Supplementary Methods

Sequencing and Assembly. The initial shotgun sequencing project generated 529,689 sequences, with 233,585 generated at TIGR and 296,104 generated at the Sanger Institute. These were assembled at TIGR using the Celera Assembler\(^1\). After multiple refinement steps, the shotgun assembly contained 971 contigs containing 29,329,984 nucleotides, with the longest contig containing 1,796,676 nucleotides. 269 of these contigs were grouped into 92 scaffolds, with the largest scaffold spanning 2.95 Mbp. The average coverage over all contigs was 10.5X. This assembly was then passed on to closure teams for directed sequencing work to produce the final assembly that contains just 11 sequence gaps and 19 contigs. Final assemblies were correlated with the optical map for further support. All areas of imperfect match between the optical map and the sequence assemblies were inspected and most were found to contain an extra restriction site in the assembly. This anomaly is due to the presence of another restriction site very close by, which is undetectable at the optical mapping resolution of 1-10 Kb. These extra sites on assemblies were verified by multiple clone coverage. In the rare instances in which the optical map had an extra site, the site was either confirmed by high coverage with multiple libraries or was found to be within telomeric repeats, rRNAs, or centromeric regions in which confirmation could not be absolute.

Coding Sequence (CDS) Prediction and Gene Identification. The assembled genomic sequence was processed through the TIGR annotation pipeline, a collection of software known as Eukaryotic Genome Control (EGC) that serves as the central data management system. EGC processes each sequence through a series of homology searches as well as algorithms for predicting genes (GlimmerM\(^2\), Exonomy\(^3\), Unveil\(^4\),
Phat$^5$ and GeneSplicer$^6$). The gene prediction algorithms were trained with a dataset of 2,633 *A. fumigatus* EST and cDNA sequences which were contributed by collaborators at the Universidad Complutense de Madrid and the University of Manchester (2,582) or downloaded from GenBank in May 2003 (51).

Homology to nucleotide and protein datasets is computed using the AAT package, which utilizes a two-step approach consisting of a fast database homology search followed by a rigorous, splice-aware local alignment$^7$. Datasets included EST and cDNA sequences from *A. fumigatus* and other Aspergillus species, the TIGR Fungal Gene Indices$^8$, and a non-redundant amino acid database filtered from public sources. Initial gene models were generated by the program Combiner, which evaluates each type of evidence (gene predictions, nucleotide alignments and protein alignments) separately for its ability to predict translation starts and stops, splice acceptor and donor sites, and protein coding regions. Gene models are constructed by merging the most likely gene model signals using statistics generated from the training set$^9$. The performance of the *ab initio* gene prediction software cited above and Combiner was evaluated against 341 gene models which were manually curated along a 1MB stretch of genomic sequence$^{10}$. The nucleotide sensitivity, or ability to correctly predict which bases are present in the coding regions of genes, ranged from 70-90% for each of the individual gene prediction algorithms, and the Combiner sensitivity was 92% versus the manually curated data set.

The Combiner gene models were manually reviewed and modifications were made to address gene merges, gene splits and missing gene models. The completeness and accuracy of the resulting *A. fumigatus* gene set, comprised of 9,926 protein-coding
genes, was evaluated using two different methods: verification against the EST set, and similarity with *A. nidulans* and *A. oryzae* gene sets.

2,771 *A. fumigatus* EST and cDNA sequences (148 additional sequences available from Genbank 11/2004) were aligned against the genomic sequence, assembled into non-redundant alignments and compared against the annotation data set using the Program to Assemble Spliced Alignments (PASA)\(^\text{11}\). 2,660 sequences were successfully aligned to the genome, resulting in 1,140 alignment assemblies mapping to 830 gene models. Of the 1,140 assemblies, 850 (75\%) were consistent with the current annotation, 190 (17\%) suggested minor gene structure updates, 60 (5\%) were in conflict with current predictions and 40 (3\%) were intergenic. Most intergenic alignment assemblies do not contain a significant open reading frame, and only 9 of the 41 contain a splice site. These may represent unlinked UTR exons, non-coding RNAs or missed genes without enough evidence to create a valid gene model.

The *A. fumigatus*, *A. nidulans* and *A. oryzae* genomes were annotated using different annotation pipelines and processes\(^\text{12,13}\). The mutual best BLASTp hits between the three predicted proteomes were grouped together into ortholog clusters\(^\text{14}\). 8,207 (83\%) of the 9,926 predicted *A. fumigatus* genes was found to have an ortholog in at least one of the other two genomes. Of the 8,367 ortholog groups identified, only 568 (7\%) consisted of only *A. nidulans* and *A. oryzae* proteins. tBLASTx searches using the corresponding genes as queries against the *A. fumigatus* genome revealed no hits with significant similarity. These results combined with verification against the EST set suggest that the annotation of *A. fumigatus* is substantially complete.
Gene product names were assigned based on significant BLASTp and domain matches. In order to organize the annotation data for further analysis, proteins were organized into putative paralogous family groupings based on conserved domain composition. Gene products were also assigned to Gene Ontology (GO) terms by transferring the GO associations of the best Saccharomyces Genome Database protein match (http://www.yeastgenome.org/). Many of these automated names and Gene Ontology assignments were manually reviewed for accuracy during analysis for this publication.

**Gene Expression Methods.** *Primer design and genomic DNA amplification:* The DNA amplicon microarray for *A. fumigatus* Af293 was fabricated as follows. For PCR targets, we selected a 700 bp region immediately upstream of the predicted stop codon from each gene. If the gene was smaller than 700 bp, we took the entire gene. Then, we included 150 bp of sequence downstream of the gene or as much as there is in the intergenic region when shorter than 150 bp. These target sequences provided a maximum of 850 bp for each gene. We conducted automated selection of PCR primer pairs by feeding the target sequences to Primer 3.0 (http://www-genome.wi.mit.edu/genome_software) with optimized design parameters that can be used to amplify greater than 5/6 of the targets. The predicted resulting PCR products were on average 710 bp in length. Using this approach we were able to design primers for 9,516 genes (96%). We amplified these target gene regions from genomic DNA. The resulting PCR products were purified and spotted in triplicate at high density on Corning (Acton, MA) UltraGAPS™ aminosilane-coated microscope slides using a robotic spotter built by Intelligent Automatic Systems (IAS) (Cambridge, MA) and cross-linked by ultraviolet illumination.
**CGH data analysis:** Hybridized slides were scanned using the Axon GenePix 4000B microarray scanner and the TIFF images generated were analyzed using TIGR Spotfinder (<http://www.tigr.org/software/>) to obtain relative transcript levels. Data from TIGR Spotfinder were stored in MAD, a relational database designed to effectively capture and store microarray data. Data was normalized using a local regression technique LOWESS (LOcally WEighted Scatterplot Smoothing) for hybridizations with RNA-based samples using a software tool MIDAS (<http://www.tigr.org/softlab>), while total intensity normalization was used for the hybridizations with genomic DNA samples. The resulting data was averaged from triplicate genes on each array, from duplicate flip-dye arrays for each experiment, and from biological replicates, taking a total of 12 intensity data points for each gene. Differentially expressed genes at the 95% confidence level were determined using intensity-dependent Z-scores (with $Z=1.96$) as implemented in MIDAS and the union of all genes identified at each time point were considered significant in this experiment. The resulting data were organized and visualized based on similar expression vectors in genes using Euclidean distance and hierarchical clustering with average linkage clustering method to view the whole data set (Fig. 1A) and k-means to group the genes in 10 clusters (Fig. 1B) with TIGR MEV (<http://www.tigr.org/software>).

All the raw microarray expression data and those presented in this manuscript are available through ArrayExpress <http://www.ebi.ac.uk/arrayexpress> with accession numbers A-MEXP-205 (array design) and E-MEXP-332, E-MEXP-333 (experimental data).
References

Ichinomiya, M., Igarashi, R., Iwashita, K., Juvvadi, P. R., Kato, M., Kato, Y.,
Kawamura, Y., Kin, T., Kiryu, H., Kokubun, A., Kojima, A., Maeda, H.,
Maeyama, N., Maruyama, J., Nakajima, T., Nierman, W. C., Nishimura, Y., Oda,
K., Okada, K., Sakamoto, K., Sawano, T., Takahashi, M., Takase, K.,
Terabayashi, Y., Wortman, J., Yamada, O., Yamagata, Y., Yu, J., Anazawa, H.,
Hata, Y., Koide, Y., Komori, T., Koyama, Y., Minetoki, T., Suharnan, S., Tanaka,
A., Isono, K., Kuhara, S., Ogasawara, N., & Kikuchi, H. Genome sequencing and
analysis of Aspergillus oryzae. SUBMITTED.


15. The Gene Ontology Consortium. Gene Ontology: tool for the unification of