ARTICLES

Natural history and evolutionary principles of gene duplication in fungi

Ilan Wapinski[1,2,3], Avi Pfeffer[3], Nir Friedman[4] & Aviv Regev[1,5]

Gene duplication and loss is a powerful source of functional innovation. However, the general principles that govern this process are still largely unknown. With the growing number of sequenced genomes, it is now possible to examine these events in a comprehensive and unbiased manner. Here, we develop a procedure that resolves the evolutionary history of all genes in a large group of species. We apply our procedure to seventeen fungal genomes to create a genome-wide catalogue of gene trees that determine precise orthology and paralogy relations across these species. We show that gene duplication and loss is highly constrained by the functional properties and interacting partners of genes. In particular, stress-related genes exhibit many duplications and losses, whereas growth-related genes show selection against such changes. Whole-genome duplication circumvents this constraint and relaxes the dichotomy, resulting in an expanded functional scope of gene duplication. By characterizing the functional fate of duplicate genes we show that duplicated genes rarely diverge with respect to biochemical function, but typically diverge with respect to regulatory control. Surprisingly, paralogous modules of genes rarely arise, even after whole-genome duplication. Rather, gene duplication may drive the modularization of functional networks through specialization, thereby disentangling cellular systems.

Gene duplication and loss are major forces of evolutionary innovation, facilitating the development of new functions and pruning of old ones[1,2]. Nonetheless, the natural history of gene duplication and loss is poorly understood. What classes of genes readily evolve through duplication and loss? Do whole-genome duplication events reshape the genome in a qualitatively distinct way? What innovations typically arise from gene duplication events? Studies addressing such questions[3–11] have been limited by the difficulty of tracing the exact evolutionary history of genes.

The growing availability of sequenced genomes enables the direct reconstruction of a genome-wide history of gene duplication and loss across species[3,7]. Here, we describe a computational method for reconstructing this history and apply it to the genomes of seventeen Ascomycota fungi spanning 300 million years of evolution[12–21]. The results suggest evolutionary principles applicable for fungi and possibly more generally.

Method for identifying orthologues and paralogues

Systematic study of gene duplication and loss requires reliable resolution of gene orthology and paralogy, a notoriously difficult problem[22–31]. We designed SYNERGY, a scalable method for resolving gene ancestry for all genes across multiple genomes (Fig. 1, Supplementary Fig. 1)[32]. The input is a species phylogeny and, for each extant species, the sequences of predicted genes and their chromosomal positions. SYNERGY partitions these genes into 'orthogroups'. Each orthogroup consists of all (and only) the genes descended from a single ancestral gene in their last common ancestral species, and is associated with a gene tree that describes the history of speciation, duplication and loss events for its genes (Methods).

An orthogroup catalogue for Ascomycota fungi

We applied SYNERGY to the complete set of 121,050 predicted protein-coding genes from seventeen genomes of Ascomycota fungi, including the model organisms *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe* (Fig. 2a, Methods). The phylogeny includes a whole-genome duplication (WGD) event[14,19,33] (Fig. 2a, red star). SYNERGY produced a catalogue of 30,110 orthogroups (Fig. 2b). Of these, 19,006 were singleton genes with no recognizable orthologues (Supplementary Note 1). We further analysed the 11,103 multigene orthogroups. The orthogroups and trees are available at http://www.broad.mit.edu/regev/orthogroups/.

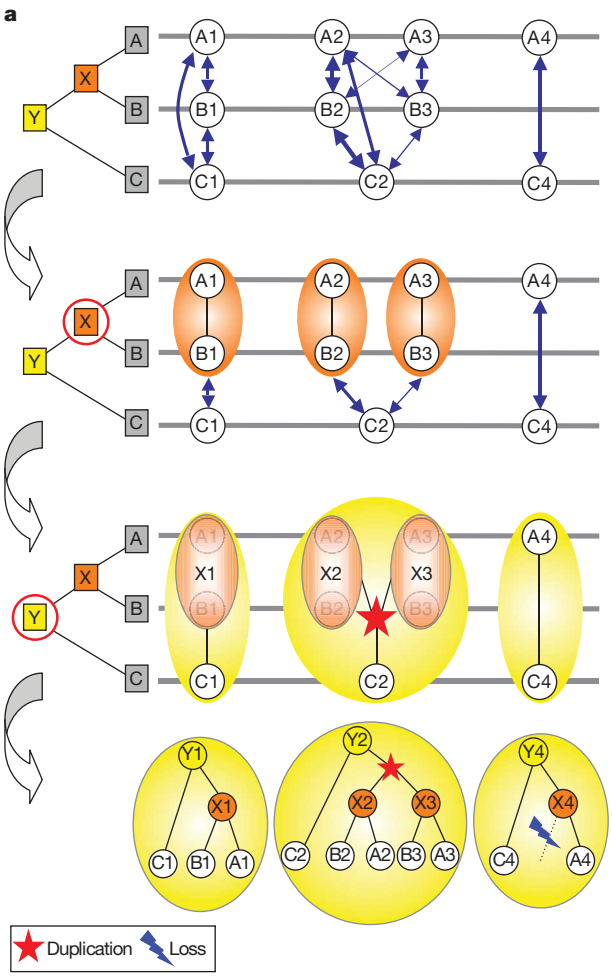
SYNERGY made high-quality predictions by several benchmarks (Methods, Supplementary Notes 2 and 3). To test sensitivity to the input quality, we applied SYNERGY to different subsets of organisms and of genes in each genome. We examined how each orthogroup was reconstructed under these perturbations, deriving four confidence measures for each orthogroup. Overall, SYNERGY was remarkably robust (Supplementary Note 2). SYNERGY's predictions also agree well with those of two independent manual assignments of orthology and paralogy[21,33] (Supplementary Note 3). Finally, SYNERGY showed high specificity and sensitivity on data attained by forward simulated evolution.

Gene duplication and loss across Ascomycota evolution

The reconstructed orthogroups show a range of evolutionary patterns. These are summarized by the extended phylogenetic profile (EPP) of each orthogroup, defined as the number of genes present in each extant and ancestral species. For example, 'uniform' orthogroups (Fig. 2c), with no duplication or loss events, have EPPs consisting only of ones. Other orthogroups exhibit duplications (Fig. 2d, red star) or losses (Fig. 2d, blue strikes) and their EPPs may consist of noughts, ones, twos, and so on (Fig. 2e). An orthogroup with at least one gene present in all species (Supplementary Fig. 2b) is 'persistent'. From the EPP, we can derive an extended copy-number variation profile that records the change in copy number at each position in the species tree (Supplementary Fig. 2).

[1]Broad Institute of MIT and Harvard, 7 Cambridge Center, Cambridge, Massachusetts 02142, USA. [2]FAS Center for Systems Biology, Harvard University, 7 Divinity Avenue, [3]School of Engineering and Applied Sciences, Harvard University, 33 Oxford Street, Cambridge, Massachusetts 02138, USA. [4]School of Computer Science and Engineering, Hebrew University, Jerusalem 91904, Israel. [5]Department of Biology, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, Massachusetts 02139, USA.

By tallying these profiles from all orthogroups, we find the numbers of genes, appearances, duplications and losses that occurred throughout Ascomycota evolution (Fig. 2f, Supplementary Fig. 3a). The 5,972 (54%) orthogroups present in the clade spanning the Hemiascomycota and Euascomycota were defined as 'ancestral', accounting for 6,047 genes in the reconstructed last common ancestor. 4,873 (84%) of *Saccharomyces cerevisiae* genes belong to these ancestral orthogroups. The ancestral orthogroups are enriched in essential *S. cerevisiae* genes; 1,008 of 1,047 genes essential for growth in rich medium are ancestral ($P < 10^{-30}$, Fischer's exact test) as are 668 of 730 genes essential only for growth in other conditions[34] ($P < 10^{-5}$). Nevertheless, 36 essential genes are not ancestral (for example, 9/80 spindle pole body proteins, $P < 10^{-6}$), suggesting that new essential functions can arise, albeit rarely (Supplementary Note 4).



Orthogroups that are not ancestral 'appear' at specific points in the phylogeny and reveal evolutionary innovations. For example, the clade spanning *S. cerevisiae* and *Kluyveromyces waltii* is marked by appearing orthogroups with *S. cerevisiae* genes related to meiosis and sporulation (51/166 sporulation genes, $P < 10^{-6}$), including the master meiosis regulator IME1[35] (Supplementary Fig. 2a). The Euascomycota clade contains many appearances (3,726 orthogroups, 72% of all appearing orthogroups); roughly half show no similarity to other orthogroups or to a more distant fungus, *Cryptococcus neoformans* (Supplementary Fig. 3d), demonstrating extensive genomic innovation within the Euascomycota[16].

We also find coordinated gene losses, indicating major changes in biological processes. For example, *Yarrowia lipolytica* has significant losses in orthogroups containing meiotic recombination genes ($P < 10^{-5}$). Interestingly, the genes lost in *Candida glabrata* extensively overlap those independently lost in the ancestor of *Candida albicans* and *Debaryomyces hansenii* ($P < 10^{-20}$), possibly reflecting the fact that these are all opportunistic or occasional human pathogens.

Copy number volatility corresponds to a functional dichotomy

The observed variation in copy number changes among orthogroups is inconsistent with random expectation (Fig. 3a, Methods). We assigned a volatility score to each orthogroup depending on the number and phylogenetic position of duplication and loss events, with 1,018 uniform orthogroups at one end of the scale and 313 'volatile' orthogroups (score >3 s.d. above the mean) at the other (Fig. 3a). Evolutionary forces have acted very differently on these two classes: the uniform and volatile orthogroups show diametrically opposed patterns in their function, regulation and essentiality in *S. cerevisiae* (Fig. 3, Supplementary Table 1).

We first tested for functional distinctions between uniform and volatile orthogroups, based on gene ontology annotations in *S. cerevisiae*[36]. Volatile orthogroups are enriched ($P < 10^{-5}$) for genes that encode peripheral transporters, receptors and cell wall proteins, and genes that participate in stress responses. In contrast, uniform orthogroups are enriched ($P < 10^{-5}$) for genes involved in essential growth processes, genes residing in the nucleus, nucleolus, mitochondrion, endoplasmic reticulum and Golgi apparatus, and genes essential for viability.

We next examined whether the evolutionary dichotomy is also aligned with the transcriptional program of *S. cerevisiae*. Using data from 1,216 gene expression profiles, we organized *S. cerevisiae* genes into a hierarchy of 163 transcriptional modules, each containing functionally related genes with a coherent expression pattern[37] (Fig. 4a, Supplementary Fig. 4, Supplementary Table 2, and

Figure 1 | The SYNERGY algorithm. a, Orthogroup construction. SYNERGY starts (top) with a collection of genes (A1, B1, C1 and so on), their chromosomal order (grey lines) and sequence distances (blue arrows; arrows of the same thickness have similar sequence distances). It then builds orthogroups as it climbs the species tree. First, it collects the genes in species A and B that share a common ancestor in species X (second panel, orange ovals). Then, it merges orthogroups formed in the previous stage with the genes in C, resulting in new orthogroups representing ancestral genes in species Y (third panel, yellow ovals). The orthogroups assembled at each stage are associated with gene trees reflecting divergence, duplication and loss events (bottom). **b**, Gene tree reconstruction and refining orthogroup assignments. An unrooted phylogeny is reconstructed for the genes and sub-orthogroups in each putative orthogroup (dashed oval). Some rootings (purple arrow) indicate that all the genes descended from a common ancestor (for example, X3, bottom left). Others (green arrow) show that a duplication occurred at the root of the gene tree (for example, X2 and X3, bottom right). In the latter case, the orthogroup is partitioned before proceeding.

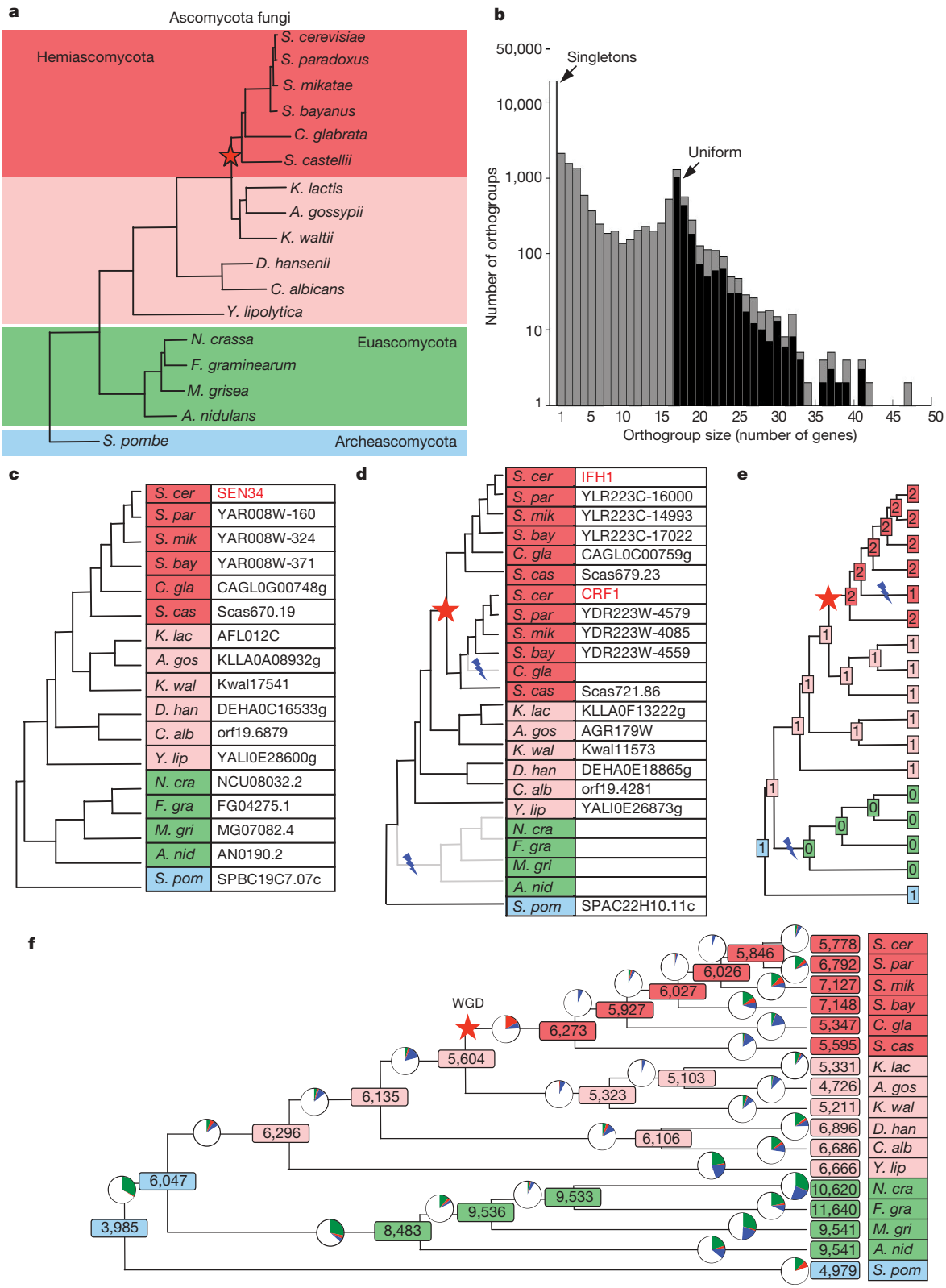
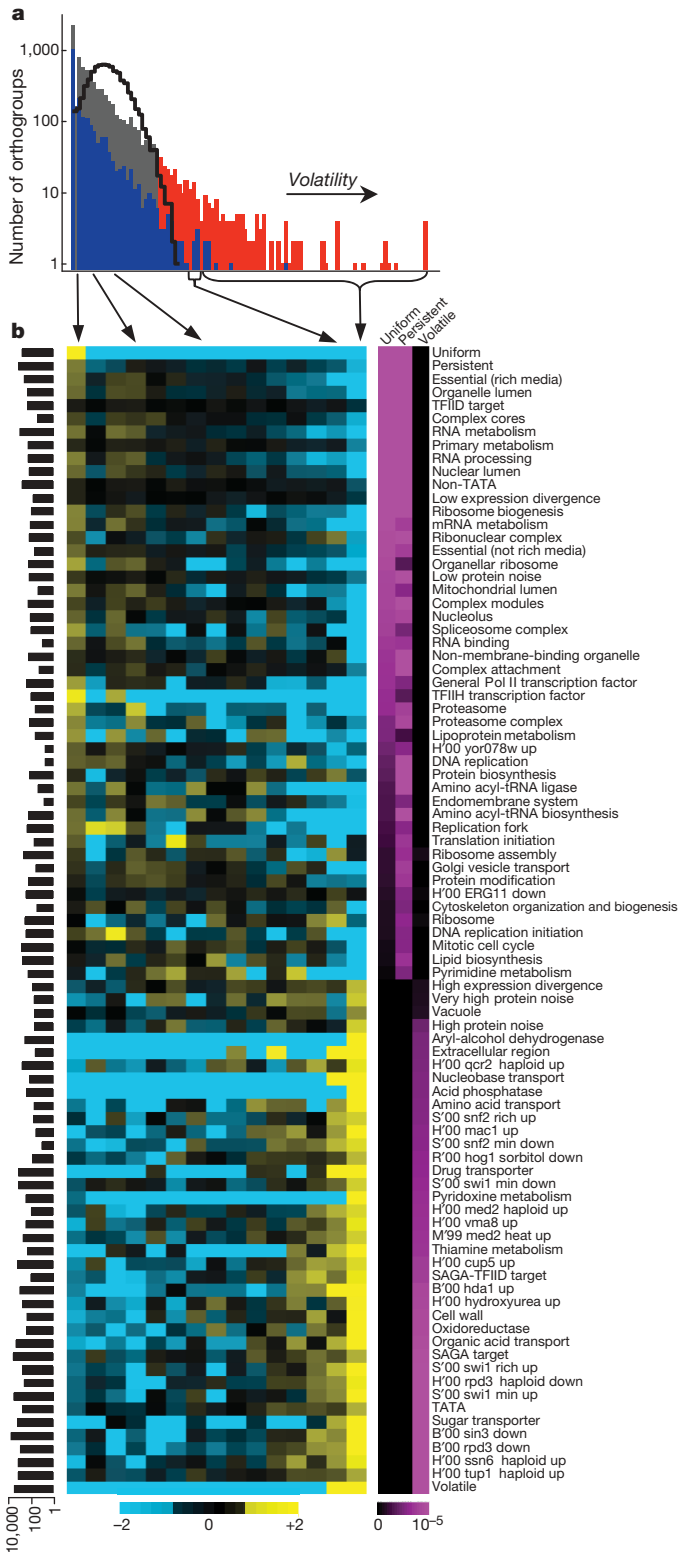


Figure 2 | A gene ancestry catalogue for Ascomycota fungi. **a**, Species tree showing the Hemiascomycota (pink), Euascomycota (green), and Archeascomycota (blue) clades, the WGD (red star), and post-WGD species (darker pink). **b**, Distribution of orthogroup sizes. Grey bars, total number of orthogroups of a certain size (number of genes). Black bars, the number of persistent orthogroups of a certain size. The uniform arrow points to orthogroups, which are persistent orthogroups with exactly 17 genes. The singleton arrow points to orthogroups with a single gene. **c**, A uniform orthogroup. The topology of the gene tree (left panel) is identical to that of the species tree. **d**, A non-uniform orthogroup with a single duplication event (red star) and two loss events (blue strikes). **e**, The extended phylogenetic profile of the orthogroup in **d** summarizes the number of genes at each extant and ancestral species in the tree (numbered boxes). **f**, Reconstructed evolutionary events and gene counts. Each species is annotated with the number of known or reconstructed genes (rectangles). Pie-charts on branches denote the fraction of appearing (green), duplicated (red), and lost (blue) genes inferred in the corresponding branch (see Supplementary Fig. 3 for exact numbers).

Methods). Analysing the volatility scores of these transcriptional modules, we see that the evolutionary dichotomy follows the regulatory branches (Fig. 4c, d, Supplementary Table 3). Modules in the 'cell cycle and meiosis' and 'fundamental processes' branches are enriched ($P < 10^{-4}$) for uniform and persistent orthogroups, whereas modules in the 'development' and 'stress and carbohydrate metabolism' branches are enriched ($P < 10^{-4}$) for volatile orthogroups.

These distinctions indicate a limit not only on the ability to lose key growth genes, but also on the ability to maintain them in duplicate. As suggested by the "gene balance hypothesis"[9], one reason for this may be that such genes often encode components of essential cellular machineries requiring stoichiometric balance. Indeed, we found that *S. cerevisiae* genes encoding core components of protein complexes[38] are enriched in uniform orthogroups (338/844 complex 'core' genes, $P < 10^{-32}$). Furthermore, uniform and persistent orthogroups are enriched for *S. cerevisiae* genes displaying haploinsufficiency[39] ($P < 10^{-4}$ and $P < 10^{-6}$, respectively). However, the dichotomy extends beyond protein complexes to many cellular processes and includes orthogroups with moderately low and high volatility scores (Fig. 3b), suggesting a general principle affecting the vast majority of genes in the genome.



To test whether differential flexibility in copy number between uniform and volatile orthogroups reflects global functional constraints, we examined the variation in their respective transcripts and proteins. We found that the volatile orthogroups are enriched in genes whose expression changes significantly in response to many single-gene knockouts[40] ($P < 10^{-5}$–10^{-22}; notably deletions of chromatin modifiers), genes with noisy levels of protein abundance within isogenic *S. cerevisiae* cells[41] ($P < 10^{-4}$), genes the transcription of which is regulated through the SAGA complex and the TATA box[42] ($P < 10^{-9}$), and genes with variable RNA expression across species[43] ($P < 10^{-11}$) (Fig. 3, Supplementary Table 1). Conversely, the uniform orthogroups are enriched in genes whose expression is largely unchanged in response to single-gene knockouts, genes whose protein levels tend to be tightly controlled ($P < 10^{-8}$), genes whose transcription is TATA-independent and regulated through TFIID[42] ($P < 10^{-24}$), and genes whose RNA expression shows less variation across species ($P < 10^{-15}$) (Fig. 3, Supplementary Table 1).

These results highlight a general bipolar principle that governs tolerance to duplications and losses. Copy-number variation in stress-responsive genes may not only be tolerable but beneficial, allowing adaptation to diverse ecological niches. In contrast, genes essential for cell growth, including those necessary for intricate complexes, cannot readily tolerate such noise and tend not to evolve by gradual duplication and loss. This evolutionary dichotomy aligns closely with a bipolarity in gene function, transcriptional program and expression noise across cells, strains and species[41,43], all reflecting similar functional constraints on the amount of gene products in the cell. Furthermore, shared functional constraints on copy-number variation also manifest in remarkably synchronized and concerted patterns of specific duplications and losses in many orthogroups harbouring functionally related or interacting genes (Supplementary Note 5, Supplementary Figs 5 and 6, and Supplementary Table 4).

Whole-genome duplication alters the nature of duplication

We next explored whether these functional principles generalize to the WGD event. We found that duplications associated with the WGD show a strikingly different pattern (Fig. 4e, Supplementary Table 3): many transcriptional modules that maintain little duplication elsewhere in the phylogeny are associated with a high level of

Figure 3 | A functional dichotomy of uniform, persistent and volatile orthogroups. a, Distribution of orthogroup volatility scores. Uniform orthogroups (leftmost blue column) are the least volatile. Orthogroups whose score is above three standard deviations from the expected mean are labelled as volatile (red columns). The remainder of the distribution is shown (grey bars) and the distribution of scores for persistent orthogroups is superimposed (blue bars). The expected distribution when sampling random orthogroups from the evolutionary model is shown as a black line. **b**, Gene class annotations that are significantly enriched among uniform, persistent or volatile orthogroups (purple, colour scale is saturated at $P < 10^{-5}$). The functional and mechanistic dichotomy between volatile and non-volatile orthogroups largely reproduces along the range of volatility scores (left columns are bins of orthogroups with similar volatility scores; rows are significant annotations). Higher (yellow) and lower (blue) relative enrichment compared to the expected enrichment in the class is shown. The colour scale is saturated at 2-fold. Class sizes are shown on the left (black bars).

volatility during the WGD. Examples include 'ribosome biogenesis' in all post-WGD lineages ($P < 10^{-5}$) and 'ER protein modification' in *Saccharomyces castellii* ($P < 10^{-3}$).

Thus, the WGD is associated with a qualitatively different pattern of duplication. The 'gene balance hypothesis'[4,9] postulates that this effect is due to post-WGD retention of paralogues for all members of a complex. Indeed, *S. cerevisiae* genes with either haploinsufficiency[39] or overexpression[44] phenotypes are enriched in orthogroups that duplicated only in the WGD ($P < 10^{-9}$). Furthermore, several modules representing essential machineries are enriched for WGD paralogues[45] (for example, 'rRNA biogenesis' $P < 10^{-5}$, and 'ribosome', $P < 10^{-36}$). However, the expanded scope of the WGD is observed beyond complexes within more volatile modules. The simultaneous duplication of all genes in a module in a WGD may permit retention of paralogues in orthogroups that are otherwise constrained against duplication, and may be a principal way in which WGD events contribute to evolutionary innovations[4,9].

Gene duplication results in limited biochemical divergence

We next explored the types of functional innovations that arise from gene duplications. In principle, both paralogues can either 'retain' the same function (Fig. 5a) or one (or both) can 'migrate' to assume a distinct function (Fig. 5b, c). Migration can either reflect the development of a novel function (neofunctionalization[11], Fig. 5b) or a division of labour, in which each paralogue assumes only some functions of the ancestral gene (subfunctionalization[11], Fig. 5c). Given the long-postulated importance of gene duplication in innovation[1,2], we hypothesized that migration would be the predominant evolutionary mode.

We quantified the extent to which paralogous gene pairs remain in or migrate from a variety of gene classes in *S. cerevisiae* (gene ontology functional categories, genes with shared regulatory mechanisms, protein complexes and transcriptional modules). We calculated the fraction of paralogous pairs that are retained within each class (Methods, and Supplementary Figs 7, 8). To avoid confounding factors, we studied only cases in which both paralogues had been annotated and the annotation had not been inferred solely from sequence similarity. Surprisingly, our analysis shows that paralogous pairs rarely migrate between functional gene ontology categories (Supplementary Figs 7–9, and Supplementary Table 5). The retention rate is highest for the 'molecular function' categories (92%) and somewhat lower for 'biological process' (85%) and 'cellular component' (81%) categories.

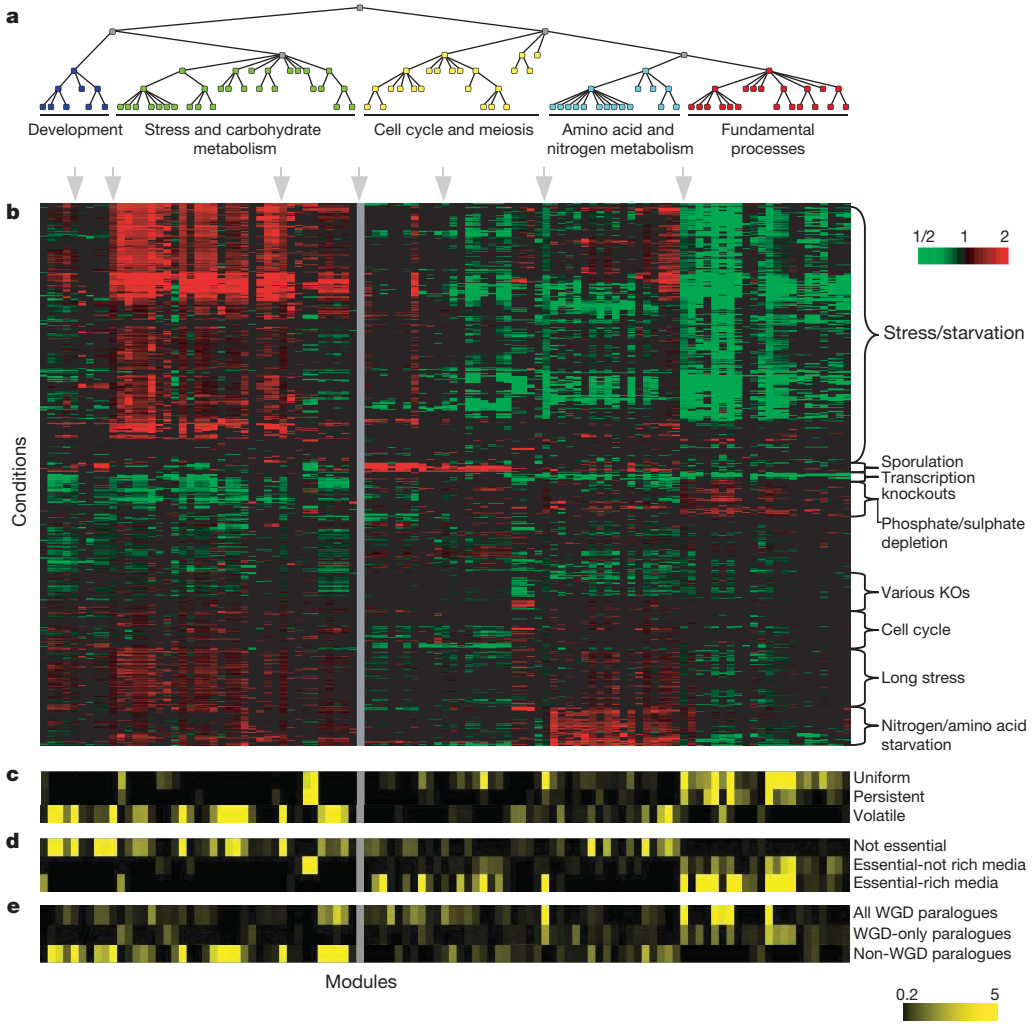


Figure 4 | Evolutionary profiles correspond to the hierarchical modular organization of the yeast transcriptional system. **a**, A functional hierarchy of *S. cerevisiae* transcriptional modules. **b**, Expression patterns of modules. Shown is the average expression of module genes (columns) in each expression array (rows, with main condition names marked on right). Red, induced; green, repressed; black, no significant change. The colour scale is saturated at 2-fold. **c**–**e**, Evolutionary characterization of modules. Enrichment significance (negative log *P*-value; yellow indicates significant; the colour scale is saturated at 10^{-5}) for the projected orthogroup classes from each module (columns) against various phylogenetic attributes (rows). **c**, Growth modules are enriched for uniform and persistent genes, with the notable exception of amino acid metabolism modules; whereas stress modules are enriched in volatile genes. **d**, Essential genes are enriched in modules from the 'growth', but not the 'stress', groups. **e**, This dichotomy is violated in the WGD.

To reveal innovation at a finer resolution than the relatively coarse functional classes, we measured the fraction of shared interaction partners for each pair of paralogues in molecular networks. We examined both biochemical networks[46] of physically interacting proteins (reflecting molecular function), and genetic networks[46] of synthetic phenotypes in double mutants (reflecting biological processes). We find that in both networks roughly half of the paralogues share a significantly high proportion of their interaction partners (136/318 pairs in the genetic network and 225/543 in the biochemical network), much more than would be expected in comparable (degree-preserving) random networks (Methods, Supplementary Fig. 10b, f). Thus, many pairs show little migration from their pre-duplication organization (Supplementary Fig. 11a), supporting the broader result at both the biochemical and the functional level.

The remaining paralogues typically share no interaction partners and may indicate neofunctionalization (Supplementary Fig. 10a, e). Indeed, paralogous pairs had more biochemical interaction partners than would be expected by chance[11] (11.64±24.73 versus 6.99±12.66; $P < 10^{-3}$, Mann–Whitney *U*-test), providing global evidence for neofunctionalization. Many of these 'disjoint paralogue pairs' are dispersed in the biochemical network (78% are separated by four or more proteins), implying divergence in molecular function (Supplementary Fig. 11b). In contrast, half of them are immediate neighbours in the genetic network (Supplementary Fig. 11c), suggesting a related biological process. This is consistent with the role of duplicate genes as either genetic 'back-ups' when their paralogues are compromised[6,47] or with division of labour through subfunctionalization (see below).

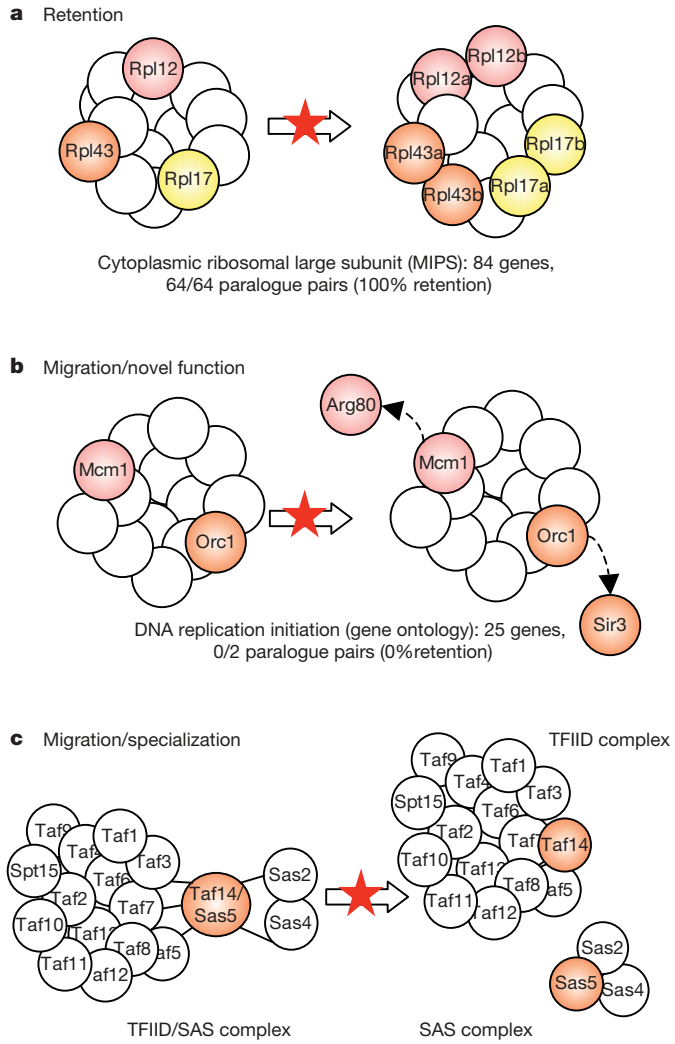


Figure 5 | Functional conservation and innovation of paralogues in classes and networks. Shown is the reconstructed functional history of several paralogous pairs of *S. cerevisiae* genes (circles with gene name). After duplication (arrow), paralogous gene pairs can be either retained within the same class (**a**), migrate to assume new functions (**b**), or specialize into distinct functions, resulting in modularization (**c**).

Gene duplication innovates through regulatory divergence

Another source of innovation is regulatory divergence. We inspected the migration of paralogous pairs with respect to gene classes representing regulatory mechanisms (genes that are targets of the same transcription factor[48] or contain the same *cis*-regulatory motif[48] or RNA-binding motif[49]) or expression patterns (transcriptional modules). We find that paralogous genes usually migrate with respect to these gene classes. In most cases (70%), regulatory gene classes contain no retained paralogy relations within them, reflecting either novel regulation or regulatory specialization between paralogues (Supplementary Figs 7f, h and 9d, e). Transcriptional modules exhibit an intermediate behaviour, with 26% of paralogous gene pairs having migrated between modules, both within and between the major branches of the module hierarchy (Supplementary Figs 7d, 8 and 9f).

Our analysis shows that paralogues diversify most frequently at the level of regulation, less frequently through changes in their cellular component, biological process or molecular interactions, and rarely in biochemical function. Although some of these differences may stem from variations in the quality and resolution of available annotations, multiple functional and regulatory data sources support this broad distinction. This highlights inherent limitations of gene duplication in accomplishing molecular innovation. It also emphasizes the overwhelming importance of regulatory divergence in driving functional divergence and reconfiguring molecular systems after duplication[5,8].

Coordinated migration of multiple paralogues is rare

When several genes in a class are duplicated they can either migrate in a coordinated manner resulting in two paralogous classes (Supplementary Fig. 12) or disperse into different classes (Fig. 5b). We expect coordinated migration after simultaneous duplications (for example, WGD[4]). To test this, we counted the number of paralogous gene pairs that connect each pair of gene classes (Supplementary Fig. 9).

Surprisingly, coordinated migration is rare: gene classes (functional, regulatory, or transcriptional) rarely share more than one or two paralogous relations, regardless of the overall proportion of paralogues retained (Supplementary Fig. 9). The few observed paralogous classes are very small and formed gradually (see, for example, Supplementary Fig. 12). Thus, paralogues overwhelmingly disperse as individuals rather than migrate in a coordinated fashion. Although theory suggests[4] that paralogous classes might form in a single concerted step (such as after a WGD), we observed little evidence of this here.

The same patterns of retention, migration and interaction are observed even among paralogues derived only from the WGD (Supplementary Fig. 8b, data not shown) and validated by manual curation[33]. Thus, while the WGD allows qualitatively different gene duplications, the subsequent patterns of innovation (or lack thereof) follow the same general trajectory for both WGD and non-WGD paralogues.

Conclusions

We set out to uncover the evolutionary potential and constraints of gene duplication and loss. We created a rich resource of evolutionary history in fungi and compared these evolutionary patterns with a wealth of functional and genomics data for *S. cerevisiae*, to uncover the principles that govern copy number changes in Ascomycota.

What is the contribution of gene duplication to system evolution and how does it affect the modularity of molecular systems? Earlier studies suggested that paralogous modules form in massive duplication events[4], but we found that paralogous modules are rare, even after a WGD. An alternative mechanism is suggested by the fact that many paralogous pairs genetically interact with each other despite having no shared physical interactions (Supplementary Fig. 11c). This may indicate a partial 'division of labour' (subfunctionalization) between two paralogous proteins that become physically or temporally separated[6]. Such specialization could modularize a molecular network by severing links within a network when duplicating a node[50]. For example, a single ancestral gene may have participated in two distinct complexes (Fig. 5c). If each of the paralogues specialized to optimally perform one of the functions and interact only with the members of one of the complexes, the resulting network will be more modular. Thus, increasing gene copy number may eventually simplify a system rather than making it more complex. Modularization could relax opposing constraints on a single component and thus set in motion further specialization and refinement[11]. As our knowledge of molecular and genetic networks improves, further studies can systematically assess this possibility.

Principles similar to those described here may apply to copy number variation in metazoan genomes. For example, evidence from vertebrates and *Arabidopsis* suggests that genes encoding signalling molecules and transcription factors are duplicated in WGD events, but rarely otherwise[3,7,9]. The reconstruction algorithm and analytical framework here make it possible to test such hypotheses in other taxa, and will facilitate other novel studies of the evolution of genes, genomes and systems.

METHODS SUMMARY

Orthology assignment and gene tree reconstruction. SYNERGY assigns orthologies in a step-wise, bottom-up fashion, solving it for each ancestral node in a species tree, starting at the leaves and concluding at the root. At each stage, SYNERGY first clusters together the genes or orthogroups from previous stages that share significant sequence similarity into new putative orthogroups (Fig. 1a). It then constructs a phylogeny of these intermediate orthogroups (Fig. 1b) using a modified neighbour-joining procedure based on the amino acid similarities that have been pre-computed and shared synteny scores. The sub-trees from each stage are based on the reconstructions at earlier stages. Each tree is rooted using a score based on sequence similarity, conserved synteny, and the inferred number of duplications and losses. Trees that invoke fewer unlikely duplication and loss events will be favoured over those that incur many such events. If the rooted tree indicates that all the orthogroups (genes) in that tree descended from a single hypothetical gene at the current stage, the cluster is defined as an orthogroup along with its tree (Fig. 1b, left). Otherwise, the orthogroup is partitioned by removing the inferred root of the gene tree (Fig. 1b, right). This process may be repeated until each orthogroup consists of genes that share a single common ancestor at the current level of reconstruction. Thus, after each stage, a complete orthology assignment and gene tree reconstruction for the genes below that node has been made. These are used as the input to subsequent stages at higher nodes in the species tree. When this procedure is completed at the root of the species tree, the genes for all species have been assigned to their orthogroups and placed in their respective trees.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 21 March; accepted 20 July 2007.

1. Ohno, S. *Evolution by Gene Duplication* (Allen and Unwin, London, 1970).
2. Lynch, M. & Conery, J. S. The origins of genome complexity. *Science* **302,** 1401–1404 (2003).
3. Blomme, T. *et al.* The gain and loss of genes during 600 million years of vertebrate evolution. *Genome Biol.* **7,** R43 (2006).
4. Freeling, M. & Thomas, B. C. Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. *Genome Res.* **16,** 805–814 (2006).
5. Gu, Z., Rifkin, S. A., White, K. P. & Li, W. H. Duplicate genes increase gene expression diversity within and between species. *Nature Genet.* **36,** 577–579 (2004).
6. Kafri, R., Bar-Even, A. & Pilpel, Y. Transcription control reprogramming in genetic backup circuits. *Nature Genet.* **37,** 295–299 (2005).
7. Maere, S. *et al.* Modeling gene and genome duplications in eukaryotes. *Proc. Natl Acad. Sci. USA* **102,** 5454–5459 (2005).
8. Makova, K. D. & Li, W. H. Divergence in the spatial pattern of gene expression between human duplicate genes. *Genome Res.* **13,** 1638–1645 (2003).
9. Papp, B., Pal, C. & Hurst, L. D. Dosage sensitivity and the evolution of gene families in yeast. *Nature* **424,** 194–197 (2003).
10. Scannell, D. R., Byrne, K. P., Gordon, J. L., Wong, S. & Wolfe, K. H. Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. *Nature* **440,** 341–345 (2006).
11. He, X. & Zhang, J. Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. *Genetics* **169,** 1157–1164 (2005).
12. Hong, E. L. B. R. *et al. Saccharomyces Genome Database* ⟨http://www.yeastgenome.org⟩ (2005).
13. Arnaud, M. B. C. M. *et al. Candida Genome Database* ⟨http://www.candidagenome.org⟩ (2006).
14. Dietrich, F. S. *et al.* The *Ashbya gossypii* genome as a tool for mapping the ancient *Saccharomyces cerevisiae* genome. *Science* **304,** 304–307 (2004).
15. Dujon, B. *et al.* Genome evolution in yeasts. *Nature* **430,** 35–44 (2004).
16. Galagan, J. E. *et al.* Sequencing of *Aspergillus nidulans* and comparative analysis with *A. fumigatus* and *A. oryzae*. *Nature* **438,** 1105–1115 (2005).
17. *Fusarium graminearum* Sequencing Project. ⟨http://www.broad.mit.edu⟩ (Broad Institute of Harvard and MIT, 2003).
18. Dean, R. A. *et al.* The genome sequence of the rice blast fungus *Magnaporthe grisea*. *Nature* **434,** 980–986 (2005).
19. Kellis, M., Birren, B. W. & Lander, E. S. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature* **428,** 617–624 (2004).
20. Kellis, M., Patterson, N., Endrizzi, M., Birren, B. & Lander, E. S. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423,** 241–254 (2003).
21. Wood, V. *et al.* The genome sequence of *Schizosaccharomyces pombe*. *Nature* **415,** 871–880 (2002).
22. Arvestad, L., Berglund, A. C., Lagergren, J. & Sennblad, B. Bayesian gene/species tree reconciliation and orthology analysis using MCMC. *Bioinformatics* **19** (Suppl. 1), i7–i15 (2003).
23. Chen, K., Durand, D. & Farach-Colton, M. NOTUNG: a program for dating gene duplications and optimizing gene family trees. *J. Comput. Biol.* **7,** 429–447 (2000).
24. Dufayard, J. F. *et al.* Tree pattern matching in phylogenetic trees: automatic search for orthologs or paralogs in homologous gene sequence databases. *Bioinformatics* **21,** 2596–2603 (2005).
25. Durand, D., Halldorsson, B. V. & Vernot, B. A hybrid micro-macroevolutionary approach to gene tree reconstruction. *J. Comput. Biol.* **13,** 320–335 (2006).
26. Fitch, W. M. Distinguishing homologous from analogous proteins. *Syst. Zool.* **19,** 99–113 (1970).
27. Jothi, R., Zotenko, E., Tasneem, A. & Przytycka, T. M. COCO-CL: hierarchical clustering of homology relations based on evolutionary correlations. *Bioinformatics* **22,** 779–788 (2006).
28. Kellis, M., Patterson, N., Birren, B., Berger, B. & Lander, E. S. Methods in comparative genomics: genome correspondence, gene identification and regulatory motif discovery. *J. Comput. Biol.* **11,** 319–355 (2004).
29. Li, H. *et al.* TreeFam: a curated database of phylogenetic trees of animal gene families. *Nucleic Acids Res.* **34,** D572–D580 (2006).
30. Remm, M., Storm, C. E. & Sonnhammer, E. L. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol.* **314,** 1041–1052 (2001).
31. Tatusov, R. L. *et al.* The COG database: an updated version includes eukaryotes. *BMC Bioinform.* **4,** article no. 41 (2003).
32. Wapinski, I., Pfeffer, A., Friedman, N. & Regev, A. Automatic genome-wide reconstruction of phylogenetic gene trees. *Bioinformatics*. doi:10.1093/bioinformatics/bmt193 (2007).
33. Byrne, K. P. & Wolfe, K. H. The Yeast Gene Order Browser: combining curated homology and syntenic context reveals gene fate in polyploid species. *Genome Res.* **15,** 1456–1461 (2005).
34. Dudley, A. M., Janse, D. M., Tanay, A., Shamir, R. & Church, G. M. A global view of pleiotropy and phenotypically derived gene function in yeast. *Mol. Syst. Biol.* **1,** 2005.0001 (2005).
35. Tzung, K. W. *et al.* Genomic evidence for a complete sexual cycle in *Candida albicans*. *Proc. Natl Acad. Sci. USA* **98,** 3249–3253 (2001).
36. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. *Nature Genet.* **25,** 25–29 (2000).
37. Segal, E., Friedman, N., Kaminski, N., Regev, A. & Koller, D. From signatures to models: understanding cancer using microarrays. *Nature Genet.* **37** (Suppl.), S38–S45 (2005).
38. Gavin, A. C. *et al.* Proteome survey reveals modularity of the yeast cell machinery. *Nature* **440,** 631–636 (2006).
39. Deutschbauer, A. M. *et al.* Mechanisms of haploinsufficiency revealed by genome-wide profiling in yeast. *Genetics* **169,** 1915–1925 (2005).
40. Hughes, T. R. *et al.* Functional discovery via a compendium of expression profiles. *Cell* **102,** 109–126 (2000).

41. Newman, J. R. *et al.* Single-cell proteomic analysis of *S. cerevisiae* reveals the architecture of biological noise. *Nature* **441,** 840–846 (2006).
42. Huisinga, K. L. & Pugh, B. F. A genome-wide housekeeping role for TFIID and a highly regulated stress-related role for SAGA in *Saccharomyces cerevisiae*. *Mol. Cell* **13,** 573–585 (2004).
43. Tirosh, I., Weinberger, A., Carmi, M. & Barkai, N. A genetic signature of interspecies variations in gene expression. *Nature Genet.* **38,** 830–834 (2006).
44. Sopko, R. *et al.* Mapping pathways and phenotypes by systematic gene overexpression. *Mol. Cell* **21,** 319–330 (2006).
45. Davis, J. C. & Petrov, D. A. Do disparate mechanisms of duplication add similar genes to the genome? *Trends Genet.* **21,** 548–551 (2005).
46. Reguly, T. *et al.* Comprehensive curation and analysis of global interaction networks in *Saccharomyces cerevisiae*. *J. Biol.* 5, article no. 11 (2006).
47. Kafri, R., Levy, M. & Pilpel, Y. The regulatory utilization of genetic redundancy through responsive backup circuits. *Proc. Natl Acad. Sci. USA* **103,** 11653–11658 (2006).
48. Harbison, C. T. *et al.* Transcriptional regulatory code of a eukaryotic genome. *Nature* **431,** 99–104 (2004).
49. Gerber, A. P., Herschlag, D. & Brown, P. O. Extensive association of functionally and cytotopically related mRNAs with Puf family RNA-binding proteins in yeast. *PLoS Biol.* 2, article no. E79 (2004).
50. Force, A. *et al.* The origin of subfunctions and modular gene regulation. *Genetics* **170,** 433–446 (2005).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements A.R. was supported by a Career Award at the Scientific Interface from the Burroughs Wellcome Fund and by NIGMS. N.F. was supported by the Israel Science Foundation. We thank E. S. Lander for discussions and D. Peer, A. Tanay and O. Rando for their comments on previous drafts of this manuscript. We are also grateful to the members of the FAS Center and the Broad Institute for their scientific and technical support, especially A. Daneau, M. Ethier and B. Mantenuto.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to A.R. (aregev@broad.mit.edu).

METHODS

Pre-processing for orthogroup reconstruction. An exhaustive sequence similarity search between all protein sequences was performed using the FASTA sequence alignment tool[51]. FASTA identifies pairs of sequence segments with significant similarity and subsequently performs a full alignment between then, thus returning a single hit per pair of similar genes. All hits whose expectation value *E* was below 0.1 were subsequently treated as potential homologues. This lenient threshold allows many homology relations to be considered. We then computed the amino acid similarity between each pair of potential homologues using the substitution model of ref. 52.

The SYNERGY procedure. SYNERGY assigns orthologies by tracing all genes in the species below a given node in the species tree to their reconstructed ancestral genes on the basis of sequence similarity and shared gene order[32]. Briefly, beginning at the leaves of the tree and traversing backwards chronologically, sets of orthogroups for two daughter species that share significant sequence similarity are assembled into candidate orthogroups at the stage pertaining to each internal node. (The orthogroups for the daughter species were identified at the previous stages; for an extant species each gene is a singleton orthogroup.) The candidate orthologues are then positioned on a phylogenetic gene tree using a modified neighbour-joining procedure that reconstructs the phylogeny of orthogroups at that stage using the amino acid similarities that have been pre-computed and shared synteny scores[53]. Synteny scores are computed by counting the fraction of neighbouring genes that are assigned to the same orthogroup according to the provisional orthology assignments. The sub-trees joined at each stage are based on the reconstruction at earlier stages. Each tree is rooted using a combined score based on sequence similarity, conserved synteny, and the inferred number of duplications and losses. Trees that invoke fewer unlikely duplication and loss events will be favoured over those that incur many such events. If the rooted tree determines that a candidate orthogroup's genes are not all descended from a single ancestral gene, the orthogroup is partitioned to two candidate orthogroups by removing the inferred root of the gene tree. This process may be repeated on the two new trees until each tree represents genes that share a single common ancestor at the current level of reconstruction. The orthology assignments and trees obtained at each stage of the algorithm are used as the input to the next stages, involving higher nodes in the species tree. The interim trees are also used to compute the distances between orthogroups by using the neighbour-joining distance update rule when nodes on the tree are joined. When this procedure is completed at the root of the species tree, the genes for all species have been assigned to their orthogroups and placed in their respective trees. Full details of the procedure are available in the companion technical manuscript[32].

Bootstrap evaluation. To assess the sensitivity of SYNERGY's orthology assignments to both the choice species and their protein-coding gene predictions, two bootstrap-based confidence measures were calculated. A 'branch-bootstrap' was performed by systematically removing branches from the species tree and counting the number of orthology assignments that changed. A 'gene-bootstrap' was similarly executed by hiding each predicted open reading frame in the data set with a fixed of probability of 0.2. For each bootstrap run, a predicted orthogroup was validated for the number of orthology assignments that were not designated during that run ('soundness'), and the proportion of the original orthologous pairs among the new predictions ('completeness'). For each pair of orthologues within an orthogroup from a bootstrap run, an assignment was considered to be inconsistent either if the pair was assigned as orthologues at a point in the species tree that deviated from the original assignment or if they were not originally assigned as orthologues at all. The average number of consistent assignments across each experiment was used as the branch- and gene-bootstrap confidence measures for an orthogroup's soundness and completeness. A detailed description of this procedure is given in Supplementary Note 1.

Genome sequences. The predicted protein sequences and their genomic locations were downloaded from the following sources. From the Saccharomyces Genome Database[12]: *S. cerevisiae* (downloaded July 2005), *S. paradoxus* (accession number AABY01000000), *S. mikatae* (accession number AABZ01000000), *S. bayanus* (accession number AACA01000000) and *S. castellii* (accession number AACF00000000). From Génolevures[15]: *C. glabrata* (accession number CR380947–CR380959), *K. lactis* (accession number CR382121–CR382126), *D. hansenii* (accession number CR382133–CR382139) and *Y. lipolytica* (accession number CR382127–CR382132). From the *Ashbya* Genome Database[14]: *A. gossypii* (accession number AE016814-AE01682). From Kellis *et al.*[19]: *K. waltii* (accession number AADM01000000). From the *Candida* Genome Database[13]: *C. albicans* (accession number AACQ00000000). From The Broad Institute Fungal Genome Initiative (http://www.broad.mit.edu/annotation/fgi/): *Aspergillus nidulans* (accession number AACD00000000), *Fusarium graminearum* (accession number AACM00000000), *Magnaporthe grisea* (accession number AACU00000000), *Neurospora crassa* (accession number AABX00000000) and *Cryptococcus neoformans* (accession number AACO01000000). From PombeDB[21]: *S. pombe* (accession numbers CU329670–CU329672).

Species tree. The species tree topology representing the phylogenetic relations between the Ascomycota fungi was compiled from several sources[33,54,55] and subsequently validated using the results from our orthogroup assembly by concatenating the multiple sequence alignments from 50 sampled uniform orthogroups and obtaining the maximum-likelihood tree topology using the Phylip package's default parameters[56]. This sampling procedure was repeated ten times, with the same tree topology resulting each time. A manual correction to this topology was introduced as previously described[33], swapping the locations of *S. castellii* and *C. glabrata* owing to chromosome-based evidence suggesting that *C. glabrata* is in fact more closely related to the *Saccharomyces* sensu stricto clade. Branch lengths for the tree presented in Fig. 1a were estimated using a maximum-likelihood approach using multiple site rate variation with the default parameters[57]. Branch lengths were not used in any of the subsequent analyses.

Functional gene classes. We compiled a total of 3,395 gene classes, obtained as follows: 1,794 from the Gene Ontology[36] hierarchy, 87 from the Kyoto Encyclopedia of Genes and Genomes[58] (KEGG), 107 from the BioCyc database[59], 1,022 from the MIPS database of manually curated protein complexes[60], 310 from a data set describing the targets genes bound by various transcription factors[48], 70 from a data set describing the target genes harbouring a given *cis*-regulatory element in their promoter[48], and 5 from a data set describing the targets of the RNA binding proteins from the PUF family[49]. These classes were used for constructing the transcription module hierarchy (see below). In addition, the following gene classes were included in various analyses, but were not included in the construction of the module hierarchy: genes controlled by the SAGA and/or TFIID transcription complexes[42], genes with and without TATA box control[43], genes with large levels of expression variation between yeast species[43], genes with high and low levels of noise in protein abundance[41], haploinsufficient genes[39], genes whose overexpression reduces fitness[44], and genes belonging to complex cores, attachments and modules based on high throughput assays[38].

Functional orthogroup classes. We used the aforementioned *S. cerevisiae*-based gene classes to define orthogroup classes by projecting the *S. cerevisiae* annotations onto the orthogroups containing *S. cerevisiae* genes.

***S. cerevisiae* protein interaction networks.** We constructed separate biochemical and genetic protein interaction networks using both manually curated and high-throughput data sources[46].

Defining orthogroup copy number variation profiles. To measure the changes of gene copy number through either duplication or loss events, we assigned a copy number variation profile to each orthogroup. The profile is defined by inspecting the extended phylogenetic profiles belonging to an orthogroup, and subtracting the number of losses observed at each index in the species tree from the number of duplications. We increment this copy number variation profile at the last common ancestor identified for the orthogroup, indicating its 'age'.

Coherence of copy number variation profiles. To assess the coherence in gene copy number variation across a class of orthogroups, we first calculated the class' centroid extended copy number variation profile (ECVP) by averaging the ECVPs from all the orthogroups belonging to a class. This centroid is then applied to estimate the degree of deviation between the orthogroups belonging to a class by summing the L1 distance from each of the class' orthogroups to it. We compare this deviation to that of 10,000 randomly assigned orthogroup classes, each containing the same number of ECVPs. The fraction of times the deviation is equal to or less than that of the orthogroup class is the measure (*P*-value) we use to evaluate the coherence of that class. Copy number variation occurs at each individual branch of the species tree, so we similarly define a coherence profile for an orthogroup class by evaluating the copy number variation coherence for each position along the species tree.

Copy number variation in protein interaction networks. To test the relation between proximity in protein interaction networks and similarity in copy number variation, we first computed the difference (using the L1 distance) between the ECVPs for each pair of proteins in the network, ignoring pairs that belong to the same orthogroup (hence sharing the same profile). Next we averaged these differences among all proteins within a given radius in the network. To determine whether these averages were significant, we repeated this computation by shuffling profile assignments to proteins in the network 1,000 times, obtaining the expected range of average differences between pairs of proteins in the network for each radius.

Statistical benchmark of orthogroup evolutionary history. To benchmark the volatility in gene copy number for each orthogroup, we used the estimated rates of duplications and losses along each branch of the species tree to calculate the log-probability of the observed number of such events in each orthogroup, assuming that they occur according to a standard Poisson distribution. This

statistic is used as a measure of volatility for each orthogroup. We compare this volatility metric to those of 10,000 hypothetical orthogroups with randomly generated duplications and losses (based on the empirical rates). In our analysis we label those orthogroups the volatility of which deviates more than three standard deviations from the mean of the random distribution as being significantly volatile.

Statistical enrichment tests. To identify the annotations in which each orthogroup class was significantly enriched, we projected the class' annotations as described above, calculated the fraction of orthogroups from that class that contained a given annotation and used the hypergeometric distribution to calculate a *P*-value for this fraction (compared with the null hypothesis of choosing the same number orthogroups at random). We corrected for multiple tests using the false discovery rate correction with a 0.01% false rate.

Construction of *S. cerevisiae* transcription module map. We constructed the hierarchy of transcriptional modules following the procedure presented by Segal *et al.*[37]. We applied this procedure to a yeast data set and followed it by manually selecting which modules to use in the hierarchy. We note that although it is not highlighted in their manuscript, this method creates modules in a hierarchical fashion. Full technical details on the construction of the map are at the end of this section.

Estimating functional divergence of paralogous genes in hierarchical annotation data. We estimated the functional divergence between a pair of paralogous genes by considering the most specific gene class in the annotation hierarchy that each gene was assigned to (for gene ontology we ignored all assignments that were attributed only to computational sequence analysis). We regarded a pair of genes as functionally diverged only if both genes are assigned to at least one annotation class and they are not both assigned to the most specific annotations of either of the two genes.

Computing degree of conserved interactions between paralogues. We used two statistics to compute the degree of conserved interactions between pairs of paralogous proteins. The first was simply the fraction of shared interactions between both proteins. For this we counted the number of interactions each protein takes part in (a_1 and a_2 for proteins 1 and 2, respectively) and the number of interactions they both share (s). The fraction of shared interactions is thus:

$$f = s/[\min(a_1, a_2)]$$

We also used the subfunctionalization index (I_{sf}) as previously described[11] to characterize how diverged a pair of paralogues' interactions are. This is calculated as:

$$I_{sf} = 1 - (s + |a_1 - a_2|)/t$$

where s is as above and t is equal to $(a_1 + a_2 - s)$. This statistic gives a reasonable estimate of the degree of subfunctionalization in the absence of neofunctionalization, because subfunctionalization would reduce the number of shared interactions. This measure considers the proportion of ancestral interactions that are no longer shared between the paralogues and the extent of subfunctionalization for these interactions.

Estimating significance of shared protein interaction neighbourhood. To estimate the significance of the shared protein interaction neighbourhood between pairs of paralogues, we first calculated the degree of conserved interactions and compared this to the degree of shared interactions between the two paralogues in a degree-preserving randomized network, obtained by swapping edges between random pairs of nodes 10^6 times. We repeated this procedure 10,000 times, and assigned a *P*-value to the shared protein interaction neighbourhood of a pair of proteins according to the number of times the fraction of shared interacting partners between paralogues is equal to or greater than the fraction in the real network.

Homology searches. To identify putative homologies between orthogroups and the *C. neoformans* genome, we first constructed a tree-assisted multiple sequence alignment from each orthogroup's protein sequences using their reconstructed gene trees and the MUSCLE alignment software[61]. We then constructed a sequence profile from this alignment, and executed a homologous protein search using the HMMSEARCH profile search software, employing an *E*-value cutoff of 1.0 (ref. 62).

DNA microarray data set. We compiled a collection of 1,216 previously published microarray experiments (Supplementary Table 6). We normalized the expression of each gene g in every data set separately as in ref. 37. For data sets generated using Affymetrix chips, we first take the log (base 2) of g's expression value in each array (truncating expression values that are below ten). For data sets generated using spotted complementary DNA chips, we use the log-ratio (base 2) between the measured sample and the control sample. In both types of data sets, we subsequently normalize the (log) expression value of gene g in each array to its average expression in all the arrays in the same data set by subtracting its average in that data set from each of its expression measurements. After this normalization, the mean value of a gene within each data set is zero.

Details of transcriptional module hierarchy construction. The transcription module hierarchy was constructed in a step-wise process as in ref. 37.

Step 1: Identifying arrays where gene classes significantly change in expression. To identify the arrays where each gene class is significantly induced (or repressed), we defined the induced (or repressed) genes in each array to be those genes whose change in expression is greater (less) than twofold. For each gene class and each array, we calculated the fraction of genes from that class that are induced (or repressed) in that array, and used the hypergeometric distribution to calculate a *P*-value for this fraction (compared to the null hypothesis of choosing the same number genes at random). We corrected for multiple tests using the false discovery rate correction with a 1% false rate.

Step 2: Identification of gene class clusters. We performed (bottom-up) hierarchical clustering of the gene classes in the matrix of all significant array–gene class pairs. We manually selected a hierarchy of gene class clusters corresponding to the cluster boundaries defined automatically, and assigned a biologically meaningful name to each cluster. We obtained a total of 163 such gene class clusters (excluding the root node). The transcriptional modules were defined from the genes in those gene class clusters (below), and organized according to the same hierarchy.

Step 3: Testing consistency of a gene with expression of a gene class. Given a class of genes G and a gene g, we test whether the expression of g is consistent with the significant changes in the expression of G using the following procedure. We first identify the subsets of arrays I and R where G is significantly induced and repressed, respectively. We then measure the extent in which the expression of g changes by more (or less) than twofold in arrays in I (or R) with the score

$$\text{Score}(g) = \sum_{\{a \in I \mid g \text{ is induced in } a\}} -\log(p_a) + \sum_{\{a \in R \mid g \text{ is repressed in } a\}} -\log(p_a)$$

where p_a is the fraction of genes in the array a that are induced (or repressed) by more than twofold for arrays in I (or in R). This score assigns more weight to induction in arrays where there are fewer induced genes (and respectively for repression).

We evaluate the significance of Score(g) with respect to the null hypothesis in which the genes in each array are randomly permuted. Under this null hypothesis, Score(g) is the sum of independent binary random variables, one for each array in I and R. The random variable corresponding to array a attains the value $-\log(p_a)$ with probability p_a and the value of 0 with probability $1 - p_a$. Because Score(g) in this model is a sum of independent random variables, its mean μ and variance σ^2 are the sum of the means and variances, respectively, of the these variables, and can be computed analytically:

$$\mu = \sum_{a \in I \cup R} -p_a \log p_a$$

$$\sigma^2 = \sum_{a \in I \cup R} p_a(1-p_a)\log^2 p_a$$

By the central limit theorem, the distribution of Score(g) under the null hypothesis can be closely approximated by a gaussian distribution with mean μ and variance σ^2. We use standard methods for computing the tail probability of a gaussian distribution to compute the probability of attaining a score as large as the observed score under the null hypothesis.

Step 4: Deriving modules from clusters of gene classes. For each cluster of gene classes, we define G to be the union of the gene classes in the cluster. We then test each gene in G for consistency (as described above). The resulting module consists of genes the expression of which is significantly consistent with the expression of the gene class (after false discovery rate correction for multiple hypotheses using a 0.01% false rate).

51. Pearson, W. R. & Lipman, D. J. Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA* **85**, 2444–2448 (1988).
52. Jones, D. T., Taylor, W. R. & Thornton, J. M. The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci.* **8**, 275–282 (1992).
53. Saitou, N. & Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406–425 (1987).
54. Kurtzman, C. P. & Robnett, C. J. Phylogenetic relationships among yeasts of the 'Saccharomyces complex' determined from multigene sequence analyses. *FEMS Yeast Res.* **3**, 417–432 (2003).
55. Kuramae, E. E., Robert, V., Snel, B. & Boekhout, T. Conflicting phylogenetic position of *Schizosaccharomyces pombe*. *Genomics* **88**, 387–393 (2006).
56. Felsenstein, J. PHYLIP—Phylogeny Inference Package (Version 3.2). *Cladistics* **5**, 164–166 (1989).

57. Ninio, M., Privman, E., Pupko, T. & Friedman, N. Phylogeny reconstruction: increasing the accuracy of pairwise distance estimation using Bayesian inference of evolutionary rates. *Bioinformatics* **23,** e136–e141 (2007).
58. Kanehisa, M. A database for post-genome analysis. *Trends Genet.* **13,** 375–376 (1997).
59. Karp, P. D. *et al.* Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. *Nucleic Acids Res.* **33,** 6083–6089 (2005).
60. Mewes, H. W. *et al.* MIPS: analysis and annotation of proteins from whole genomes in 2005. *Nucleic Acids Res.* **34,** D169–D172 (2006).
61. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32,** 1792–1797 (2004).
62. Eddy, S. HMMER: Hidden Markov models for sequence profile analysis. ⟨http://hmmer.janelia.org/⟩ (2003).