Supplementary material for:

Pereira-Leal and Teichmann (2004) "Novel specificities emerge by step-wise duplication of functional modules"

Supp. 1. False negatives

In order to answer the question whether module duplication is a feature of biological systems, we sacrificed sensitivity in our search algorithm as it was more important to avoid false positives than to find all cases of modular duplication. Our estimates of the degree of modular duplication thus provide a lower bound, and the actual contribution of this type of mechanism for functional specialization is likely to be higher. In fact, detailed analysis of the "non-significant" hits in the manually curated data set reveals at least one more case of putative homology that failed to accomplish a significant score because the number of shared components was higher than the number of similar proteins. These are the RNA polymerases I, II and III, multi-protein complexes, which like the other cases discussed in the text, show similar function with different specificities: transcribing different sets of genes in this case. A complete and detailed list of all similarities detected in each data set are provided on the authors' website:

http://www.mrc-lmb.cam.ac.uk/genomes/jleal/modules/module_duplication.html

Supp 2. Which modules duplicate?

We investigated whether there is a bias for duplication of complexes in a particular sub-cellular localization or biological process, and also whether duplicated modules tend to have co-expressed components.

In order to assign a sub-cellular localization or biological process to a complex, we used a simple majority-voting scheme based on the annotations of the individual components of the complex. For each component, we used GO functional annotations[1] at the biological process level, and experimentally derived sub-cellular localization[2] (see methods). We first analysed the cellular localization of complexes. In all three data sets, protein complexes assigned to nucleus and cytoplasm are the most frequent, which reflects the relative abundance of proteins in these categories. However some compartments are enriched with complexes with duplicates in comparison with the level of duplication in the whole data set. In all data sets, this is true of the early Golgi. In both MIPS and TAP data the late Golgi and the cellular periphery, and in the HMS-PCI data, the bud and bud neck, are also enriched in complexes with duplicates, as shown in Figure S1.

This trend for compartments associated with membrane trafficking to contain complexes with duplicates is supported by the observation that the biological process ‘vesicle mediated transport’ is over-represented in duplicated complexes in all three data sets, as shown in Figure S2. In the MIPS data set, this is due to the COP and AP complexes. We conclude that there is an increased propensity for duplication of modules associated with membrane trafficking, but because the numbers in each compartment and biological process are small, it is difficult to assess whether this is statistically significant.

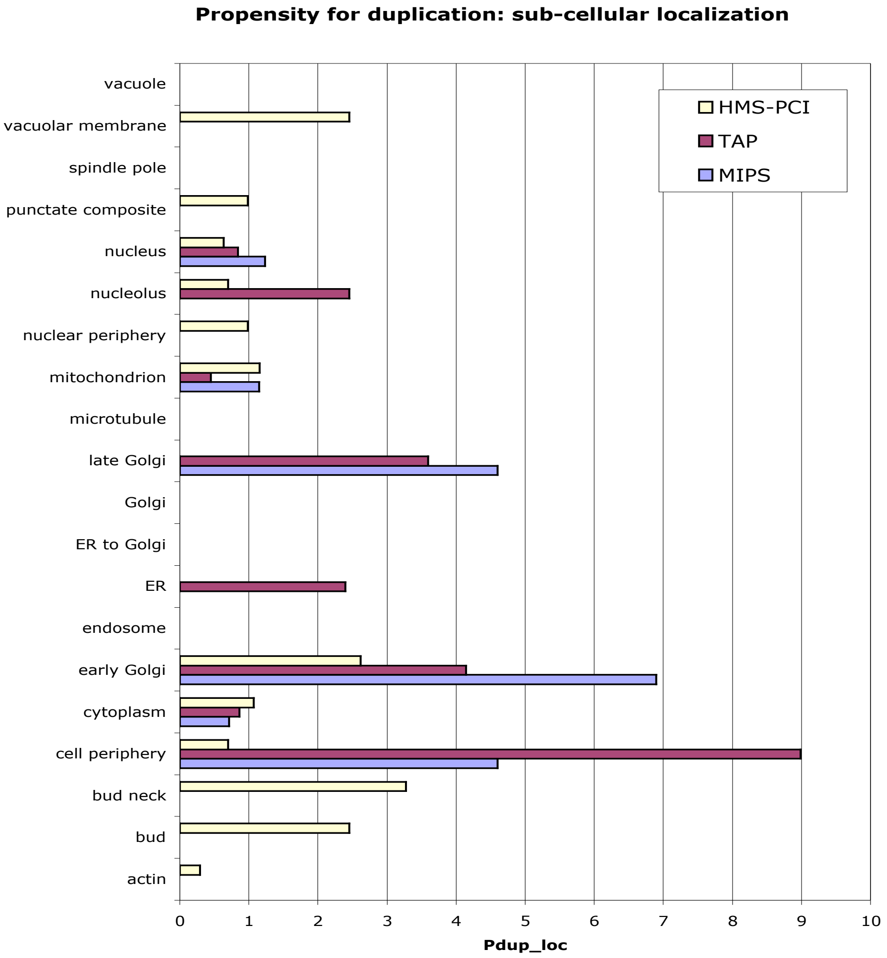

Figure S1 – Propensity for duplication of complexes in particular cellular compartments. We calculated the bias in complex duplication as follows: $P_{dup\_loc} = \dfrac{f_{loc}}{P_{dup}}$, where f_{loc} represents the relative frequency of complexes with duplicates in the compartment and P_{dup} is the probability of duplication in the whole data set.

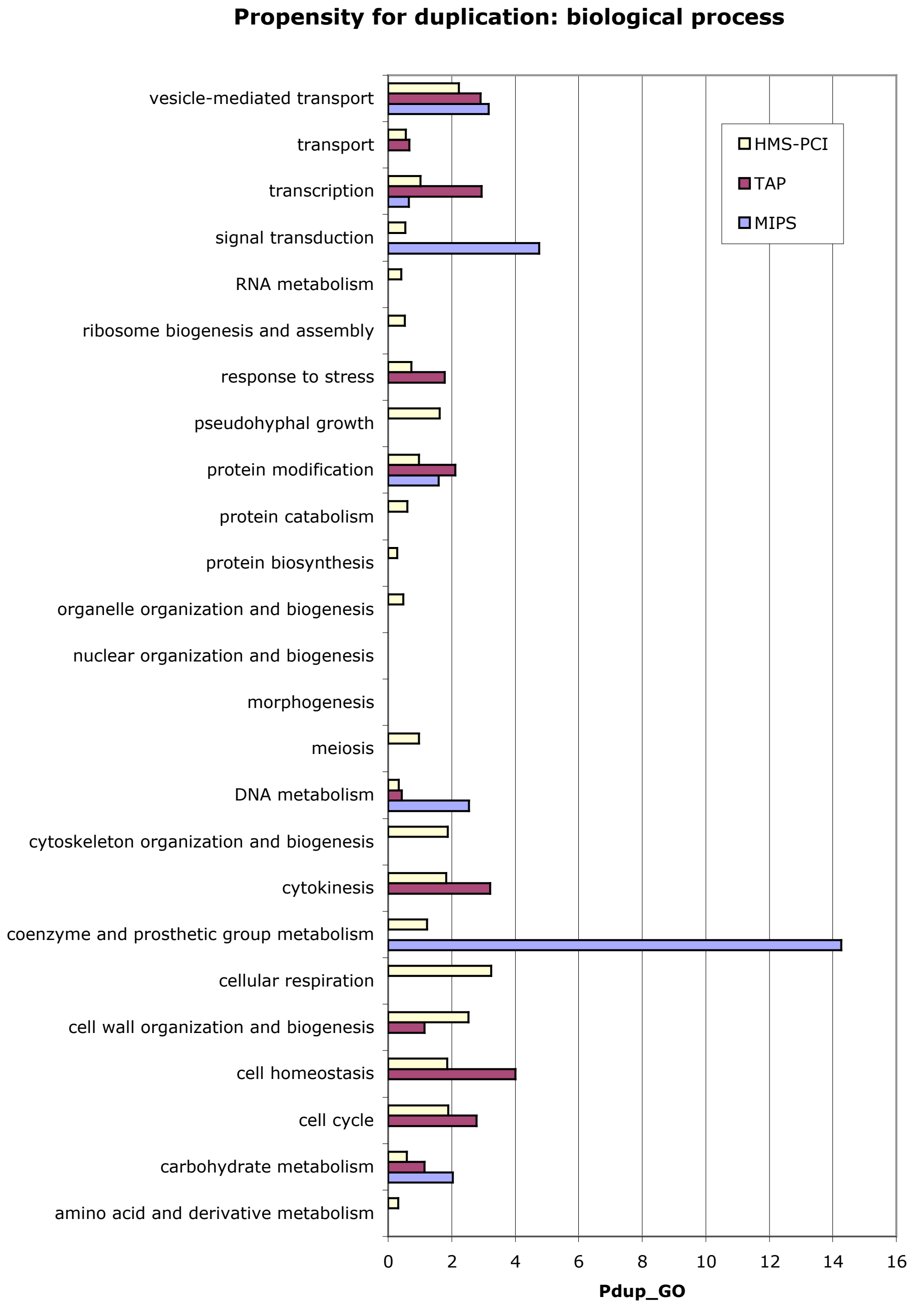

Figure S2 – Propensity for duplication of complexes in particular biological processes. We calculated the propensity of duplication in each biological process as follows: $P_{dup\_GO} = \frac{f_{GO}}{P_{dup}}$, where f_{GO} represents the relative frequency of complexes with duplicates in the functional class and P_{dup} is the probability of duplication in the whole data set.

In order to investigate a different aspect of the function of duplicated modules, we compared the degree of mRNA co-expression within duplicated and singleton complexes across cell cycle[3]. We find a small tendency for singletons to have co-expressed components more frequently than duplicated complexes, as shown in Table S1. The reason for this may be that most duplicated complexes share components with other complexes.

Table S1 – Protein complexes with co-expressed components during cell cycle

	No. (and %) of complexes with duplicates with co-expressed component	No. (and %) of singletons with co-expressed components
MIPS	0 (0%)	29 (14%)
TAPs	4 (12%)	93 (17%)
HMS-PCI	9 (13%)	95 (17%)

Supp 3. Are duplicated complexes coded by genes adjacent on the yeast chromosomes?

We tested for adjacency of the genes coding for the proteins in a complex within 30 Kb[5] and also within 500 Kb. We chose a length of 500Kb, because recent work suggests that spontaneous segmental duplications of this size may be frequent in the genome of the budding yeast[6]. As shown in Figure S3, there is no particular tendency for genes within a duplicated complex to be closer to each other compared to genes in complexes in general, or genes within singleton complexes. This is true for both distances, 30Kb and 500Kb.

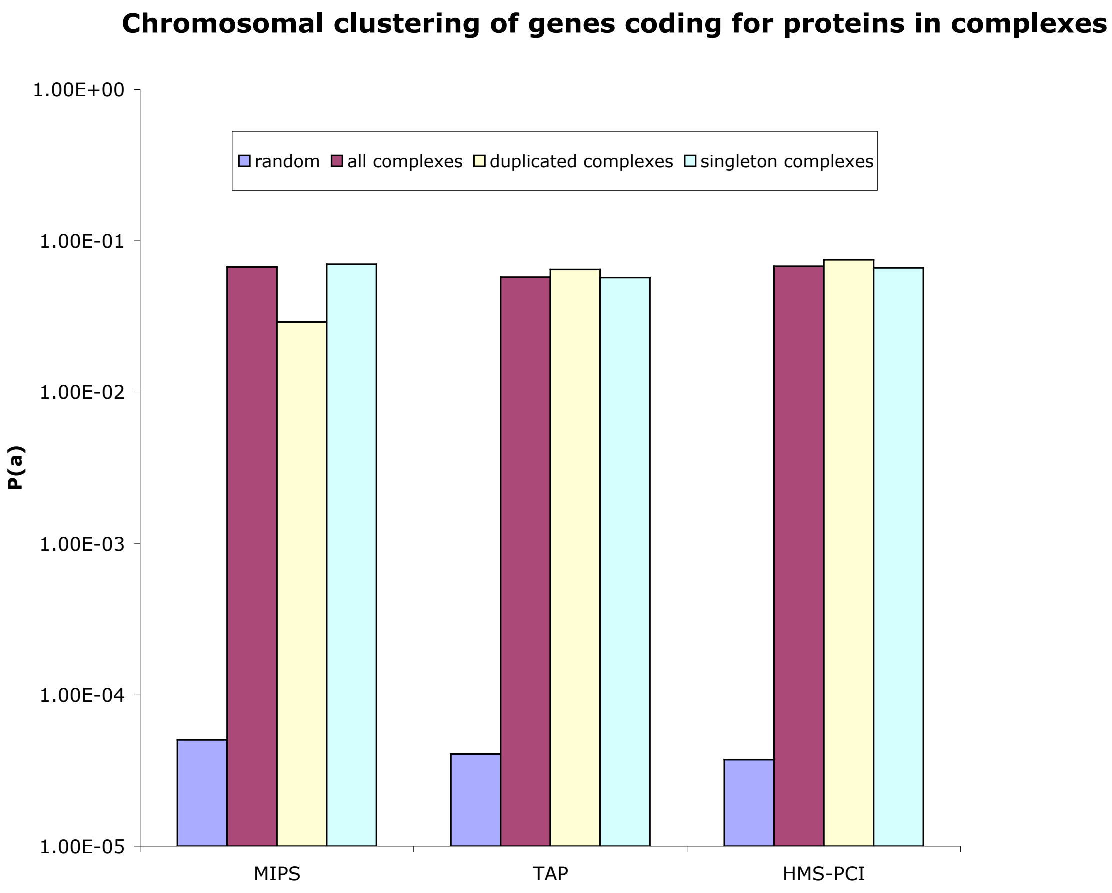

Figure S3 - Frequency of pairs of genes in the same complex being within 30Kb on a yeast chromosome. Distance between genes is measured between the centres of each gene ($\frac{start - end}{2}$). P(a) is the fraction of gene pairs within complexes within 30Kb out of all possible pairs of genes within a complex. The random expectation was calculated for the proteins in each of the three data sets, by 10,000 random samplings of gene locations. The genes of proteins within complexes are much more frequently within 30Kb of each other than randomly located genes as pointed out in Teichmann & Veitia[4], but

there is essentially no difference between duplicated complexes and singleton complexes for any of the three data sets. The same trends hold using a threshold of 500Kb.

Supp 4. Supplementary material references

1. Ashburner, M. et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25-9 (2000).
2. Huh, W. K. et al. Global analysis of protein localization in budding yeast. *Nature* **425**, 686-91 (2003).
3. Cho, R. J. et al. A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol Cell* **2**, 65-73 (1998).
4. Jansen, R., Greenbaum, D. & Gerstein, M. Relating whole-genome expression data with protein-protein interactions. *Genome Res* **12**, 37-46 (2002).
5. Teichmann, S. A. & Veitia, R. A. Genes encoding subunits of stable complexes are clustered on the yeast chromosomes. *Genetics* **In press** (2004).
6. Koszul, R., Caburet, S., Dujon, B. & Fischer, G. Eucaryotic genome evolution through the spontaneous duplication of large chromosomal segments. *Embo J* **23**, 234-43 (2004).