Supporting Online Material

Table of Supplementary Figures

Table of Supplementary Tables

Evolution of transposons

Tf1 and Tf2 are long terminal repeat (LTR)-retrotransposons found in *Schizosaccharomyces pombe* (*1*). They belong to the family of elements that includes the Gypsy transposon of Drosophila melanogaster and the Ty3 transposon of *Saccharomyces cerevisiae*. Gypsy family retrotransposons have structures and sequences that are closely related to retroviruses, such as HIV and MLV (*2*).

The lab strain 972 of *S. pombe* encodes 11 full length copies of Tf2 and two tandem elements that are separated by a single LTR (*3*). However, 972 contains no copies of Tf1, the highly active cousin of Tf2 that was isolated from NCYC132, a wild strain of *S. pombe*. In addition, Strain 972 contains 174 copies of single LTRs derived from Tf1, Tf2, and other related elements (*3*). No other transposons have been detected in 972, although it contains the transposon-like *wtf* family of repeats.

There are 15 full length LTR-retrotransposons in *S. japonicus* and one in *S. cryophilus* (see Materials and Methods). We found no transposons in *S. octosporus*. All the transposon sequences we have identified belong to the gypsy family of retrotransposons; no sequences similar to the *S. pombe wtf* elements were found in other fission yeast. The 15 full length retrotransposons in *S. japonicus* belong to 10 distinct families we designated Tj1 through Tj10 (Table S6). The only transposon found in *S. cryophilus*, designated Tcry1, has only one full length copy (Table S6). Some characteristics of the Tjs and Tcry1 are shown in Table S6. Identical LTRs and target site duplications are indications of relatively recent integration. A stop codon or frame shift between *gag* and *pol* reflect common strategies of retrotransposons that translate higher amounts of capsid relative to polymerase proteins. For Tj1 and Tj4, there are stop codons between *gag* and *pol.*

Unlike the case in *S. pombe*, the LTR-retrotransposons in *S. japonicus* have extensive diversity (Fig. S5). We conducted a phylogenetic analysis based on the most conserved sequences of RT. The retrotransposons found in the species of fission yeast fall into two major lineages. Tj1, Tj4, Tj6 and Tcry1 are in the same clade as Tf1 and Tf2. Tj2, Tj3, Tj5, Tj7, Tj8, Tj9 and Tj10 are in the same clade as Ty3, a retrotransposon of budding yeast. It is known that Ty3, gypsy, Tom and Yoyo retrotransposons use host tRNAs to prime their reverse transcription, while, Tf1, Tf2, Maggy, Skippy and Cft1 perform a self primed reverse transcription in which a sequence in the 5'-LTR functions as the primer by annealing to the primer binding sequence (PBS). We analyzed the putative PBS of the newly identified retrotransposons and found TGG in the PBSs of the retrotransposons in the Ty3 clade (Table S27). The TGG sequence is indicative of tRNA priming because it anneals to the highly conserved CCA at the 3' end of all tRNAs. For the members in the Tf1 clade, we found sequences in the LTRs of Tj1, Tj4 and Tj6 that complement their PBSs suggesting that these elements are self-primed. Tcry1 is in the clade of self primed reverse transcription, but we did not find a complement sequence to its PBS in the LTR. This situation could be due to the accumulation of mutations.

The integration of Ty3 is noted for being within two nucleotides of a polIII transcript, usually a tRNA. We searched the positions of LTRs for all the Tj transposons and found they were not tightly associated with tRNA genes suggesting that these elements used a different targeting mechanism than that of Ty3.

Usually, transposons occupy a small fraction of the genomes in unicellular organisms. This situation could be because fast duplication rates generate evolutionary pressure for small sized genomes. In the strain 972 of *S. pombe,* there are no full length copies of Tf1, and only solo copies of its LTRs (*3*). This observation indicates that yeast species can lose transposons. Hence, the scarcity of transposons in *S. cryophilus* and *S. octosporus* is not so surprising. We did find there are many repeat sequences in these two genomes, which could be the remnants of transposons. But because they are not related to known transposon sequences, it is hard to determine their origin.

No DNA transposons were found in any of the four *Schizosaccharomyces* species. This is consistent with the lack of DNA transposons in *S. cerevisiae*. There are however retrotransposons of the copia family, such as Ty1, Ty2, Ty4 and Ty5, in budding yeast (*4*). There are no copia retrotransposons found in the *Schizosaccharomyces* species.

mat **loci structure**

All of the known *cis*-acting regulatory sequences required for epigenetic imprinting and recombinational switching are conserved in *S. octosporus* and *S. cryophilus* (Fig. S10) (*5-7*). The *mat1* locus is not present in the *S. japonicus* genome assembly, although it is detectable by physical mapping (Fig. S11). In addition, it genetically maps to centromere-proximal location of *mat2* and *mat3*. We speculate that *mat1* may be in the unassembled peri-centrosomal repeats, but further physical mapping will be required to test that hypothesis. The lack of *mat1* makes it impossible to say if the epigenetic imprinting and recombinational switching motifs are conserved in *S. japonicus*. However, the donor-loci *mat2-P* and *mat3-M* are flanked by homologous sequences required for recombinational switching. Furthermore, the epigenetically-programmed genomic mark associated with mat loci (*8*) is detectable in *S. japonicus* and *S. cryophilus* (*9*) (Fig. S11).

Phylogenetic profiles of the proteomes

Using the topology of the phylogenetic tree in Fig. 1A (see Materials and Methods), a phylogenetic profile of the proteomes from each species was determined using the MCL protein clusters at inflation parameter 1.5. Comparing the distribution of phylogenetic profiles in the pie diagram of each species it became clear that a greater percentage of the *Taphrinamycotina* proteomes had an *Ascomycota* profile relative to the other two subphyla (5-6% versus 1-3%, Fig. S22).

The *Saccharomycotina* species had 191 clusters in the *Ascomycota* node and the *Taphrinamycotina* species had 248 clusters designated with an *Ascomycota* node profile. Interestingly, within this profile the *Taphrinamycotina* species shared more protein clusters with the *Pezizomycotina* species (204) than did the *Saccharomycotina* species (147). For any of the other deeper phylogenetic node profiles (Ophistokont, Fungi and Dikarya) the *Taphrinamycotina* always shared more clusters with the *Pezizomycotina* as one would expect from a sister subphylum.

The most common pattern (absence/presence of species in a cluster) encountered under the *Ascomycota* profile was clusters that included all *Ascomycota* (containing 45 clusters with 295 proteins from the *Taphrinamycotina*). However, the second most common pattern excluded all species in the *Saccharomycotina* but included all species in the *Taphrina*- and *Pezizomycotina* (containing 31 clusters of which 161 proteins from the *Taphrinamycotina*). The third most common pattern excluded all *Saccharomycotina* except the earliest diverging yeast of this subphylum, *Yarrowia lipolytica* and included all species in the *Taphrina*- and *Pezizomycotina* (containing 18 clusters of which 103 proteins from the *Taphrinamycotina*).

The frequency of these *Ascomycota* node patterns suggests that the *Ascomycota*-specific proteins in the *Schizosaccharomyces* species had more similarity to *Pezizomycotina* proteins even though these species are more distantly related. Since the 2nd and 3rd most common pattern excluded all the *Saccharomycotina* (except in the 3rd most common pattern which included *Y. lipolytica*), it is likely that selection pressure and adaptation may have resulted in a loss or considerable change of the proteins in the fast evolving *Saccharomycotina*. It looks like these shared characters between the *Pezizomycotina* and *Taphrinamycotina* have a common ancestral origin. Convergent evolution seems less likely. However, further characterization and evolutionary studies of these protein clusters will reveal more detail.

Eukaryotic protein kinases in the *Schizosaccharomyces*

Eukaryotic protein kinases (ePKs) comprise the largest protein superfamily in eukaryotes, and regulate most aspects of cell biology. We therefore identified and classified the sets of ePKs of the four *Schizosaccharomyces* species (Table S28), and compared them to each other, and to the well characterized set of ePKs from *S. cerevisiae* (*10*). Like *S. cerevisiae*, which contains 118 ePKs, the *Schizosaccharomyces* have compact sets. Their total number of ePKs range from 106 - 111 (Fig. S23), far fewer than are encoded in the genomes of simple, non-parasitic eukaryotes such as *D. discoideum* and *T. thermophila*, which possess 255 and 1029 ePKs, respectively (www.kinase.com). Also, like *S. cerevisiae*, the *Schizosaccharomyces* lack the entire tyrosine kinase-like (TKL) group, which occurs in the *Basidiomycetes*, and widely throughout the eukaryotes.

S. japonicus, which diverged from the other *Schizosaccharomyces* earliest, has the largest number of ePKs (Fig. S23). Kinases found in by *S. japonicus* but not *S. pombe*, *S. octosporus* and *S. cryophilus* include additional members of the CAMKK, CAMK1, SRPK, STE11 and PAKA families, and notably, a member of the CAMK/CAMKL/MARK family, which occurs in eukaryotes as diverse as *S. cerevisiae*, *T. thermophila*, and *H. sapiens*. In *S. pombe* and *S. cerevisiae* the core function of MARK kinases in the regulation of microtubule interactions has been transferred to the kinases of the fungal-specific Kin1 subfamily (Par1 in *S. pombe*, Kin1 and Kin2 in *S. cerevisiae*) (*11*), so it is not surprising that the distribution of MARK kinases in fungi is inconsistent.

While the number of ePKs possessed by the *Schizosaccharomyces* is smaller than that of *S. cerevisiae* (Fig. S23), the *Schizosaccharomyces* have a greater family and subfamily diversity (Table S29). The YANK, PITSLRE, DYRK2, PRP4, MEK1, and HisK-MAK families and subfamilies are represented in the *Schizosaccharomyces* but not *Saccharomyces*. All of these families and subfamilies except HisK-MAK are also present in *H. sapiens*, underscoring the utility of *S. pombe* as a model organism. The HisK-MAK family kinases Mak2 and Mak3 are involved in the oxidative stress response (*12*) and contain both ePK domains and domains for histidine kinase activity. This architecture is conserved outside the metazoa, but the budding yeast histidine kinase, Sln1, lacks the ePK domain, as do the *Schizosaccharomyces* Mak1 histidine kinases. Although *S. cerevisiae* has fewer ePK families then do the *Schizosaccharomyces*, it has more ePKs due to a family of unknown function, IKS, that is widely-conserved outside the metazoa but is missing from the *Schizosaccharomyces*, and small expansions in its GSK, ERK and HAL families. The most striking results from comparison of the sets of ePKs from the *Schizosaccharomyces* and *S. cerevisiae* is their small size in comparison with sets from other simple, non-parasitic eukaryotes (www.kinase.com), and their similarity to each other in terms of size and content (Fig. S23).

Additional protein kinases generated by the whole-genome duplication event affecting the *Saccharomyces* have been effectively removed, resulting in a sets that differ by a only few percent. This observation suggests that within the lineage leading to *Schizosaccharomyces* and *Saccharomyces* there is either selective pressure toward minimal regulatory complexity, or little pressure toward additional complexity.

The availability of RNA-Seq transcription data provides additional insights into the function and regulation of *Schizosaccharomyces* protein kinases. The function of YANK kinase is currently unknown. Expression of this kinase increases in all four *Schizosaccharomyces* species under glucose depletion and early stationary conditions (average increases of 4.2-fold and 2.8-fold, respectively), suggesting a role for it in the response to nutrient limitation.

The loci of twenty nine, or 8.2%, of *Schizosaccharomyces* ePKs have more antisense than sense transcription under one or more of the experimental conditions tested (Tables S21, S28); the rate of conservation in two or more species of excess antisense transcription is 83%. The corresponding rates for all proteins are 5.2% and 51%, respectively. The significantly larger rates of excess antisense, and antisense conservation for the ePKs supports the hypothesis that antisense regulation plays an important role in *Schizosaccharomyces* biology. Eighteen of the ePK genes with excess antisense transcription have been annotated as having roles in meiosis (http://old.genedb.org/genedb/pombe/), and include *mde3*, *mug27*, *mek1*, *ppk31* and *spo4*, which is highlighted in the main text.

The set of ePKs with excess antisense transcription from the *Schizosaccharomyces* was compared with a related set from budding yeast (*13*) in order to detect antisense transcription conserved over large phylogenetic distances. While conservation of antisense transcription was not observed for direct orthologs, there is an instance in which the same signaling network appears to be regulated by antisense. The locus for the replication stress mitotic checkpoint kinase Mik1 has excess antisense transcription in early stationary phases of *S. octosporus* and *S. pombe* (and also the glucose depletion and heat shock phases of *S. octosporus*), while antisense transcription is reported for the cell-size mitotic checkpoint kinase Swe1 in the log phase of *S. cerevisiae*. Mik1 and Swe1 are both members of the WEE family and phosphorylate the same substrate, Cdc2 (*14*). In addition, the locus for Chk1, a kinase involved in the mitotic checkpoint (*15*), has excess antisense transcription in the glucose depletion and early stationary phases of *S. japonicus*.

S. japonicus exhibits species-specific antisense transcription for two protein kinases in addition to Chk1. The loci for Ppk24, a HAL-family kinase, and SJAG_05663, a species-specific SRPK paralog, display 6-fold and 17-fold excess antisense transcription for combined conditions, respectively. *S.*

japonicus displays species-specific absence of excess antisense transcription at the *mek1* and *mug27* loci under the conditions tested.

Gene gain, loss and duplication

Of 5973 orthologous groups in the fission yeast clade, 4218 are 1:1:1:1 orthologs (Tables 3, S12). 1484 (85%) of the changes involve the gain of species- or clade-specific genes. These genes fall into three general categories: transcriptional regulation, cell wall/cell membrane function and meiosis (Table S14). The rapid evolution of transcriptional regulation and cell surface proteins likely has allowed fission yeast to adapt to changing environmental conditions, while that of meiotic genes may be involved in speciation events within the clade. Because gene gain is defined by sequence similarity, much of the observed gene gain is likely to be due to rapidly evolving gene families. Indeed, genes involved in both transcriptional regulation and meiosis are significantly enriched in the fastest evolving gene families (Tables S30, S31). It is also the case that these classes of genes are more easily recognized across evolutionary distance – transcription factors by their modular nature, membrane proteins by their topology and meiotic genes by their expression patterns, allowing them to be functionally annotated more extensively among the novel genes. In addition to the evolution of new genes, there are also several gene families, including those involved in chromatin structure, RNAi and the Cop9 signalosome, that appear to be novel in fungi when compared to others in a portfolio of 76 genomes from across the fungal kingdom (Table S32), but are conserved in other eukaryotes (Table S14). These families presumably have been lost from, or rapidly diverged in, the rest of the fungi.

In addition to the gain of new orthologous gene groups, there are 101 instances of expansions of orthologous group by gene duplications. These duplications are significantly enriched for genes with functions in membrane and cell-wall biosynthesis and membrane transport (Tables S12, S14) and they tend to cluster in subtelomeric regions, as is common in other eukaryotes.

Gene loss is less common than gene gain. There are 407 examples of lineage-specific gene loss in the clade (Table 1). As for the duplicated genes, the lost genes are significantly enriched for those with functions membrane and cell-wall biosynthesis and membrane transport (Table S14), consistent with similar sporadic losses in other *Ascomycota* (*16*).

Of 3788 orthologous groups of genes conserved in at least 15 of 20 other sequenced *Ascomycetes* (Table S32), only 745 (20%) are missing from all fission yeast (Table S33). This group of genes lost from fission yeast contains many that encode transcription factors, membrane proteins and meiotic proteins, as well as proteins involved in budding. However, the only significantly enriched categories of lost genes are involved in sub-pathway of fatty acid biosynthesis and glycogen metabolism (Table S34; see below). Therefore, with these exceptions, fission yeast appear to maintain a core complement of fungal genes (Fig. S22).

Gene content changes associated with glucose metabolism

The convergent evolution of aerobic fermentation in the fission yeast and the post-WGD budding yeast allows a comparison of different approach the two clades have taken to glucose metabolism. The following six examples of evolutionary changes to central genes in carbon metabolism shed light on the distinct metabolic capacities of fission yeast. i) Fission yeast lack the glyoxylate cycle, preventing the use of ethanol as a sole carbon source (*17, 18*). ii) Although budding yeast use both glycogen and trehalose as internal carbohydrate stores, fission yeast have lost the glycogen biosynthetic pathway and have duplicated the *tps2* gene for trehalose-phosphate synthase. iii) Fission yeast have fewer pairs of paralogs within glycolysis genes than budding yeast: although post-WGD budding yeast species have paralogous gene pairs for enzymes catalyzing seven of ten glycolysis reactions, fission yeast have paralogous gene pairs for enzymes catalyzing only three reactions (glyceraldehyde-3-phosphate dehydrogenase, enolase, and hexokinase), and most notably lack a paralog for the *pfk1* gene (which is duplicated in both pre- and post-WGD *Saccharomycotina* species). iv) Fission yeast have a reduced capacity to generate glucose from other carbon sources. In particular, they lack phosphoenolpyruvate carboxykinase *(pck1)*, which is required for gluconeogenesis from ethanol. This loss further limits their ability to use ethanol as a primary carbon source. *S. japonicus* also lacks *fbp1*, which is required for gluconeogenesis from non-sugar carbon sources, including glycerol. v) Fission yeast have not expanded the ADH genes, as has happened in the post-WGD budding yeast. In contrast to the four *ADH1* paralogs in post-WGD species, there is a single *adh1* ortholog in *S. pombe, S. octosporus* and *S. cryophilus*, and two *adh1* paralogs in *S. japonicus*. The fission yeast ADH1 homologs are similar to *S. cerevisiae ADH1*, and to the ancestrally reconstructed ADH, both of which are kinetically tuned for ethanol production, rather than consumption (*19*). vi) Fission yeast lack some, but not all, of the critical regulatory proteins known to control glucose repression or activation of respiratory functions in the absence of glucose in *S. cerevisiae*. Most notably, they lack Cat8 and Sip4, known to activate gluconeogenesis genes in *S. cerevisiae*. All of these adaptations are consistent with the inability of fission yeast to consume ethanol as a sole carbon source.

Gene evolution rates

We used the relative degree of amino acid sequence divergence among 1:1:1:1 orthologs to calculate relative evolution rates with the fission yeast clade. We observed several classes of protein-coding genes that have elevated or reduced mutations rates. For example, proteins localized in the cellular membranes have a tendency to be fast-evolving ($P = 1.39X10^{-4}$, K-S test), including those at the cell tip ($P = 0.0043$) and those involved in septum formation ($P = 2.17x10^{-7}$). Several categories of gene sets involved with meiosis are also enriched with fast-evolving proteins ($P = 4.50x10^{-6}$), including proteins related to chromosome segregation during meiotic cell division ($P = 2.64x10^{-7}$), and proteins involved with meiotic horsetail chromosome movement ($P = 8.42x10^{-4}$). Other cell division and chromosome separation proteins are also faster evolving, including the kinetochore complex ($P = 2.11x10^{-4}$), the spindle pole body ($P = 0.0014$), double-strand break repair ($P = 0.0026$), and sister chromatid cohesion ($P = 0.0033$). Other fast-evolving functional categories include factors involved with the regulation of mRNA transcription ($P = 3.57x10^{-7}$), the Golgi apparatus ($P = 5.39X10^{-4}$), peroxisome biogenesis ($P = 9.87x10^{-5}$), and pheromone response ($P = 0.0082$).

In contrast to these fast-evolving gene sets, many other functional categories are evolve slower than expected, suggesting greater constraints against accumulating mutations. These include proteins encoding the cytosolic ribosome ($P = 7.58x10^{-4}$), proteins required for ribosome assembly and biogenesis ($P = 1.73x10^{-6}$), cell-cycle regulation ($P = 5.01x10^{-13}$), DNA replication genes in the MCM complex ($P = 2.41x10^{-5}$), and RNA polymerase (pol-II) complex genes (transcription initiation (P =

1.01×10^{-07}), DNA-directed polymerase activity (P = 8.46×10^{-08}), RNA polymerase II transcription factor (P = 7.88×10^{-05})). A host of biosynthetic and metabolic pathways are also more slow-evolving than the background: glycolysis, TCA cycle, gluconeogenesis, glutamate metabolism, leucine, isoleucine, methionine, arginine, lysine, ergosterol, threonine, purine, histidine, and others (Table S35).

The large number of sequenced genomes among the *Ascomycota* fungi allow us to compare the substitution rates within the *Schizosaccharomyces* clade to those of their orthologs within other clades of species. This comparative evolutionary rates analysis should permit us identify genes and classes of genes whose selective pressures may have changed across the evolution of this phylum. We identified many functional classes that are under differential selection when comparing substitution rates between the three *Schizosaccharomyces* species and four *Saccharomyces sensu stricto* species. Several classes of genes are evolving more rapidly within the fission yeasts than in the *Saccharomyces*, including those involved in membrane synthesis and transport, translation, glycolysis and respiration (Tables S30, S31). Of the gene classes that evolve relatively slowly in fission yeast, the most striking example is the pre-mRNA splicing machinery (Table S35) which may reflect the conservation of complex exon structures in fission yeast in contrast to the significant simplification of exon structures and the splicing machinery in budding yeast (*20*).

Global conservation of expression programs within fission yeasts

To examine the evolution of gene expression in the fission yeast clade, we compared the patterns of differential gene expression between the species by correlating the levels of gene expression in each condition. As expected, in *S. pombe*, expression levels in the stress conditions of glucose depletion, early stationary phase and heat shock are more similar to each other (r = 0.71 - 0.80) than any of them are to log phase (r = 0.25 - 0.63). Surprisingly, we found that expression levels in the other species (with the exception of *S. octosporus* early stationary phase) showed much less difference between conditions (r = 0.78 - 0.97) and clustered with the *S. pombe* log-phase sample (Figure S25). These results suggest that the other species show a less robust transcriptional response to environmental conditions, as tested under standard lab conditions.

Despite these differences, the global organization of expression is conserved across species, with dominant programs for ‘growth’ and ‘stress’. For a more detailed look at gene regulation, we analyzed expression patterns across the four conditions and between the four fission yeast by phylogenic clustering to identify conserved modules of gene expression (Figure 4). In general, we find that the pattern of gene expression between species grown in similar conditions is similar, with two dominant patterns of gene expression: one associated with growth (log and heat shock – heat-shock clusters with log-phase growth in this analysis because on the 15-minute timescale used here, a relatively small number of genes is affected) and another associated with stress (glucose depletion and early stationary phase). Moreover, similar expression clusters are enriched for similar gene annotations across the species. These results show that although the amplitude of expression changes is significantly higher in *S. pombe* than the other species, the overall patterns of expression are conserved.

Although patterns of transcriptional regulation during glucose depletion are similar across the fission yeast, the clade displays a number of differences when compared to other *Ascomycetes* (Table

S24). In response to glucose depletion, fission yeast specifically up regulate genes involved in mitosis, including those involved in the kinetocore, the spindle pole body and the anaphase-promoting complex. In contrast, fission yeast down regulate genes involved in other aspects of the cell cycle, including genes involved in cell-wall biosynthesis, S-phase DNA damage checkpoints and repair, transcription and splicing, all consistent with imminent withdrawal from the cell cycle. Further, they up regulate genes involved in meiosis, and in meiotic recombination in particular. None of these genes are significantly regulated in glucose depletion in *S. cerevisiae* (*21*). The apparently paradoxical up-regulation of mitotic genes and down-regulation of other cell-cycle genes reflects the fission yeast cell cycle, in which G2 is the longest phase of the cycle, so that most cells are in G2. This configuration contrasts with the budding yeast cell cycle, in which most cells are in G1. In response to starvation, fission yeast stop growing and withdraw from the cell cycle, but only after executing final mitotic divisions that reduces their size and allows them to arrest in G1 (*22*). The up regulation of meiotic genes in response to glucose depletion reflects the fission yeast haploid life style; fission yeast diploids are unstable and are induced to enter meiosis in response to even mild nutrient limitation.

Antisense transcription of meiotic genes

Previous work has concluded that meiotic genes in *S. pombe* are regulated by intron retention (*23, 24*). However, these studies did not used strand-specific techniques, making it impossible to distinguish an unspliced sense transcript from an antisense transcript. Furthermore, experiments to localize meiotic regulatory elements have used both 5' and 3' non-coding sequences (*24*) or used the 3' non-coding sequence from the *nmt1* gene (*23*), which also shows strong antisense transcription (Table S21). In log-phase cultures of *S. pombe*, we find few sense reads, but abundant antisense reads, for most of the genes reported to be regulated by intron retention (Table S21 and Fig. S20). In the cases where there is sufficient sense coverage to reconstruct a transcript, we see the predicted splicing events, not intron retention (Fig. S20).

Although meiotic genes are highly enriched among genes with more antisense than sense transcription in log-phase growth, they are not enriched among genes with greater than 5% antisense transcription but less than 100% antisense transcription (p = 0.47, hypergeometric test). This observation is consistent with a stoichiometric mechanism of regulation in which any leaky transcription from meiotic genes during log growth is bound by its antisense transcript and possibly destroyed by double-strand-specific RNases. In *S. pombe*, the double stranded RNA produced by bidirectional transcription is a substrate for RNAi machinery (*25*). Indeed, RNAi-dependent heterochromatin has been observed at several meiotic genes in *S. pombe* (*26*), although we see no evidence for systematic enrichment of siRNA, Dcr1, Ago1 binding or H3K9 methylation (from published datasets) at antisense-transcribed genes (*27-29*). Moreover, deletions of *dcr1* or *ago1* show no significant effect on sense or antisense transcript levels at several antisense-transcribed loci (Fig. S20). However, in the absence of Dcr1, double-strand RNA may be destroyed by Pac1, an essential RNase-III-type double-strand-specific RNase.

Materials and Methods

Strains

The strains used are listed in Table S36. Unless otherwise noted, all strains were grown in YES at 30˚C, except for *S. cryophilus*, which was grown at 25˚C. yFS275 was derived from ATCC10660 by isolation of a single spore from a completely viable octad. Likewise, yFS286 was derived from ATCC4206-U by isolation of a single spore from a completely viable octad.

Genome Sequencing and Assembly

A summary of all DNA sequencing data types used in genome assemblies is provided in Table S1. Sequence data were generated at the Broad Institute unless otherwise specified.

S. octosporus

The genome assembly incorporated a number of data types, including standard Sanger sequencing of several insert sizes in several vectors, and unpaired 454 fragment reads. In addition, because of the apparent cloning bias observed in initial assemblies of standard plasmid sequence data, a reduced AT bias plasmid library was generated in which care was taken to maximize representation of sequences with very low %GC composition. This method is based on the method for generation of WGS libraries used previously (*30*), with the exception that the resulting DNA was ligated into a low copy vector rather than linker ligated for PCR amplification. Briefly, modifications to the published protocol involved: i) replacing of the heat-inactivation steps after the end repair and linker ligation reactions with phenol:chloroforom extraction followed by ethanol precipitation; ii) Ligation into the pJAN low-copy plasmid vector (which has a pBR322 replication origin); iii) After ligation into the plasmid, the ligase is inactivated by incubation at 65˚C in the presence of 300 mM NaCl to stabilize AT-rich sequences; iv) After agarose gel size selection, DNA fragments were electroeluted from agarose slices in dialysis tubing and concentrated with using the Microcon YM-100 Centrifugal Filter Device (Millipore) according to manufacturer's specifications.

Assembly was carried out using HybridAssemble <http://www.broadinstitute.org/crd/wiki/index.php/HybridAssemble>, which is a module of the Arachne assembler (*31*) that takes input for multiple sequence data types. 454 data were first assembled with Newbler version MapAsmResearch-03/15/2010 <http://454.com/products-solutions/analysis-tools/gsde-novo-assembler.asp> (*32, 33*) before being input to HybridAssemble. Anchored sequence data from the physical map (see below) were used to order and orient 11.2 Mb of scaffolds into 3 chromosome-length ultra-scaffolds. All scaffolds were anchored to the physical map, with the single exception of a 39 kb scaffold that appeared to be comprised entirely of collapsed centromeric repeat sequences.

S. japonicus

The genome assembly incorporated a number of data types, including standard Sanger sequencing of several insert sizes in several vectors, unpaired 454 fragment reads and unpaired Illumina data. To create the BAC library, genomic DNA of *S. japonicus* (NIG2017) was prepared from cell embedded in agarose plugs. The DNA was partially digested with HindIII and separated by PFGE. Appropriately sized DNA was extracted from the gel, ligated with pCC1BAC vector (EPICENTER) and transformed into E. coli

DH10B. Average length of inserted DNA fragments was about 80 kb. The BAC paired-end Sanger data and Illumina read data were generated at the National Institute of Genetics, Japan. The 454 read data were generated at the GenePool, University of Edinburgh, UK.

Illumina data were assembled with ABySS (*34*), and the output was included with the Sanger and 454 data as input to HybridAssemble, as above. Anchored sequence data from the genetic map (see below) were used to order and orient 11.2 Mb of scaffolds into 6 chromosome ultra-scaffolds each representing a chromosome arm. The largest unanchored scaffold was 168 kb in length.

S. cryophilus

Paired and unpaired 454 data were generated by Roche/454 under contract from the Stowers Institute. Data were assembled with Newbler as above (*32, 33, 35*).

S. pombe

We used the Jan 2007 assembly of the *S. pombe* genome, downloaded from the Sanger Center <http://www.sanger.ac.uk/Projects/S_pombe/download.shtml> (*36, 37*). Illumina data for NCYC132 and SPK1820 (*38*) were generated at the Broad Institute and SNPs were identified with the manufactures software.

Identification and assembly of telomere repeats.

To identify telomere subunit motifs, we employed the tandem repeat detection tool mreps (*39*). mreps was run on all available sequence read data for each of the three genomes. We analyzed and assessed the mreps output, looking at tandem unit copy number, position in reads, motif complexity, frequency in the dataset, and the ratio of copy number to frequency. Using these metrics, we identified one or more putative subunit motif candidates for each genome, and aligned them to their putative *ter1* loci for confirmation.

For every chromosomal end, we assembled a pool of reads starting with the set of reads containing the identified motif and any mate pairs thereof. We identified assembled reads that had unplaced mates that that would likely place past the edge of the assembled chromosome based on the average insert size of the library, and added these reads to the pools. With chromosome ends as anchors, we used HybridAssemble (*31*) with each associated read set to perform localized assisted assemblies. We then aligned the remaining unplaced reads from the original assembly to the newly generated consensuses, identified reads that aligned, and added them to their corresponding pools. New localized assemblies were then performed using the boosted read pools and extended chromosome ends. The alignment and assembly steps were repeated until no further extension occurred.

S. japonicus appears to have a more complex telomere structure than seen in the other species. The template for the dominant telomere motif (GTCTTA) was found at the putative *ter1* locus. However, while analyzing small assembled telomeric contigs, we found another telomere-associated motif (GGGTTTA). Furthermore, this motif assembled in tandem with the dominant motif in a third putative (GTCTTAGGGTTTA). One possible explanation for two telomeric repeats would be two *ter* RNAs.

Physical mapping of *S. octosporus*

The orientations of the 10 largest supercontigs in assembly SO2 of the *S. octosporus* genome were determined by restriction mapping. Cells were grown to mid-log phase in YES and then washed in CSE (20 mM citrate/phosphate pH 5.6, 40 mM EDTA, 1.2 M Sorbitol), resuspended at 8x10^8/ml in CSE containing lyticase (Sigma) at 100U/ml and incubated for three hours at 37°C. An equal volume of 1.25% low gelling temperature agarose in CSE was then added and the mixture dispensed into moulds of 120 µl in volume. The mixture was allowed to set on ice and then individual plugs were extruded into NDS (0.5 M EDTA, 1% sodium lauroyl sarcosine, 0.05M Tris HCl, pH 9.5) containing pronase (Roche) at 1 mg/ml. The mixture was incubated for 24-48 hours at 55°C prior to restriction enzyme digestion. Individual plugs were rinsed three times in 100 volumes of 10mM Tris- HCl, 1mM EDTA, pH 8.0 (TE), treated for twenty minutes with TE containing 1mM phenyl methyl sulphonyl fluoride and rinsed with distilled water. DNA embedded in agarose was then digested for eight hours using the conditions recommended by the manufacturer prior to size fractionation using a rotating plate gel electrophoresis system (*40*). DNA was transferred to nylon membranes (Genescreen) using standard protocols. Membranes were cut in to 20 strips and individually hybridized overnight to a different ^{32}P-dATP-labeled probe. Strips were washed using standard protocols and exposed to phosphorimager plates (Fuji). Probes were designed to the last unique sequence at each end of 10 the super contigs. The probes we amplified by PCR using *S. octosporus* genomic DNA as template and the primers listed in Table S37. The resulting chromosome assemblies are independent of the *S. pombe* genome assembly.

Genetic mapping of *S. japonicus*

The 7 largest supercontigs in assembly SJ1 were genetically marked at their ends by integration of a *ura4:kanMX6* double marker flanked by 500 bp genomic sequences of target locus (Table S3) (*41*). For *ade6*, the *domE* allele was used (*41*); for *mat*, mating type was used. Genetic mapping was used to link the supercontig ends to each other as described (Fig. S3A) (*41*). Genetic distances were derived from the formula (cM = 100 x (TT+6NPD) / 2 (TT+NPD+PD)). Centromeric distances were calculated from the formula (cM = 100 x TT / 2(TT+NPD+PD)) The segregation of the markers *smc3*, *tel1*, *ade6*, *mat*, 1F, 3E and 6A confirms these loci are centromere linked (Fig. S3B, Table S2). The resulting chromosome assemblies are independent of the *S. pombe* genome assembly.

Comparative annotation

We made significant improvements to the gene structure and gene product names of the four fission yeast genomes through the use of RNA-Seq-base transcript models (see below) and the clustering of protein sequences, followed by extensive manual review and adjustments. The gene product names were also reviewed and updated based on blast hits to SwissProt and JCVI Hmmer equivalog hits.

Protein-coding genes

The PASA assemblies (transcript models) built from RNA-Seq data were used to updated the gene structures of protein-coding genes from previous annotation efforts. The resulting gene models were further improved through comparative annotation using all four *Schizosaccharomyces* genomes. Briefly, we first ran OrthoMCL (*42, 43*) to build protein clusters using PASA-updated genes from all four species and made protein sequence alignments for each orthologous group using ClustalW (*44*). We then analyzed the protein clusters and cluster alignments to identify discrepancies in gene models among the four genomes. These discrepancies include missing genes, missing or over-extended 5'- or 3'-ends, missing or incorrect internal exons. The flagged gene models were manually reviewed and fixed when appropriate.

We also added new protein coding genes from the PASA assemblies. New genes were filtered to not overlap with annotated CDS regions and to have one of the following 3 ORF characteristics: an ORF of at least 100 codons, ORF of at least 70 codons if it is the first and longest ORF of the transcript or an ORF of at least 30 codons that is conserved in at least two genomes. We found 89 new protein-coding genes in *S. pombe*, 53 of which are conserved (Table S7 and Fig. S13). These newly annotated protein-coding genes had been previously missed because they have short ORFs (the median ORF length for all 89 genes is 90 amino acids compared to 379 for the full gene set) and short exons (the average length of exons for the 37 multi-exonic genes is 115 nt compared to 1001 for the multi-exonic genes in the full gene set).

Gene symbols and gene product names for *S. pombe* were obtained from GeneDB (http://old.genedb.org/genedb/pombe/). These were assigned to genes in the other three genomes if a gene cluster contains one *S. pombe* gene and 1 or 0 gene from each of the other three genomes.

Protein kinase identification and classification

Fission yeast proteins were searched for eukaryotic protein kinase domains with an E-value of 100 using a hidden Markov model (http://hmmer.janelia.org/) built from an alignment of the full complement of ePKs from *D .discoideum* (*45*). Divergent protein kinases were distinguished from false positives by identifying protein kinase subdomain motifs (*46*) in sequence alignments generated by the hmmer hmmsearch algorithm. Protein kinases were classified using a BLAST search (*47*) against the set of curated protein kinases maintained at www.kinase.com; in order to classify a sequence, its three best BLAST from www.kinase.com were required to agree in order to minimize over-classification errors. For *S. japonicus*, *S. octosporus* and *S. cryophilus* protein kinases were named using the controlled vocabulary maintained at www.kinase.com, and gene symbols from GeneDB (http://old.genedb.org/genedb/pombe/) were mapped to the set of protein kinases when possible. Inconsistencies (holes) in classification were filled or confirmed using orthology information, and comparison with the core set of ePK families and subfamilies from www.kinase.com, where the core set is conserved in *D. discoideum*, *S. cerevisiae*, *C. elegans*, *D. melanogaster* and *H. sapiens*. Occurrences of antisense transcription for kinase loci were verified manually.

Non-coding RNA genes

Non-coding RNA (ncRNA) genes were identified from orphan PASA assemblies that have less than 10% overlap with annotated UTRs on the same strand. The orphan PASA assemblies are further filtered by transcript length (>200 nt) and expression level (combined FPKM > 2), and the remaining orphan PASA assemblies are classified into antisense ncRNA if the PASA overlaps with a CDS region on the opposite strand more than 30%, and intergenic ncRNA if the PASA does not overlap with annotated UTRs on the same strand or the overlap is less than 10%. For all ncRNAs that overlap with UTRs on the same strand (0-10%) were annotated as potential alternative UTRs. Because the difficulty of defining the extent of ncRNAs and because many of the GeneDB ncRNA annotations are not strand-specific, we did not try to modify the existing ncRNA annotations. Instead, we gave each ncRNA a new unique name and noted which overlap with previous ncRNA annotations (Table S18).

Phylogenetic analysis

Using the HAL pipeline (*48, 49*), a super alignment was constructed from 440 homologs identified from *Drosophila melanogaster* and 15 fungi (*Aspergillus fumigatus*, *Saccharomyces cerevisiae*, *Debaryomyces hansenii*, *Ashbya gossypii*, *Yarrowia lipolytica*, *Schizosaccharomyces pombe*, *Neurospora crassa*, *Schizosaccharomyces japonicus*, *Schizosaccharomyces cryophilus*, *Phanerochaete chrysosporium*, *Schizosaccharomyces octosporus*, *Stagonospora nodorum*, *Rhizopus oryzae*, *Schizosaccharomyces pombe*, *Ustilago maydis*). The alignment consisted of 98,765 amino acid characters. Using the program RAxML, a maximum likelihood phylogenetic analysis was conducted using a WAG+I+G model and 100 bootstraps. A maximum parsimony analysis with PAUP was also used to analyze the data set of which 60,485 were found parsimony-informative. Both analyses produced the same topology with 100% bootstrap support for all nodes (Fig. 1A).

Identification of transposons

Using BLAST, we searched the supercontigs of the *S. japonicus, S. octosporus, and S. cryophilus* (assemblies SJ1, SO3 and SCY2, respectively) for amino acid sequences similar to known transposons (including RT and IN of Tf, Ty and retroviruses, protein of LINE, zorro, dhp1, Hermes etc.) using tblastn and blastx. We then searched for repeat regions upstream or downstream of coding regions to identify the LTRs (or other terminal repeats). We also predicted the start and the stop codons of the ORFs. To identify the complete set of full length transposon and LTR sequences, we used all identified transposon sequences to search the supercontigs with BLAST. Subsequent *S. japonicus* genome assemblies, in particular SJ4, have disrupted transposon ORFs, possible due to the inclusion of a large amount of short-read data causing mis-assembly of highly repetitive sequences.

Gene orthology and evolutionary rates

Orthology assignments within the four species were obtained using the Synergy algorithm (*16*). Briefly, Synergy performs a bottom-up traversal of a species tree, identifying orthologs between the species below each ancestral species in the tree. Synergy uses sequence similarity and gene order to generate putative orthology assignments, and employs a modified Neighbor-Joining procedure to reconstruct gene tree topologies at each intermediate stage of the algorithm. It refines orthology assignments according to the resulting tree structure. This method generates a genome-wide catalog of orthology assignments and their corresponding gene trees.

In order to detect potential sequence homology across more distant species, we employed BLASTp sequence similarity searches (*50*) for each predicted protein in the four *Schizosaccharomyces* species across a compendium of 76 annotated fungal genomes using an E-value threshold of 0.01 (Table S32).

If any sequence pertaining to a group of orthologs was found to have a hit to any sequence in any of the outgroup genomes, then the group of orthologs was deemed to not be *Schizosaccharomyces*-specific. Protein sequences from all 4128 1:1:1:1 groups of orthologs within the *Schizosaccharomyces* clade were aligned using the MUSCLE alignment program using the species taxonomic tree for guidance (*51*). Gene tree branch lengths were then calculated using the PAML software package with the JTT amino acid substitution matrix (*52*). The program's estimated tree length was used as a measure each group of orthologs' evolutionary rates.

In order to compare these rates across disparate phylogenetic clades, this procedure was repeated for four *Saccharomyces sensu stricto* species and the orthologous rates for the corresponding 2,715 orthologous gene relations between these two clades (*37*) were compared. In order to account for the vast difference in evolutionary timescales between the *Schizosaccharomyces* and the *Saccharomyces* clades, the rates within each clade were first normalized by its median rate.

To test whether a particular category of functionally related genes was evolving comparatively faster or slower within the *Schizosaccharomyces* species, we performed a Kolmogorov-Smirnof test to determine if the rates within a functional category are significantly different from the overall background set of rates.

Enrichment of GO categories among set of fission yeast genes was calculated using the GO annotations of the *S. pombe* orthologs (*37*) with custom scripts, which are available upon request.

In order to identify examples of potential horizontal gene transfer events, each species' proteins were BLASTed against the UniRef90 protein database using a threshold of e-10 (*50, 53*). The resulting matches were filtered for self-same hits and the taxonomic information was then extracted to identify the sequence's origin. Cases where the most significant hits originated from bacterial genomes were considered putative HGT events. Those genes with top matches to bacterial proteins were further examined in a phylogenetic context to support or refute lateral gene transfer. The top 30 matching UniRef 90 proteins were aligned with the *Schizosaccharomyces* protein using MUSCLE (*51*), and a tree was constructed from the multiple alignment using FastTree (*54*). The trees were uploaded to ITOL (*55*) and manually examined. *Schizosaccharomyces* proteins branching in clades containing other eukaryotic proteins were refuted. Those proteins branching in clades with proteins of bacterial origin or restricted to bacterial protein homologs within the top 30 UniRef matches were classified as potential examples of lateral transfer. In one case in which three orthologs were examined and one was refuted based on tree topology, the other two were refuted by association.

In order to identify putative orthologs of species-specific, singleton genes, we systematically inspected each candidate's immediate upstream and downstream genes to determine whether they contained orthologs in the other genomes that were also neighboring singleton genes on the same strand. Most species-specific genes are in sub-telomeric region with low conservation of gene order, complicating this analysis. However, by ClustalW2 alignment of all groups of cognate singletons, we were able to recover potential true orthologs for a number of fast-evolving genes (Table S16) (*44*).

Motif discovery

To study evolution of regulatory mechanism, we used a catalog of over a dozen motifs that can be detected both in three of the four *Schizosaccharomyces* species (*S. japonicus, S. octosporus,* and *S. pombe*) and in 20 other *Ascomycota*. For each species, the motifs were learned by taking previously described motifs from *S. cerevisiae* (*56-58*), and using MEME (*59*) on the promoters of their putative targets in a different species. All motifs are associated with known transcription factors, based on their known DNA binding preferences in *S. cerevisiae* (*56-58*). For each motif in the catalog, we determined a set of conserved target genes within each clade of species (such as the *Schizosaccharomyces* clade, the *Candida* clade, and the *Saccharomyces* clade), and further associated it with a concise description of the functional roles of various gene modules in each clade, based on known functional annotations of those genes (*60-65*) (Fig. 5). In particular, the stress-response genes were taken from published expression analysis (*66*) and the antisense genes were taken from Table S21. Finally, we examined each case, to identify those examples where the motif is associated with distinct targets and functions in *Schizosaccharomyces*.

Comparative genomics analysis of conservation and coding potential

We first used MULTIZ (*67*) to generate a whole-genome multiple alignment of *S. pombe*, S. *cryophilus*, *S. octosporus* and *S. japonicus*. We then generated a nucleotide-level conservation track based on this alignment, using an analysis similar to the LRT method in phyloP (*68*). We also used PhyloCSF (*69*) to specifically assess protein-coding evolutionary signatures in conserved regions. For the transcripts representing novel loci, we used PhyloCSF to search for high-scoring windows of at least 25 codons, and then manually inspected the alignments of high-scoring candidates to distinguish those likely to represent complete, start-to-stop open reading frames and those that may represent partial or fragmentary transcript models. We also categorized novel transcripts containing highly conserved regions, but no apparent protein-coding signatures, as candidate conserved non-coding RNAs, while the remaining novel transcripts are apparently not conserved across these species. For the transcript models suggesting revisions to annotated coding genes, we compared the PhyloCSF scores for the annotated and revised open reading frames to show that the revisions are likely to improve the overall quality of the annotation (Fig. S13).

Sequencing of small RNAs

Two independent sRNA libraries were cloned from total, size-selected RNA prepared from logarithmically growing *S. japonicus* cultures. In one approach, RNAs in the 10-40 nt size range were enriched from 40 µg of total RNA using flashPAGE (Ambion). The adapters (Small RNA Oligo Only, Illumina) were ligated sequentially using T4 RNA Ligase (Promega). The ligation of the 5' adapter was dependent on the 5'-monophosphate on sRNA and the ligation of the 3' adapter was dependent on the 3'-OH on sRNA. Following ligation of each adapter, reaction products were purified by denaturing 7M urea-PAGE (15% 19:1 acrylamide:bisacrylamide), eluted from the gel slice and precipitated. RNA was reverse-transcribed using Superscript II (Invitrogen) and the library was amplified by PCR using Pfusion Hot Start High Fidelity DNA Polymerase (Finnzymes). PCR products were precipitated and purified by native PAGE. The sRNA library was subject to parallel sequencing on an Illumina Genome Analyzer.

Alternatively, total RNA was extracted from 25 ml mid-log cells with a 1:1 mixture of TES and hot phenol and sRNA in the 18-28 nt size range were purified from 100 mg of total RNA using denaturing 15% polyacrylamide gel electrophoresis, followed by FlashPage (Ambion). Purified sRNAs samples were prepared for Solexa-based sequencing as previously described (*70*). The resulting reads from both libraries were parsed using Perl scripts, and selected based on the presence of adapter sequences, mapped to the *S. japonicus* genome. The reads from the two data sets both have a modal size of 23 nucleotides and show similar genomic distributions. They were combined in to a single data set for analysis.

Deep sequencing of cDNA from polyA-enriched RNA

RNA preparation

S. pombe (SPY73), *S. japonicus* (yFS275), *S. octosporus* (yFS286) and *S. cryophilus* (OY26) were grown in yeast extract (1.5%), peptone (1%), dextrose (2%), SC amino acid mix (Sunrise Science) 2 grams per liter, adenine 100 mg/L, tryptophan 100 mg/L, uracil 100 mg/L, at 200 RPM in an New Brunswick Scientific air-shaker. This medium was chosen to minimize cross-species variation in growth (*71*).

To obtain log, diauxic-shift and early stationary phase cultures, overnight cultures for each species were grown to saturation in 3 ml rich medium. From the 3 ml overnight cultures, 300 ml of rich media was inoculated at the OD600 corresponding to 1x106 cell/ml for each species and grown in New Brunswick Scientific shaking water baths. Culture density was monitored by OD600 (Fig. S24). Glucose levels were monitored using the 2700 Select Bioanalyzer (YSI). Cells were harvested at mid-log, glucose depletion, and after growth plateau by quenching them in 60% liquid methanol at -40°C that was later removed by centrifugation at -9°C and stored overnight at -80°C. Harvested cells were later washed in RNase-free water and archived in RNAlater (Ambion) for future preparations.

To obtain heat-shock cultures, overnight cultures of each species were grown in 650 ml of media at 22°C (except *S. cryophilus*, which was grown at 20°C) to between $3x10^7$ and $1x10^8$ cell/ml OD600 = 1.0. The overnight culture was split into two 300 ml cultures and cells from each were collected by removing the media via vacuum filtration (Millipore). The cell-containing filters were re-suspended in pre-warmed media to either control (22°C, or 20°C for *S. cryophilus*) or heat-shock temperatures (37°C), except *S. cryophilus* that was heat shocked at 34°C. Density measurements were taken approximately one minute after cells were re-suspended to ensure that concentrations did not change during the transfer from overnight media. 60 ml of culture were harvested at 15 minutes after re-suspension by quenching them in 60% liquid methanol at -40°C that was later removed by centrifugation at -9°C and stored overnight at -80°C. Harvested cells were later washed in RNAse-free water and archived in RNAlater (Ambion) for future preparations. Cells were also harvested from cultures just before treatment for use as controls.

Total RNA was isolated using Qiagen RNeasy kit following manufacture's protocol for mechanical lysis using 0.5 mm zirconia/silica beads (Biospec). PolyA-enriched RNA was isolated from total RNA using Poly(A) purist kit (Ambion) or Dynabeads mRNA purification kit (Invitrogen). Total RNA and polyA-enriched RNA were treated with Turbo DNA-free (Ambion) as described. The integrity of the RNA was confirmed using the Agilent 2100 Bioanalyzer and quantified using RNA Quant-it assay for the Qubit Fluorometer (Invitrogen).

Strand-specific cDNA library

We created dUTP second-strand libraries starting from 200 ng of Turbo DNase treated, polyA-enriched RNA, using a previously described method (*72*) with the following modifications. We fragmented RNA in 1x fragmentation buffer (Affymetrix) at 80°C for 4 min, purified and concentrated it to 6 µl after ethanol precipitation. In addition, we added an 8-base barcode to each library to enable pooling of these libraries. The adaptor ligation step was done with 1.2 µl of index adaptor mix and 4000 cohesive end units of T4 DNA Ligase (New England Biolabs) overnight at 16°C in a final volume of 20 µl. Finally, we generated libraries with an insert size ranging from 225 to 425 bp.

Transcript reconstruction from RNA-Seq and incorporation into gene annotations

Transcript reconstruction from RNA-Seq reads was performed using a hybrid approach leveraging genome alignments coupled with *de novo* RNA-Seq read assembly. RNA-Seq reads were first aligned to genome sequences using TopHat (*73*) requiring a minimum intron length of 25 bases and a maximum intron length of 1 kb. Unmapped reads were identified and realigned to the genome using the more sensitive BLAT (*74*) alignment tool. Based on genome alignment coordinates and strand-specificity, reads were partitioned into disjoint strand-specific ‘coverage regions’, and neighboring coverage regions within 1 kb were merged into larger regions. The reads corresponding to alignments in the merged coverage regions were subjected to *de novo* assembly as part of the transcript reconstruction process described below.

Prior to *de novo* assembly, operations were performed to mitigate the errant fusion of adjacent and minimally overlapping transcripts by considering the evidence for transcription contiguity based on read pairing support. A pair contiguity sensor was devised as follows: A 200 base window was slid across each covered region with a step size of 33 bases. The RNA-Seq fragments (defined by either read of pairs derived from single fragments) overlapping the first 50 bases (segment A) and the last 50 bases (segment B) of the 200 base window were identified as fragment sets (A, B). The contiguity of fragment coverage was computed as a Jaccard similarity coefficient: $(A \cap B)/(A \cup B)$. Similarity coefficient values at or below 0.05 trigger a search for a covered region clip point. First, the smallest coefficient value within 100 bases is defined as a candidate clip point. If regions of greater transcript read pair contiguity exist within 300 bases of both sides of the candidate clip point, the covered region is clipped at the candidate site, partitioning the reads into disjoint read sets and ultimately precluding the co-assembly of reads that are divided by the dissection. Read sequences corresponding to alignments within the resulting disjoint coverage regions were *de novo* assembled using the Inchworm RNA-Seq assembly tool (http://inchworm.sourceforge.net/), using a k-mer length of 25 and requiring a minimum contig length of 100 bases.

We enhanced the PASA annotation pipeline (*75*) to leverage RNA-Seq for transcript reconstruction and genome annotation by leveraging strand-specific Inchworm RNA-Seq assemblies similarly to expressed sequence tags (ESTs) as follows. Inchworm-assembled transcript sequences were aligned to the genome using GMAP (*76*), reporting only the single best-scoring alignment for each transcript.

Valid transcript alignments were required to align to the genome across at least 90% of their length and with at least 95% sequence identity. Spliced transcript alignments were required to include consensus (GT-AG, GC-AG, AT-AC) dinucleotide splice pairs at intron boundaries. The read-derived k-mer coverage of each inchworm-assembled transcript (a proxy for transcript expression) exists as an attribute of each sequence, incorporated into the accession of the fasta-formatted Inchworm-assembled transcript. The PASA pipeline piles the inchworm transcripts along each strand of the genome with pile height prioritized by the k-mer coverage value, adding transcripts with greater coverage (expression) to piles first. Tentative RNA-Seq assembly artifacts, corresponding to transcripts with >= 30% overlap (by shortest length) of a dominant (prioritized) transcript and at most 10% the k-mer coverage of the dominant transcript were discarded. Remaining valid Inchworm-generated transcript alignments were treated identically to EST alignments in the PASA pipeline and integrated into gene annotations as previously described (*75*).

Validation of antisense transcription

Strand-specific q-rtPCR

RNA was extracted from 10 ODs of log-phase cells homogenized by vortexing with glass beads in 1 ml TRIzol Reagent (Invitrogen) following the manufacturer's instructions. The integrity of the purified RNA was checked by gel analysis (1.2% agarose, 1% 10X MOPS buffer, 5% formaldehyde gel run in 1X MOPS buffer). Contaminating DNA was removed by DNase I treatment (NEB). The complimentary DNA of each RNA transcript was then generated by using the SuperScriptIII First-Strand Synthesis System for RT-PCR (Invitrogen) and gene specific primers (Table S38). The concentration of cDNA was measured by NanoDrop 1000 spectrophotometry. 20 µl reactions contained 8 ng cDNA, 500 nM each of forward primer and reverse primer, and 10 µl SsoFast EvaGreen Supermix (Bio-Rad). The raw data was fit to a sigmoidal curve (*77*), normalized within each sample to the level of the *ade4* sense transcript and then between each sense-antisense pair to the level of the experimental sense transcript.

Strand-specific northern blot analysis

Synchronized meiosis was induced by shifting an asynchronous diploid *pat1-114* culture to the restrictive temperature for the indicated times. 10 µg total RNA was separated on 1% agarose gel containing 2.2M formaldehyde and 1x MOPS. Gel was rinsed with excess water and 20x SSC prior to over night capillary transfer with 20x SSC onto nylon membrane (Hybond-XL, Amershan). Template for ^{32}p-labeled RNA probe was PCR product cloned into pSC-A vector (Stratagene) that has T3 and T7 binding site flanking the inserted sequence using primers listed in Table S38. RNA probes for detecting specific genes were synthesized using MAXIscript (Ambion) with ^{32}P-αUTP according to manufacture's instructions. The probes were purified using Microcon YM-30 (Millipore) and eluted in 50 µl water. 10 µl probe (usually 0.5-1 x 10^6 CPM) was hybridized at 68°C over night in ULTRAhyb buffer (Ambion), washed 3 times with 1x SSC, 0.1% SDS for 10 minutes at 68°C, 3 times with 0.1x SSC, 0.1% SDS for 20 minutes, and exposed to a storage phosphor screen.

PCR-based splicing assay

Total RNA was primed with random hexamers or anchored gene specific primers (P1-gsp). The anchor is a unique sequence (named P1, see Table S38) at the 5' of each gene specific primers, so that only the cDNA primed with anchored gene specific primers would have the anchor sequence and cDNA primed other ways (such as primed with bits of DNA or RNA fragments in the RNA preparation or with RNA folding/random annealing) would lack the anchor sequence. We also added Actinomycin D into the reverse transcription reaction to inhibit the usage of DNA as template (Ruprecht et al., 1973). The cDNA was digested with RNaseA and RNAseH and then purified to remove all of the unused anchored gene specific primers. The purified cDNA was used for splicing assays with gene specific primer 5' of the intron and anchor sequence (P1) 3' of the intron, so that only the cDNA made from the sense strand can serve as template.

Chromatin immunoprecipitation

ChIP was performed as described previously for *S. pombe* (*78*), using anti-H3K9me2 mAb 5.1.1.

Clustering RNA-Seq expression of 4 Schizosaccharomyces species

We compared the patterns of differential gene expression between the fission yeast species by correlating the levels of gene expression in each condition. As expected, in *S. pombe*, expression levels in glucose depletion, early stationary phase and heat shock, all of which are stress conditions that inhibit cellular growth, are more similar to each other (r = 0.71 - 0.80) than any of them are to log phase (r = 0.25 - 0.63). Surprisingly, we found that expression levels in the other species (with the exception of *S. octosporus* early stationary phase) showed much less difference between conditions (r = 0.78 - 0.97) and clustered with the *S. pombe* log-phase sample (Fig. S25). These results suggest that the other species show a less robust transcriptional response to environmental conditions.

Despite these differences, the global organization of expression is conserved across species, with dominant programs for growth and stress. For a more detailed look at gene regulation, we clustered the RNA-Seq expression data by grouping together genes with similar expression profiles in each species, while minimizing the number of phylogenetic differences between species, using the 4-species phylogenetic tree. The clustering model within each species is a Gaussian mixture and is learned using Expectation Maximization (*79*). After the algorithm converges, we have a discrete probability distribution over cluster assignments for each gene in each species. A gene is assigned to the cluster with the highest probability of generating the gene's expression profile. We identify the number of clusters k, using penalized log likelihood of hold out data followed by manual inspection of cluster profiles learned for different values of *k*.

Identifying expression clusters with similar gene content

The outputs of the algorithm are *k* expression clusters for each of the input species that are mapped across the species by the cluster ID. Expression clusters that have the same ID across the four species tend to have similar expression profiles. However, the genes within these expression clusters are not necessarily orthologs. To assess the extent of gene content conservation we compared the gene content of the k clusters in one species to the gene content of the k clusters in another species based on a hyper-geometric test of gene set overlap. We consider a pair of clusters to be of similar gene content if the clusters have the most significant overlap with each other compared to their overlaps with any other cluster.

GO enrichment for expression clusters

To assess the biological coherence of our expression clusters we performed a hyper-geometric test based enrichment analyses of genes within our clusters. We performed both GOslim and the general GO process enrichment. Only those terms that had an FDR-corrected *p*-value < 0.05 were selected.

Supplemental References

1. H. L. Levin, D. C. Weaver, J. D. Boeke, *Mol Cell Biol* **10**, 6791 (1990).

2. X. Gao, D. F. Voytas, *Trends Genet* **21**, 133 (2005).

3. N. J. Bowen, I. K. Jordan, J. A. Epstein, V. Wood, H. L. Levin, *Genome Res* **13**, 1984 (2003).

4. C. Neuveglise, H. Feldmann, E. Bon, C. Gaillardin, S. Casaregola, *Genome Res* **12**, 930 (2002).

5. M. Kelly, J. Burke, M. Smith, A. Klar, D. Beach, *EMBO J* **7**, 1537 (1988).

6. B. Arcangioli, A. J. Klar, *EMBO J* **10**, 3025 (1991).

7. S. Sayrac, S. Vengrova, E. L. Godfrey, J. Z. Dalgaard, *PLoS Genet* **7**, e1001328 (2011).

8. D. H. Beach, *Nature* **305**, 682 (1983).

9. s. S. O. Material,

10. G. Manning, G. D. Plowman, T. Hunter, S. Sudarsanam, *Trends Biochem Sci* **27**, 514 (2002).

11. D. Matenia, E. M. Mandelkow, *Trends Biochem Sci* **34**, 332 (2009).

12. J. Quinn *et al.*, *Antioxid Redox Signal* (2010).

13. M. Yassour *et al.*, *Genome Biol* **11**, R87 (2010).

14. K. Lundgren *et al.*, *Cell* **64**, 1111 (1991).

15. N. Walworth, S. Davey, D. Beach, *Nature* **363**, 368 (1993).

16. I. Wapinski, A. Pfeffer, N. Friedman, A. Regev, *Nature* **449**, 54 (2007).

17. P. de Jong-Gubbels, J. P. van Dijken, J. T. Pronk, *Microbiology* **142**, 1399 (1996).

18. J. Piskur, E. Rozpedowska, S. Polakova, A. Merico, C. Compagno, *Trends Genet* **22**, 183 (2006).

19. J. M. Thomson *et al.*, *Nat Genet* **37**, 630 (2005).

20. N. F. Kaufer, J. Potashkin, *Nucleic Acids Res* **28**, 3003 (2000).

21. J. L. DeRisi, V. R. Iyer, P. O. Brown, *Science* **278**, 680 (1997).

22. G. Costello, L. Rodgers, D. Beach, *Current Genetics* **11**, 119 (1986).

23. A. Moldon *et al.*, *Nature* **455**, 997 (2008).

24. N. Averbeck, S. Sunder, N. Sample, J. A. Wise, J. Leatherwood, *Mol Cell* **18**, 491 (2005).

25. T. A. Volpe *et al.*, *Science* **297**, 1833 (2002).

26. H. P. Cam *et al.*, *Nat Genet* **37**, 809 (2005).

27. I. Djupedal *et al.*, *EMBO J* **28**, 3832 (2009).

28. M. Buhler, N. Spies, D. P. Bartel, D. Moazed, *Nat Struct Mol Biol* **15**, 1015 (2008).

29. K. J. Woolcock, D. Gaidatzis, T. Punga, M. Buhler, *Nat Struct Mol Biol* **18**, 94 (2011).

30. M. R. Henn *et al.*, *PLoS One* **5**, e9083 (2010).

31. D. B. Jaffe *et al.*, *Genome Res* **13**, 91 (2003).

32. M. Margulies *et al.*, *Nature* **437**, 376 (2005).

33. N. L. Quinn *et al.*, *BMC Genomics* **9**, 404 (2008).

34. J. T. Simpson *et al.*, *Genome Res* **19**, 1117 (2009).

35. R. M. Helston, J. A. Box, W. Tang, P. Baumann, *FEMS Yeast Res* (2010).

36. V. Wood *et al.*, *Nature* **415**, 871 (2002).

37. M. Aslett, V. Wood, *Yeast* **23**, 913 (2006).

38. G. Singh, A. J. Klar, *Genetics* **162**, 591 (2002).

39. R. Kolpakov, G. Bana, G. Kucherov, *Nucleic Acids Res* **31**, 3672 (2003).

40. E. M. Southern, R. Anand, W. R. Brown, D. S. Fletcher, *Nucleic Acids Res* **15**, 5925 (1987).

41. K. Furuya, H. Niki, *Yeast* **26**, 221 (2009).

42. F. Chen, A. J. Mackey, C. J. J. Stoeckert, D. S. Roos, *Nucleic Acids Res* **34**, D363 (2006).

43. L. Li, C. J. J. Stoeckert, D. S. Roos, *Genome Res* **13**, 2178 (2003).

44. M. A. Larkin *et al.*, *Bioinformatics* **23**, 2947 (2007).

45. J. M. Goldberg *et al.*, *PLoS Genet* **2**, e38 (2006).

46. S. K. Hanks, T. Hunter, *FASEB J* **9**, 576 (1995).

47. S. F. Altschul *et al.*, *Nucleic Acids Res* **25**, 3389 (1997).

48. B. Robbertse, J. B. Reeves, C. L. Schoch, J. W. Spatafora, *Fungal Genet Biol* **43**, 715 (2006).

49. B. Robbertse, R. J. Yoder, A. Boyd, J. Reeves, J. W. Spatafora, *PLoS Curr* **3**, RRN1213 (2011).

50. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, *J Mol Biol* **215**, 403 (1990).

51. R. C. Edgar, *Nucleic Acids Res* **32**, 1792 (2004).

52. Z. Yang, *Comput Appl Biosci* **13**, 555 (1997).

53. U. Consortium, *Nucleic Acids Res* **39**, D214 (2011).

54. M. N. Price, P. S. Dehal, A. P. Arkin, *Mol Biol Evol* **26**, 1641 (2009).

55. I. Letunic, P. Bork, *Bioinformatics* **23**, 127 (2007).

56. K. D. MacIsaac *et al.*, *BMC Bioinformatics* **7**, 113 (2006).

57. V. Matys *et al.*, *Nucleic Acids Res* **34**, D108 (2006).

58. C. Zhu *et al.*, *Genome Res* **19**, 556 (2009).

59. T. L. Bailey, C. Elkan, *Proc Int Conf Intell Syst Mol Biol* **2**, 28 (1994).

60. M. Ashburner *et al.*, *Nat Genet* **25**, 25 (2000).

61. T. R. Hughes *et al.*, *Cell* **102**, 109 (2000).

62. M. Kanehisa, S. Goto, *Nucleic Acids Res* **28**, 27 (2000).

63. M. Kanehisa *et al.*, *Nucleic Acids Res* **34**, D354 (2006).

64. H. W. Mewes *et al.*, *Nucleic Acids Res* **39**, D220 (2011).

65. E. Segal *et al.*, *Nat Genet* **34**, 166 (2003).

66. D. Chen *et al.*, *Mol Biol Cell* **14**, 214 (2003).

67. M. Blanchette *et al.*, *Genome Res* **14**, 708 (2004).

68. K. S. Pollard, M. J. Hubisz, K. R. Rosenbloom, A. Siepel, *Genome Res* **20**, 110 (2010).

69. M. F. Lin, I. Jungreis, M. Kellis, *Nature Precedings* (2010).

70. M. Tanurdzic *et al.*, *PLoS Biol* **6**, 2880 (2008).

71. A. M. Tsankov, D. A. Thompson, A. Socha, A. Regev, O. J. Rando, *PLoS Biol* **8**, e1000414 (2010).

72. J. Z. Levin *et al.*, *Nat Methods* **7**, 709 (2010).

73. C. Trapnell, L. Pachter, S. L. Salzberg, *Bioinformatics* **25**, 1105 (2009).

74. W. J. Kent, *Genome Res* **12**, 656 (2002).

75. B. J. Haas *et al.*, *Nucleic Acids Res* **31**, 5654 (2003).

76. T. D. Wu, C. K. Watanabe, *Bioinformatics* **21**, 1859 (2005).

77. R. G. Rutledge, *Nucleic Acids Res* **32**, e178 (2004).

78. A. Kagansky *et al.*, *Science* **324**, 1716 (2009).

79. A. P. Dempster, N. M. Laird, D. B. Rubin, *Journal of the Royal Statistical Society. Series B (Methodological)* **39**, 1 (1977).

80. M. Segurado, A. de Luis, F. Antequera, *EMBO Rep* **4**, 1048 (2003).

Supplemental Figure Legends

Figure S1 *S. octosporus* physical map

A physical map of a previous assembly (SO3) of the *S. octosporus* genome. The linkages between the scaffold ends were inferred by identifying restriction fragments that hybridize to probes from the ends of adjacent scaffolds. NS, AS and SS refer to the blot (NotI, SfiI and AscI, respectively) that informed the connection. All scaffold pairings show conservation of gene order with *S. pombe*. The non-centromeric scaffold gaps were closed in subsequent assemblies (SO4 and SO5).

Figure S2 Genetic linkage between the *S. japonicus* supercontigs

A) *S. japonicus* chromosomes inferred from genetic linkage within the 10 largest supercontigs in assembly SJ1 (Table S3 and Table S2). The gap between 7A and 5D was closed in the final assembly SJ4.

B) A schematic representation of centromeric linkage between the three centromeric regions.

Figure S3 An ultrametric phylogeny of the *Ascomycota*

The dates of species divergences were estimated by standard approaches. Estimated dates in millions of years are shown at each node. The branch on which the budding yeast whole-genome duplication occurred is indicated by a star.

Figure S4 Haplotype structure within *S. pombe*

A) The density of SNPs in NCYC132 and SPK1820, two divergent strains of *S. pombe*, relative to the reference strain 972 are displayed in 2.5 kb windows along their chromosome lengths. Note that SPK1820 is a mosaic of regions with a SNP pattern similar to 972 and regions similar to NCYC132.

B) An enlargement of the left subtelomere of Chromosome 2 showing that in regions of high SNP density, the pattern between NCYC132 and SPK1820 is similar. Note also the abrupt change from one SNP pattern to the other, the degradation of the SPBC1348.12 into a pseudogene and the loss of the neighboring pseudogene SPBC1348.11 from SPK1820.

Figure S5 Fission yeast transposon phylogeny

A phylogenetic tree of gypsy LTR-retrotransposons. Sequences of highly conserved domains of RTs were aligned by Clustal W. A neighbor-joining phylogenetic tree was then generated using MEGA 4.1. The tree was rooted to Ty4, a retrotransposon of copia family. The bootstrap values are labeled on branches.

Figure S6 The *cbp1* gene family

The 3 *S. pombe* and 6 *S. octosporus cbp1* orthologs were clustered with ClustalW. The fact that all of the nodes in the dendrogram are near the root suggests that the family arose near the divergence of *S. pombe* and *S. octosporus* and may have been involved in the speciation event that created their clade.

Figure S7 Enrichment of histone 3 lysine-9 dimethylation at centromeres, telomeres and mating-type loci

H3K9me2 ChIPs were analyzed by duplex PCR and ethidium bromide-stained agarose gels using primer pairs specific for putative centromere, telomere and mating type regions, in conjunction with primers for a control locus. Positions of PCR products are indicated; primer sequences are listed in Table S39. Enrichment of specific regions in ChIPs was quantified relative to the control. In cases in which the control band was too weak to quantify, level of enrichment was estimated by inspection and is indicated by + symbols.

Figure S8 Conservation of gene order at centromeres

Gene order is conserved across all three centromeres of *S. pombe* and *S. octosporus*, and on the presumptive centromere-proximal scaffold ends of *S. cryophilus*. Genes are shown as turquoise boxes. Collinear orthologous gene pairs are connected by pink (direct) or blue (inverted) bands. Duplicated genes are connected to both of their paralogs.

Figure S9 siRNAs are produced from transposons in *S. japonicus* but not *S. pombe*

Several different feature classes were extracted from the *S. pombe* and *S. japonicus* genomes: protein coding sequence, transposon sequence, and, for *S. pombe*, centromeres. The frequency of siRNA reads was calculated over sequential 20 bp windows across these features. Frequency was normalized to the maximum frequency within each feature class to compensate for the different class sizes.

Figure S10 Fission yeast mating-type loci

Schematic representations of the mating-type loci of the four fission yeast. *cis*-acting sequences involved in mating-type switching - H1, H2 and H3 - are indicated in red and orange. *cis*-acing sites involved, or putatively involved, in silencing are shown in green. Inverted repeats potentially involved in limiting the extent of the silenced domains are shown in magenta.

Figure S11 Physical mapping of mating-type loci

Southerns blots of genomic DNA probed for *matP* or *matM*. DNA was digested XbaI and BclI (*S. octosporus*), SacI and KpnI (*S. cryophilus*) or StuI and BstS17I (*S. japonicus*). The P and M probes identify *mat2-P* and *mat3-M*, respectively, as well as *mat1*, indicating that the strains switch mating type alleles at *mat1*, as *S. pombe* does. In addition, they identify bands (indicated by asterisks) associated with the epigenetically-programmed mark at *mat1*, which is manifest as a double-strand break.

Figure S12 Reannotation of the *S. pombe* genome

This example depicts the reannotation of exon structure, the identification of 5'- and 3'-UTRs, the reannotation of non-coding RNAs and the identification of new genes. The first exon of SPBPB21E7.02c in the previous annotation is not transcribed, nor is it conserved in other fission yeast. The first exon in the new annotation is transcribed and conserved in all four fission yeast. SPBPB21E7.04c shows the addition of 5'- and 3'-UTRs to the annotation. The extent of the non-coding transcript SPNCRNA1303 is better defined. SPBPB21E7.11c was annotated as a non-coding transcript, but is now annotated as coding. SPBPB21E7.10 is identified as a new coding gene. Above and below the chromosome coordinates are the previous annotations from GeneDB.org in light blue and the new annotations in dark blue. Above and below these are the strand-specific RNA-Seq read densities on a 0-1000 scale; signal above 1000 is truncated to make the low amplitude signal visible.

Figure S13 RNA-Seq-based reannotation improves the conserved coding capacity of the predicted gene models

389 *S. pombe* gene models that were changed after RNA-Seq-based reannotation were tested for changes to conserved coding capacity. For each gene, the gene models from before and after the reannotation were analyzed by PhyloCSF (*69*), which scores genes from related species on the basis of the conservation of coding capacity. The PhyloCSF 'before' scores are plotted versus the 'after' scores. 249 of the reannotations (64%) had no significant effect, often because the change did not effect conserved code regions. 122 (31%) improved the conserved coding capacity of the gene model. 18 (5%) reduced the conserved coding capacity. These gene models were manually inspected and corrected.

Figure S14 Alternately spliced transcripts

A) An example of intron retention. The splicing map shows the two alternative splicing patterns. The isoforms below show the products of the spliced and unspliced transcripts. Red indicates translated sequence; black indicates untranslated sequence. Retention of the intron truncates the 3'-coding sequence. Nonetheless, this alternative splice form is present in *S. octosporus* and *S. japonicus,* as well.

B) As in A, except the retention of the intron does not disrupt the coding sequence. The position of the intron is conserved in all four fission yeasts, but the alternative splicing is not, nor is the sequence of the intron.

C) As in B, but in this case the spliced isoform is the minor isoform. This intron is not conserved in other fission yeast and the coding sequence of the retained intron is conserved in all four fission yeasts.

D) As in A, except the unspliced isoform maintains the conserved ORF. The spliced isoform is not found in other species.

Figure S15 PhyloCSF analysis of alternately-spliced isoforms

300 *S. pombe* gene that display intron retention were tested for changes to conserved coding capacity. For each gene, the gene models from spliced and unspliced isoform were analyzed by PhyloCSF (*69*). The PhyloCSF 'spliced' scores are plotted versus the 'retained' scores. 273 (98%) of the spliced isoforms had equivalent or better coding capacity that their unspliced cognate (Table S10). 7 of the unspliced isoforms had the high conserved coding capacity of the gene model, suggesting that, in these rare cases, the unspliced isoform may be the biologically active species.

Figure S16 Intergenic polyadenylated transcription is enrich at origins

Two different feature classes were extracted from the *S. pombe* genome intergenic sequence (nucleotides between UTRs of protein coding genes) and origins (intergenic regions identified as origins of DNA replication(*80*)). The frequency of RNA-Seq reads was calculated over sequential 20 bp windows across these features; for coding sequence, the frequency of antisense reads was also calculated. Frequency was normalized to the maximum frequency within each feature class to compensate for the different class sizes.

Figure S17 Antisense transcription is frequently associated with apparent read-through transcription

A) Examples of antisense transcription. SPAC11D3.14c, SPAC11D3.15, SPAC11D3.16c all show some level of antisense transcription. The antisense transcription of SPAC11D3.14c is annotated as a non-coding gene SPNCRNA.609; this ncRNA is annotated as a possible alternate UTR because it overlaps with the 3'-UTR of SPAC11D3.13. The antisense transcription of SPAC11D3.15 is annotated as a non-coding gene SPNCRNA.610; the antisense transcription of SPAC11D3.16c is unannotated. Note that in the cases of both SPAC11D3.15 and SPAC11D3.16c, the antisense transcription is in the 5' region of a convergently transcribed gene. This situation is frequently seen. Above and below the chromosome coordinates are the gene annotations on the top and bottom strand, respectively. Above and below these are the strand-specific RNA-Seq read densities on a 0-100 scale; signal above 100 is truncated to make the low amplitude signal visible.

B) As in a. Both *hem14* and *lhs1* show antisense transcription apparently associated with read through of the other's transcriptional stop site. In the case of *lhs1*, the read-through transcription is sufficiently strong to be annotated as a long 3'-UTR.

Figure S18 Annotation of non-coding transcripts

A) An example of a distinct, conserved transcript with no apparent significant protein coding capacity. A nucleotide sequence conservation track for the four species (see Materials and Methods) is shown above the chromosome coordinates. Above and below these are the transcript annotations on the top and bottom strand, respectively, and the strand-specific RNA-Seq read densities on a 0-100 scale; signal above 100 is truncated to make the low amplitude signal visible.

B) As in a, except there is significant the opposite-strand overlap with the 3'-UTR of the downstream gene.

C) As in a, except that the sequence of the transcript is not conserved. Note the significant antisense transcription of the upstream *lea1* gene.

Figure S19 Enrichment of GO annotations among antisense-transcribed fission yeast genes

Enrichment of GO annotations among the 1:1:1:1 orthologs that have more antisense than sense transcription in at least one species are displayed. They are grouped by the number of species in which they are antisense transcribed. For example, the 1 of 4 gene set includes all orthologs for which an ortholog is antisense-transcribed in one of the four species. The species enrichments are for all antisense-transcribed genes in each indicated species. The *S. pombe* enrichments differ slightly from Figure 3B because only 1:1:1:1 orthologs across the fission yeast clade are included.

Figure S20 Validation of antisense-transcription of meiotic genes

A) Total RNA from log-phase cells was reverse transcribed with strand-specific primers and analyzed by q-rtPCR. Signal in each sample was normalized to an internal control (*ade4*) and then each pair was normalized to the amount of sense transcript.

B) Synchronized meiosis was induced by shifting *pat1-114ts* diploid cells to their restrictive temperature; The 0 hours time point contains non-synchronized vegetative cells. Total RNA was analyzed by northern blotting with strand-specific RNA probes. Note that antisense transcription of some, but not all, genes decreases as cells enter meiosis. In particular, The antisense transcription of *mde2* does not go down, and the antisense transcription of *spo4* only goes down an 6 hours, after the sense transcription has already come up.

C) Total RNA from synchronized meiotic cultures was reverse transcribed into standard (right) or strand-specific (left) cDNA and analyzed by intron-spanning PCR. Bands amplified from spliced (S) and unspliced (U) cDNA are indicated. *res2* is not an antisense-transcribed gene and is inefficiently spliced in both vegative and meiotic growth.

D) AS in A, using wt (yFS101), *dcr1*Δ (yFS779) or *ago1*Δ (yFS780) cells.

Figure S21 Loss of carbon-metabolism genes restricts fission yeast carbon sources

The central pathways for glucose metabolism are shown. Genes and pathways missing in fission yeast are shown in gray, *fbp1* missing just from *S. japonicus* and shown with green fill. Genes for which *S. pombe* is missing a paralog found in *S. cerevisiae* are shown with gray fill. Genes that are subject to glucose-dependent transcriptional regulation in *S. cerevisiae* but not *S. pombe* are show in blue.

Figure S22 A phylogenetic profile of the fission proteomes

A phylogenetic profile of the proteomes from each species was determined as described in the Methods. The *Schizosaccharomycetes* proteomes have a significantly greater percentage of content with an *Ascomycota* profile than the *Saccharomycotina* (5-6% versus 1-3%), as indicated with asterisks.

Figure S23 The fission yeast kinomes

The number of eukaryotic protein kinases (ePKs) occurring in *S. japonicus* (*Sjap*), *S. octosporus* (*Soct*), *S. pombe* (*Spom*), *S. cryophilus* (*Scry*), and *S. cerevisiae* (*Scer*). The ePK sequences are classified into eight major groups that occur widely in eukaryotes based on sequence similarity (*46*), six of which (AGC, CAMK, CK1, CMGC, Other and STE) are represented in *Schizosaccharomyces* and *Saccharomyces*. The 'Other' group consists of a collection of smaller families rather than a single cluster.

Figure S24 Growth curves of fission yeast cultures

Cultures were grown in well-aerated shaking flasks and monitored for optical density and glucose content. Samples were taken for RNA preparation at the indicated times.

Figure S25 Clustering of fission yeast expression patterns

Relative density of reconstructed cDNA fragments from paired-end RNA-Seq reads, normalized to gene length (fragments per kilobase per million reads, FPKM), was calculated for each annotated protein coding gene in each genome. The pair-wise Pearson correlations between expression patterns for each growth condition in each species were calculated from the FPKM values for all 1:1:1:1 orthologs (Table S40). Expression levels were clustered by r values.