Supplementary Note S5

Overrepresentation of slowly evolving genes and genes involved in highly conserved biological processes among genes that underwent RGL.

In Table S5.1 we show that genes involved in various aspects of RNA metabolism, especially ribosome biogenesis and maturation are overreprepresented amongst RGL loci. We also show that genes that underwent RGL are slower evolving on average than those in other gene loss classes (S5.2). Finally, we use partial correlations to show that these effects are independent (S5.3).

S5.1. Overrepresentation of genes involved in RNA related processes among loci that underwent RGL.

Overrepresentation of genes implicated in RNA related processes among RGL loci was assessed by Fisher's exact tests against control sets of genes. RGL for snoRNA genes was defined exactly as described for protein coding genes (see Methods in main text) and determined by searching genomic DNA with *S. cerevisiae* snoRNA gene sequences downloaded from ftp://genome-ftp.stanford.edu/pub/yeast/sequence/genomic_sequence/rna (May 2004). For the snoRNA comparison the control was the proportion of RGL among snoRNA genes versus the proportion of RGL among protein coding genes in Classes 3 and 4 (i.e., 12/52 versus 219/2176). For all other annotations in Table S5.1, Class 3 loci (none of which have undergone RGL) were compared to Class 4 loci (all of which have undergone RGL) as defined in Fig. 2 in the main text.

Annotation	RGL loci			Non-RGL loci			Significance
	With Annotation	Total	Proportion	With Annotation	Total	Proportion	
YPD: All RNA Terms§	40	173	23%	215	1438	15%	5×10^{-3}
RNA binding complexes†	36	219	16%	142	1956	7%	2×10^{-5}
Nucleolar Localization¶	22	170	13%	57	1451	4%	7×10^{-6}
snoRNA genes¥	12	-	-	40	-	-	3×10^{-4}

§ Genes annotated by the Yeast Proteome Database whose annotations contain the term 'RNA', compared to all others except those annotated as 'Biological process unknown'.
† Membership of an RNA-binding complex as determined by Krogan *et al.*[1], compared to genes that were not found to be members of an RNA-binding complex.
¶ Nucleolus-localized proteins as determined by Huh *et al*.[2] compared to all other proteins with a known alternative subcellular localization.
¥ The proportion of RGL loci among snoRNA genes was compared to the proportion of RGL among protein coding genes in Classes 3 and 4 as described above.

References

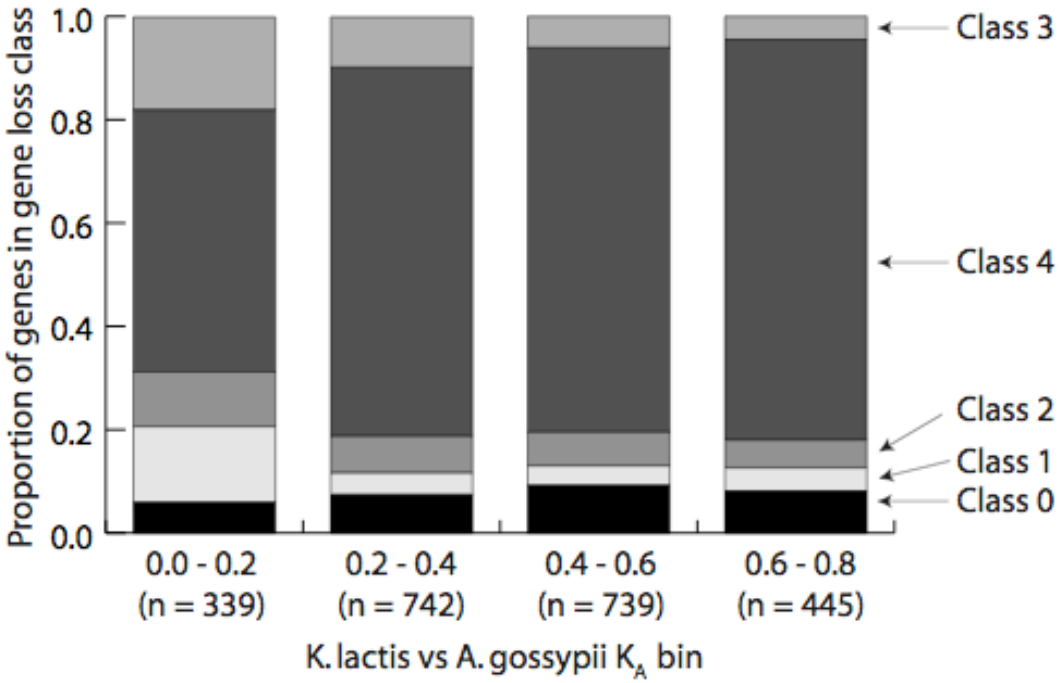
1. Krogan, N. J. et al. High-definition macromolecular composition of yeast RNA-processing complexes. Mol Cell 13, 225-39 (2004).
2. Huh, W. K. et al. Global analysis of protein localization in budding yeast. Nature 425, 686-91 (2003)

S5.2. Evidence that slowly evolving genes are overrepresented among RGL loci.

Representative KA values[1] were calculated for each ancestral locus using *K. lactis* and *A. gossypii* orthologs. Estimates were generated using yn00 in the PAML package on ungapped alignments. Each ancestral locus is represented once regardless of whether or not it is still duplicated.

Proportion of loci in *K. lactis vs A. gossypii* KA bins accounted for by Gene Loss Classes

Gene Loss Class	Bin 0.0 - 0.2	0.2 - 0.4	0.4 - 0.6	0.6 - 0.8	Median K_A for class
0	0.059	0.074	0.092	0.081	0.471
1	0.147	0.042	0.038	0.045	0.358
2	0.106	0.071	0.065	0.054	0.412
3	0.180	0.098	0.061	0.045	0.331
4	0.507	0.714	0.744	0.775	0.477
Number of Loci	339	742	739	445	



The contribution of Class 3 (RGL) loci declines with increasing rate class.
The comparatively large contribution of Class 1 loci in the slowest rate bin is due to enrichment for ribosomal proteins. See Supplemental Information S1 for details.

References

1. Davis, J. C. & Petrov, D. A. Preferential duplication of conserved proteins in eukaryotic genomes. PLoS Biol 2, E55 (2004).

S5.3. Evolutionary rate and functional class contribute independently to the pattern of gene loss.

Non-parametric partial correlations (described below) were used to investigate the relationship between the following factors:

"RGL status": Whether a locus has undergone RGL or not (coded as 1 or 0).

"Ka": Extent of nonsynonymous substitution in the same locus compared between *K. lactis* and *A. gossypii* .

"RNA": Locus is involved in RNA-related biological processes according to YPD annotation, or not (coded as 1 or 0).

Nonparametric correlation				Nonparametric partial correlation		
Factor 1	Factor 2	Spearman's rho	P	Controlling for	Partial correlation	P
RGL status	Ka	0.17	4.97E-11	RNA	0.17	9.17E-11
RGL status	RNA	-0.10	1.17E-04	Ka	-0.10	2.18E-04

The correlation between RGL status and Ka does not change when gene involvement in RNA-related functions is controlled for. Likewise, the correlation between RGL status and RNA-related gene functions does not change when Ka is controlled for. The dataset is 1417 loci of which 171 are RGL loci and 1246 are Class 4.

We also examined the relationship between RGL status, Ka, and protein abundance. Here, "Exp" is protein abundance data for S. cerevisiae from Ghaemmaghami et al., Nature 425:737-741 (2003).

Nonparametric correlation				Nonparametric partial correlation		
Factor 1	Factor 2	Spearman's rho	P	Controlling for	Partial correlation	P
RGL status	Ka	0.14	1.73E-06	Exp	0.09	3.62E-04
RGL status	Exp	-0.14	5.08E-06	Ka	-0.08	1.55E-03

The correlations of RGL status with Ka and protein abundance are not independent.
The dataset is 1086 loci of which 132 are RGL loci and 956 are Class 4.