungal genomes on hand, representing yeast and filamentous fungi.

A highly useful tool for comparative genomics of fungi, FungalBlast, has recently been developed at the Saccharomyces Genome Database (SGD; www.yeastgenome.org/) (Balakrishnan et al. 2005). This takes advantage of all completely or partially sequenced fungal genomes, representing at the time of writing 38 species, and allows parallel searches in these for protein or DNA sequences similar to the query. Other tools, such as the Fungal Alignment (displaying amino acid sequence homologies) and the Synteny Viewer (displaying the gene arrangement around corresponding gene loci) exploit the genome sequences of closely related *Saccharomyces* species. These and other bioinformatics devices developed explicitly for comparisons between fungal genomes have quickly become popular among molecular biologists.

1.2 Relationships between sequenced fungal genomes

From a *Saccharomyces cerevisiae*-centric perspective, the presently sequenced fungal genomes represent a sliding scale from sibling species to quite distant relatives. There is first a set of closely related *Saccharomyces* species (*S. paradoxus, S. mikatae, S. bayanus, S. kudriavzevii*). These are estimated to have diverged between 5 and 20 Myr ago. Extensive studies have been invested into the *Hemiascomycetes* as a whole, comprising the vast majority of known ascomycetous yeast species. Thus, genome sequences are available from *S. castellii, Candida albicans, C. glabrata, Yarrowia lipolytica, Debaryomyces hansenii, Kluyveromyces lactis, K. waltii, Hansenula polymorpha,* and *Ashbya gossypii*. By virtue of its relatedness on the sequence level, *A. gossypii* is classified with the *Hemiascomycetes* despite its predominantly filamentous mode of growth. A summary of what has been observed from genome comparisons within the *Hemiascomycetes* is found in this volume, Chapter 8 by Bolotin-Fukuhara et al.

Small gene families, most often consisting of two to three members, are quite common in many hemiascomycetes. The *S. cerevisiae* genome sequence revealed that the organisation of such duplications was such that blocks of genes often could be mapped to corresponding blocks of seemingly duplicated genes elsewhere in the genome. This prompted the suggestion that a series of large duplication and recombination events were key in shaping of the budding yeast genome (Philippsen et al. 1997; Wolfe and Shields 1997). Direct confirmation of this prediction came recently with the sequences of genomes from fungi that split off from the *Saccharomyces* branch before these duplications took place, namely *Ashbya gossypii* (Dietrich et al. 2004), *Kluyveromyces lactis* (Dujon et al. 2004), and *K. waltii* (Kellis et al. 2004a). Here, it is possible to find relationships between long syntenic blocks of genes in *Saccharomyces* vs. these other non-duplicated species on a 2:1 basis (see Chapter 4 by Wong and Wolfe in this volume). Beside the basic rule that extensive gene duplications are a distinctive feature of the *Saccharomyces sensu lato* group, there are cases where a gene has been duplicated independently in two branches of the fungal tree. Thus, an investigation comparing *S. cerevisiae* and *Sz. pombe* revealed 56 such duplications (Hughes and Friedman 2003). Other examples are pyruvate decarboxylase genes, which have been independently duplicated in *S. cerevisiae* and *S. kluyveri* (Møller et al. 2004), and genes encoding mitochondrial ADP/ATP carriers in *S. cerevisiae* and *Y. lipolytica* (Mentel et al. 2005). Obviously, assignment of orthologous relationships is often ambiguous in such cases (see section on orthologue mapping in Chapter 10 by Wood).

Representatives of two more subclasses of *Ascomycetes* have been fully sequenced. The fission yeast *Schizosaccharomyces pombe* (see Chapter 10 by Wood), a widely used experimental organism, belongs to the *Archiascomycetes*. The genus *Schizosaccharomyces* has only three characterised species, and no other close relatives are known. The fission yeasts are thought to lack many of the special evolutionary adaptations of the *Hemiascomycetes*. Several ascomycetous filamentous fungi, classified in *Euascomycetes*, (*Aspergillus nidulans*, *Giberella zeae* [a.k.a. *Fusarium graminearum*], *Magnaporthe grisea*, *Neurospora crassa*) have been fully sequenced. Some of these (*A. nidulans*, *N. crassa*) are important genetic model organisms with a long scientific history. The larger complexity of the filamentous lifestyle is reflected in a gene number about twice as high as in the typical yeast (Table 1). In contrast to the recently diverged *Saccharomyces* species, the split between these three branches of *Ascomycetes* (*Hemiascomycetes*, *Archiascomycetes*, and *Euascomycetes*) took place as long as 0.3 – 1 GYr ago (Maddison 1997; Sipiczki 2000; Heckman et al. 2001; Douzery et al. 2004), thus comparable to the distance separating vertebrates and arthropods.

Even further away on the evolutionary scale are the basidiomycetes. Genome sequences are available from *Phanerochaete chrysosporium* (a filamentous fungus causing white-rot of wood) and *Ustilago maydis* (a maize pathogen with a multicellular as well as a unicellular, yeast-like, life phase). This is also the basidiomycete with the best-studied genetics. The full genome sequence is available from *Cryptococcus neoformans*, a yeast pathogenic for humans. It should be noted that the concept of yeasts is operational, since unicellular fungi occur both among ascomycetes and basidiomycetes. The predominant theory for the evolution of ascomycetous yeasts is by evolution from filamentous ancestors (Liu and Hall 2004); for the basidiomycetous yeasts, such a tracing of evolutionary history is less apparent. *Coprinopsis cinerea*, a free-living mushroom that can be cultivated in defined medium and for this reason has permitted genetic analysis, has also been extensively sequenced.

Finally, there is one sequenced representative of *Microsporidia*, for which the relationships to other major classes of fungi have long remained unresolved, that of the intracellular parasite *Encephalitozoon cuniculi*. The impact of genomic sequencing on the phylogeny of fungi is treated in Chapter 2 in this volume by Piškur and Kurtzman. On the other hand, phylogenetic advances can impact genome sequencing by suggesting new species to be sequenced. For instance, can we map more closely the point where a whole-genome duplication event took place within the *Hemiascomycetes*?

1.3 Properties of sequenced fungal genomes

Compared to those of higher plants and animals, the presently sequenced fungal genomes are compact; the gene density exceeds 0.26 per kb (Table 1). Consequently, intergenic regions are short. Introns, which are a rare commodity in budding yeast genomes, are more frequent in other fungal genomes. However, fungal introns tend to be short even where they are numerous (Table 1). Repetitive sequences, which make up a major fraction of vertebrate DNA, are low in abundance. Transposable elements, both DNA transposons and retroelements, are found in all branches of the fungal kingdom (Daboussi 1997). The genome of the fully sequenced basidiomycetes, *Cryptococcus neoformans*, reveals a considerably higher abundance of both introns and transposable elements than seen in ascomycetes (Loftus et al. 2005).

Table 1. Key features of some sequenced and annotated fungal genomes

	Genome size (Mb)	Gene number	Genes per kb	Average gene length	Fraction coding DNA	Average intergenic length	Introns per gene	Average length introns
Saccharomyces cerevisiae	12.5	5777	0.46	1460	0.71	515	0.05	216
Kluyveromyces waltii	10.6	5230	0.49					
Candida albicans	14.9	6419	0.43		0.73		0.42	
Debaryomyces hansenii	12.2	6906	0.57		0.79			
Ashbya gossypii	8.7	4718	0.54		0.80	341	0.05	
Schizosaccharomyces pombe	14.1	4973	0.35	1430	0.57	952	0.95	78
Gibberella zeae	35.6	11640	0.33	1744	0.50	1301	2.22	93
Magnaporthe grisea	38.8	11108	0.29	1683	0.41	1503	1.89	143
Neurospora crassa	38.9	10082	0.26	1673	0.38	1953	1.70	135
Aspergillus nidulans	30.1	9457	0.31	1882	0.51	1266	2.69	101
Cryptococcus neoformans	19.1	6572	0.34	2195	0.51	916	5.30	67
Ustilago maydis	19.7	6522	0.33	1935	0.61	1040	0.75	127
Encephalitozoon cuniculi	2.5	1996	0.80		0.86			

Numbers were derived from the following sources: *Saccharomyces cerevisiae*, Chapter 10 by Wood (this volume); *Kluyveromyces waltii*, Kellis et al. (2004b); *Candida albicans*, Jones et al. (2004); *Debaryomyces hansenii*, www.ebi.ac.uk/2can/genomes/eukaryotes/Debaryomyces_hansenii.html; *Ashbya gossypii*, Dietrich et al. (2004); *Schizosaccharomyces pombe*, Chapter 10 by Wood (this volume); *Gibberella zeae* (=*Fusarium graminearum*) http://www.broad.mit.edu/annotation/fungi/fusarium/; *Magnaporthe grisea*, www.broad.mit.edu/annotation/fungi/magnaporthe/; *Neurospora crassa*, www.broad.mit.edu/annotation/fungi/neurospora_crassa_7/index.html; *Aspergillus nidulans*, www.broad.mit.edu/annotation/fungi/aspergillus/; *Cryptococcus neoformans*, Loftus et al. (2005); *Ustilago maydis* www.broad.mit.edu/annotation/fungi/ustilago_maydis/; *Encephalitozoon cuniculi*, Katinka et al. (2001)

Throughout, there is a clear trend that unicellular organisms have smaller and more compact genomes than multicellular organisms. Among the presently sequenced fungal genomes, there are representatives for both free-living unicellular and filamentous species. Also here, the obvious tendency is for the unicellular organisms (the yeasts) to have the more compact genomes; the average genome size