Supporting Information

Leonard and Richards 10.1073/pnas.1210909110

SI Materials and Methods

***fdf*BLAST.** To our knowledge, there is no standardized automated method to compare genome datasets (i.e., predicted proteomes) for the purpose of identifying differentially distributed gene fusions. To fill this gap, we designed a bioinformatic analysis pipeline to identify differentially distributed gene fusions between whole-genome datasets. This approach is not an exhaustive one but allows the identification of a large number of candidate gene fusions. The five-step bioinformatics pipeline uses a series of Perl scripts available at https://github.com/guyleonard/fdfBLAST.

As illustrated in Fig. S1, step 1 performs serial all-against-all BLASTp comparisons of predicted proteome datasets. In step 2 all BLAST search hits at or above a specified e-value threshold are counted to identify differential hit patterns. In step 3 reciprocal BLAST searches are used to confirm differential distributed BLAST hits. The program allows the e-value threshold for step 3 to be adjusted so users can control how differential hits are sampled. For example, a user may want to adjust these e-value thresholds to compare closely related or distantly related species and to manage noise in the form of false-positive BLAST hits. Differential hit patterns are the target datasets (e.g., 1-to-2 or 3-to-4, and so forth); all other patterns are excluded (e.g., 1-to-1, 3-to-3, 1-to-0, 3-to-0, and so forth). The differential hit patterns identified in steps 1–3 theoretically encompass gene families that have diversified under a number of different evolutionary scenarios: (*i*) differential gene duplication, (*ii*) differential gene loss, (*iii*) differential tandem exon duplication, (*iv*) inconsistent recovery of homologs because of differing rates of sequence variation in the gene family, and (*v*) gene fusions. The remaining two steps of the *fdf*BLAST pipeline (described below) generate a series of images that are designed to allow users to curate the results and eliminate datasets consistent with scenarios 1–4 and identify gene fusions (scenario 5).

Step 4 (Fig. S1) involves multiple rank and sorting processes used to distinguish putative fused and unfused pairs. By using the hit-range information from each set of BLAST results, the position of each match can be compared with the query sequence's start (amino acid position zero), end (the number of the last amino acid), and middle (length divided by two). Hits then can be classified (sorted) as left of the middle, right of the middle, or spanning the middle based on the value of the hit's start and end position. Specifically, if the hit's end position is higher than middle, it is recorded as right-sided; if it is lower than the middle value, it is classified as left-sided. The process is repeated using the hit's start position.

Hits that have a mixed results, i.e., a left-sided start position and a right-sided end position, are potential full-length hits. Hits that span the middle and that have a length (defined by the user, currently fixed to 90%) similar to the query full length are excluded and are not shown on the graphical output, because they are likely to represent complete homologous genes and therefore are unlikely to be differentially distributed gene fusions. Shorter hits that span the middle are shown in the graphical output according to the ranking procedures described below.

Hits that produce consistent results (i.e., both start and end are classified as left or as right) are retained and shown in the graphical output. Note that the program is set up so that hits in which ≤10% of the region of similarity spans the middle are not classified as "middle" and are retained and displayed in the output figures.

The final set of split sequences then is ranked in reference to the query sequence in two additional ways. First, each is given a percentage score based on the number of amino acid bases matched to the query sequence. Then each potential unfused ORF alignment is illustrated by a color: 80–100%, green; 70–80%, light blue; 60–70%, purple; 40–60%, dark red; and <40%, gray. This color scheme is reflected in the cartoons of the final gene-to-gene alignment (Fig. S1, step 4). Second, a ratio is calculated based on the remaining ORFs matched to the query sequence: The lengths of the matched ORFs (left and right matches separately) are ordered from shortest to longest, and a ratio score for each pair of the matched ORFs is calculated. The highest end value from the left match is divided by the lowest start value from the right side, providing a proxy for the distance of the two partial hits relative to the query sequence. A value of 1 suggests that the left and right matches are adjacent when aligned against the query sequence, and a value of 0.1 means that the left and right matches are relatively far apart when aligned against the query sequence. Overlapping hits are removed at an earlier stage in the pipeline and are not shown on the output figure. All combinations are out put by *fdf*BLAST and organized into folders ranging (in incremental steps of 0.1) from 0.1–1.0 so the graphical results can be searched systematically.

These two rank-and-sort methods, although seemingly complex, make the data produced by *fdf*BLAST accessible for curation. Because gene-fusion events can be considered the product of the union of multiple domains, it is advantageous (at least programmatically) to categorize the location of matched split ORFs to the potentially "fused"-ORF state. This categorization helps with the manual curation and identification of candidate split ORFs. For example, if all the matched ORFs for one fused ORF are similar in length (and span the whole putative fused ORF), they can be identified as potential complete-gene-length homologs and can be discarded. Similarly, if all the matched ORFs appeared to be one-sided (i.e., match only one half of the putative fused ORF), the putative gene-fusion prediction is likely to be an artifact.

Step 5 (Fig. S1) involves comparison of the candidate sequences with the PFAM database. The sequences representing the fused ORF and the two best unfused ORFs from each set of candidate gene fusions are passed to a program to map conserved functional domains on to the alignment diagrams. The program HMMER (http://hmmer.org) (1) is used to search sequence databases of homologous protein sequences using profile hidden Markov models. The data output from HMMER then is displayed as an overlay on the alignment diagrams (domain overlays). We use this step to remove putative gene fusions that do not contain PFAM conserved domains. Although this step may remove a number of gene fusions of domains not represented in the PFAM database, we believe that this approach is important to remove noise created by more frequently occurring false-positive hits (i.e., differential matches for regions of low complexity).

Pipeline for Preliminary Fusion Domain Phylogenetic Analysis. The preliminary phylogenies were calculated from taxon sampling using a MySQL database (www.mysql.com) of predicted proteomes containing a diversity of opisthokont taxa available at the Department of Energy Joint Genome Institute, the National Center for Biotechnology (NCBI) GenBank database, and the Broad Institute (Table S1). For this analysis, we also made use of the *Blastocladiella emersonii* genome assembly produced in our laboratory in collaboration with Suely Gomes (Universidade de São Paulo, São Paulo, Brazil). All sequence data are available in the form of unmasked and masked alignments (see below) (2). Each candidate sequence was compared against sequences in the

database using BLASTp (3), and the best-similarity hits from each species were extracted (using the e-value 1e-10 gathering threshold). These sequences were aligned using MUSCLE (4), conserved regions from this alignment were sampled using trimAL (5), and phylogenetic trees were constructed using FastTree (6) with the options SLOW and BIONJ and the default substitution model. Topology support was assessed using the SH-like aLRT branch support values.

All fusion gene component phylogenies were inspected manually to check the phylogeny for resolution. If inspection of tree and alignment suggested the domain phylogeny was unlikely to demonstrate useful levels of resolution in and around the fusion branches, no further analysis was conducted, and the dataset was not analyzed further. This process often required several rounds of manual alignment checks and reanalysis for confirmation. For all fusion gene domains we then performed a series of BLAST searches focusing on additional sampling from the GenBank nonredundant (nr) database and the GenBank EST database. Additional sequences were added to the alignments as required. This process was facilitated by using the sequence management for phylogeny programs REFGEN and TREENAMER (7). Each alignment then was edited manually and masked to remove gaps and ambiguous alignment positions using the alignment program SEAVIEW (2). All gene alignments and sequence data are available at http://gna-phylo.nhm.ac.uk/content/leonard_and_richards_2012.

In some cases, genes had large sections of the amino acid sequence missing relative to the alignment, most likely because of incomplete assembly or poor gene prediction (specifically intron/exon boundaries during automatic annotation of the genomes). When the presence of these putatively incomplete sequences did not alter the taxonomic representation significantly relative to the clade of fusion genes under investigation, incomplete sequences were excluded from the alignment. When putatively incomplete sequences were important for evolutionary analysis of the gene fusion, the sequence data were checked manually as described below.

For each domain alignment we identified the optimal model for phylogenetic analysis using MODELGENERATOR (8). For the models used, see Table S3. Then PHYML (9) analysis was used to assess the tree topology using the model parameters identified using MODELGENERATOR. Statistical support was evaluated with 100 bootstrap replicates.

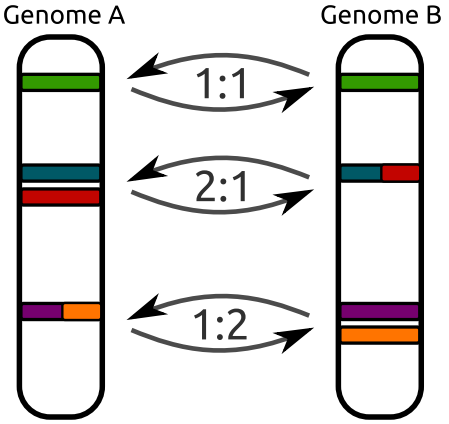
Checks of Gene-Fusion Genome Annotations in Taxa Branching Around Identified Gene Fusions. For taxa with unfused genes or partial sequences that branch close to gene fusions in our domain phylogenies, we checked specifically for cases of misprediction of these genes in individual genome assemblies. Using publicly available genome browsers on the Broad Institute, NCBI, and Department of Energy Joint Genome Institute websites (which allow visual inspection of the genome assembly as tracks and display information about contigs, supercontigs/scaffolds, and their associated gene predictions), we identified genes up- and downstream from the location of the unfused gene. This step allowed us to identify the direction in which and the contig on which the gene occurs. If, for example, the two genes we are interested in occur on different contigs, are in opposite orientation, or are both flanked and separated by other genes, we can confirm that the two domains form separate, unfused genes. However, if two separate genes that branch close to the gene fusion on our phylogenetic analyses are next to each other on a genome contig and occur in the same direction, we suggest these genes have been misannotated as separate genes when they should be fused, and so we tentatively annotate this pair of genes as a gene fusion. These alterations can be found in Dataset S1; genes confirmed as separate are marked with a red X in the *SI Appendix*, and genes corrected to putative fusions are marked with green ticks in the *SI Appendix*. We note that, lacking experimental transcription and proteomic data, these annotations are not definitive results; consequently, the relative rate of fission in this dataset may be underestimated, or the position of a gene fission marked on Fig. 2 may be misplaced.

To investigate further evidence in support of each gene fusion, we searched the GenBank nr EST database for sequences that verified that the candidate gene fusion is transcribed as a gene fusion. In 18 of the 63 gene fusions we could identify evidence that the gene fusion was transcribed as a gene pair (gene fusion) using EST data (Table S3).

1. Eddy SR (2011) Accelerated profile HMM SEARCHES. *PLOS Comput Biol* 7(10):e1002195.
2. Galtier N, Gouy M, Gautier C (1996) SEAVIEW and PHYLO_WIN: Two graphic tools for sequence alignment and molecular phylogeny. *Comput Appl Biosci* 12(6):543–548.
3. Altschul SF, et al. (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res* 25(17):3389–3402.
4. Edgar RC (2004) MUSCLE: A multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5:113.
5. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T (2009) trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25(15): 1972–1973.
6. Price MN, Dehal PS, Arkin AP (2010) FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS ONE* 5(3):e9490.
7. Leonard G, Stevens JR, Richards TA (2009) REFGEN and TREENAMER: Automated sequence data handling for phylogenetic analysis in the genomic era. *Evol Bioinform Online* 5:1–4.
8. Keane TM, Creevey CJ, Pentony MM, Naughton TJ, Mclnerney JO (2006) Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified. *BMC Evol Biol* 6:29.
9. Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* 52(5):696–704.

Step 1 *Automated Serial BLASTp Analysis*

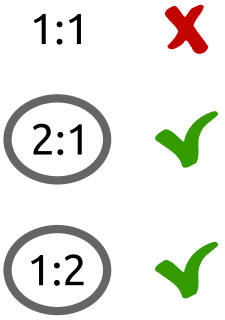
Genomes of interest are collated and then subjected to NCBI's local BLAST tools, "formatdb" and "blastall".



An all against all analysis is carried out, producing the standard BLAST output. For example; three genomes A, B and C are analysed in this way: A to A, A to B, A to C and B to B, B to A, B to C and C to C, C to A, C to B.

Step 2 *Comparative Hit Counts.*

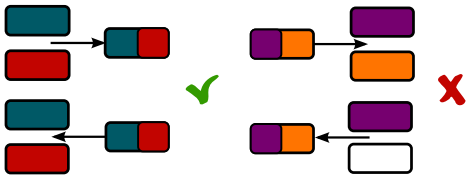
Genes with differential hit patterns are identified and passed on for further analyses by parsing the previous BLAST output using BioPerl.



Program includes user adjustable e-value threshold so that multiple comparisons, with different cut-offs, can be performed.

Step 3 *Reciprocal Hit Matching*

For each gene that has displayed evidence of differential hit patterns the reciprocal (e.g. A to B and B to A) analysis is queried to see if the differential pattern is preserved.

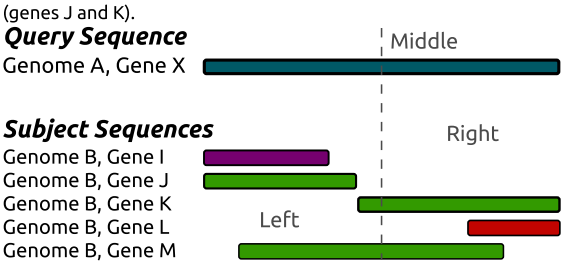


Genes with differential hits that display hits in both directions are selected.

If the second gene is not present at the selected e-value cut-off it is not considered a complete reciprocal hit.

Step 4 *Ranking and Sorting*

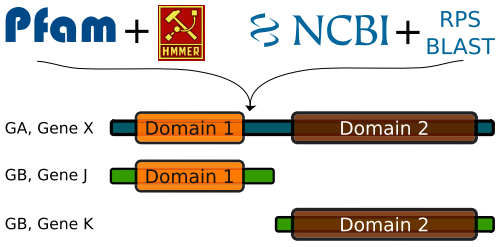
Sorting: The subject ORFs are sorted by their 'location' compared to the query sequence's length; placing them left, right or spanning the middle. This helps remove 'complete full length' homologues (gene M) and identify potential split domains (genes J and K).



Ranking: Each ORF is given a score based on the number of bases matched to the query sequence divided by the total length. These are coloured; green (80-100), blue (70-80), purple (60-70), red (40-60) and black (<40) based on %-identity in fdfBLAST's output.

The resulting images only include two candidate unfused ORFs, unlike the above image which represents the internal program data structure. Genes I, L and M are discarded.

Step 5 *Conserved Domains*



Candidate gene fusions are scanned against the Pfam and/or CDD databases using HMMER and RPS-BLAST respectively.

Conserved domains are then mapped on to the previous images in order to help manual confirmation, further narrowing the list of predicted putative gene fusion events.

Fig. S1. Cartoon illustrating the *fdf*BLAST analysis pipeline. The figure includes notes on the processes coded in the pipeline scripts. See *SI Materials and Methods* for more details. All pipeline scripts are available at https://github.com/guyleonard/fdfBLAST.

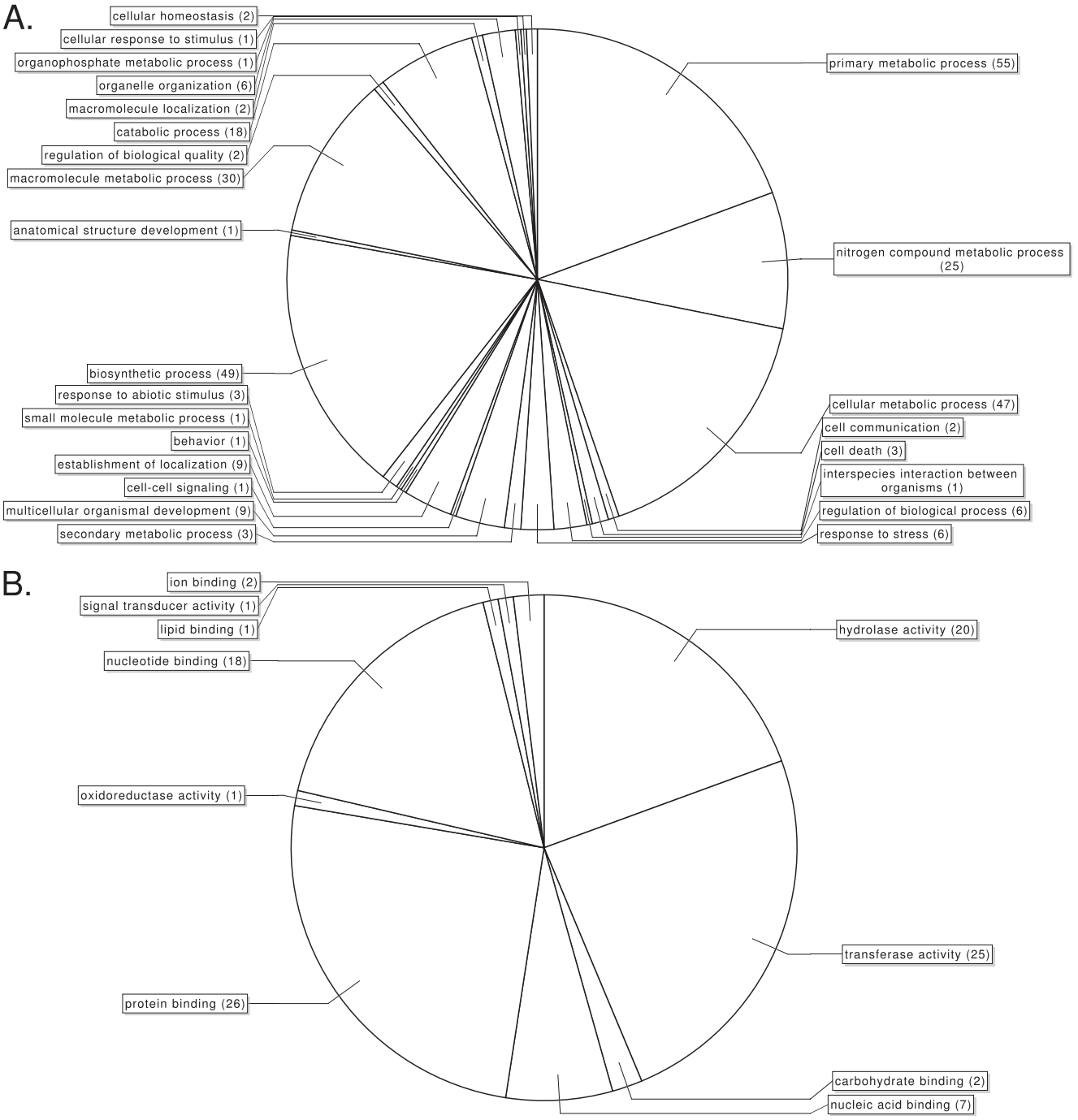


Fig. S2. BLAST2GO annotation categories represented by the component domains in the 63 gene fusions. For each fusion the individual sequences representing each domain were collated into one FASTA file. This file then was used with the program BLAST2GO to identify gene ontologies. One feature of the program BLAST2GO is the generation of a combined graph that summarizes the functional annotation content of a dataset. Here we show two pie charts of functional content at BLAST2GO level 3: biological process (*A*) and molecular function (*B*). We used this analysis to investigate whether the gene fusions were specific to a functional category or if they were classified into a diversity of categories. (*A*) Level 3 biological process annotations suggest that a large number of the domains function in primary metabolic processes, nitrogen compound metabolism, cellular metabolism, biosynthetic processes, and macromolecule metabolic processes. (*B*) Molecular function annotation (level 3) demonstrates that a large number of domains function in hydrolase activity, transferase activity, nucleic acid binding, nucleotide binding, or protein binding.

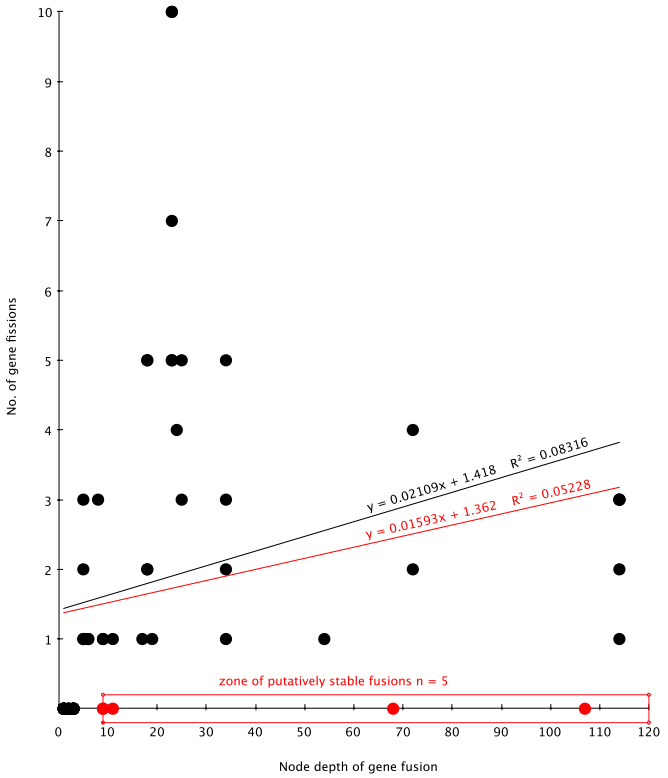


Fig. S3. Chart showing the relationship between node depth and gene fission, demonstrating that the number of fissions increases with the phylogenetic depth of the fusion. This analysis identified a cohort of gene fusions (n = 5) for which, given a depth of eight or more nodes (nine derived genomes), the fusion appeared fixed (red circles). Two lines of best fit are shown. The red line includes putatively fixed fusions. The black line excluded putatively fixed fusions.

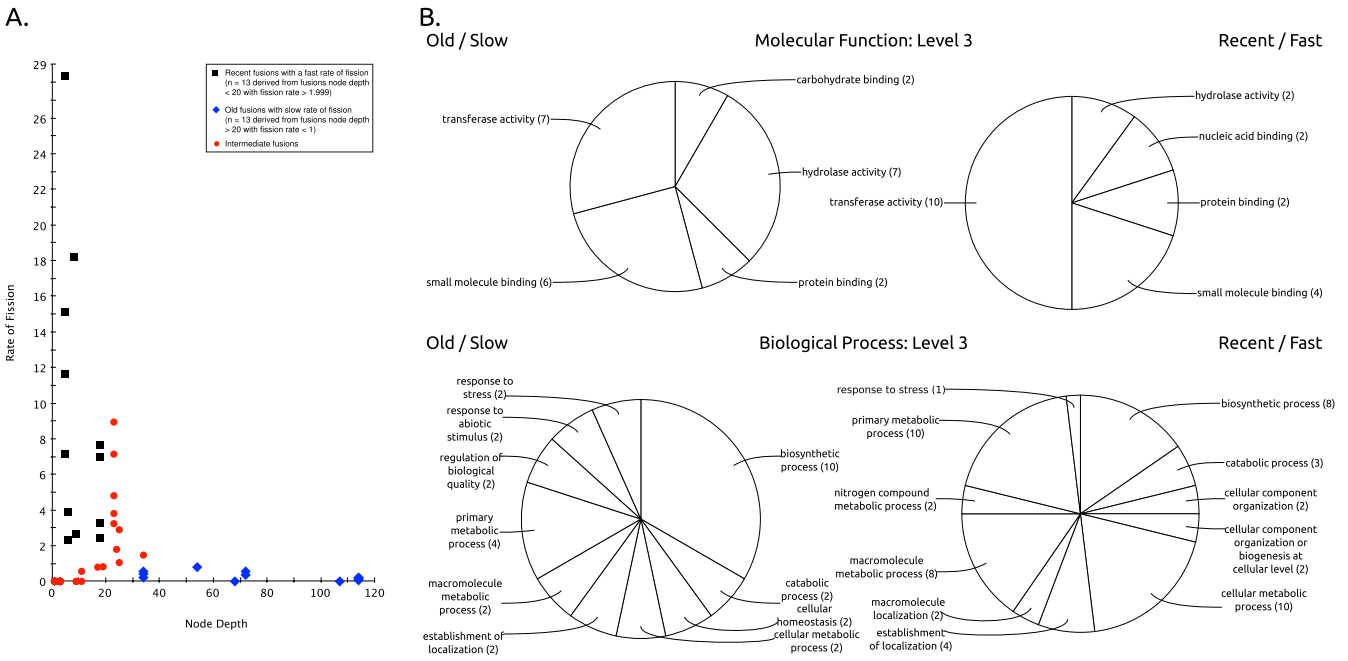


Fig. S4. (*A*) Chart showing the relationship between node depth and the relative rate of gene fission. This analysis demonstrated that the gene fusions generally are divided between recent gene fusions that show a high relative rate of fission (black squares) and old gene fusions that show a low relative rate of fission (blue diamonds). (*B*) Trends in putative function among recent gene fusions with a high rate of fission and old fusions with a low rate of fission. Using the BLAST2GO approach described in Fig. S2 and focusing on the subsets of 13 gene fusions representing each trend, we investigated the putative functional trends in each category. This analysis demonstrated no clear functional trend among old/slow and recent/fast fusion genes.

Table S1. Genomes used for comparative fusions analyses and phylogeny

Genome sampled for domain phylogeny	Genome sampled for 67-gene fungal phylogeny?
Acremonium alcalophilum	Yes
Agaricus bisporus	Yes
Allomyces macrogynus ATCC 38327	Yes
Alternaria brassicicola	Yes
Ashbya gossypii	Yes
Aspergillus aculeatus	Yes
Aspergillus carbonarius	Yes
Aspergillus clavatus	Yes
Aspergillus flavus	Yes
Aspergillus fumigatus	Yes
Aspergillus nidulans	Yes
Aspergillus niger	Yes
Aspergillus oryzae	Yes
Aspergillus terreus	Yes
Auricularia delicata	Yes
Batrachochytrium dendrobatidis	Yes
Baudoinia compniacensis	Yes
Bjerkandera adusta	Yes
Blastocladiella emersonii	Yes
Blastomyces dermatitidis	Yes
Botrytis cinerea	Yes
Branchiostoma floridae	Not included as not fungi
Caenorhabditis elegans	Not included as not fungi
Candida albicans SC5314	Yes
Candida caseinolytica	Yes
Candida glabrata	Yes
Candida tenuis	Yes
Capitella sp. I	Not included as not fungi
Capsaspora owczarzaki	Not included as not fungi
Ceriporiopsis subvermispora	Yes
Chaetomium globosum	Yes
Ciona intestinalis	Not included as not fungi
Coccidioides immitis	Yes
Coccidioides posadasii	Yes
Cochliobolus heterostrophus	Yes
Cochliobolus sativus	Yes
Coniophora puteana	Yes
Coprinus cinereus	Yes
Cryphonectria parasitica	Yes
Cryptococcus neoformans	Yes
Dacryopinax sp.	Yes
Danio rerio	Not included as not fungi
Daphnia pulex	Not included as not fungi
Debaryomyces hansenii	Yes
Dichomitus squalens	Yes
Dothistroma septosporum	Yes
Drosophila melanogaster	Not included as not fungi
Fomitiporia mediterranea	Yes
Fomitopsis pinicola	Yes
Fusarium graminearum	Yes
Fusarium oxysporum	Yes
Fusarium verticillioides	Yes
Gallus gallus	Not included as not fungi
Ganoderma sp.	Yes
Gloeophyllum trabeum	Yes
Hansenula polymorpha NCYC 495 leu1.1	Yes
Helobdella robusta	Not included as not fungi
Heterobasidion annosum	Yes
Histoplasma capsulatum	Yes
Homo sapiens	Not included as not fungi
Hysterium pulicare	Yes
Laccaria bicolor	Yes
Leptosphaeria maculans	Yes
Lipomyces starkeyi	Yes

Table S1. Cont.

Genome sampled for domain phylogeny	Genome sampled for 67-gene fungal phylogeny?
Lottia gigantea	Not included as not fungi
Magnaporthe grisea	Yes
Malassezia globosa	Yes
Melampsora laricis-populina	Yes
Microsporum canis	Yes
Microsporum gypseum	Yes
Monosiga brevicollis	Not included as not fungi
Mucor circinelloides	Yes
Mus musculus	Not included as not fungi
Mycosphaerella fijiensis	Yes
Mycosphaerella graminicola	Yes
Nectria hematococca	Yes
Nematostella vectensis	Not included as not fungi
Neosartorya fischeri	Yes
Neurospora crassa	Yes
Neurospora tetrasperma	Yes
Paracoccidioides brasiliensis	Yes
Phanerochaete carnosa	Yes
Phanerochaete chrysosporium	Yes
Phlebia brevispora	Yes
Phlebiopsis gigantea	Yes
Phycomyces blakesleeanus	Yes
Pichia membranifaciens	Yes
Pichia stipitis	Yes
Pleurotus ostreatus PC15	Yes
Pleurotus ostreatus PC9	Excluded from 67-gene phylogeny because represented by other *Pleurotus* genome
Pneumocystis carinii	Excluded from 67-gene phylogeny because of long-branch artifact
Podospora anserina	Yes
Postia placenta	Yes
Puccinia graminis	Yes
Punctularia strigosozonata	Yes
Pyrenophora teres	Yes
Pyrenophora tritici-repentis	Yes
Rhizopus oryzae	Yes
Rhodotorula graminis	Yes
Rhystidhysteron rufulum	Yes
Saccharomyces cerevisiae	Yes
Schizophyllum commune	Yes
Schizosaccharomyces cryophilus	Yes
Schizosaccharomyces japonicus	Yes
Schizosaccharomyces octosporus	Yes
Schizosaccharomyces pombe	Yes
Sclerotinia sclerotiorum	Yes
Septoria musiva	Yes
Septoria populicola	Yes
Serpula lacrymans	Yes
Setosphaeria turcica	Yes
Spathaspora passalidarum	Yes
Sphaeroforma arctica jp610	Not included as not fungi
Spizellomyces punctatus daom br117	Yes
Sporobolomyces roseus	Yes
Sporotrichum thermophile	Yes
Stagonospora nodorum	Yes
Stereum hirsutum	Yes
Strongylocentrotus purpuratus	Not included as not fungi
Takifugu rubripes	Not included as not fungi
Thecamonas trahens atcc 50062	Not included as not fungi
Thielavia terrestris	Yes
Trametes versicolor	Yes
Tremella mesenterica	Yes

Table S1. Cont.

Genome sampled for domain phylogeny	Genome sampled for 67-gene fungal phylogeny?
Trichoderma atoviride	Yes
Trichoderma reesei	Yes
Trichoderma virens	Yes
Trichophyton equinum	Yes
Trichoplax adhaerens	Not included as not fungi
Uncinocarpus reesii	Yes
Ustilago maydis	Yes
Verticillium albo-atrum	Yes
Verticillium dahliae	Yes
Wallemia sebi	Yes
Wickerhamomyces anomalus	Yes
Wolfiporia cocos	Yes
Xenopus tropicalis	Not included as not fungi
Yarrowia lipolytica	Yes

Table S2. Summary of data output from Mesquite analysis

Fusion	Node depth of fusion	Model used for likelihood evaluation of character distribution (fused/unfused) in Mesquite (1)*	Forward rate (fusion)	Reverse rate (fission)	Proportional likelihood of branching position of fusion/s shown in Fig. 2†	No. of fissions	Proportional likelihood of fissions‡	Notes (See *SI Appendix* for diagrammatic outputs from Mesquite analysis summarizing the distribution of fused, unfused, and absent characters used to calculate proportional likelihoods)
1‡	114	MK1	—	0.124	0.999	2	0.994, 1	Fused before Fungi
2	11	MK1	0.071	—	0.986	0		Evidence of horizontal gene transfer (HGT) (2). Some fusions were recorded as absent so HGT does not mislead Mesquite character analyses
3‡	72	Asymm. 2	0.088	0.361	0.973	2	0.929, 1	Evidence of multiple paralogs, but domain trees show mixed resolution. To use Mesquite, we excluded anything that is not clearly an ortholog
4	11	Asymm. 2	0.07	0.565	0.944	1	0.902	
5	1	MK1	0.06	—	0.981	0		
6	1	MK1	0.059	—	0.993	0		
7	1	MK1	0.059	—	0.993	0		
8‡	114	MK1	—	0.213	0.999	3	1, 0.959. 0.999	Fused before Fungi. Some orthologs missing, so we recoded some characters as missing to avoid false patterns of fusion/fission during Mesquite character analyses
9‡	114	MK1	—	0.184	0.999	3	0.993, 0.999, 1	Fused before Fungi
10	54	Asymm. 2	0.077	0.8	0.988	1	1	Some orthologs missing, so we recoded some characters as missing to avoid false patterns of fusion/fission during Mesquite character analyses
11	34	Asymm. 2	0.081	0.415	0.996	2	1, 1	
12‡	23	Asymm. 2	0.249	7.138	0.596	10	1, 1, 1, 1, 0.911, 1, 1, 1,	Proportional likelihood for ancestral fusion at node shown (Fig. 2) is weak. However, this solution is favored, because, even if one assumes multiple convergent fusions, the data still require multiple fissions
13‡	34	Asymm. 2	0.083	0.557	0.996	3	1, 1, 1	
14	34	Asymm. 2	0.069	0.216	0.997	1	1	
15‡	34	Asymm. 2	0.098	1.475	0.991	5	1, 1, 1, 1, 1	
16‡	23	Asymm. 2	0.178	8.94	0.989	10	0.695, 1, 1, 1, 1, 1, 0.75, 0.79, 1, 1	
17‡	114	MK1	—	0.199	0.999	3	1, 1, 0.998	Fused before Fungi
18	3	MK1	0.059	—	0.997	0		
19	3	MK1	0.059	—	0.997	0		
20	3	MK1	0.059	—	0.997	0		
21	3	MK1	0.059	—	0.997	0		
22	3	MK1	0.059	—	0.997	0		
23	3	MK1	0.059	—	0.997	0		
24	2	MK1	0.06	—	0.994	0		
25	3	MK1	0.059	—	0.997	0		
26	3	MK1	0.059	—	0.997	0		
27	1	MK1	0.06	—	0.981	0		
28	1	MK1	0.06	—	0.981	0		
29	1	MK1	0.06	—	0.981	0		
30	3	MK1	0.059	—	0.997	0		
31‡	25	Asymm. 2	0.097	2.898	0.892	5	1, 1, 0.493, 1, 0.825	
32A	6	Asymm. 2	0.072	2.32	0.97	1	1	Fusion appears as two distinct domain architectures so counted as two fusion events (32A and 32B)

Table S2. Cont.

Fusion	Node depth of fusion	Model used for likelihood evaluation of character distribution (fused/unfused) in Mesquite (1)*	Forward rate (fusion)	Reverse rate (fission)	Proportional likelihood of branching position of fusion/s shown in Fig. 2[†]	No. of fissions	Proportional likelihood of fissions[‡]	Notes (See *SI Appendix* for diagrammatic outputs from Mesquite analysis summarizing the distribution of fused, unfused, and absent characters used to calculate proportional likelihoods)
32B[‡]	23	Asymm. 2	0.098	4.809	0.994	5	1, 1, 1, 1, 0.913	Fusion appears as two distinct domain architectures so counted as two fusion events (32A and 32B)
33	68	MK1	0.127	—	0.909	0		
34[‡]	72	Asymm. 2	0.211	0.557	0.709	4	0.969, 1, 0.872, 1	Proportional likelihood for ancestral fusion at node shown (Fig. 2) is weak. However, this solution is favored, because, even if one assumes multiple convergent fusions, the data still require multiple fissions
35	107	MK1	0.169	—	0.979	0		
36[‡]	23	Asymm. 2	0.107	3.24	0.999	7	1, 1, 1, 1, 1, 1, 0.924	
37	5	Asymm. 2	0.19	7.126	0.349	2	1, 1	
38[†]	2/1	MK1	0.123	—	0.957, 1	0		Low proportional likelihood for single ancestral fusion suggests separate convergent fusion. Counted as two separate convergent fusions.
39[†]	19	Asymm. 2	0.065	0.85	0.989	1	1	
40[‡]	8	Asymm. 2	0.271	18.196	0.708	3	1, 0.974, 0.605	Proportional likelihood for ancestral fusion at node shown (Fig. 2) is weak. However, this solution is favored, because, even if one assumes multiple convergent fusions, the data still require multiple fissions.
41	17	Asymm. 2	0.063	0.796	0.999	1	1	
42	5	Asymm. 2	0.178	15.091	0.952	1	1	
43	9	MK1	0.059	—	0.995	0		
44[‡]	18	Asymm. 2	0.074	2.46	0.997	2	1, 1	
45[‡]	25	Asymm. 2	0.085	1.055	0.842	3	1, 1, 1	
46	6	Asymm. 2	0.081	3.876	0.998	1	1	
47	9	MK1	0.059	—	0.995	0		
48	9	Asymm. 2	0.076	2.67	0.955	1	1	
49[‡]	24	Asymm. 2	0.08	1.797	0.985	4	0.826, 1, 1, 1	
50[‡]	18	Asymm. 2	0.079	3.262	0.997	2	1, 0.996	
51	1	MK1	0.059	—	0.994	0		
52	2	MK1	0.058	—	0.999	0		
53	3	MK1	0.06	—	0.992	0		
54[‡]	18	Asymm. 2	0.12	7.68	0.994	5	1, 0.992, 1, 1, 1	
55	5	Asymm. 2	0.362	28.375	0.782	3	1, 1, 1	
56[‡]	23	Asymm. 2	0.095	3.809	0.999	5	1, 0.444, 1, 1, 0.786	
57	1	MK1	0.059	—	0.997	0		
58	2	MK1	0.059	—	0.998	0		
59	5	Asymm. 2	0.141	11.641	0.998	1	1	
60	9	Asymm. 2	0.073	2.674	0.996	1	0.995	
61[‡]	18	Asymm. 2	0.111	6.96	0.995	5	1, 1, 0.993, 1, 0.983	
62	114	MK1	—	0.06	0.999	1	0.991	Fused before Fungi

—, Rate value absent as only one type of character transition was identified (fission or fusion).

*When only fusions or fissions were present, the MK 1 model was used. When both fusions and fissions were present, the Asymm 2 param. model was used.

[†]In some cases proportional likelihood analyses of fusion states favored separate convergent gene fusions, in these cases two values are listed. The proportional likelihood values which correspond to each fusion are labeled on the Mesquite output trees in *SI Appendix*.

[‡]Proportional likelihood values for multiple fissions are listed. The proportional likelihood values that correspond to each fission are labeled on the Mesquite output trees in *SI Appendix*.

1. Maddison WP, Maddison DR (2011) Mesquite: A modular system for evolutionary analysis. Version 2.75. Available at http://mesquiteproject.org.
2. Slot JC, Rokas A (2010) Multiple GAL pathway gene clusters evolved independently and by different mechanisms in fungi. *Proc Natl Acad Sci USA* 107(22):10136–10141.

Table S3. EST data providing support that the gene is transcribed as a gene fusion and listing model used for domain phylogenies

ID*	EST support [yes (GenBank ID)/absent (blank)]	Domain 1	Model†	I‡	G§	Domain 2	Model¶	I‡	G§
1	GR748940, CU897583	PRA-CH~PRA-PH	LG+I+G	0.128	0.625	Histidinol_dh	LG+I+G	0.178	1.167
2	DB663361, DB663361	Epimerase	LG+I+G	0.228	1.018	Aldose_epim	LG+I+G	0.227	1.029
3		SurE	LG+I+G	0.042	1.068	TTL	LG+I+G	0.046	1.096
4		PseudoU_synth_2	LG+I+G	0.08	1.14	dCMP_cyt_deam_1	LG+I+G	0.096	1.17
5		PAP2	LG+I+G	0.08	1.14	dCMP_cyt_deam_1	LG+I+G+F	0.078	1.068
6	JK213075	URO-D	LG+G	—	0.76	Porphobil_deam~ Porphobil_deamC	LG+I+G	0.066	0.93
7		Iso_dh	LG+G	—	0.672	Aconitase~Aconitase_C	LG+I+G	0.042	0.943
8	GT899320, EX842960, FP690749, GE931171, EX842960, FP690749, JK213037	GATase~IGPS	LG+I+G	0.072	0.961	PRAI	LG+I+G	0.166	0.855
9	GE294957	Indigoidine_A	LG+I+G	0.032	0.974	PfkB	LG+I+G	0.068	0.992
10	GH346002	Peroxidase	WAG+I+G+F	0.057	2.093	WSC	—	—	—
11	EY995910, EC046483	Allantoicase~Allantoicase	LG+I+G	0.073	1.124	Ureidogly_hydro	LG+I+G	0.052	1.261
12		Pex2_Pex12	LG+G	—	0.848	SPX~Ank_2~GDPD	LG+I+G+F	0.017	1.362
13	EX789001, JK212765	Spermine_synth	LG+I+G	0.064	1.118	Saccharop_dh	LG+G	—	0.53
14	DY845282	Cys_Met_Meta_PP	LG+I+G	0.139	0.76	GHMP_kinases_N~ GHMP_kinases_C	LG+I+G	0.085	1.048
15		SET	LG+I+G	0.065	1.508	dCMP_cyt_deam_1	LG+G	—	0.657
16		Thiolase_N~Thiolase_C	LG+I+G	0.098	1.15	KH_2, Ribosomal_S3_C	LG+G	—	0.424
17	DY892051	FolB~FolB	LG+I+G+F	0.042	1.665	HPPK~Pterin_bind	LG+I+G	0.123	1.212
18		NUDIX	LG+I+G	0.084	1.599	TPK_catalytic~TPK_ B1_binding	LG+I+G	0.041	1.442
19		Rsm22	LG+I+G	0.044	1.324	CtaG_Cox11	LG+I+G	0.121	0.839
20		Hydrolase	LG+I+G	0.04	0.933	CDP-OH_P_transf	LG+I+G+F	0.086	1.155
21	FY125691	COX15-CtaA	LG+I+G+F	1.039	0.143	Fer2	LG+I+G	0.162	1.118
22	FY173519	Palm_thioest	LG+I+G	0.036	1.146	PAP2	LG+I+G+F	0.079	1.109
23	FY173519	Aconitase~Aconitase_C	RtREV+G+F	—	1.97	Ribosomal_L21p	LG+I+G	0.155	0.665
24		Methyltransf_16	LG+I+G	0.07	1.096	dCMP_cyt_deam_1	LG+I+G+F	0.053	1.294
25		FSH1	LG+I+G	0.064	0.841	DHFR_1	LG+I+G	0.062	1.644
26		adh_short	CpREV+I+G+F	0.049	1.648	PIG-F	LG+I+G+F	0.033	1.507
27		Peptidase_M1~Leuk-A4-hydro_c	LG+I+G+F	0.052	1.011	IPPT	LG+I+G	0.057	1.134
28	JK211039	PGAM	LG+I+G	0.055	1.155	Thymidylat_synt	LG+I+G	0.198	0.833
29		Rm1D_sub_bind	LG+I+G	0.03	0.966	PX~Vps5	LG+I+G	0.025	1.927
30		Flavoprotein	LG+I+G	0.176	0.875	Thymidylat_synt	LG+G	—	0.853
31		TPR_2 Repeats	LG+I+G	0.035	1.666	Ribosomal_S7e	LG+G	—	0.804
32		WW and FF Repeats	—	—	—	HATPase_c~HSP90	LG+I+G+F	0.048	0.925
33	GW365491	Glyoxal_oxid_N	LG+I+G	0.035	1.4	DUF1929	—	—	—
34	EB044201	Na_H_Exchanger	—	—	—	Nha1_C	LG+I+G+F	0.054	1.052
35	HS540726	WSC Repeats	—	—	—	Glyoxal_oxid_N~DUF1929	LG+I+G	0.05	1.691
36		AAA	LG+I+G	0.242	0.354	HATPase_c ~HATPase_c	LG+G		0.872
37		Adh_short	LG+I+G+F	0.032	1.455	SelP_N~SelP_N~DUF3716	LG+I+G	0.06	1.523
38		ALG3	LG+I+G	0.155	1.623	2OG-FeII_Oxy	—	—	—
39		Allantoicase~Allantoicase	LG+I+G	0.085	0.966	Ank_2~DIL	—	—	—
39		Ank_2~DIL	—	—	—	Allantoicase~Allantoicase	LG+I+G	0.085	0.966
40		Biotin_lipoyl~E3_binding~ 2-oxoacid_dh	LG+I+G+F	0.129	1.112	Sec20	LG+I+G+F	0.021	1.344
41		BTB	LG+G	—	1.94	Bromodomain	LG+I+G	0.074	1.364
42		KH_1	LG+G	—	0.88	Aconitase~Aconitase_C	LG+I+G	0.132	0.614
43		DUF298	LG+I+G	0.024	1.897	Ribosomal_L32e	Dayhoff+I+G	0.081	0.995
44		Flavodoxin_1~ FAD_binding_ 1~NAD_binding_1	LG+I+G+F	0.047	1.107	ETF	LG+I+G	0.046	1.002
45		GTP_EFTU	LG+G	—	0.804	Calreticulin	WAG+I+G	0.112	1.201
46		Methyltransf_16	LG+G	—	0.871	PhyH	LG+I+G	0.109	0.986
47		MIF4G~MA3	LG+I+G+F	0.175	0.868	Pyr_redox_2~ Pyr_redox_dim	LG+I+G+F	0.175	0.868
48		MSC	LG+I+G	0.035	1.313	Ribosomal_L44	RtREV+I+G	0.24	0.418
49	FL604979	Ribosomal_L32e	LG+G	—	0.827	Memo	LG+I+G	0.086	1.013

Table S3. Cont.

ID*	EST support [yes (GenBank ID)/absent (blank)]	Domain 1	Model†	I‡	G§	Domain 2	Model¶	I‡	G§
50		Rogdi_lz	LG+G+F	—	1.113	Rsm22	LG+I+G	0.029	1.505
51		Rsm22~Rsm2	LG+I+G	0.016	1.345	Pantoate_transf~ Pantoate_trans	LG+I+G	0.07	1.002
52		Seipin	LG+I+G+F	0.021	1.701	SurE~TTL	LG+I+G	0.069	1.036
53		Sulfate_transp	LG+I+G+F	0.024	0.991	STAS	LG+I+G	0.069	1.308
54		TGT	LG+G	—	0.423	Thymidylat_synt	LG+I+G	0.227	0.923
55		Tom37~Tom37_C	LG+G	—	2.179	DUF1208	LG+I+G	0.139	0.667
56		TPR_2 Repeats	LG+I+G	0.021	1.414	Ribosomal_S7e	LG+G	—	0.782
57		tRNA-synt_1b	LG+I+G	0.117	1.122	SAICAR_synt	LG+I+G+F	0.151	1.09
58		tRNA-synt_2b~ HGTP_anticodon	LG+I+G	0.142	0.874	Maf	LG+I+G	0.079	1.219
59		Tyrosinase	LG+I+G	0.056	1.185	FAD_binding_3	LG+I+G	0.012	1.812
60		UAA	LG+G+F	—	0.732	PMT~MIR~DUF334	LG+I+G	0.092	0.987
61		Glyco_hydro_71	WAG+I+G+F	0.003	1.222	Peptidase_M18	LG+I+G	0.054	1.233
62		tRNA-synt_1b	LG+I+G	0.117	1.122	tRNA-bind	—	—	—

— indicates not selected by MODELGENERATOR analysis.

*Gene fusion number 1–62.

†Substitution model used for domain 1 phylogeny, selected by MODELGENERATOR analysis. LG/WAG/RTREV/CPREV represent alternative amino acid replacement matrices.

‡Invariant sites.

§Alpha parameter for the gamma distribution.

¶Substitution model used for domain 2 phylogeny, selected by MODELGENERATOR analysis. LG/WAG/RTREV/CPREV represent alternative amino acid replacement matrices.

Table S4. Protein domains used for fungal species phylogeny

PFAM	NCBI Saccharomyces	Domain name
PF00022.14	AAA34391.1	Actin
PF00709	CAA88590.1	Adenylosuccinate synthetase
PF05856	NP_012912.1	ARP2/3 complex 20 kDa subunit (ARPC4)
PF04045	NP_014433.1	Arp2/3 complex, 34 kDa subunit p34-Arc
PF02374	AAT93183.1	Ars operon
PF04729	NP_012420.1	ASF1 like histone chaperone
PF03477_..._PF02867	NP_010993.1	ATP cone _ Ribonucleotide reductase
PF01813	NP_010863.1	ATP synthase subunit D
PF08145	NP_013764.1	BOP1
PF05291	NP_009806.3	Bystin
PF04054	NP_010017.2	CCR4-Not complex component, Not1
PF04078	NP_014111.1	Cell differentiation family, Rcd1-like
PF01394.15	EDN61921.1	Clathrin
PF07718	EDN60571.1	Coatamer beta C-terminal region
PF08767	NP_011734.3	CRM1 C-terminal
PF06418	CAA37941.1	CTP synthase N terminus
PF04442	NP_015193.1	Cytochrome *c* oxidase assembly protein CtaG/Cox11
PF02167	NP_014708.1	Cytochrome C1
PF02628	EDN63118.1	Cytochrome oxidase assembly protein
PF01916	EDN62305.1	Deoxyhypusine synthase
PF00940	EDN59114.1	DNA-dependent RNA polymerase
PF08351_PF05127	NP_014267.1	Domain of unknown function (DUF1726)
PF04034	NP_014648.1	Domain of unknown function (DUF367)
PF04037	NP_013967.1	Domain of unknown function (DUF382)
PF01912	Q12522.1	eIF-6 family
PF03587	NP_013287.1	EMG1/NEP1 methyltransferase
PF02919_PF01028	NP_014637.1	Eukaryotic DNA topoisomerase I
PF03332	NP_116609.1	Eukaryotic phosphomannomutase
PF08644	NP_011308.1	FACT complex subunit (SPT16/CDC68)
PF01125	NP_009990.1	G10 protein
PF00342	EDV11919.1	Glucose-6-phosphate isomerase
PF00953	NP_009802.3	Glycosyl transferase family 4
PF00009.22	EDN61207.1	GTP-binding elongation factor family, EF-Tu/EF-1A subfamily
PF00012.15	AET14830.1	Hsp70
PF00183.13	P02829.1	HSP90
PF01875	EDN63335.1	Memo-like protein
PF07994	NP_012382.2	Myo-inositol-1-phosphate synthase
PF00063.16	AAA34810.1	Myosin head (motor domain)
PF01233_PF02799	P14743.1	Myristoyl-CoA:protein *N*-myristoyltransferase, N and C-terminal domain
PF01592	NP_014869.3	NifU-like N-terminal domain
PF04981	AAA74491.1	NMD3 family
PF04065	EDN61456.1	Not1 N-terminal domain, CCR4-Not complex component
PF06732	NP_011617.1	Pescadillo N terminus
PF10559_PF00344	NP_009842.1	Plug domain of Sec61p
PF01379_PF03900	CAA77804.1	Porphobilinogen deaminase
PF00490	P05373.2	Porphobilinogen synthase
PF08082_..._PF08084	NP_012035.1	PRO8NT (NUC069), PrP8 N-terminal domain
PF06777	NP_011098.3	Protein of unknown function (DUF1227)
PF06026	NP_014738.1	Ribose 5-phosphate isomerase A (phosphoriboisomerase A)
PF01775	NP_013969.3	Ribosomal L18ae/LX protein domain
PF01294	NP_013862.1	Ribosomal protein L13e
PF00828	CAA64550.1	Ribosomal protein L18e/L15
PF01092	NP_015235.1	Ribosomal protein S6e
PF01251	NP_014739.1	Ribosomal protein S7e
PF00833	NP_013688.1	Ribosomal S17
PF01015	NP_013648.1	Ribosomal S3Ae family
PF09416	NP_013797.1	RNA helicase (UPF2 interacting domain)
PF04563.10	CAA99357.1	RNA polymerase beta subunit
PF04997.7	EEU08500.1	RNA polymerase Rpb1, domain 1
PF07780	NP_009877.1	Spb1 C-terminal domain
PF03531	NP_013642.1	Structure-specific recognition protein (SSRP1)
PF00118.19	P19882.1	TCP-1/cpn60 chaperonin family
PF00303	AAA60940.1	Thymidylate synthase

Table S4. Cont.

PFAM	NCBI Saccharomyces	Domain name
PF00091.20	AAA35181.1	Tubulin/FtsZ family, GTPase domain
PF00091.20	CAA24603.1	Tubulin/FtsZ family, GTPase domain
PF01209	P49017.1	UbiE/COQ5 methyltransferase family
PF03690	NP_011083.3	Uncharacterized protein family (UPF0160)

Fifty-seven of these 67 proteins are derived from ref.1. The 10 remaining proteins (gray shading) are gene families we favor for multigene phylogeny of the Fungi.

1. Torruella G, et al. (2012) Phylogenetic relationships within the Opisthokonta based on phylogenomic analyses of conserved single-copy protein domains. *Mol Biol Evol* 29(2):531–544.

Dataset S1 (XLS)

Data showing genome checks of gene fusion/fission annotations.

SI Appendix (PDF)

Collated phylogenetic data and genome analysis of fusions 1–63. This PDF file contains all 63 final accepted gene fusions, with the numbering scheme used in Fig. 2 (32 contains two fusions labeled "32a" and "32b"). The file therefore is split into 62 sections, each containing the gene sequences (in FASTA format) of the separate domains that form the gene fusion, a note on how the tree was constructed, a note explaining the absence of one domain tree (if absent), and the trees annotated with PFAM domains for each sequence given. Bootstrap supports are given (in red). Putative genome annotation corrections on the presence and absence of gene-fusion characters are labeled with ticks or crosses (where appropriate). The data supporting these annotation corrections can be found in Dataset S1. Trees that indicate additional fusions are labeled with the appropriate number and can be found elsewhere in the file as indicated by the number given. Diagrammatic results from Mesquite analysis also are included for each fusion.