ONLINE SUPPLEMENTARY MATERIAL

Maintenance of duplicate genes and their functional redundancy by reduced expression

Wenfeng Qian, Ben-Yang Liao, Andrew Y.-F. Chang and Jianzhi Zhang

Corresponding author: Zhang, J. (jianzhi@umich.edu)

Supplementary Methods

We identified one-to-one orthologs, two-to-one orthologs, and many-to-one orthologs between *S. cerevisiae* and *S. pombe* by using the Fungal Orthogroups database (http://www.broadinstitute.org/regev/orthogroups/). *S. cerevisiae* negative genetic interactions between duplicate genes were obtained from [1-2]. Gene expression levels in *S. cerevisiae* and *S. pombe* measured by RNA-Seq were obtained from [3-4]. We multiplied the read numbers in *S. cerevisiae* by 1.33 to equalize the mean expression levels of the 1597 one-to-one orthologous genes of *S. cerevisiae* and *S. pombe*. Genes with small numbers of sequencing reads are likely to have large estimation errors in their expression levels. We thus excluded genes with less than 20 sequencing reads in our analysis. Use of different cutoffs in the number of sequencing reads did not change our conclusions. Protein sequences of *S. cerevisiae* were downloaded from the *Saccharomyces* genome database (SGD, http://www.yeastgenome.org/). We computed d_N/d_S between *S. cerevisiae* duplicate genes for two-to-one orthologs using CODEML in the PAML package with default parameters [5]. We obtained the yeast protein complex data from SGD (ftp://genome-ftp.stanford.edu/pub/yeast/data_download/literature_curation/go_protein_complex_slim.tab), which contained 358 complexes comprising 1792 genes. Spearman's rank correlation coefficient between d_N/d_S and the average expression level of *S. cerevisiae* duplicates is -0.18 (P =0.002, one-tail t test) for the set of two-to-one orthologs considered in this work.

In a comparison of expression levels between 181 two-to-one and 891 one-to-one orthologs of *S. cerevisiae* and *S. pombe*, we calculated P values based on a chi-square test and then estimated Q values. We considered genes with Q values lower than 5% to have significant expression differences between the two yeasts. Among the one-to-one orthologs, 382 have significantly lower expressions in *S. cerevisiae* and 316 have significantly higher expressions in *S. cerevisiae*. Among the two-to-one orthologs, 99 have significantly lower expressions in *S. cerevisiae* and 35 have significantly higher expressions in *S. cerevisiae*. The ratio between the number of expression reduction genes to that of expression increase genes is 1.2 and 2.8, respectively, for one-to-one and two-to-one orthologs and their difference is highly significant ($P = 2\times10^{-5}$, Fisher's exact test).

A total of 16,027 human-mouse orthologous sets, of which 956 contained paralogs that arose from human or mouse lineage-specific duplications, were obtained from Ensembl Compara v56 [6]. Human and mouse RNA-Seq data from the brain, liver, and muscle were from previous publications [7-8]. Raw *n*-mer (n = 32 for human and 25 for mouse) RNA-Seq reads were mapped to the human (Ensembl v56 GRCh37) or mouse (mm9) genome coordinates by SeqMap [9]. Number of mapped RNA-Seq reads per gene was divided by the number of unique *n*-mer's per gene to yield the normalized expression level (R) for each gene [10]. The Z-score of log_2R for each gene was calculated by the formula $Z = (log_2R - T_M)/T_{SD}$, where T_M is the mean of log_2R for all genes and T_{SD} is the standard deviation of log_2R. Only human-mouse orthologous sets that contained at least one human or one mouse gene with RNA-Seq reads were kept for subsequent analysis. The data from the three tissues were analyzed separately. Because of the remaining distributional difference in Z score between the two species (Fig. S3B), we also ranked genes in each species according to their expression levels (highly expressed genes have high ranks). We found that human/mouse gene expression rank ratio is negatively correlated with human/mouse gene number ratio (ρ =-0.13, $P < 10^{-15}$ for brain; ρ =-0.13, $P < 10^{-15}$ for liver; ρ =-0.11, $P < 10^{-15}$ for muscle).

A comprehensive gene list of member proteins of manually annotated human and mouse protein complexes was obtained from the CORUM database (http://mips.helmholtz-muenchen.de/genre/proj/corum) [11]. For each orthologous set, its member genes were searched against the protein complex gene list. As long as at least one member gene of an orthologous set is either a member of a human or mouse protein complex, the orthologous set is considered to belong to the "complex" group. Otherwise, it belongs to the "non-complex" group.

References

1. Vavouri, T., *et al.* (2008) Widespread conservation of genetic redundancy during a billion years of eukaryotic evolution. *Trends Genet* 24, 485-488
2. Costanzo, M., *et al.* (2010) The genetic landscape of a cell. *Science* 327, 425-431
3. Nagalakshmi, U., *et al.* (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* 320, 1344-1349
4. Wilhelm, B.T., *et al.* (2008) Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature* 453, 1239-1243
5. Yang, Z. (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci* 13, 555-556
6. Hubbard, T.J., *et al.* (2007) Ensembl 2007. *Nucleic Acids Res* 35, D610-617
7. Pan, Q., *et al.* (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet* 40, 1413-1415
8. Mortazavi, A., *et al.* (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* 5, 621-628
9. Jiang, H. and Wong, W.H. (2008) SeqMap: mapping massive amount of oligonucleotides to the genome. *Bioinformatics* 24, 2395-2396
10. Sultan, M., *et al.* (2008) A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science* 321, 956-960
11. Ruepp, A., *et al.* (2010) CORUM: the comprehensive resource of mammalian protein complexes-2009. *Nucleic Acids Res* 38, D497-D501

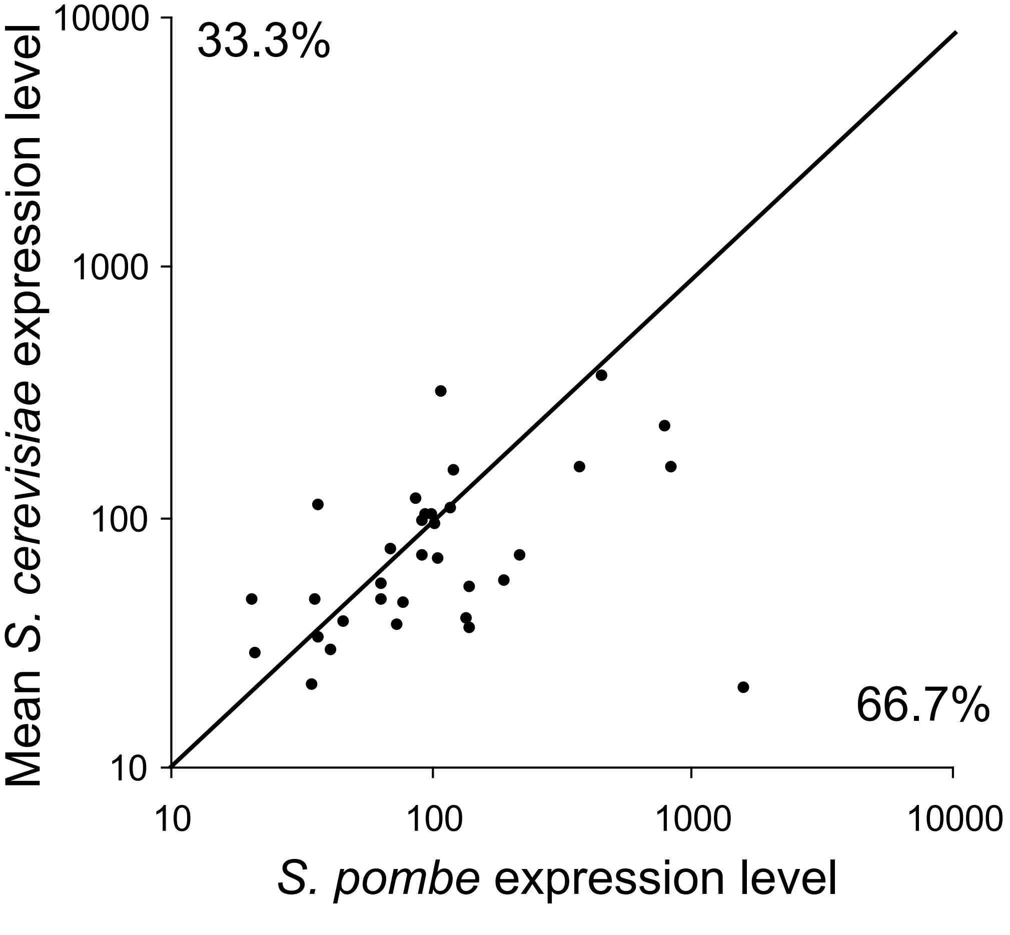

Figure S1. Expression levels of yeast many-to-one orthologs. Each dot represents a many-to-one ortholog. The expression level of the single gene in *S. pombe* and the mean expression of the multiple paralogs in *S. cerevisiae* are presented. The fraction of dots below the diagonal is significantly greater than expected ($P = 0.05$). We estimated that in this group of *S. cerevisiae* duplicate genes, an excess of 30.0% experienced mean expression reduction. The median expression ratio (*S. cerevisiae*/*S. pombe*) is 0.80 for many-to-one duplicates, significantly lower than that (0.94) for one-to-one orthologs ($P = 0.02$).

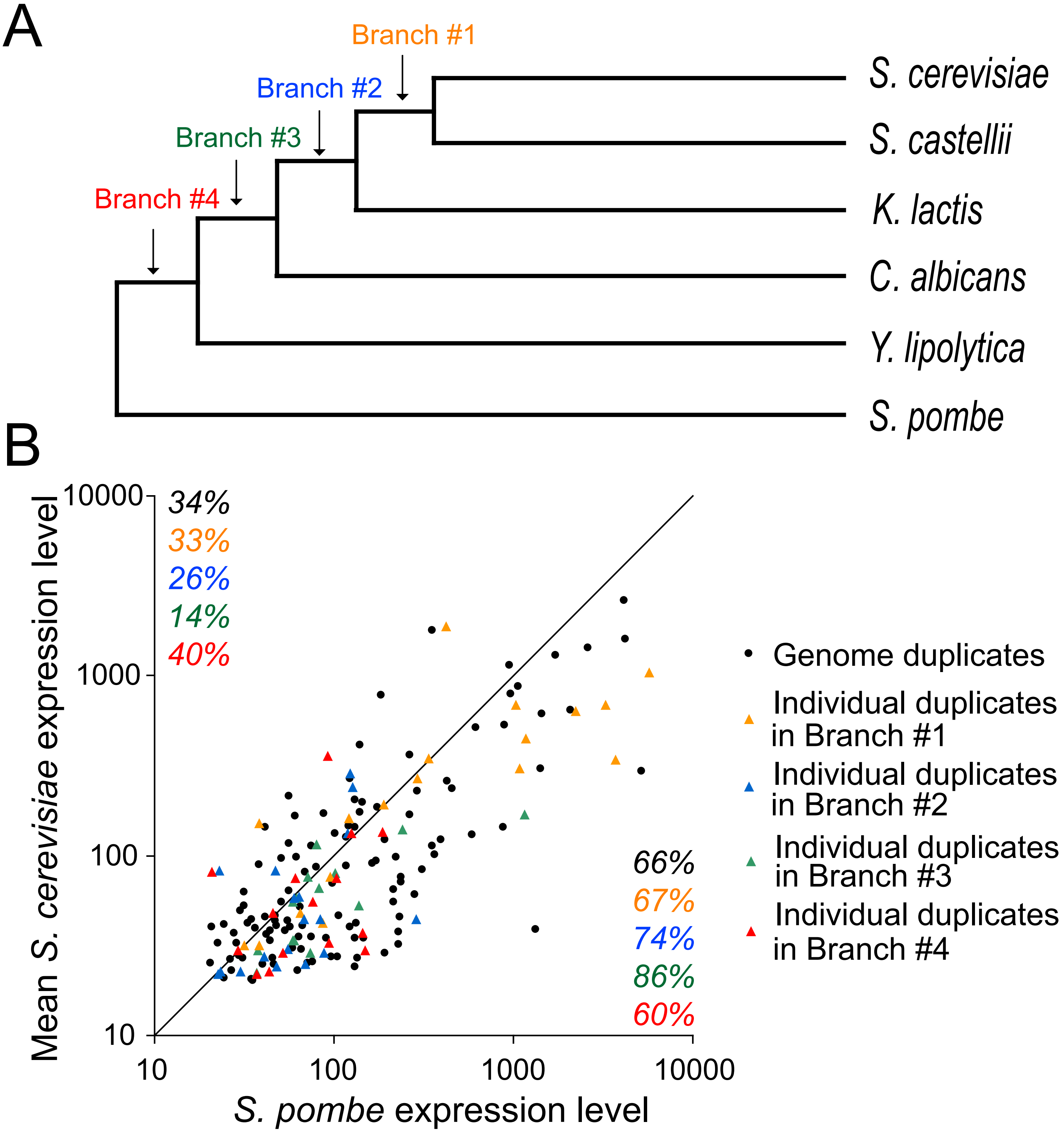

Figure S2. Expression levels of all two-to-one orthologs in *S. cerevisiae* and *S. pombe*. **(A)** The fungal phylogeney shows the branches on which gene duplications occurred. **(B)** Expression levels of different groups of two-to-one orthologs. Each dot represents a two-to-one ortholog, where the black color indicates genome-wide duplication and other colors indicate individual gene duplications with different colors corresponding to different age groups shown in panel A. The percentages of dots below and above the diagonal are indicated for each group.

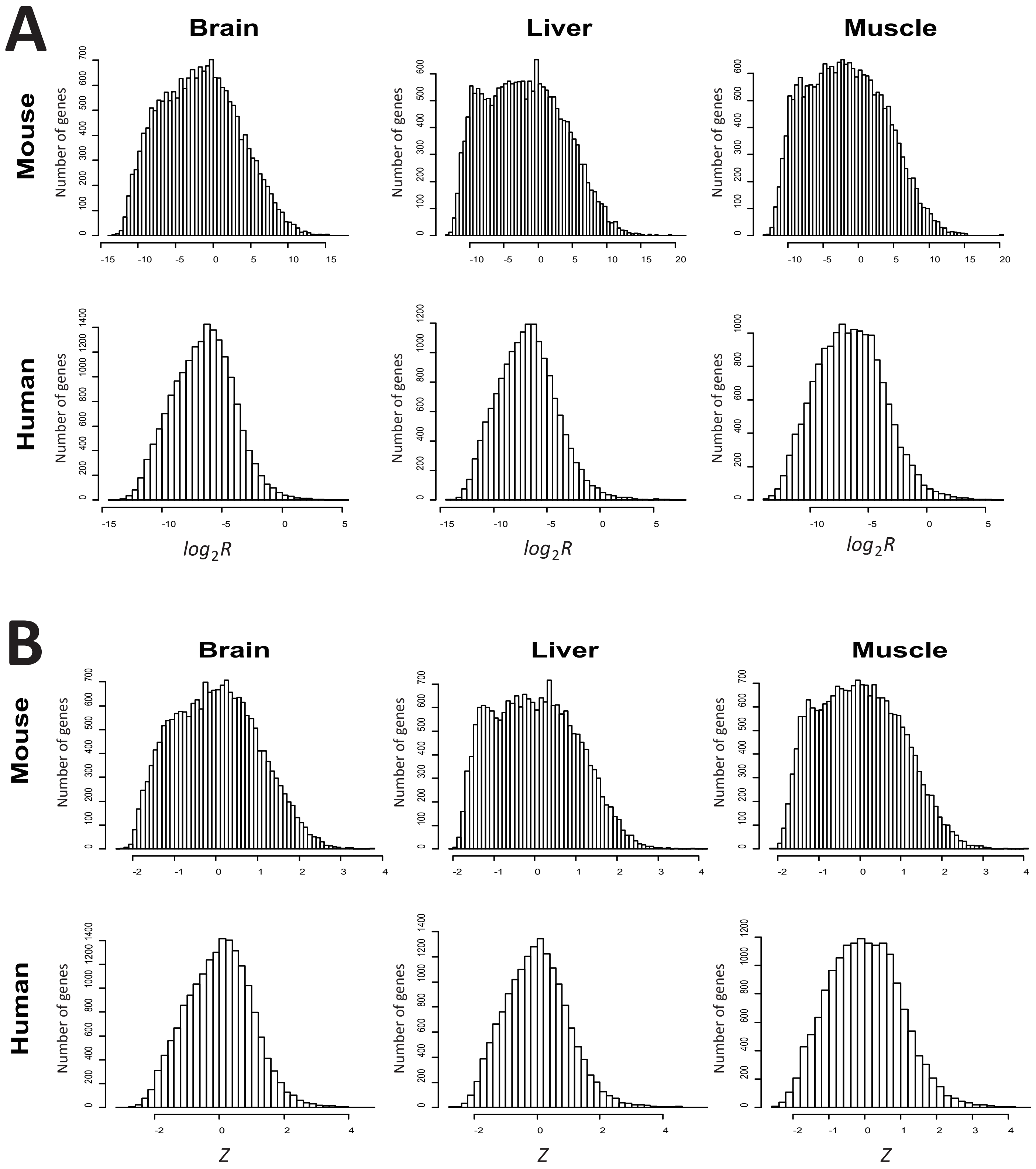

Figure S3. Distributions of human and mouse gene expressions in three tissues. **(A)** Distributions of *log_2R*, where *R* is the effective number of RNA-Seq reads per nucleotide for a gene. Note the different X-axis scales between human and mouse. **(B)** Distributions of the *Z* score of *log_2R*.