Lawrence Berkeley National Laboratory

Recent Work

Title

Fungal Genomic Annotation

Permalink

<https://escholarship.org/uc/item/9m98z4pk>

Author

Salamov, Asaf A.

Publication Date

2006

Fungal Genomic Annotation

Igor V. Grigoriev¹, Diego A. Martinez² and Asaf A. Salamov¹

¹US Department of Energy Joint Genome Institute, Walnut Creek, CA 94598

(ivgrigoriev@lbl.gov, AASalamov@lbl.gov); ²Los Alamos National Laboratory Joint Genome Institute, P.O. Box 1663 Los Alamos, NM 87545 (admar@lanl.gov).

Sequencing technology in the last decade has advanced at an incredible pace. Currently there are hundreds of microbial genomes available with more still to come. Automated genome annotation aims to analyze this amount of sequence data in a high-throughput fashion and help researchers to understand the biology of these organisms. Manual curation of automatically annotated genomes validates the predictions and set up 'gold' standards for improving the methodologies used. Here we review the methods and tools used for annotation of fungal genomes in different genome sequencing centers.

1. INTRODUCTION

In recent years the power of DNA sequencing has dramatically increased, with dedicated centers running 24 hours a day 7 days a week able to produce as much as 2 gigabases of raw sequence or more a month. The researchers who work on a variety of fungi are fortunate, as most fungal genomes are under 50 megabases and produce high-quality draft assembly almost as easily as bacteria. This feature of fungal genomes is a key reason that the first sequenced eukaryotic genome was of the ascomycete *Saccharomyces cerevisiae* (Goffeau et al. 1996). As of the submission of this chapter, one can obtain draft sequences of more than 100 fungal genomes (Table 1) and the list is growing. While some are species of the same genus (e.g., *Aspergillus* has three members and more coming), there still remains a height of data that could confuse and bury a researcher for many years.

Large-scale fungal genome annotation and analysis started after the sequencing of the yeast *S. cerevisiae* was completed (Goffeau et al. 1996), followed by another yeast *Schizosaccharomyces pombe* (Wood et al. 2002). This period also saw the first filamentous fungi *Neurospora crassa* (Galagan et al. 2003), the first basidiomycete genome of *Phanerochaete chrysosporium* (Martinez et al. 2004) and, through the Phytophthora Genome Initiative (Waugh et al. 2000), the first oomycetes, *Phytophthora sojae* and *Phytophthora ramorum*, were sequenced (genome.jgi-psf.org/sojae1 and genome.jgi-psf.org/ramorum1). Large genome sequencing centers have begun to focus some of their sequencing capacity on the fungal kingdom. One such center, the Joint Genome Institute (JGI) (www.jgi.doe.gov), started the sequencing and annotation of fungi with the whiterot genome (*P. chrysosporium*) over two years ago and now has approximately 20 genomes in various stages of the sequencing and annotation pipeline. The JGI has

also hosted three fungal annotation *jamborees* (see Section 5.0) for *P. chrysosporium*, *Trichoderma reesei*, and the two *Phytophthora* genomes. Both the Broad Institute and the JGI are set to sequence members of the zygomycetes and the chytridiomycetes. In the 1990s there was a call for many other fungal genomes to be sequenced, and to heed this call, the Fungal Genome Initiative (FGI) (www.broad.mit.edu/annotation/fungi/fgi/) started a coordinated effort on targeted sequencing fungal genomes in a kingdom-wide manner; that is, by selecting a set of fungi that maximizes the overall value through a comparative approach. Currently, from the list of about 40 genomes, 20 were sequenced at the Broad Institute and gene models are available for 7 of those genomes. Unlike the Broad Institute's FGI, the JGI is sequencing individual fungi proposed by researchers world-wide and selected through the Community Sequencing Program (www.jgi.doe.gov/CSP/index.html) on the basis of the organism's scientific and economic importance and through the Department of Energy's microbial genomics program (microbialgenome.org).

The Génolevures Consortium is another large initiative on fungal genomics, focused on large-scale comparative genomics between *S. cerevisiae* and 14 other yeast species representative of the various branches of the Hemiascomycetous class. The consortium sequenced and manually curated the complete genome sequences of four yeast species: *Debaryomyces hansenii*, *Kluyveromyces lactis*, *Candida glabrata*, and *Yarrowia lipolytic*, as well as a number of random genomic libraries (Table 1) (Dujon et al. 2004, Sherman et al. 2004).

To combat the initial problem of making sense of the incredible amount of data, many sequencing centers offer resources to make genomic information more accessible and assist in stimulating research. Collectively these resources are termed annotation. In the field of genomics, the term annotation refers to two types of annotation. The first type, which is performed after assembly, is to locate genes and describe gene structure. This is often termed *structural annotation* or *gene modeling*. In bacteria, this process is relatively straightforward as prokaryotes utilize almost all of their DNA for coding. Of the prokaryotic genomes listed at NCBI (www.ncbi.nlm.nih.gov), the average percentage of coding DNA is 85.5% (C Stubben, personal communication, 2005). For eukaryotes of even small to medium genome sizes, this task can be quite challenging because of the complexity of eukaryotic gene structure and the amount of noncoding DNA. For comparison, the percentage of coding DNA in the whiterot Basidiomycete *P. chrysosporium* is approximately 45%. The second type of annotation is called *functional annotation*. Once the genes have been identified, an attempt is made to identify what the gene does for the cell in a biochemical, structural, signaling, etc. context. This discovery method relies largely on an analysis of the resulting protein.

Table 1. Non-exhaustive list of Genomes and Respective Sequencing Centers important to biotechnology. Also shown are current status and availability of information. For a complete list of genomes, please visit the GOLD database (www.genomesonline.org).

†Also available in the MIPS Pedant genome database.

*Indicates more than one strain from this species has been sequenced. This includes strains sequenced at the same institution.

Sequencing Center and Genome	Sequenced	Annotated	Published	References
European Consortium				
<i>Saccharomyces cerevisiae</i> [†]	+	+	+	(Goffeau et al. 1996)
Sanger Center				
<i>Schizosaccharomyces pombe</i> [†]	+	+	+	(Wood et al. 2002)
Broad Institute/ German Consortium				
<i>Neurospora crassa</i> [†]	+	+	+	(Galagan et al. 2003)
US DOE Joint Genome Institute				
<i>Phanerochaete chrysosporium</i> [†]	+	+	+	(Martinez et al. 2004)
<i>Phytophthora sojae</i>	+	+		
<i>Phytophthora ramorum</i> [†]	+	+		
<i>Trichoderma reesei</i>	+	+		
<i>Pichia stipitis</i>	+			
<i>Laccaria bicolor</i>	+			
<i>Nectria haematococca</i>	+			
<i>Glomus intraradices</i>	+			
<i>Postia placenta</i>	+			
<i>Aspergillus niger</i>	+			
<i>Mycosphaerella graminicola</i>	+			
<i>Sporobolomyces roseus</i>				
<i>Mycosphaerella fijiensis</i>				
<i>Piromyces</i> sp.				
<i>Melampsora larici-populina</i>				
<i>Batrachochytrium dendrobatidis</i>				
<i>Phycomyces blakesleenus</i>				
<i>Xanthoria parietina</i>				
<i>Trichoderma virens</i>				
<i>Phytophthora capsici</i>				
Broad Institute				
<i>Aspergillus nidulans</i> [†]	+	+	+	(Dean et al. 2005)
<i>Chaetomium globosum</i>	+	+		
<i>Fusarium graminearum</i> [†]	+	+		
<i>Magnaporthe grisea</i> [†]	+	+	+	

<i>Stagonospora nodorum</i>	+	+
<i>Ustilago maydis</i> [†]	+	+
<i>Phytophthora infestans</i>		
<i>Botrytis cinerea</i>	+	
<i>Candida guilliermondii</i>	+	
<i>Candida lusitanae</i>	+	
<i>Candida tropicalis</i>	+	
<i>Coccidioides immitis</i> [†]	+	+
<i>Coprinus cinereus</i> [†]	+	+
<i>Cryptococcus neoformans</i> (A) ^{†*}	+	+
<i>Cryptococcus neoformans</i> (B) [*]	+	
<i>Fusarium verticillioides</i>	+	
<i>Rhizopus oryzae</i>	+	
<i>Saccharomyces cerevisiae</i> RM11-1a	+	
<i>Saccharomyces paradoxus</i> [†]		+
<i>Sclerotinia sclerotiorum</i>	+	
<i>Uncinocarpus reesii</i>	+	

Washington University, St Louis

<i>Alternaria brassicicola</i>		
<i>Saccharomyces kudriavzevii</i> [†]	+	+
<i>Saccharomyces mikatae</i> [†]	+	+
<i>Saccharomyces castellii</i> [†]	+	+

TIGR/ Univ. of Manchester / Sanger Centre / Institut Pasteur / Univ. of Salamanca / Nagasaki Univ

<i>Aspergillus fumigatus</i> [†]	+	+	+
---	---	---	---

TIGR/ Stanford Genome Sequencing Center

<i>Cryptococcus neoformans</i> (D)	+	+	+	(Loftus 2003)
------------------------------------	---	---	---	---------------

Stanford Genome Sequencing Center

<i>Candida albicans</i>	+	+	+	(Braun et al. 2005, Jones et al. 2004)
-------------------------	---	---	---	--

Japanese Consortium

<i>Aspergillus oryzae</i>	+	+	+
---------------------------	---	---	---

Genoscope

<i>Kluyveromyces thermotolerans</i> [†]	+	+
--	---	---

<i>Kluyveromyces marxianus</i> var <i>marxianus</i> [†]	+	+	
<i>Kluyveromyces lactis</i> ^{†*}	+	+	
<i>Saccharomyces exiguus</i> [†]	+	+	
<i>Saccharomyces exiguus</i> [†]	+	+	
<i>Saccharomyces servazzi</i> [†]	+	+	
<i>Zygosaccharomyces rouxi</i> [†]	+	+	
<i>Debaryomyces hansenii</i> var <i>hansenii</i> ^{†*}	+	+	
<i>Yarrowia lipolytica</i> ^{†*}	+	+	
<i>Pichia angusta</i> [†]	+	+	
<i>Pichia sorbitophila</i> [†]	+	+	
<i>Candida glabrata</i> [†]	+	+	
<i>Candida tropicalis</i> [†]	+	+	
Broad Institute/Genoscope			
<i>Saccharomyces bayanus</i> ^{†*}	+	+	
Washington University, St Louis/Genoscope			
<i>Saccharomyces kluyveri</i> ^{†*}	+	+	
Syngenta Biotechnology Inc			
<i>Ashbya gossypii</i> [†]	+	+	+

(Dietrich et al. 2004)

2. Gene Discovery in the Fungi

With more genomes, computational methods for genome annotation have evolved and different research groups and centers have developed various gene prediction methods and tools. Nevertheless, it appears that there are no completely automated methods to predict gene models in eukaryotes. Most of the eukaryotic gene predictors have been developed for the human genome or other higher eukaryotes and cannot be used for the annotation of a “random” genome without carefully tuning the parameters for gene prediction. Furthermore, gene modeling algorithms made for complex vertebrate genomes show a marked decrease in accuracy even when applied to other vertebrate genomes (Burset and Guigo 1996) and therefore will likely perform poorly on fungal genomes. Guigo et al. have also shown that gene prediction accuracy drops significantly for draft sequences (Burset and Guigo 1996). The methods that rely on open reading frame ORF compatibility across exons (e.g., Fgenesh (Salamov and Solovyev 2000)) suffer most. Others, such as Grail (Xu et al. 1997) and GeneWise (Birney et al. 2004), allow frameshifts, but then produce a mixture of real genes damaged by sequencing errors and potential pseudogene candidates. This is, however, a useful feature for finding pseudogenes (see Section 3.3).

2.1 Gene Modelers

Genes in eukaryotic genomes can be predicted using a variety of different approaches, including *ab initio*, homology-based, EST-based, and synteny-based methods, the first two of which are the most used approaches, especially in the absence of ESTs or sequences of other closely related genomes. Overall, performance of *ab-initio* gene finding algorithms greatly depends on which species gene structures were used in the generation of modeling parameters. In general, the predicted models will be highly inaccurate if the genome that the gene finding algorithm is applied to is different in gene structure than the genome that the algorithm was trained on (Korf 2004, Salamov 2005). Therefore, one seeks to train a modeling algorithm on as much data from the genome that it is going to be run on.

Gene-specific parameters are generally subdivided into *content-based* and *signal-based*. Content-based parameters describe oligonucleotide compositions of coding, intronic and intergenic sequences, and also such characteristics as distributions of exon and intron lengths specific to a given genome, average number of exons per gene, etc. Many programs, such as GeneMark (Lukashin and Borodovsky 1998), Genscan (Burge and Karlin 1997), and Fgenesh (Salamov and Solovyev 2000), use 5th order Markov chain probabilities for describing oligonucleotide preferences of genomic sequences. Signal-based parameters describe the specific patterns of splice sites, branch points, polypyrimidine tracts and other functional signals that are important for mechanisms of splicing and transcription. They can be modeled by position weight matrices, weight array matrices (generalized multipositional weight matrices) or by some combined features of sequences, implemented for example through neural nets, discriminant functions and other techniques (Solovyev 2002).

Gene modeling parameters are tuned based on a collection of known gene structures for annotated genome. For genomic information, there should be at least several pieces of relatively large (> 50kb) genomic contig sequences, and this is usually available from early stages of genomic sequencing. All known genes from GenBank, full length cDNA, and EST data are then mapped to the genomic sequences, providing coding, intronic and information about splice sites. Exploratory data analysis is then performed, for example removing redundancy in sequences, removing some questionable EST mappings and estimating if enough data is available to make reliable parameter values. A subset of the above information is usually set aside to form a test set from known genes, where prediction accuracies with various methods and parameters can be obtained. From the above it is obvious that the quality of the parameters greatly depends on the number of available known gene structures for a given genome. For example, if the number of known genes is too small for the reliable estimation of the oligonucleotide composition parameters, it is better to use the parameters from other related species from which they were calculated, or at least from organisms with comparable GC content. For some functional signals, such as the TATA box, signal peptides, polyA signals and transcription start sites (TSS), often little species-specific information is known, and thus it is difficult to train them for specific genomes and

only general available data may be used. The investigation of these elements is usually left to the end-user.

If a given genome has a sufficient number of known genes or full-length cDNAs, then all these parameters can be efficiently computed and implemented through existing gene-finding algorithms. This presents a problem for many newly sequenced genomes, including new fungal genomes, where there is a scarcity of high-quality information about gene structures. In such a situation, some glimpses about particular gene structures prevalent in a given genome can be inferred from EST data. EST collections are a significant source of data for annotation (Loftus 2003). They can be either mapped directly, or used in EST-based gene predictors like GrailEXP (Xu et al. 1997), Exonerate by ENSEMBL (Slater and Birney 2005) and EST_MAP (softberry.com). Another source of known genes comes from homology-based gene modeling programs such as GeneWise (Birney et al. 2004) or Fgenesh+ (softberry.com). Homology-based programs rely on close protein homologs, which retain similar exon-intron structures.

In recent years, there has been a trend to sequence and annotate genomes of closely related organisms, some even in the same genus. This rapid increase in the number of complete genomes of closely related organisms allows us to effectively use synteny-based gene prediction methods that predict genes in one genome on the basis of comparison with gene models in another. In the last few years a number of such methods have been developed ((Kellis et al. 2003), in yeast). Although in general they provide a reasonable quality of predicting exons, large-scale genome prediction suffers from chimerism, i.e., linking neighbor models into one long model. Therefore, application of these methods is often limited to correction of gene models. For example, in the annotation of *P. sojae* and *P. ramorum* genomes, Fgenesh2 (softberry.com) was used to correct orthologous gene models predicted by other methods if coverage of the alignment between the orthologs was higher in one protein than in another (Tyler et al., in preparation). Other examples of successful use of these methods include the annotation of two *Aspergillus* genomes by TIGR using TWAIN (Majoros et al. 2005) in combination with TigrScan (Majoros et al. 2004) and annotation of different serotypes of *Cryptococcus neomorphans* genomes using TwinScan (Flicek et al. 2003, Korf et al. 2001) followed by RT-PCR validation (Tenney et al. 2004).

Each gene prediction method has its own advantages and disadvantages. A number of benchmarks of different gene prediction methods on different sets of data have been published. Combining different methods can improve overall quality of gene models. Methods to select entire gene models (e.g., Bayesian framework (Pavlovic et al. 2002)) or assemble model fragments into *de novo* models (e.g., Combiner (Allen et al. 2004)) have been proposed. Annotation pipelines at JGI and the Broad Institute employ the first approach to combine several gene predictors, each of which by itself already maximizes use of available evidence.

2.2 Fungal Gene Structure

The G+C content of genomes is a feature of genomic organization that affects codon usage and other oligonucleotide preferences. Most gene modelers predict more

accurately in low GC regions because they strongly rely on hexamer frequencies to discriminate between coding and noncoding regions (Burset and Guigo 1996). In fungal genomes the G+C content varies greatly from 33% for *Candida albicans* to 57% in *P. chrysosporium*. The number of exons per gene also varies greatly among diverse fungi, from the largely single-exon gene structure of *S.cerevisiae* to the high proportion of multi-exon genes in *C. neoformans*. However, in comparison with metazoan genes, fungal genes have relatively short introns. For example, in *C. neoformans*, preliminary analysis has shown that introns have a very tight distribution around 68bp and therefore, when annotating this genome, authors explicitly coded this 'spiked' intron length distribution in the TWINSKAN program instead of the default geometric distribution used in the original program (Tenney et al. 2004). Kupfer et al. (Kupfer et al. 2004) provided the first comprehensive analysis of introns and splicing sites in five diverse fungi, which included the yeasts *S. cerevisiae* and *S. pombe*; two well-studied Ascomycetes, *A. nidulans* and *N. crassa*; and one Basidiomycete, *C. neoformans*. Based on EST data they found that for all studied fungi more than 98% of all splice sites have the canonical 5'GT ... AG3' donor-acceptor pairs in agreement with vertebrate splice sites. On the other hand, they found that polypyrimidine tracts between the intron 3' end and the branch point are absent in a large fraction (31%–72%) of introns across all studied genomes. Their results also suggest that for some short introns, absent polypyrimidine tracts may be compensated by poly(T) tracts upstream of the branch point.

2.3 Validation of gene predictions

Validation of predicted gene models is an important part of automated annotation. It is not sufficient to determine an average accuracy of gene predictors on the test set of genes. Divergence of fungal genomes makes it impossible to use the same parameters for different genomes and therefore accuracy also varies from genome to genome.

Predicted gene models can be normally validated through either their expression or conservation. Evidence of gene expression can be collected from ESTs/cDNAs overlapping with a gene model, oligonucleotide probes placed on microarrays, or peptides from mass-spectrometry experiments aligned against genomic sequence. Conservation can be inferred from homology of a predicted protein and proteins from other organisms in either hand curated datasets like SwissProt (Boeckmann et al. 2003) or all the proteins in Genbank (www.ncbi.nlm.nih.gov). In addition, the percentage coverage of the alignment of the predicted protein and its best homolog serves as a measure of completeness of the predicted gene model especially in alignments between the orthologs. Independent of gene prediction, the alignment between genomic sequences of two or more closely related organisms can reveal islands of DNA conservation and suggest or confirm location of exons and nonconserved functionally important regions. For this reason the VISTA genome analysis tool (Mayor et al. 2000) became a standard feature of JGI genome annotation.

While the number of gene models supported by either of the aforementioned types of evidence describes overall quality of gene models, knowing the quality of every individual gene model is important for many biologists. Based on the same lines of

evidence all genes are divided into more or less reliable predictions using gene-naming conventions. While the naming conventions vary from place to place, all genes can be divided into three major categories by their functional assignment: (1) higher confidence assignment based on strong homology to protein from GenBank or SwissProt (e.g., TIGR: "known/putative", Broad Institute: "known/conserved hypothetical/hypothetical, similar to"), (2) lower confidence assignment supported by ESTs (e.g., TIGR: "expressed") or weak homology (Broad Institute: hypothetical), and (3) *ab initio* gene predictions without homology or EST support (e.g., TIGR: "hypothetical," Broad Institute: "predicted").

Analysis of the aforementioned lines of evidence may help to elucidate an overpredicted portion of a gene set, i.e. *ab initio* gene models, without any additional support. On the other hand, a conservative approach to genome annotation can cause gene underprediction, which can be assessed given a "core" reference set of genes/functions. However, this is a challenging task. First, generation of such a set requires analysis of large collection of diverse genomes. Second, a lack of a "core" gene in a genome does not necessarily mean underprediction because of (1) the draft nature of genome sequence and a good chance of finding the gene in gaps or unassembled DNA reads, or (2) nonhomologous gene substitution, i.e., recruitment of a different protein to perform the same or similar function. Both of these tasks for the moment can be only addressed by a human curator.

3. FUNCTIONAL ANNOTATION

The promise of genomics to biology is not only to find genes but also to describe the function of each resulting protein. While this set of goals was originally that of the fields of genetics and cell and molecular biology, in the genomic era it takes on a new scope. Of the genomes from Table 1, 40 have been through the gene-modeling process, and several have at least preliminary functional annotations. While many biologists feel that manual annotation is best, and will volunteer to examine the staggering numbers of gene models that are predicted for their organism of interest, (e.g., the manual annotation of *C. albicans*) (Braun et al. 2005) and the continued annotation by the Munich Information Center for Protein Sequences (MIPS) (Mewes et al. 2004), there appears to be a need for a reliable automated functional annotation. The *N. crassa* genome alone contains 4,140 (40%) completely unknown genes. Automated annotation, however, has its problems. In Koonin and Galperin (Koonin and Galperin 2002) there are several humorous examples of automated annotation, of which we should be aware. Finally, we must also ask the question, "can we assign protein function by computational methods?"

3.1 Automated Methods

Most sequencing centers have turned to some form of first pass automated annotation to deal with the numbers of genomes that are being sequenced. This data is usually used by the community to attempt to find a function. We present here various approaches to discover gene function that are used in whole genome projects.

3.1.1 Homologous Relationships and Gene Identity

The attempt to transfer gene function from a known protein to an unknown protein can be a difficult task, as evolution can change the context of what a gene does depending on the environment (Francino 2005) that the organism has been in since the time of speciation.

The general approach is to tease out evolutionary relationships by discovering orthologous and paralogous relationships between protein sequences in whole genomes. Orthologs are genes originating from a single ancestral gene in the last common ancestor of the compared genomes (Fitch 1970, Koonin 2005). Paralogs are genes within the same genome that arose from duplications. While conserved function of the proteins is not a part of the definition of orthology, it would reason that the amino acid conservation is due to functional conservation (Koonin 2005, Storm and Sonnhammer 2002). Such an approach is useful because it is less likely that paralogous genes that have fixed in the population have retained the same function and may have been recruited (Lynch and Conery 2000), thus making their function ambiguous.

The most widely accepted method for inferring orthology is through the analysis of phylogenetic trees. Many robust phylogenetic methods exist for recovering the orthologous relationships between genes from different organisms. These are especially useful for understanding more complicated relationships among groups of related genes, such as paralogs, which may appear as many one-to-one orthologs depending on the time of speciation since duplication. This is, however, usually a manually if not computationally intensive method for understanding related genes. Automation is thus required to efficiently process the quantity of sequences found in whole genomes. There has been some headway in automating phylogenetic analyses (Storm and Sonnhammer 2002, Zmasek and Eddy 2002), but they are still limited because of the complexities involved in building phylogenetic trees.

Because of the complexities and manual analysis involved in phylogenetics, most people use a method that relies on a sequence similarity method often called "mutual best hits" or "bidirectional best hits" to identify putative orthologs. This relationship is calculated with all the proteins in the genome. The logic in performing this is as follows: in two genomes A and B containing genes Xa and Xb, respectively, Xa and Xb are potential orthologs if there is no better alignment to Xa from genome B than Xb, and there is no better alignment to Xb from genome A than Xa (Lee et al. 2002, Overbeek et al. 1999). COGs (Tatusov et al. 2001) extends this approach by requiring that orthologs be from three genomes ("triangles" of proteins termed BeTs) to be considered orthologous, thus ensuring that the gene has persisted through time.

There is an unfortunate caveat to the usefulness of such techniques. In all genomes, there is a large fraction of genes whose function is unknown, for example, in the well-studied filamentous fungi there are 4,140 (41%) genes with no similarity to any protein in GenBank (Galagan et al. 2003). It is immediately apparent that there is a need to develop techniques to identify the function of many thousands of genes in a high-throughput manner.

3.2.1 Annotation in Fungi With Experimental Data

With a dramatic increase in the number of unknown and hypothetical genes being produced from whole genome projects, there is a need to integrate the data from high-throughput experiments into the annotation process. The database for this organism is in the *Saccharomyces* Genome Database (Balakrishnan et al. 2005). One can access the data from a variety of microarray information for many of the approximately 6,000 genes predicted in this yeast. An approach of integrating data in the fashion of SGD will drive fungal research and assist in the search for the function of all the genes in a genome.

With transcriptomics and proteomics we are able to understand under what conditions and times mRNAs accumulate in the cell. The types of studies that appear in the literature for fungi are particularly useful for annotation, as they are often under conditions that are unique to the organism, and likely will give clues to many of the species or fungal-specific genes that are common in databases (Lorenz 2002, Rementeria et al. 2005). It is also possible to create a probe for every exon in the genome, so that the predicted structure of a gene can be verified with useful suggestions on how to correct some gene models (Sims et al. 2004).

Because most functioning genes create proteins it is also possible to describe them with proteomics. In fungi this is often identifying what proteins are secreted, as fungi are important degraders of biomass (Medina et al. 2005, Medina et al. 2004, Vanden Wymelenberg et al. 2005) have symbiotic relationships with roots of agriculturally important plants (Bestel-Corre et al. 2004) and protect plants from other soil-borne microbes (Grinyer et al. 2005, Grinyer et al. 2004). The majority of these studies are again targeting biological niches that are dominated by fungi, and are expected to involve fungal-specific genes.

3.3 Pseudogene Annotation

In all studied genomes, eukaryotic and prokaryotic, there are remnants of genes that are no longer transcriptionally active. These inactivated genes are called *Pseudogenes*, often preceded with the greek letter *psi*. There are two types of pseudogenes that are named for how they arise: *processed* and *nonprocessed*. Processed pseudogenes occur when a normal gene is transcribed, introns removed, and a DNA copy is made from the gene by the reverse-transcriptase enzyme of a retrotransposon. Processed pseudogenes usually do not appear to have introns or regulatory elements and can often have poly-A tails. In addition, this type of pseudogene usually contains disablements over the length, such as frameshifts and stop codons in the coding frame. The second type, nonprocessed pseudogenes, were once genes or were duplications of genes. Like processed pseudogenes they contain disablements; however, nonprocessed pseudogenes often have features that make them appear to be genes. This makes nonprocessed pseudogenes more difficult to identify and they can be listed erroneously as a transcribed gene.

In fungi there are previously described pseudogenes (Borsuk et al. 1988, Fink 1987, Gniadkowski et al. 1991, Metzenberg et al. 1985) which were discovered before the genomic era. The determination of pseudogenization was done by manual analysis. In the postgenomic era however, few researchers have the luxury to analyze the average 10,000 or so genes that may contain the hallmarks of pseudogenes. To keep up with the barrage of genomic data in fungi, it will be necessary to apply automated analyses in discovering pseudogenes. Such techniques have already been developed for humans (Zhang et al. 2003). In the yeast genomes, *S. cerevisiae* and *Schizosaccharomyces pombe*, there are 221 for the former (Harrison et al. 2002) and 33 (Wood et al. 2002) for the latter. For the larger filamentous fungus, *P. chrysosporium* (Martinez et al. 2004) no analysis of pseudogenes has been provided because of ambiguity in their discovery. This is also the case for *N. crassa*, *Magnaporthe grisea* (Dean et al. 2005), and *C. albicans* (Braun et al. 2005, Jones et al. 2004) likely because of the ambiguity of stop codons in draft genomes.

One of the key features of pseudogenes is the appearance of stop codons and frameshifts in the coding region. This is usually found by using GeneWise (Birney et al. 2004) which performs a sensitive alignment to a known gene in order to create a gene model, placing an "X" in the predicted amino acid sequence where a frame shift is likely to have occurred, thus allowing the extension of the gene model beyond what could be a sequencing error. There are other criteria (Zhang and Gerstein 2004); however, the stop appears to be the strongest signal. This is the primary difficulty in finding pseudogenes for many genome projects. The data in whole genome shotgun is of the highest quality of sequencing; the error rate is usually 1 in 10,000 (Martinez et al. 2004) for draft genomes. This means that several hundred genes in each genome could contain frame shifts caused by sequencing error alone. Recently however, Torrents et al. (Torrents et al. 2003) has devised a novel technique in verifying pseudogenes that does not rely on the presence of stops. This method applies the Ka/Ks ratio test (rate of nonsynonymous vs. synonymous substitutions) to decide whether a gene is really a pseudogene. In a recent technique comparison from Zhang and Gerstein (Zhang and Gerstein 2004), with some alteration of parameters, the Torrents technique is able to predict the approximately 14,000 pseudogenes in the human genome that other methods were able to find. With the application of this technique, it now may be possible to identify pseudogenes in draft genomes.

4.0 Annotation Pipelines

The centers involved in fungal annotation use a system of steps in order to produce a final set of gene models and annotation, collectively called a *pipeline*. With this broad variety of methods and tools available for gene prediction it is interesting to understand the practical solutions that have been developed by these centers (Table 1). The overall workflow is similar between the different pipelines and includes a few major steps common to all. These common steps are (1) repeat masking, (2) mapping ESTs/known genes, (3) homologs, (4) gene modeling using different methods sequentially or in parallel and then combining them (see Figure 1), and (5) annotating produced sets of gene models using various domain prediction and homology searches.

Figure 1. Annotation pipeline workflow diagram

The JGI and the Broad Institute both use a similar basic set of gene predictors (Fgenesh (Salamov and Solovyev 2000), Fgenesh+ (softberry.com), and GeneWise (Birney et al. 2004)), but in order to produce a nonredundant set of genes they combine them in a slightly different way. Broad Institute uses a prioritization system weighting various gene predictors on the amount and quality of information that exists and the performance of each algorithm. This system gives first priority to GeneWise models with >90% amino acid identity to the translated genome, the second to Fgenesh+ models with identity between 80% and 90%, and then selects the one with the best homology among Fgenesh, Fgenesh+ and GeneWise predictions. This is a sequential gene prediction procedure. JGI predicts all models independently, utilizing ESTs to correct and expand predicted gene models and add UTR regions, and fixes incomplete models by analysis of local genomic regions. The JGI treats all models equally (except known genes that have a higher weight). The JGI selection procedure analyzes each cluster or locus of overlapping models. The final gene model is chosen according to a

hierarchy of criteria: (1) homology to other proteins, (2) EST support, and (3) length and completeness.

After gene models are predicted, each of them is translated and the predicted proteins are functionally analyzed in terms of functional domains and homologs. Functions are automatically assigned on basis of the best homology hit. Comparison with the specialized databases (e.g., KEGG (Kanehisa et al. 2004)) and functional classification allows one to map the predicted proteins onto metabolic pathways, Gene Ontology and KOG (Tatusov et al. 2003) categories provide the user with multiple entry points into the annotation data. Although implementation of these steps varies, most of the pipeline utilizes Blast or Smith-Waterman searches to find all potential homologs, InterProScan (Mulder et al. 2005) or various domain-search methods to predict domains, and public software (e.g., TMHMM (www.cbs.dtu.dk/services/TMHMM/), SignalP (Bendtsen et al. 2004), and TargetP (Emanuelsson et al. 2000)) for more specialized analysis.

In the CAAT-box package (Frangeul et al. 2004) used for annotation of yeast genomes (Dujon et al. 2004, Sherman et al. 2004), gene prediction and functional annotation are integrated with assembly process. However, genes in CAAT-box are identified simply as ORFs (similar to bacterial gene prediction) and while is acceptable for yeasts with low number of exons (a similar approach was taken for *S.cerevisiae* (Goffeau et al. 1996)) it cannot be used broadly for all yeast genomes or especially fungi in general. Even for yeasts the package was used as a first-pass tool combined with the use of GeneMark in the intragenic regions. A similar combination of tools was used in the annotation of *C. albicans* (Braun et al. 2005)

MIPS (Mewes et al. 2004) provides both structural and functional annotation for many of the genomes listed in Table 1. For all genomes housed at MIPS the automated functional annotation system Pedant (Frishman et al. 2003) is used. The Pedant system performs Blast against known proteins from GenBank and the Funcat database (Ruepp et al. 2004), as well as predicting domains using Interpro (Mulder et al. 2005) and other domain-specific databases. For the genomes *S. cerevisiae*, *N. crassa*, *Ustilago myadis* and *Magnaporthe grisea*, MIPS performs in-depth manual curation and verification of both gene structure (provided by the sequencing centers) and gene function.

5.0 Manual Curation: It Takes A Village

Automated annotation and functional genomics methods have reduced the amount of work needed to turn the data in whole genome projects into useful information. There is however still some amount of error in the results in both automated functional and structural annotation (Bork 2000). To verify the calls made by automatic methods and to add the value of personal knowledge to the information presented, volunteers will manually curate the data. Such a resource currently exists or is under development for all known fungal genomes.

Community annotation usually begins with a conference, often termed “Jamboree,” so named for the original *Drosophila melanogaster* genome annotation conference (Pennisi 2000). The jamboree serves several purposes. The volunteers that will be

manually curating the information are trained how to use the specialized tools. Groups of genes are assigned to individuals, and they will then become the curator of that family of genes or pathways. The group of curators will then proceed to manually verify both automated gene calls as well as automated functional data using custom interfaces that connect to a relational database, usually via the web through a web browser.

Several of the fungal genomes listed in Table 1 are currently being curated or have been curated in this manner. The JGI uses custom software for functional annotation and the Apollo editor (Lewis et al. 2002) for updating gene structure features. The results can be viewed on the web, and include the genomes of the basidiomycete *P. chrysosporium*, the oomycete *Phytophthora* species *sojae* and *ramorum*, and the ascomycete *T. reesei*. The genome of *S. cerevisiae* has one of the oldest databases available on line, the Saccharomyces Genome Database (www.yeastgenome.org). The Broad Institute, an important center for fungal genomes, is in the process of creating an interface for community annotation; however, their automated annotations are available (www.broad.mit.edu/annotation/). Other fungal genomes have employed the community annotation model, such as the *Aspergillus* (www.cadre.man.ac.uk) and *C. albicans* (Braun et al. 2005) communities.. The *Aspergillus* site uses the Ensembl (Hubbard et al. 2005) system, while the *C. albicans* annotation project used the Artemis system (Berriman and Rutherford 2003).

5.0 CONCLUSIONS

The genome of the yeast *S. cerevisiae* was completed and published nearly a decade ago. Further improvements in sequencing technology will provide a rapid explosion in the number of fungal genomes, which will result in a critical mass of data for fungal genomes and is essential for changing annotation strategy, as more genome sequences will provide a better understanding of the individual genomes. It is quite possible that someday soon acquiring the genome of the organism you wish to study will be another tool in the biology lab, akin to a centrifuge. Creating resources and perfecting methods to make sequence information accessible is key to making it useful. There exists a need, however, to be able to compare multiple fungal genomes at one time. Despite a number of rich information resources for individual species there is not a unified fungal genomics resource that allows one to quickly compare a newly sequenced genome against others and get an understanding of commonalities and specifics on all levels from individual genes to families and pathways to whole genome organization. On this front collaboration from all centers and researchers involved need to address the need to create a common interface and work together to produce the best available fungal genomic resource possible.

Acknowledgements: This work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program, and Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48, Lawrence Berkeley National Laboratory under contract No. DE-AC02-05CH11231, and Los Alamos National Laboratory under contract No. W-7405-ENG-36. We would like to thank our colleagues Gary Xie, Jean Challacombe, and Monica Misra for their critical review of this work.

REFERENCES

- Allen JE, Pertea M and Salzberg SL (2004). Computational gene prediction using multiple sources of evidence. *Genome Research* 14 (1):142-148.
- Balakrishnan R, Christie KR, Costanzo MC, Dolinski K, Dwight SS, Engel SR, Fisk DG, Hirschman JE, Hong EL, Nash R, Oughtred R, Skrzypek M, Theesfeld CL, Binkley G, Lane C, Schroeder M, Sethuraman A, Dong S, Weng S, Miyasato S, Andrada R, Botstein D and Cherry JM (2005). *Saccharomyces Genome Database*. <http://www.yeastgenome.org/>
- Bendtsen JD, Nielsen H, von Heijne G and Brunak S (2004). Improved prediction of signal peptides: SignalP 3.0. *Journal of Molecular Biology* 340 (4):783-795.
- Berriman M and Rutherford K (2003). Viewing and annotating sequence data with Artemis. *Briefings in Bioinformatics* 4 (2):124-132.
- Bestel-Corre G, Dumas-Gaudot E and Gianinazzi S (2004). Proteomics as a tool to monitor plant-microbe endosymbioses in the rhizosphere. *Mycorrhiza* 14 (1):1-10.
- Birney E, Clamp M and Durbin R (2004). GeneWise and genomewise. *Genome Research* 14 (5):988-995.
- Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S and Schneider M (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Research* 31 (1):365-370.
- Bork P (2000). Powers and pitfalls in sequence analysis: The 70% hurdle. *Genome Research* 10 (4):398-400.
- Borsuk P, Gniadkowski M, Bartnik E and Stepień PP (1988). UNUSUAL EVOLUTIONARY CONSERVATION OF 5S RIBOSOMAL RNA PSEUDOGENES IN ASPERGILLUS-NIDULANS SIMILARITY OF THE DNA SEQUENCE ASSOCIATED WITH THE PSEUDOGENES WITH THE MOUSE IMMUNOGLOBULIN SWITCH REGION. *Journal of Molecular Evolution* 28 (1-2):125-130.
- Braun BR, van het Hoog M, Enfert C, Martchenko M, Dungan J, Kuo A, Inglis DO, Uhl MA, Hogues H, Berriman M, Lorenz M, Levitin A, Oberholzer U, Bachewich C, Hargus D, Marcil A, Dignard D, Iouk T, Zito R, Frangeul L, Tekaiia F, Rutherford K, Wang E, Munro CA, Bates S, Gow NA, Hoyer LL, Hler G, Morsch, user J, Newport G, Znaidi S, Raymond M, Turcotte B, Sherlock G, Costanzo M, Ihmels J, Berman J, Sanglard D, Agabian N, Mitchell AP, Johnson AD, Whiteway M and Nantel A (2005). A Human-Curated Annotation of the *Candida albicans* Genome. *PLoS Genetics* 1 (1):e1.
- Burge C and Karlin S (1997). Prediction of complete gene structures in human genomic DNA. *Journal of Molecular Biology* 268 (1):78-94.
- Burset M and Guigo R (1996). Evaluation of gene structure prediction programs. *Genomics* 34 (3):353-367.
- Dean RA, Talbot NJ, Ebbole DJ, Farman ML, Mitchell TK, Orbach MJ, Thon M, Kulkarni R, Xu J-R, Pan H, Read ND, Lee Y-H, Carbone I, Brown D, Oh YY, Donofrio N, Jeong JS, Soanes DM, Djonovic S, Kolomiets E, Rehmeier C, Li W, Harding M, Kim S, Lebrun M-H, Bohnert H, Coughlan S, Butler J, Calvo S, Ma L-J, Nicol R, Purcell S, Nusbaum C, Galagan JE and Birren BW (2005). The genome sequence of the rice blast fungus *Magnaporthe oryzae*. *Nature* 434 (7036):980-986.
- Dietrich FS, Voegeli S, Brachat S, Lerch A, Gates K, Steiner S, Mohr C, Pohlmann R, Luedi P, Choi SD, Wing RA, Flavier A, Gaffney TD and Philippissen P (2004). The *Ashbya gossypii* Genome as a Tool for Mapping the Ancient *Saccharomyces cerevisiae* Genome. *Science* 304 (5668):304-307.
- Dujon B, Sherman D, Fischer G, Durrens P, Casaregola S, Lafontaine I, de Montigny J, Marck C, Neuveglise C, Talla E, Goffard N, Frangeul L, Aigle M, Anthouard V, Babour A, Barbe V, Barnay S, Blanchin S, Beckerich JM, Beyne E, Bleykasten C, Boissrame A, Boyer J, Cattolico L, Confaniolieri F, de Daruvar A, Despons L, Fabre E, Fairhead C, Ferry-Dumazet H, Groppi A, Hantraye F, Hennequin C, Jauniaux N, Joyet P, Kachouri R, Kerrest A, Koszul R, Lemaire M, Lesur I, Ma L, Muller H, Nicaud JM, Nikolski M, Oztas S, Ozier-Kalogeropoulos O, Pellenz S, Potier S, Richard GF, Straub ML, Suleau A, Swennen D, Tekaiia F, Wesolowski-Louvel M, Westhof E, Wirth B, Zeniou-Meyer M, Zivanovic I, Bolotin-Fukuhara M, Thierry A, Bouchier C, Caudron B, Scarpelli C, Gaillardin C, Weissenbach J, Wincker P and Souciet JL (2004). Genome evolution in yeasts. *Nature* 430 (6995):35-44.
- Emanuelsson O, Nielsen H, Brunak S and von Heijne G (2000). Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *Journal of Molecular Biology* 300 (4):1005-1016.
- Fink GR (1987). PSEUDOGENES IN YEAST? *Cell* 49 (1):5-6.
- Fitch WM (1970). DISTINGUISHING HOMOLOGOUS FROM ANALOGOUS PROTEINS. *Systematic Zoology* 19 (2):99-113.
- Fliscek P, Keibler E, Hu P, Korf I and Brent MR (2003). Leveraging the mouse genome for gene prediction in human: From whole-genome shotgun reads to a global synteny map. *Genome Research* 13 (1):46-54.
- Francino MP (2005). An adaptive radiation model for the origin of new gene functions. *Nature Genetics* 37 (6):573-577.

- Frangoul L, Glaser P, Rusniok C, Buchrieser C, Duchaud E, Dehoux P and Kunst F (2004). CAAT-Box, contigs-Assembly and Annotation Tool-Box for genome sequencing projects. *Bioinformatics* (Oxford) 20 (5):790-NIL_0758.
- Frishman D, Mokrejs M, Kosykh D, Kastenmuller G, Kolesov G, Zubrzycki I, Gruber C, Geier B, Kaps A, Albermann K, Volz A, Wagner C, Fellenberg M, Heumann K and Mewes HW (2003). The PEDANT genome database. *Nucleic Acids Research* 31 (1):207-211.
- Galagan JE, Calvo SE, Borkovich KA, Selker EU, Read ND, Jaffe D, FitzHugh W, Ma LJ, Smirnov S, Purcell S, Rehman B, Elkins T, Engels R, Wang SG, Nielsen CB, Butler J, Endrizzi M, Qui DY, Ianakiev P, Pedersen DB, Nelson MA, Werner-Washburne M, Selitrennikoff CP, Kinsey JA, Braun EL, Zelter A, Schulte U, Kothe GO, Jedd G, Mewes W, Staben C, Marcotte E, Greenberg D, Roy A, Foley K, Naylor J, Stabge-Thomann N, Barrett R, Gnerre S, Kamal M, Kamvysselis M, Mauceli E, Bielke C, Rudd S, Frishman D, Krystofova S, Rasmussen C, Metznerberg RL, Perkins DD, Kroken S, Cogoni C, Macino G, Catcheside D, Li WX, Pratt RJ, Osmani SA, DeSouza CPC, Glass L, Orbach MJ, Berglund JA, Voelker R, Yarden O, Plamann M, Seller S, Dunlap J, Radford A, Aramayo R, Natvig DO, Alex LA, Mannhaupt G, Ebbole DJ, Freitag M, Paulsen I, Sachs MS, Lander ES, Nusbaum C and Birren B (2003). The genome sequence of the filamentous fungus *Neurospora crassa*. *Nature* 422 (6934):859-868.
- Gniadkowski M, Fiett J, Borsuk P, Hoffmannzacharska D, Stepień PP and Bartnik E (1991). STRUCTURE AND EVOLUTION OF 5S RIBOSOMAL RNA GENES AND PSEUDOGENES IN THE GENUS *ASPERGILLUS*. *Journal of Molecular Evolution* 33 (2):175-178.
- Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, Louis EJ, Mewes HW, Murakami Y, Philippsen P, Tettelin H and Oliver SG (1996). Life with 6000 genes. *Science* 274 (5287):546-&.
- Grinyer J, Hunt S, McKay M, Herbert BR and Nevalainen H (2005). Proteomic response of the biological control fungus *Trichoderma atroviride* to growth on the cell walls of *Rhizoctonia solani*. *Current Genetics* 47 (6):381-388.
- Grinyer J, McKay M, Nevalainen H and Herbert BR (2004). Fungal proteomics: Initial mapping of biological control strain *Trichoderma harzianum*. *Current Genetics* 45 (3):163-169.
- Harrison P, Kumar A, Lan N, Echols N, Snyder M and Gerstein M (2002). A small reservoir of disabled ORFs in the yeast genome and its implications for the dynamics of proteome evolution. *Journal of Molecular Biology* 316 (3):409-419.
- Hubbard T, Andrews D, Caccamo M, Cameron G, Chen Y, Clamp M, Clarke L, Coates G, Cox T, Cunningham F, Curwen V, Cutts T, Down T, Durbin R, Fernandez-Suarez XM, Gilbert J, Hammond M, Herrero J, Hotz H, Howe K, Iyer V, Jekosch K, Kahari A, Kasprzyk A, Keefe D, Keenan S, Kokocinski F, London D, Longden I, McVicker G, Melsopp C, Meidl P, Potter S, Proctor G, Rae M, Rios D, Schuster M, Searle S, Severin J, Slater G, Smedley D, Smith J, Spooner W, Stabenau A, Stalker J, Storey R, Trevanion S, Ureta-Vidal A, Vogel J, White S, Woodward C and Birney E (2005). Ensembl 2005. *Nucleic Acids Research* 33 (January 1):D447-D453.
- Jones T, Federspiel NA, Chibana H, Dungan J, Kalman S, Magee BB, Newport G, Thorstenson YR, Agabian N, Magee PT, Davis RW and Scherer S (2004). The diploid genome sequence of *Candida albicans*. *Proceedings of the National Academy of Sciences of the United States of America* 101 (19):7329-7334.
- Kanehisa M, Goto S, Kawashima S, Okuno Y and Hattori M (2004). The KEGG resource for deciphering the genome. *Nucleic Acids Research* 32 (Database Issue):D277-D280.
- Kellis M, Patterson N, Endrizzi M, Birren B and Lander ES (2003). Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* 423 (6937):241-254.
- Koonin EV (2005). Orthologs, Paralogs, and Evolutionary Genomics. *Annual Review of Genetics* 39 (0):309-338.
- Koonin EV and Galperin MY (2002). Information Sources for Genomics. In: ed. *Sequence - Evolution - Function*. Norwell, Massachusetts. pp. 51-110
- Korf I (2004). Gene finding in novel genomes. *BMC Bioinformatics* 5 59.
- Korf I, Flicek P, Duan D and Brent MR (2001). Integrating genomic homology into gene structure prediction. *Bioinformatics* 17 (90001):140S-148.
- Kupfer DM, Drabenstot SD, Buchanan KL, Lai HS, Zhu H, Dyer DW, Roe BA and Murphy JW (2004). Introns and splicing elements of five diverse fungi. *Eukaryotic Cell* 3 (5):1088-1100.
- Lee Y, Sultana R, Pertea G, Cho J, Karamycheva S, Tsai J, Parvizi B, Cheung F, Antonescu V, White J, Holt I, Liang F and Quackenbush J (2002). Cross-referencing eukaryotic genomes: TIGR Orthologous Gene Alignments (TOGA). *Genome Research* 12 (3):493-502.
- Lewis SE, Searle SMJ, Harris N, Gibson M, Iyer V, Richter J, Wiel C, Bayraktaroglu L, Birney E, Crosby MA, Kaminker JS, Matthews BB, Prochnik SE, Smith CD, Tupy JL, Rubin GM, Misra S, Mungall CJ and Clamp ME (2002). Apollo: a sequence annotation editor. *Genome Biology* 3 (12):research0082.0081 - 0082.0014.

- Loftus B (2003). Genome sequencing, assembly and gene prediction in fungi. In: D. K. Arora and G. G. Khachatourians, ed. *Applied Mycology and Biotechnology*. v. 3: Fungal Genomics. Amsterdam. pp. 65-81
- Lorenz MC (2002). Genomic approaches to fungal pathogenicity. *Current Opinion in Microbiology* 5 (4):372-378.
- Lukashin AV and Borodovsky M (1998). GeneMark.hmm: New solutions for gene finding. *Nucleic Acids Research* 26 (4):1107-1115.
- Lynch M and Conery JS (2000). The evolutionary fate and consequences of duplicate genes. *Science (Washington D C)* 290 (5494):1151-1155.
- Majoros WH, Pertea M and Salzberg SL (2004). TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics (Oxford)* 20 (16):2878-2879.
- Majoros WH, Pertea M and Salzberg SL (2005). Efficient implementation of a generalized pair hidden Markov model for comparative gene finding. *Bioinformatics (Oxford)* 21 (9):1782-1788.
- Martinez D, Larrondo LF, Putnam N, Gelpke MDS, Huang K, Chapman J, Helfenbein KG, Ramaiya P, Detter JC, Larimer F, Coutinho PM, Henrissat B, Berka R, Cullen D and Rokhsar D (2004). Genome sequence of the lignocellulose degrading fungus *Phanerochaete chrysosporium* strain RP78. *Nature Biotechnology* 22 (6):695-700.
- Mayor C, Brudno M, Schwartz JR, Poliakov A, Rubin EM, Frazer KA, Pachter LS and Dubchak I (2000). VISTA: Visualizing global DNA sequence alignments of arbitrary length. *Bioinformatics (Oxford)* 16 (11):1046-1047.
- Medina ML, Haynes PA, Breci L and Francisco WA (2005). Analysis of secreted proteins from *Aspergillus flavus*. *Proteomics* 5 (12):3153-3161.
- Medina ML, Kiernan UA and Francisco WA (2004). Proteomic analysis of rutin-induced secreted proteins from *Aspergillus flavus*. *Fungal Genetics and Biology* 41 (3):327-335.
- Metzenberg RL, Stevens JN, Selker EU and Morzykawroblewska E (1985). IDENTIFICATION AND CHROMOSOMAL DISTRIBUTION OF 5S RIBOSOMAL RNA GENES IN *NEUROSPORA-CRASSA*. *Proceedings of the National Academy of Sciences of the United States of America* 82 (7):2067-2071.
- Mewes HW, Amid C, Arnold R, Frishman D, Gueldener U, Mannhaupt G, Muensterkoetter M, Pagel P, Strack N, Stuempflen V, Warfsmann J and Ruepp A (2004). MIPS: Analysis and annotation of proteins from whole genomes. *Nucleic Acids Research* 32 (Database Issue):D41-D44.
- Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bradley P, Bork P, Bucher P, Cerutti L, Copley R, Courcelle E, Das U, Durbin R, Fleischmann W, Gough J, Haft D, Harte N, Hulo N, Kahn D, Kanapin A, Krestyaninova M, Lonsdale D, Lopez R, Letunic I, Madera M, Maslen J, McDowall J, Mitchell A, Nikolskaya AN, Orchard S, Pagni M, Pointing CP, Quevillon E, Selengut J, Sigrist CJA, Silventoinen V, Studholme DJ, Vaughan R and Wu CH (2005). InterPro, progress and status in 2005. *Nucleic Acids Research* 33 (January 1):D201-D205.
- Overbeek R, Fonstein M, D'Souza M, Pusch GD and Maltsev N (1999). The use of gene clusters to infer functional coupling. *Proceedings of the National Academy of Sciences of the United States of America* 96 (6):2896-2901.
- Pavlovic V, Garg A and Kasif S (2002). A Bayesian framework for combining gene predictions. *Bioinformatics (Oxford)* 18 (1):19-27.
- Pennisi E (2000). Ideas fly at gene-finding jamboree. *Science* 287 (5461):2182-+.
- Rementeria A, Lopez-Molina N, Ludwig A, Vivanco AB, Bikandi J, Ponton J and Garaizar J (2005). Genes and molecules involved in *Aspergillus fumigatus* virulence. *Revista Iberoamericana de Micologia* 22 (1):1-23.
- Ruepp A, Zollner A, Maier D, Albermann K, Hani J, Mokrejs M, Tetko I, Guldener U, Mannhaupt G, Munsterkotter M and Mewes HW (2004). The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. *Nucleic Acids Research* 32 (18):5539-5545.
- Salamov AA (2005). unpublished observations.
- Salamov AA and Solovyev VV (2000). Ab initio gene finding in *Drosophila* genomic DNA. *Genome Research* 10 (4):516-522.
- Sherman D, Durrens P, Beyne E, Nikolski M and Souciet J-L (2004). Genolevures: Comparative genomics and molecular evolution of hemiascomycetous yeasts. *Nucleic Acids Research* 32 (Database Issue):D315-D318.
- Sims AH, Gent ME, Robson GD, Dunn-Coleman NS and Oliver SG (2004). Combining transcriptome data with genomic and cDNA sequence alignments to make confident functional assignments for *Aspergillus nidulans* genes. *Mycological Research* 108 (Part 8):853-857.
- Slater GSC and Birney E (2005). Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* 6 31.
- Solovyev VV (2002). Structure, Properties and Computer Identification of Eukaryotic genes. In *Bioinformatics from Genomes to Drugs*. Germany: Wiley-VCH. pp
- Storm CEV and Sonnhammer ELL (2002). Automated ortholog inference from phylogenetic trees and calculation of orthology reliability. *Bioinformatics (Oxford)* 18 (1):92-99.

- Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov S, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ and Natale DA (2003). The COG database: An updated version includes eukaryotes. *BMC Bioinformatics* 4 (41):
- Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, Rao BS, Kiryutin B, Galperin MY, Fedorova ND and Koonin EV (2001). The COG database: New developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Research* 29 (1):22-28.
- Tenney AE, Brown RH, Vaske C, Lodge JK, Doering TL and Brent MR (2004). Gene prediction and verification in a compact genome with numerous small introns. *Genome Research* 14 (11):2330-2335.
- Torrents D, Suyama M, Zdobnov E and Bork P (2003). A genome-wide survey of human pseudogenes. *Genome Research* 13 (12):2559-2567.
- Vanden Wymelenberg A, Sabat G, Martinez D, Rajangam AS, Teeri TT, Gaskell J, Kersten PJ and Cullen D (2005). The *Phanerochaete chrysosporium* secretome: Database predictions and initial mass spectrometry peptide identifications in cellulose-grown medium. *Journal of Biotechnology* 118 (1):17-34.
- Waugh M, Hraber P, Weller J, Wu YH, Chen GH, Inman J, Kiphart D and Sobral B (2000). The *Phytophthora* Genome Initiative database: Informatics and analysis for distributed pathogenomic research. *Nucleic Acids Research* 28 (1):87-90.
- 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, Collins M, Connor R, Cronin A, Davis P, Feltwell T, Fraser A, Gentles S, Goble A, Hamlin N, Harris D, Hidalgo J, Hodgson G, Holroyd S, Hornsby T, Howarth S, Huckle EJ, Hunt S, Jagels K, James K, Jones L, Jones M, Leather S, McDonald S, McLean J, Mooney P, Moule S, Mungall K, Murphy L, Niblett D, Odell C, Oliver K, O'Neil S, Pearson D, Quail MA, Rabinowitsch E, Rutherford K, Rutter S, Saunders D, Seeger K, Sharp S, Skelton J, Simmonds M, Squares R, Squares S, Stevens K, Taylor K, Taylor RG, Tivey A, Walsh S, Warren T, Whitehead S, Woodward J, Volckaert G, Aert R, Robben J, Grymonprez B, Weltjens I, Vanstreels E, Rieger M, Schafer M, Muller-Auer S, Gabel C, Fuchs M, Fritz C, Holzer E, Moestl D, Hilbert H, Borzym K, Langer I, Beck A, Lehrach H, Reinhardt R, Pohl TM, Eger P, Zimmermann W, Wedler H, Wambutt R, Purnelle B, Goffeau A, Cadieu E, Dreano S, Gloux S, Lelaure V, Mottier S, Galibert F, Aves SJ, Xiang Z, Hunt C, Moore K, Hurst SM, Lucas M, Rochet M, Gaillardin C, Tallada VA, Garzon A, Thode G, Daga RR, Cruzado L, Jimenez J, Sanchez M, del Rey F, Benito J, Dominguez A, Revuelta JL, Moreno S, Armstrong J, Forsburg SL, Cerrutti L, Lowe T, McCombie WR, Paulsen I, Potashkin J, Shpakovski GV, Ussery D, Barrell BG and Nurse P (2002). The genome sequence of *Schizosaccharomyces pombe*. *Nature (London)* 415 (6874):871-880.
- Xu Y, Mural RJ and Uberbacher EC (1997). Inferring gene structures in genomic sequences using pattern recognition and expressed sequence tags. *Fifth International Conference on Intelligent Systems for Molecular Biology*. Halkidiki, Greece, p.344-353
- Zhang ZL and Gerstein M (2004). Large-scale analysis of pseudogenes in the human genome. *Current Opinion in Genetics & Development* 14 (4):328-335.
- Zhang ZL, Harrison PM, Liu Y and Gerstein M (2003). Millions of years of evolution preserved: A comprehensive catalog of the processed pseudogenes in the human genome. *Genome Research* 13 (12):2541-2558.
- Zmasek CM and Eddy SR (2002). RIO: Analyzing proteomes by automated phylogenomics using resampled inference of orthologs. *BMC Bioinformatics* 3 (14):

This work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48, Lawrence Berkeley National Laboratory under contract No. DE-AC02-05CH11231 and Los Alamos National Laboratory under contract No. W-7405-ENG-36.