Annotation of Fungal Genomes

Diego Martinez, Igor Grigoriev, Asaf A. Salamov*
Department of Energy, Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA 94598, USA

We describe the annotation process for fungal genome sequencing projects, with special emphasis on protein-coding genes. The characteristics of gene structures in various fungal phyla and strength and weaknesses of the available gene prediction programs are discussed. The automated pipelines used for annotation of fungal genomes at various large-scale sequencing centers are also reviewed.

Keywords: *gene prediction, fungal genomes, automatic annotation pipelines*

INTRODUCTION

Fungal large-scale genome annotation and analysis started after the sequencing of the yeast *Saccharomyces 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 *N.crassa* (Galagan et al., 2003), the first basidiomycete genome of *P.chrysosporium* (Martinez et al., 2004) and through the Phytophthora Genome Initiative (Waugh, 2000) the first oomycetes were sequenced, from the economically important genus *Phytophthora*, *P.sojae* and *P.ramorum* (Tyler et al., 2006). Both the Broad and the JGI are also begun to sequence members of the zygomycetes and the chytridiomycetes. In the 1990's there was a call for many other fungal genomes to be sequenced, and to heed this call, the Fungal Genome Initiative (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 50 genomes, 27 were sequenced at Broad/MIT.

Other large genome sequencing centers have begun to focus some of their sequencing capacity on the fungal kingdom. One such center, the Joint Genome Institute (www.jgi.doe.gov), started the sequencing and annotation of fungi with the whiterot genome (*P.chrysosporium*) over five years ago and now has approximately 20 genomes at various stages of the sequencing and pipeline and has hosted several fungal annotation *jamborees*. Unlike the Broad Center'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 basis of their scientific and economic importance.

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

Many sequencing centers offer resources to make genomic information more accessible and assist in stimulating research, collectively termed annotation. In the field of genomics, the term annotation refers to two types of annotation. The first type, gene modeling, is performed after assembly, to locate genes and describe gene structure. For fungi, as for other eukaryotes this task can be quite challenging due to the complexity of eukaryotic gene structure and the amount of non-coding DNA. The second phase of annotation is called *functional annotation*. It is based largely on an analysis of the resulting protein.

Features of Fungi gene structures

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 non-coding regions. In fungal genomes the G+C content varies greatly from 33% for *C.albicans* to 57% of *P.chrysosporium.* The number of exons per gene also varies greatly among diverse fungi, from the largely single-exon gene structure of *S.cerevisae* to the high number of multi-exon genes in *C.neoformans*. However, in comparison with metazoan genes, fungal

*E-mail: asaf.salamov@gmail.com

genes have relatively short introns. In addition many fungal introns lengths fall into a narrow range of 50 to 70 bp. For example in *C.neoformans*, preliminary analysis have shown that introns have a very tight distribution around 68 bp and therefore for annotating this genome, authors explicitly coded this 'spiked' intron length distribution in the TWINSCAN program, instead of the default geometric distribution used in original program (Tenney et al., 2004). Kupfer et al. (2004) provided the first comprehensive analysis of introns and splicing sites in five diverse fungi, which included yeasts *S.cerevisae* and *S.pombe*, 2 well-studied Ascomycetes: *A.nidulans* and *N.crassa* and one Basidmycete: 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 3'ss and the branch point are absent in 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.

Gene finding in Fungi

Genes in eukaryotic genomes can be predicted using a variaty of different approaches, including *ab initio*, homology-based, EST-based, and synteny-based methods, and the first two of which are the most used approaches especially in 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 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; A.S. unpublished observations). 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 on content-based and signal-based. Content-based parameters describe oligonucleotide compositions of coding, intronic and intergenic sequences and also such characteristics as distribitions of exon and intron lengths specific for a given genome, average number of exons per gene, etc. Many programs, like HmmMark, Genscan and Fgenesh 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, polypirimidine tracts and other functional signal, important for mechanisms of splicing and transcription. They can be modeled by position weight matrices, weight array matrices (generalized multi-positional weight matrices) or by some combined features of sequences, implemented for example through neural nets, discriminant functions and other techniques. Last year brought a new generation of gene predictions methods based on conditional random fields (CRF), which promise to improve *de nove* gene finding (Brent, 2008).

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. They can be either mapped directly, or used in EST-based gene predictors like GrailEXP or Exonerate. Since most gene predictors predict only CDS, JGI uses ESTs for gene model extension (e.g., adding UTR) and validation. Similarly protein data can be used for finding genes where relatively little gene information is known. This data set usually comes from close protein homologs, using homology-based gene-predictors such as GeneWise (Birney and Durbin, 2000) or Fgenesh+ (www.softberry.com).

Gene modeling parameters are tuned based on a collection of information. For genomic information, there should be at least several pieces at least some relatively large (> 50kb) genomic contig sequences, and this is usually available from early stages of genomic sequencing. All known genes from Genbank, FL 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 the reliable values of the parameters needed. 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 estimated. From the above it is obvious that the quality of estimated parameters greatly depends on the number of available known gene structures for a given genome. For example, if number of known genes is quite small for reliable estimation of oligonucleotide composition, it is better to use those parameters from other related species, for which they were calculated, or at least from organisms with comparable GC content. For some functional signals, like the TATA-box, signal peptides, polyA signals and transcription start sites (TSS), little species-specific information is known, and is thus difficult to train them for specific genomes and only general available data might be used, and is usually left to the end-user of the information to find.

In recent years there has been a trend to sequence and annotate genomes of closely related organisms, some even in the same genus. This rapidly increasing 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 basis of comparison with models in another. In the last few years a number of such methods have been developed (Manolis et al., 2005). 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 *Phytophthora sojae* and *Phytophthora ramorum* genomes, Fgenesh2 (www.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., 2006). Other examples of successful use of these methods include the annotation of two *Aspergillus* genomes by TIGR using TWAIN (Majoros and Salzberg, 2005) in combination with TigrScan and annotation of different serotypes of *Cryptococcus neomorphans* genome using TwinScan, followed by RT-PCR validation (Cawley, 2001).

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 (Guigo et al., 1996; Yao, 2005). Combining different methods can improve overall quality of gene models. There are two traditional ways to do it. One is to combine different types of evidence and assemble models from signals coming from these different types of evidence (f.e. Eugene (Foissac et al., 2003) and Combiner (Allen et al., 2004)). Another is to choose a model predicted by one of multiple methods without changing model structure (e.g., Bayesian framework (Pavlovic et al., 2001)). The former is better in situations when most of evidence comes from experimental data (ESTs, homology, etc.). The latter wins when someone tries to combine different gene predictors, almost each of which already maximally utilizes available evidence.

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. Divergency of fungal genomes makes it impossible to use the same parameters for different genomes and therefore accuracy also varies from a genome to genome.

Two types of evidence can be used to computationally validate predicted gene models (ii) gene/protein conservation and (i) indication of gene/protein expression. Homology of a predicted protein to proteins from other organisms – either hand curated datasets like SwissProt, or all proteins in NCBI non-redundant set – described for example in terms of alignment coverage for both predicted protein and its best homolog can serve as a measure of completeness of predicted gene model, especially, when we consider alignments between the orthologs. In the case of two or more closely related genomes, independently of gene prediction, comparison of DNA sequence can support predicted genes, regions of conservation in these DNA alignments indicate location of exons and non-coserved functionally important regions. VISTA genome conservation became the standard feature of JGI genome annotation (Mayor et al., 2000). Gene expression can be described in terms of EST/cDNA coverage, microarray oligos, or peptides from mass-spec experiments aligned against genomic sequence.

While number of gene models supported by either of above mentioned types of evidence describes overall quality of gene models, knowing quality of every individual gene model is important for a biologist. 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 varies from place to place, all genes can be divided into 3 major categories by their functional assignment (i) 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'), (ii) lower confidence assignment supported by ESTs (TIGR: 'expressed') or weak homology (Broad Institute: 'hypothetical'), and (iii) *ab initio* gene predictions without homology or EST support (TIGR: 'hypothetical', Broad Institute: 'predicted').

Analysis of above mentioned lines of evidence may help to elucidate an overpredicted portion of genes set, i.e. *ab inito* gene models without any additional support. On the other hand, conservative approach to genome annotation can cause gene underprediction, which can be assessed given a 'core' reference set of genes/functions. This is however a challenging task. First, generation of such a set requires analysis of large collection of diverse genomes. Second, lack of a 'core' gene in a genome does not necessary means underprediction because of (i) draft nature of genome sequence and a good chance to find the gene in gaps or unassembled DNA reads, or (ii) non-homologous gene substitution, i.e. recruitment of a different protein to perform same or similar function. Both of these tasks for the moment can be only addreassed by human curator.

Automated functional Annotation in Fungi

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

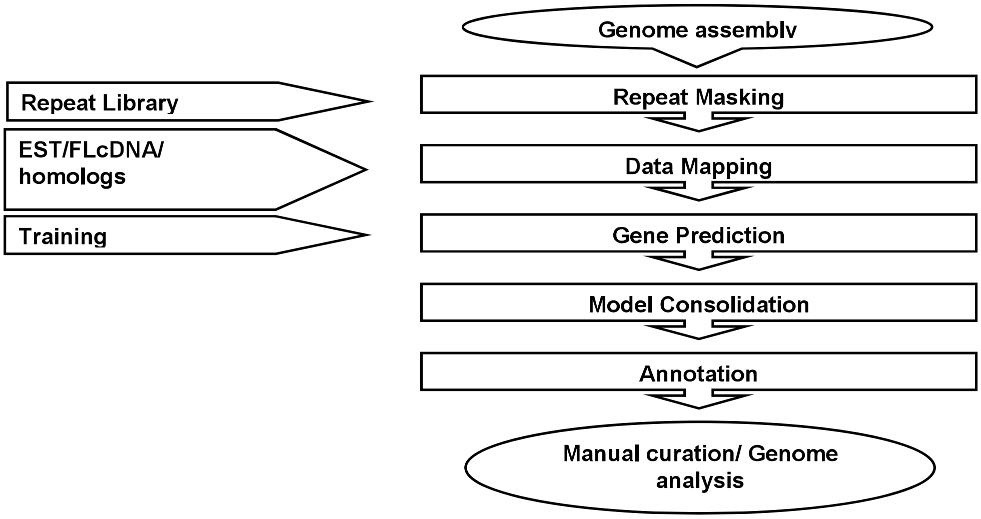

Figure 1. Annotation pipeline workflow

One method is to tease out evolutionary relationships by distinguishing orthologs and paralogs from sequences with some amount of identity in whole genomes. Functional annotation by the identification of orthologs is a step in the direction of high quality annotations. Orthologs are defined as a genes descended from common ancestor, as first defined by Fitch (1970). Alignments to other proteins in the genome are possibly also related, however, they are likely originated from inter-genome duplications. Such genes are termed 'paralogs'. There are several ways of finding orthologous and paralogous relationships. While function of the protein is not necessarily a part of the definition of orthology, it would reason that the conservation is due to a conserved function preserved through time (Storm and Sonnhammer, 2002; Koonin, 2005).

Methods of inferring orthology are mostly based on the analysis of phylogenetic trees. There are have been some progress in automated ortology/paralogy discovery (Storm and Sonhammer 2002; Zmasek and Eddy, 2002), but they are still limited because of complexities in building phylogenetic trees.

Experimental support of annotation

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. In organisms such as yeast, this technique has been used to identify possible function/process involvement of many unknown genes (Uetz et al., 2000; Hazbun and Fields, 2001; Ito et al., 2001; Gavin et al., 2002; Ho et al., 2002). This approach has gained wide acceptance in other organisms, including the filamentous fungi. Data derived from these techniques, so-called transcriptomics and proteomics, is valuable information in the field of annotation.

With transcriptomics we are able to understand under what conditions and times mRNAs accumulate in the cell. The power of this technique is apparent in Sims et al. (2004a). This also includes the approach to create a probe for every predicted exon, so that it is possible to verify the automated gene-structure prediction, with useful suggestions on how to correct some gene models. Transcriptomic studies in pathogenesis are important as it is possible that genes that are organism or fungal specific (thus listed as hypothetical and function unknown) may be involved in processes that are unique to the organism or fungi, providing rare information (Rementeria, 2005).

Since most functioning genes create proteins it is also possible to describe them with proteomics. In fungi this is usually understanding what proteins are secreted, as fungi are important degraders of biomass, have symbiotic relationships with roots of agriculturally important plants (Martin et al., 2008) and protect plants from other soil-borne microbes (Grinyer, 2005). A search of the available literature will lead one to the realization that what is studied with the above techniques is usually an investigation into the unique abilities and processes of fungi. This is a key benefit to annotation, as one might expect some of the unknown and hypothetical genes to be involved in these unique fungal abilities.

Pseudogenes

In all studied genomes, eukaryotic and prokaryotic, there are remnants of genes that are no longer transcriptionally active. These inactivated genes are called *pseudogenes*. There are two types of pseudogenes that are named for how they arise, *processed* and *non-processed*. Processed pseudogenes occurr 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, regulatory elements and can often have poly-A tails. In addition, this type of pseudogene usually contains disablements over the length, such as stop codons in the coding frame. The second type, *non-processed pseudogenes*, was once genes or was duplications of genes. Like processed pseudogenes they contain disablements, however, pseudogenes of the non-processed type often have features that make them appear to be genes. This makes non-processed psuedogenes can be more difficult to identify and can be listed erroneously as a transcribing gene.

Pseudogene Discovery

Finding pseudogenes can be difficult. One of the key features of pseudogenes is the appearance of stop codons 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, thus allowing the extension of the gene model beyond what could be a sequencing error. There are other criterions (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 target quality for Whole Genome Shotgun projects is phred q20 minimum. This means that the minimum quality is an assembly error of 1 bases every 10,000. While the probability is low that the error will result in a mutation, it is still difficult to tell if the likely error is a legitimate stop or a sequencing error.

Recently, 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 synonymous vs. non-synonymous substitutions) to decide whether a gene is really a pseudogene. In a recent technique comparison from Zhang and Gerstein (2004), with some alteration of parameters the technique is able to predict the standard 14,000 or so pseudogenes in the human genome. Application of this technique to draft genomes, gives promise to correct identification of pseudogenes.

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*.

The overall workflow is similar between the different pipelines and includes a few major steps common to all (Figure 1). These common steps are repeat masking, mapping ESTs/known genes, homologs, gene modeling using different methods sequentially or in parallel and then combining them, and finally annotating produced sets of gene models using various domain prediction and homology searches.

The JGI and the Broad Institute both use a similar basic set of gene predictors (Fgenesh (Salamov and Solovyev, 2000), Fgenesh+ (www.softberry.com), and GeneWise (Birney et al., 2004)), but in order to produce a non-redundant 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 select the one with the best homology among Fgenesh, Fgenesh+ and GeneWise predictions. And 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, fix incomplete models by analysis of local genomic regions and then treat 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: (i) homology to other proteins, (ii) EST support, and (iii) 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 databased (e.g., KEGG) and functional classification allows one to map the predicted proteins onto metabolic pathways, Gene Ontology and KOG categories, that provide user with multiple entry points into the annotation data. Although implementations of these steps varies, most of the pipeline utilize Blast or Smith-Waterman searches to find the list of homologs, InterproScan or various domain-search methods to predict domains, and use public software (like TMHMM, SignalP, TargetP, etc.) for more specialized analysis.

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. 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. Community annotation usually begins with a conference, often termed "Jamboree". The jamboree serves several purposes. The volunteers that will be manually curating the information are trained how to use the specialized tools. 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.

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 Berkeley National Laboratory under contract No. DE-AC02-05CH11231 and Los Alamos National Laboratory under contract No. W-7405-ENG-36.

REFERENCES

Allen J.E., Pertea M., Salzberg S.L. (2004) Computational gene prediction using multiple sources of evidence. Genome Res. **14:** 142-148.

Birney E., Clamp M., Durbin R. (2004) Genewise and genomewise. Genome Res. **14:** 988-995.

Brent M. (2008) Steady progress and recent breakthroughs in the accuracy of automated genome annotation. Nat. Rev. Gene **9:** 62-73.

Burset M.G. (1996) Evaluation of gene structue prediction programs. Genomics **34:** 353-367.

Dujon B., Sherman D., Fischer G. et al. (2004) Genome evolution in yeasts. Nature **430:** 35-44.

Fitch W. (1970) Distinguishing homologous from analogous proteins. System. Zool. **19:** 19-113.

Foissac S., Bardou P., Moisan A. et al. (2003) EUGENE'HOM: A generic similarity-based gene finder using multiple homologous sequences. Nucleic Acids Res. **31:** 3742-3745.

Galagan J.E., Calvo S.E., Borkovich K.A. et al. (2003) The genome sequence of the filamentous fungus *Neurospora crassa*. Nature **422:** 859-868.

Gavin A.-C., Bosche M., Krause R. et al. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature **415**: 141-147.

Goffeau A., Barrell B.G., Bussey H. et al. (1996) Life with 6000 genes. Science **274**: 563-567.

Grinier J., McKay M., Herbert B.R. et al. (2005) Proteomic response of the biological control fungus *Trichoderma atrovide* to growth on the cell walls of *Rhizoctonia solani*. Curr. Genet. **47:** 381-388.

Hazbun T.R., Fields S. (2001) Networking proteins in yeast. Proc. Natl. Acad. Sci. USA **98**: 4277-4278.

Ho Y., Gruhler A., Heilbut A. et al. (2002) Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. Nature **415**: 180-183.

Howe K., Ito T., Chiba T., Ozawa R., Yoshida M. et al. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc. Natl. Acad. Sci. USA **98:** 4569-4574.

Jeffries T.W., Grigoriev I.V., Grimwood J. et al. (2007) Genome sequence of the lignocellulose-bioconverting and xylose-fermenting yeast *Pichia stipitis*. Nat. Biotechnol. **25:** 319-326.

Kellis M., Patterson N., Endrissi M. et al. (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature **423:** 241-254.

Koonin E.V. (2005) Orthologs, paralogs, and evolutionary genomics. Ann. Rev. Genet. **39:** 309-338.

Korf I. (2004) Gene finding in novel genomes. BMC Bioinformatics **5:** 59.

Kupfer D.M., Drabenstot S.D., Buchanan K.L. Introns and splicing elements of five diverse fungi. Eukaryotic Cell **3:** 1088-1100.

Martin F., Aerts A., Ahren D. et al. (2008) The genome of *Laccaria bicolor* provides insights into mycorrhizal symbiosis. Nature **452:** 88-92.

Martinez D., Larrondo L.F., Putnam N. et al. (2004) Genome sequence of the lignocellulose degrading fungus *Phanerochaete chrysosporium* strain RP78. Nat. Biotechnol. **22:** 695-700.

Majoros W.H., Pertea M., Salzberg S. (2005) Efficient implementation of a generalized pair hidden Markov model for comparative gene finding. Bioinformatics **20:** 2878-2879.

Mayor C., Brudno M., Schawartz J.R. et al. (2000) VISTA: Visualising global DNA sequence alignments of arbitrary length. Bioinformatics **16:** 1046-1047.

Pavlovic V., Garg A., Kasif S. (2002) A Bayesian framework for combining gene predictions. Bioinformatics **18:** 19-27.

Remeteria A., Lopez-Malina N., Ludwig A. et al. (2005) Genes and molecules involved in *Aspergillus fumigatus* virulence. Revista IberoAmericana de Micologia **22:** 1-23.

Salamov A.A., Solovyev V.V. (2000) *Ab initio* gene finding in *Drosophila* genomic DNA. Genome Res. **10:** 516-522.

Sherman D., Durrens P., Beyne E. et al. (2004) Genolevures: Comparative genomics and molecular evolution of hemiascomycetous yeasts. Nucleic Acids Res. **32:** 315-318.

Sims A.H., Gent M.E., Robson G.D. et al. Combining transcriptome data with genomic and cDNA sequence alignments to make confident functional assignments for *Aspergillus nidulans* genes. Mycol. Res. **108:** 853-857.

Storm C.E., Sonnhammer E.L. (2002) Automated ortolog inference from phylogenetic trees and calculation of orthology reliability. Bioinformatics **18:** 92-99.

Tenney A.E., Brown R.H., Vaske G. et al. (2004) Gene prediction and verification in a compact genome with numerous small introns. Genome Res. **14:** 2330-2335.

Torrents D., Suyama M., Zdobnov E., Bork P. (2003) A genome-wide survey of human pseudogenes. Genome Res. **12:** 2559-2567.

Tyler B.M., Tripathy S., Zhang X. et al. (2006) *Phytophthora* genome sequences uncover evolutionary origins and mechanisms of pathogenesis. Science **313:** 1261-1266.

Uetz P., Giot L., Cagney G., Mansfield T.A. et al. (2000) A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. Nature **403:** 623–627.

Wood V., Gwilliam R., Rajandream M.A. (2002) The genome sequence of *Schizosaccharomyces pombe*. Nature **415:** 871-880.

Yao H., Guo L., Fu Y. et al. (2005) Evaluation of five *ab initio* gene prediction programs for the discovery of maize genes. Plant Mol. Biol. **57:** 445-460.

Zhang Z.L., Gerstein M. (2004) Large-scale analysis of pseudogenes in the human genome. Curr. Opin. Genet. Dev. **14:** 328-335.

Zmasek C.M., Eddy S.R. (2002) RIO: Analysing proteomes by automated phylogenomics using resampled inference of orthologs. BMC Bioinformatcs **3(1):** 14.