1.

מכון ויצמן למדע
WEIZMANN INSTITUTE OF SCIENCE

Thesis for the degree
Doctor of Philosophy

חבור לשם קבלת התואר
דוקטור לפילוסופיה

By
Orna Man

מאת
אורנה מן

מעורבותה של בקרת ביטוי חלבונים בהתפצלות מינים

The involvement of regulation of protein expression in species divergence

Regular Format

Advisors
Prof. Joel L. Sussman
Dr. Yitzhak Pilpel

מנחים
פרופ' יואל ל. זוסמן
דר' יצחק פלפל

August 2007

אלול תשנ"ז

Submitted to the Scientific Council of the
Weizmann Institute of Science
Rehovot, Israel

מוגש למועצה המדעית של
מכון ויצמן למדע
רחובות, ישראל

Acknowledgements

Now that my PhD has come to an end I feel indebted to many people for helping me get to this point. First and foremost, I would like to thank both my advisors – Joel Sussman and Tzachi Pilpel. Joel who, was my mentor from the beginning of my PhD, continued to have faith in me even when my initial project came to a dead end, and above all was always willing to listen and help. Tzachi, who was on my PhD advisory board, and came to my rescue, offering to be my supervisor, when it was concluded that I would have to change my PhD project. Switching projects turned out to be a positive turning point in my PhD, and I feel I owe much of the success of this turning point to Tzachi, who has contributed to my success both through our fruitful discussions and through his constant encouragement.

I would also like to thank my collaborators on the olfactory receptor project – Gustavo Glusman and Yoav Gilad. I especially want to thank Yoav who is not only my collaborator but has also been my friend for many years, and with whom I could consult on practically any subject. I am also grateful to my PhD advisory board. To Debbie Fass I am grateful for agreeing to sit on an advisory board of a PhD that is far from the subjects researched in her lab, thus allowing me her unique outlook on my research. Debbie has also gone beyond her call of duty, and it is her interest in my progress that led me to switch PhD projects on time. I am also grateful to Hanah Margalit who agreed to join my advisory board in the middle of my PhD, and who especially came from Jerusalem to my advisory board meetings - her advice during these meetings was most helpful.

I would like to thank the members of the Pilpel lab who were always willing to discuss my work with me, especially Zohar Bloom, Orna Dahan and Amir Mitchell. I would also like to thank Aviv Paz and Tzviya Zeev-Ben-Mordehai, my friends from the Sussman lab, for the fun lunches and moral support.

Finally, I would like to thank my family. There is no doubt that much of who I am today is due to my upbringing. So, I would like to thank my parents for encouraging my curiosity and teaching me to be tenacious. Last, but definitely not least, I am grateful to Idan, my husband, for his love, support and encouragement.

Summary

Differences in protein levels may result in different phenotypes among species. These variations in protein levels are due to differences in the various levels that control protein expression, which range from the presence/absence of a gene from the genome to more subtle genetic differences that affect a protein's localization or post-translational modification. The aim of my thesis studies was to expand the picture of mechanisms underlying phenotypic divergence by examining the involvement of differences in coding sequences, which affect protein expression, in species phenotypic divergence. The thesis is composed of two projects; the first consists of analysis of yeast species and the second of mammals.

In the first project, I aimed to investigate the relationship between differential translation efficiency of orthologous genes and phenotypic differences between yeast species. To this end, I focused on the extent of adaptation of the codon usage of a gene to the available tRNA pool, as a surrogate measure for the efficiency of the peptide elongation process. To assess translation effects on divergence, I analyzed ~2,800 orthologous genes in nine yeast genomes, predicting for each gene in each species its translation efficiency. Mining this data set, I found many groups of functionally-related genes with correlated patterns of translational efficiency across the species. Among these, I found cases where the patterns of translation efficiency for the group of genes was in concordance with a phenotype. For instance, the genes required for respiration were found to be translated more efficiently by obligate aerobic species than by species that prefer fermentation over respiration even under aerobic conditions, whereas the glycolytic enzymes displayed a complementary pattern. Also, the genes required for mRNA splicing are translated more efficiently by species with a higher demand for splicing. I also found many cases where statistically significant patterns cannot be explained by known phenotypic differences and may open avenues to further investigation into the physiology of the species analyzed. My results indicate that synonymous codon choices may be under strong selection, adapting the codons to the tRNA pool to different extents, depending on the gene's function and organism's needs. I found a relatively constant tRNA pool in all species, indicating that co-evolution of protein-coding genes and tRNAs takes place mainly in

a distributive fashion at the protein-coding gene level. The same gene in different species adapts itself to different extents to an essentially unmodified tRNA repertoire. I conclude that like factors such as transcription regulation, translation efficiency affects and is affected by the process of species divergence.

In the second project, I compared the olfactory receptor repertoires of human and chimpanzee that are assumed to have different sensory needs. The coding sequence characteristics I aimed at in this project were open reading frame disruptions (nonsense mutations or frameshifts) that transform a gene into a pseudogene that is transcribed but cannot form a functional protein, as well as differences in the ratio of synonymous to non-synonymous mutations that imply that selective pressures on the gene differ among the species. Confirming previous predictions, I found that, compared to chimpanzees, the human genome contains a significantly larger proportion of nonfunctional OR pseudogenes, implying that the human sense of smell is deteriorating at a greater rate than that of chimpanzee. Through a comparison of pseudogenes that became nonfunctional before the divergence of these two species with those that became nonfunctional after the divergence I was able to reject the hypothesis that humans have been accumulating OR pseudogenes at a constant neutral rate since their divergence from chimpanzees. The comparison of the two repertoires revealed that most chimpanzee OR genes have a clear human ortholog, so that the overall structure of the repertoire is conserved. Nonetheless, I found two chimpanzee-specific OR subfamily expansions and three expansions specific to humans. The comparison also suggested that a subset of OR genes are under positive selection in either the human or the chimpanzee lineage. Thus, although overall there is relaxed constraint on human olfaction relative to chimpanzee, species-specific sensory requirements appear to have shaped the evolution of the functional OR gene repertoires.

In conclusion, both projects have revealed differences among the coding sequences of the species analyzed that may take part in their phenotypic divergence. It is my hope that future experimental studies will examine whether the differences identified in my studies indeed underlie phenotypic differences among yeast species and sensory differences between humans and chimpanzees.

2. List of Abbreviations and Symbols

bp – base pairs

CAI – codon adaptation index

CRP – cytosolic ribosomal protein

Dn/Ds – ratio of nonsynonymous to synonymous divergence

f(Xg) – function predicting the effective number of codons from silent GC content

FDR – false discovery rate

GO – gene ontology

HMM – hidden Markov model

LRT – likelihood ratio test

MRP – mitochondrial ribosomal protein

MY – million years

MYA – million years ago

Nc – effective number of codons

OR – olfactory receptor

ORF – open reading frame

SF – synonymous family

SGD – *Saccharomyces* Genome Database

tAI – tRNA adaptation index

TCA – tricarboxylic acid

Xg – silent GC content

3. Introduction

4. The genetic basis of phenotypic differences

A major challenge in comparative genomics is to understand how phenotypic differences among species are encoded within their genomes. The most obvious way, perhaps, that a difference in phenotype may be generated among species is through differential gene repertoires, such that only the genomes of species that possess the trait of interest contain the relevant genes. As an example, Avidor-Reiss et al. (Avidor-Reiss et al. 2004) identified a group of close to 200 genes that are present in the genomes of ciliated species but absent from the genomes of non-ciliated species. These authors suggested that members of this group of genes, which contained more than 80% of the genes previously implicated in the formation of cilia, are involved in cilia biogenesis and function.

Changes in phenotype may also be attained through mutations in the coding sequences of genes that alter their function. For instance, the long wavelength-sensitive opsin (LWS) protein of haplochromatic cichlid fish species from Victoria Lake differs in its peak value of light absorption at long (red) wavelength between populations, in a manner that is correlated with the transparency of the waters these populations inhabit (Terai et al. 2006). This difference in protein function was shown to be the result of differences in the binding site of this protein.

Differential regulation of transcript levels of shared genes with similar functions may also provide a basis for phenotypic divergence. The involvement of differential transcription regulation by transcription factors in phenotypic divergence has been widely studied (c.f. (Powers and Schulte 1998; Ihmels et al. 2005; Prud'homme et al. 2007; Simpson 2007)), with examples including the generation of various wing pigmentation and bristle patterns in *Drosophila* species through differential regulation of the *yellow* and *scute* genes, respectively (Simpson 2007). Other modes of gene transcript regulation that undergo selection, and, therefore, may potentially diverge to create different phenotypes include: chromatin modifications, as well as mRNA degradation, splicing, polyadenylation and localization. Despite the potential for involvement of these factors in phenotypic divergence, a literature search revealed only one example that found an association between changes in one of these factors (alternative splicing) and differences in phenotype. This is the case of differential alternative splicing patterns of the *hagoromo* gene that were found to be associated

with divergent patterns of coloration in cichlid fishes in the lakes of East Africa (Terai et al. 2003)).

Finally, phenotypic divergence may be generated through differential translation regulation, as well as through changes in protein localization, modification and degradation. For the former, no evidence for involvement in phenotypic variation was available at the time I began my study. Since then, a study associating between the translation efficiency of genes and species' lifestyles has been published, and will be addressed in the Discussion section. For the latter mechanisms of gene regulation there are still no known examples of phenotypic divergence that entails changes in them.

In this study I aimed to complement and expand the picture by examining the involvement of differences in coding sequences, which potentially affect protein expression, in species divergence. This work is divided into two chapters, each examining a different aspect of coding sequence evolution. In the first chapter, I aimed to investigate the relationship between differential translation efficiency of orthologous genes and phenotypic differences between yeast species. In the second chapter, a joint project with Dr. Yoav Gilad and Dr. Gustavo Glusman, we compared the olfactory receptor repertoires of human and chimpanzee that are assumed to have different sensory needs. In both cases I found evidence for the involvement of differences in coding sequences in the phenotypic divergence of the species analyzed.

5. Translation efficiency

The translation efficiency of a coding sequence is commonly defined as the extent of the adaptation of its codon usage to the tRNA cellular pools (Sharp and Li 1987; dos Reis et al. 2004), which serves as a surrogate measure for its speed of translation. This definition stems from an early observation of a trend of increasing codon usage bias with increasing gene expression levels in a sample of *E. coli* genes (Sharp and Li 1986), and that tRNA concentrations are rate limiting in the elongation of nascent peptides (Varenne et al. 1984). The same trend of codon usage bias was also observed in several other organisms, including: *S. cerevisiae* (Sharp et al. 1986), *C. elegans*, *D. melanogaster* and *A. thaliana* (Duret and Mouchiroud 1999). In the cases where data regarding the abundance of the various tRNA species were available it could be shown that the codons "preferred" by highly expressed genes were those

corresponding to the most abundant tRNAs (Sharp et al. 1986; Moriyama and Powell 1997).

The term translational selection refers to natural selection acting to maintain high translation efficiency of the coding sequences of certain genes. It is of note that translational selection is not the only factor shaping codon usage patterns in the genomes of species. Other factors that affect codon usage include directional mutational pressure and strand-specific mutational patterns (i.e. a difference in G versus C between the leading and lagging strands). Codon usage is typically shaped by several such factors, and some may be more dominant than others. Thus, the trend of increased translational efficiency in highly expressed genes was not found in the genomes of some organisms such as those of *M. luteus* (Ohama et al. 1990), and the spirochaetes *B. burgdorferi* and *T. pallidum* (Lafay et al. 1999), perhaps because factors other than translational selection were more dominant in shaping codon usage.

5.1. Measures of translation efficiency

Many measures have been proposed over the years to evaluate the codon usage of coding sequences in terms of translation efficiency. A representative set of these measures will be reviewed here. These measures can roughly be divided into two categories.

The measures in the first category measure the deviation of the codon usage of genes from the equal use of synonymous codons. If genes that are known, or assumed, to be highly expressed in the genome obtain scores that are associated with high bias then it is assumed that selection for translation efficiency is a major force shaping codon usage in the particular genome. One may then draw conclusions regarding the expression levels of other genes in the genome for which no prior information regarding their expression is available, based on their codon usage bias scores. A disadvantage of this methodology is that it requires knowledge of a set of highly expressed genes within the genome. If no expression studies have been performed on the genome in question then the set of highly expressed genes may be inferred by orthology. For this purpose, it is usually assumed that genes involved in protein translation, protein folding, and glycolysis are highly expressed (c.f. (Carbone et al. 2003)). However, it is not clear how widely applicable this assumption is.

The most commonly used index for codon bias in this first category is the effective number of codons (Nc) (Wright 1990). This index scores coding sequences according to their tendency for equal usage of synonymous codons, in a manner that is independent of the genome the sequence originates from. Theoretically, a sequence which uses synonymous codons indiscriminately would obtain a score of 61, indicating that it uses all 61 sense codons, whereas a highly biased coding sequence which uses a single codon for each amino acid would obtain a score of 20 (but see Methods section for deviations from these theoretical boundaries).

The measures in the second category measure the conformance of a sequence's codon usage to a 'translationally optimal' codon usage. Ideally, one would derive the translational optimality of codons from the abundances of the tRNAs in the cellular pool (taking into account wobble interactions). However, experimental data about the concentrations of the various tRNA types in the cell are available for very few species (Ikemura 1982; Kanaya et al. 1999). This drawback has been addressed in two ways. The first solution was to obtain a reference set of genes that are known to be highly expressed, and hence assumed to have optimal codon usage in terms of translation. All remaining genes belonging to the same genome are then scored according to the similarity of their codon usage to that of the reference set. A disadvantage of this solution is that, like the methods in the first category of translation indices, knowledge of a set of highly expressed genes is necessary. An alternative solution is to rely on a surrogate measure for the cellular abundances of tRNAs. It has been observed that the *in vivo* concentration of a tRNA bearing a certain anticodon is highly proportional (r=0.91 for *S. cerevisiae*) to the number of gene copies coding for this tRNA type (Percudani et al. 1997; Kanaya et al. 1999). This is in line with a recent study that showed that in *S. cerevisiae* the promoters of many of the tRNA genes have a low predicted affinity to the nucleosome, suggesting a constitutive expression with little transcriptional regulation capacity (Segal et al. 2006). Thus, for fully sequenced genomes, the relative concentrations of the various tRNAs in the cell, and therefore the optimality of the various codons in terms of translation, can be approximated using the respective tRNA gene copy numbers in the genome.

The fraction of optimal codons, F_{op} (Ikemura 1981), the earliest index of translation efficiency, is a member of this second category of translation efficiency measures. This index entails the determination of the most optimal codons for each amino acid (in a species-dependent manner), and then calculating the fraction of

codons in a coding sequence that are optimal (i.e. the values for this index range from 0 (no bias) to 1 (maximum bias)). This index results in an obvious loss of information, since for those amino acids that have more than two possible codons, the sub-optimal codons are not necessarily equal in terms of their translation efficiency.

The codon adaptation index, CAI (Sharp and Li 1987), a measure developed slightly later, is the most widely used index of translation efficiency. This index assigns weights to the various codons according to their frequency of occurrence in the coding sequences of a training set of highly expressed genes. The original training set used for CAI in the genome of *S. cerevisiae* constituted 24 genes, encoding 16 ribosomal proteins, one elongation factor and seven glycolytic enzymes. Coding sequences are then scored by combining the derived weights of their individual codons, resulting in values ranging from 0 (no bias) to 1 (maximum bias). More recently, a variation on the original CAI was proposed by Carbone et al. (Carbone et al. 2003). Under this variation, a reference set of genes that are most representative of the dominant codon bias in the genome is determined iteratively. If the resultant reference set contains genes involved translation, protein folding and glycolysis it is concluded that the genome analyzed is subject to translational selection. Codon weights and individual sequence scores are then computed as in the original CAI, using the reference set as a basis for calculations.

The tRNA adaptation index, tAI (dos Reis et al. 2004), is a relatively recent index of translation efficiency that uses the copy numbers of tRNA genes in the genome as a means to calculate the translation efficiency weights of the various codons, taking into account wobble interactions. The weights of the codons in a coding sequence are then combined in a manner that is identical to that of CAI to obtain a translation efficiency score for the sequence, which ranges from 0 (no bias) to 1 (maximum bias).

5.2. Measures of translation efficiency as predictors of gene expression levels

Several studies have examined the relationship between the codon usage bias values and the corresponding gene expression levels (Coghlan and Wolfe 2000; Jansen et al. 2003; Friberg et al. 2004; Goetz and Fuglsang 2005; Supek and Vlahovicek 2005). The expression values used in these studies were mainly from *E. coli* and *S. cerevisiae*, the two model organisms for which selection for translation

efficiency was first established and for which large-scale expression data exists. Since a large scale data set of protein expression levels in *S. cerevisiae* was published only in 2003, most of the studies used mRNA expression values or small-scale protein expression values obtained from 2D-gel experiments. The comparison of codon usage, which is assumed to be selected for translation efficiency, with mRNA levels was justified by the general correspondence of high mRNA levels with high protein levels. Overall, all studies found the correlation between codon usage bias and expression levels to be highly statistically significant. However, the values of the correlation coefficients are not high, and are similar to the correlations observed between mRNA and protein expression levels (Jansen et al. 2003). In fact, for lowly-expressed transcripts there is almost no relationship between mRNA levels and codon usage bias (Coghlan and Wolfe 2000). These studies therefore suggest the use of codon usage bias as a rule of thumb, rather than a highly reliable predictor of expression levels. It was also found that codon usage bias correlated better with mRNA expression levels measured in growth conditions matching the organism's natural habitat than with those measured under defined growth medium (Goetz and Fuglsang 2005; Supek and Vlahovicek 2005). A possible explanation for this observation is that the effect of codon usage, a static property of a gene, is negligible when mRNA levels are close to zero, and would therefore be strongest when a gene reaches its maximal mRNA level (Goetz and Fuglsang 2005).

Finally, measures of translation efficiency have been shown to be potentially useful in functional inferences. Two studies investigated the predicted expression levels of proteins, as approximated by their codon usage bias, in closely related yeast species of the genus *Saccharomyces* (Fraser et al. 2004) and in many other unicellular organisms (Lithwick and Margalit 2005). Both studies found that for functionally related protein pairs, the predicted protein expression levels tend to correlate across species for functionally related protein pairs, implying that the expression levels of these protein pairs co-evolve.

6. Ascomycotic fungi

The *Ascomycota*, also known as sac fungi, is a monophyletic group of species, which accounts for about 75% of described fungi. The ascomycotic fungi include some extensively-researched model organisms, such as the *Saccharomyces cerevisiae*

and *Schizosaccharomyces pombe*, as well as many yeasts noted for their involvement in disease (e.g. *Candida albicans*) or their utility in industry (e.g. *Yarrowia lipolytica*). Over the last ten years the genomes of a large number of members of *Ascomycota* have been sequenced, and many more are in the process of being sequenced. These sequencing projects have improved our understanding of the evolutionary relationships among members of this group, and revealed large scale genome dynamics in these organisms, such as a whole genome duplication (Wolfe and Shields 1997; Kellis et al. 2004). Comparative analyses among the sequenced genomes allowed the elucidation of regulatory motifs for extensively-researched model organisms (Cliften et al. 2003; Kellis et al. 2004), as well as the annotation of genes in the genomes of less researched organisms. In spite of this, by and large biological knowledge regarding phenotypes and their genetic basis in these sequenced species is still scarce.

In the first part of my research, I aspired to gain more knowledge regarding the molecular basis of phenotypes in ascomycotic species through the analysis of the translation efficiency of their genes. The research was inspired by a recent study, which showed that differences in the way hemiascomycotic species (a monophyletic subset of *Ascomycota*) prefer to metabolize glucose is connected to the regulation of transcription of the cytosolic vs. the mitochondrial ribosomal proteins (CRPs and MRPs, respectively) (Ihmels et al. 2005). My aim was to investigate whether differential translational efficiency of orthologous genes could also be related to phenotypic divergence of species. In particular I wanted to know whether patterns of differential translation efficiency could explain known physiological differences among related fungal species, as well as suggest new avenues of research of the species.

7. Phenotypic divergence between humans and chimpanzees

The chimpanzee (*Pan troglodytes*), together with the bonobo (*Pan paniscus*), is our closest extant evolutionary relative, and has diverged from humans ~5-7 million years ago (MYA) (Chen and Li 2001; Brunet et al. 2002). Chimpanzees share with us many similarities, some of them behavioral, such as tool use and group aggression (Goodall 1964; Whiten et al. 1999). However, in many respects humans are unique, relative to chimpanzees and other "great apes". These include habitual bipedalism, a

greatly enlarged brain and complex language (Whiten et al. 1999). Despite many physiological similarities between humans and chimpanzees, there are also important differences in the incidence and severity of several major human diseases (Olson and Varki 2003). As an example, humans are susceptible to malaria caused by *P. falciparum*, whereas chimpanzees are resistant to this form of malaria (Escalante et al. 1995; Ollomo et al. 1997). A comprehensive list of phenotypic traits which are confirmed or have been claimed to be different between humans and great apes (as a group) can be found in (Varki and Altheide 2005).

A comparison of the draft of the chimpanzee genome sequence with the human genome sequence revealed ~1% nucleotide substitutions between the two genomes, as well as that insertion and deletion events (indels) result in ~1.5% of the sequence in each species being lineage-specific ("The Chimpanzee Sequencing and Analysis Consortium" 2005).This low sequence divergence was already observed in earlier comparisons that used partial sequences, and has led to the formulation of several hypotheses regarding the genetic basis of the phenotypic differences among humans and chimpanzees. The earliest of these hypotheses was proposed in 1975 by Mary-Claire King and Alan Wilson (King and Wilson 1975): based on the paucity of protein sequence differences among the two species, these two researchers suggested that the phenotypic divergence among humans and chimpanzees was likely to be due to changes in the gene regulation of the two species. More recently, Olson formulated the 'less-is-more' hypothesis, which emphasizes the importance of loss-of-function mutations in the generation of human-specific phenotypes (Olson 1999). Two other genetic mechanisms of divergence that have been proposed to explain phenotypic differences between humans and chimpanzees are gene duplication and divergence, as well as non-synonymous substitutions in single-copy genes (Olson and Varki 2003). These hypotheses have been serving as a guide in the comparative analyses of the two species. However, studies conducted so far could not point to a single mechanism that is dominant in generating the phenotypic differences between humans and chimpanzees. Rather, all the above proposed mechanisms probably contributed to the divergence in physiologies: divergence in expression patterns (Gilad et al. 2006), loss-of-function mutations (e.g. (Puente et al. 2005; Wang et al. 2006)), non-synonymous mutations (e.g. (Bustamante et al. 2005; Bakewell et al. 2007)), and gene duplication (Demuth et al. 2006).

8. Olfactory receptors

Olfactory receptor (OR) genes provide the basis for the sense of smell, and with >1000 genes, are the largest gene superfamily in mammalian genomes (Buck and Axel 1991; Zhang and Firestein 2002). The completion of the human genome enabled the identification of the entire human OR gene repertoire (Glusman et al. 2001). At the time we began this study the number of OR genes identified in human was 862, 56% of them carrying one or more coding-region disruptions, and hence annotated as nonfunctional pseudogenes. In mouse, and presumably in other mammals as well, although these pseudogenes are transcribed to form full-length transcripts, they do not produce functional OR proteins (Serizawa et al. 2003). The mouse and dog total OR repertoires are roughly the same size, and ~20% larger than that of human (Young et al. 2002; Zhang and Firestein 2002; Quignon et al. 2003; Olender et al. 2004). Moreover, the proportion of pseudogenes in mouse and dog is only ~20% (Young et al. 2002; Quignon et al. 2003). Thus, the number of putatively functional OR genes is three times larger in mouse and dog relative to human. However, when only apparently intact (putatively functional) OR gene repertoires of the three species are contrasted, it appears that although humans have a sharply reduced functional OR repertoire, more than 60% (150 out of 250) of the different OR gene subfamilies are shared by all three species (Quignon et al. 2003; Godfrey et al. 2004; Olender et al. 2004). These observations led to the suggestion that humans sense a repertoire of odors that is comparable to that of dog and mouse, albeit with a diminished resolution (Godfrey et al. 2004; Malnic et al. 2004).

Prior to the current study, Gilad et al. (Gilad et al. 2003b) analyzed the coding sequences of 50 OR genes in five different primates and found that the human lineage accumulated OR pseudogenes almost four times more rapidly than any nonhuman primate lineage. As a result, apes and Old World monkeys have many fewer OR pseudogenes than humans. Nonetheless, the proportion of OR pseudogenes in these nonhuman primates is still significantly higher than those of dog or mouse (Rouquier et al. 2000; Gilad et al. 2003b). Taken together, the data suggest that a deterioration of the olfactory repertoire occurred during primate evolution, with a particularly steep decline in the human lineage.

The second part of my research, performed in collaboration with Dr. Yoav Gilad (then at Yale University, New Haven, Connecticut, USA) and Dr. Gustavo Glusman (Institute of Systems Biology, Seattle, Washington, USA), was inspired by the completion of the genome of the chimpanzee. Our primary aim was to characterize the chimpanzee OR gene repertoire in comparison to the human repertoire. Modern humans and chimpanzees lead vastly different lifestyles, and one can assume these lifestyles require different olfactory capacities. Such a comparison might uncover the genetic basis for the presumed differing capacities. More specifically, we aimed to provide a better estimate of the fraction of OR pseudogenes in chimpanzee (previous estimates were based on only 50 (Gilad et al. 2003b) and 20 (Gilad et al. 2003a) OR genes), as well as identify subfamilies that have expanded or have been reduced in the human or chimpanzee lineages since their divergence. In addition, we wanted to investigate which human OR genes evolved under positive selection, in order to highlight genetic differences that might underlie differing olfactory capacities among humans and chimpanzees. Finally, we aimed to provide an estimate of the point in time when humans started the more rapid accumulation of pseudogenes (relative to other apes).

9. Methods

1. Methods used for the comparative analysis of translation efficiency in yeast species

1.1. Species analyzed

For this study I used ascomycotic species whose genomes were completely assembled according to NCBI (http://www.ncbi.nlm.nih.gov/) at the time I began the study, and for which I could infer the tRNA gene repertoire reliably: *Saccharomyces cerevisiae*, *Candida glabrata*, *Ashbya gossypii*, *Kluyveromyces lactis*, *Debaryomyces hansenii*, *Yarrowia lipolytica* and *Schizosaccharomyces pombe*. I also included three additional species: *Candida albicans*, an important fungal pathogen, for which a high-quality gene collection (including tRNA genes) was available (Braun et al. 2005); *Saccharomyces bayanus*, a *Saccharomyces* sensu stricto species (i.e. a species that is closely related to *S. cerevisiae*) that diverged from *S. cerevisiae* ~20 MYA, for which the overwhelming majority of ORFs are available (Kellis et al. 2004); and *Aspergillus nidulans*, a filamentous fungus with a high quality sequence. Since most of the species used are yeasts, I collectively refer to them in the text as yeast species.

1.2. tRNA gene copy numbers

For all species except *C. albicans* and *S. bayanus* the tRNA gene copy numbers were obtained by applying the tRNAscan-SE software version 1.1 (Lowe and Eddy 1997), which uses a hidden Markov model (HMM)-based approach, to the genome sequences. For *A. nidulans* the genome sequence was obtained from http://www.broad.mit.edu/annotation/genome/aspergillus_nidulans; chromosome sequences for the remaining seven species were obtained from GenBank (http://www.ncbi.nlm.nih.gov/Genbank).

For *S. bayanus* I used the tRNA gene copy numbers of the closely related *S. cerevisiae*. Although a list of the tRNA genes for this species is available (Kellis et al. 2004), the low total number of protein-coding genes available for it and the other two sensu stricto *Saccharomyces* species sequenced in the same project (less than 5,000 in each of the three species, compared to close to 6,000 in *S. cerevisiae*), indicates that the quality of the genome sequence may not be high enough to reliably determine the copy numbers of tRNA genes. The strong conservation of synteny (i.e. the preserved

order of genes on chromosomes) between *S. bayanus* and *S. cerevisiae* (Kellis et al. 2004) and the relatively short time that has passed since their divergence (~20 million years) makes the use of the tRNA gene copy numbers from *S. cerevisiae* a conservative choice. For *C. albicans* I extracted the tRNA gene counts from the *Candida* Genome Database (CGD) (Arnaud et al.) on August 14 2005. See Appendix 1 for the data on gene copy numbers of tRNAs that decode sense codons in each of the species.

1.3. Protein and coding sequences

The *C. albicans* protein and coding sequences included in assembly Ca19 (Braun et al. 2005), were downloaded from http://candida.bri.nrc.ca on August 14 2005. This gene set corresponds to the haploid genome of *C. albicans*. *S. cerevisiae* and *S. bayanus* protein and coding sequences were downloaded from the *Saccharomyces* Genome Database (SGD) (Balakrishnan et al.) on June 16 2005. For *S. bayanus* several sequences may correspond to different fragments of the same ORF. I used the annotation of Kellis et al. (Kellis et al. 2004) to merge such fragments. Protein, gene sequences and gene structures from release 4 of the *A. nidulans* genome sequence (Galagan et al. 2005) were downloaded from http://www.broad.mit.edu/annotation/genome/aspergillus_nidulans. Using the gene sequences and the corresponding gene structures I obtained coding sequences for the *A. nidulans* genes.

For the remaining six species, files in both fasta and UniProt formats were downloaded from Integr8 (Pruess et al. 2005). The UniProt format files were used to construct a dictionary, linking accessions of nucleotide sequences to their corresponding proteins. I downloaded, in EMBL format, all entries of the EMBL database (Kanz et al. 2005) that corresponded to the species in question and contained a "CDS" feature. A perl script utilizing BioPerl (Stajich et al. 2002) was then used to go over the EMBL format file to extract coding sequences of accessions corresponding, according to the dictionary, to sequences in the protein fasta file. Coding sequences were used only if their length was at least three times the length of the protein sequence. If the coding sequence was longer than this length, I assumed that this was due to an alternative initiation site in the sequence, and used the last N nucleotides, where N is the expected length for the coding sequence. For *S. pombe*

there were many instances in which I couldn't obtain the coding sequences of genes using the above-mentioned method. For those genes that had orthologs in one of the other genomes examined (see below) I obtained the coding sequence manually from the *S. pombe* section of GeneDB (Hertz-Fowler et al. 2004).

The protein set of *C. neoformans*, which was used as an outgroup to establish orthology relationships (see below), was downloaded in fasta format from Integr8 (Pruess et al. 2005).

Finally, I removed mitochondrially-encoded sequences from all sequence sets.

1.4. Calculation of the tRNA adaptation index (tAI) for coding sequences

The tRNA adaptation index is described in detail in (dos Reis et al. 2004). Briefly, the method entails calculating a weight for each of the sense codons, derived from the copy numbers of all the tRNA types that recognize it (including wobble interactions). For a given coding sequence, the tAI value is then the geometric mean of the weights of all its sense codons (stop codons were ignored when encountered). The tAI of a coding sequence ranges from 0 to 1, with high values corresponding to high levels of translational efficiency. To calculate the tAI for coding sequences I used the codonR script supplied by (dos Reis et al. 2004), downloaded from http://people.cryst.bbk.ac.uk/~fdosr01/tAI/, which I modified to include the first codon, as well as other methionines.

1.5. Calculation of the codon adaptation index (CAI) for coding sequences

The codon adaptation index (CAI) is described in detail in (Sharp and Li 1987)). This index derives codon optimality weights from the coding sequences of a reference set of highly expressed genes. For a given coding sequence, the CAI value is then the geometric mean of the weights of all its individual codons. The CAI of a coding sequence ranges from 0 (low translation efficiency) to 1 (high translation efficiency). CAI values for all *S. cerevisiae* ORFs were calculated using the CodonW software (Peden JF, unpublished; http://www.molbiol.ox.ac.uk/cu), selecting "Saccharomyces cerevisiae Sharp and Cowe (1991) Yeast 7:657-678" as the CAI type.

1.6. Calculation of Nc and the pathologies of this index

The effective number of codons (Nc (Wright 1990)) is a measure of how far the codon usage of a coding sequence departs from equal usage of synonymous codons. In theory this index should yield values in the range 20 (for highly biased genes using only one codon per amino acid) to 61 (for genes that display equal usage of synonymous codons). However, in practice the equation for this index may yield values that are greater than 61. In order to understand how such 'illegal' values are obtained one must examine the definition of the index. For simplicity, the definition assumes a sequence coded by the 'universal' genetic code (although it can be easily modified to accommodate alternative codes). Under this code there are 2 amino acids with only one codon choice, 9 with two, 1 with three, 5 with four, and 3 with six. These represent five synonymous family (SF) types, designated as SF type 1, 2, 3, 4, and 6 according to the respective number of synonymous codon. For each amino acid we define its homozygosity (F) as follows:

$$\hat{F} = \frac{n\sum_{i=1}^{k} p_i^2 - 1}{n-1}, \quad (1)$$

where n is the total count for the amino acid within the gene, k is the SF type of the amino acid, and p_i are the frequencies of the codons coding for the amino acid ($p_1 + \ldots + p_k = 1$). The effective number of codons for the amino acid, which should in principle range from 1 to k, is then given by

$$\hat{N}_e = 1/\hat{F} \quad (2)$$

The effective number of codons for the whole sequence is then calculated using the following equation:

$$\hat{N}_c = 2 + \frac{9}{\hat{F}_2} + \frac{1}{\hat{F}_3} + \frac{5}{\hat{F}_4} + \frac{3}{\hat{F}_6} \quad (3)$$

where $\hat{F}_i$ is the average homozygosity for amino acids of SF type i, taken only over those amino acids that are present within the sequence and are abundant enough so that both the numerator and the denominator in equation (1) are greater than 0. A notable exception is isoleucine: if this amino acid is missing or not abundant enough, then $\hat{F}_3$ is computed as the average of $\hat{F}_2$ and $\hat{F}_4$. If any of the other SF types is completely missing or not abundant enough then $\hat{N}_c$ should not be computed.

In order to understand how values of $\hat{N}_c$ can exceed their theoretical bounds I examine the homozygosity and corresponding effective number of codons for a single amino acid belonging to SF type k (k>1).

If only one codon is used for the amino acid then, there exists 1≤i≤k, such that $p_i=1$, and for all j≠i $p_j=0$. Thus, we obtain $\hat{F}=\frac{n\cdot 1-1}{n-1}=1$, and therefore the effective number of codons for the amino acid will be $\hat{N}_e=1$, as expected.

On the other hand, if the synonymous codons coding for the amino acid are used indiscriminately, then for all 1≤i≤k $p_i=\frac{1}{k}$. Therefore, we obtain $\hat{F}=\frac{n\cdot k/k^2-1}{n-1}=\frac{n/k-1}{n-1}=\frac{n-k}{k\cdot(n-1)}<\frac{n-1}{k\cdot(n-1)}=\frac{1}{k}$ (assuming k>1), which leads to $\hat{N}_e>k$, thus exceeding the theoretical boundary. It is of note that $\hat{N}_e>k$ may be obtained also when synonymous codon usage is very close to uniform (but not exactly uniform).

Wright (Wright 1990) recognized that the values of $\hat{N}_c$ may exceed 61, but claimed that this would rarely occur, and would be due to a very extreme amino acid composition (where many amino acids are missing) or a very short gene. He suggested that in such cases the value of $\hat{N}_c$ should be revised to 61. However, in reality the tendency of the homozygosity to exceed 1/k when synonymous codon usage is uniform, or close to uniform, affects not only those sequences that obtain a value greater than 61 for $\hat{N}_c$. In fact, Marashi and Najafabadi (Marashi and Najafabadi 2004) claim that for an organism of intermediate GC content $\hat{F}<1/k$ in about 25% of cases. Therefore, rather than accepting Wright's suggestion of re-adjusting those values that exceed 61, I chose to use the values obtained from the original equation.

Nc was calculated with a modified version of the codonW program, supplied by (dos Reis et al. 2004), which avoids the re-adjustment of values exceeding 61. The program was downloaded from http://people.cryst.bbk.ac.uk/~fdosr01/tAI/. This version of codonW was further modified to accommodate the alternative yeast nuclear code used by *D. hansenii* (Tekaia et al. 2000) and *C. albicans* (Sugita and Nakase 1999).

1.7. Calculation of $f_1(Xg)$-Nc

$f_1(Xg)-Nc$ is a function describing the amount of codon bias that is not explained by Xg, the silent GC content, i.e. the GC content at the third codon position ($f_1(Xg)$ is the value of Nc that is expected based on Xg). The value of this function was calculated using a script downloaded from http://people.cryst.bbk.ac.uk/~fdosr01/tAI/.

1.8. Calculation of the significance of the observed correlation between $f_1(Xg)$-Nc (or Nc) and tAI

The significance of the observed correlation of $f_1(Xg)-Nc$ (or Nc) with tAI was calculated by permuting the tAI weights of the sense codons 1000 times. Each such permutation was then used to compute the correlation of $f_1(Xg)-Nc$ (or Nc) with the tAI values calculated using the randomized weights. The significance of the observed correlation was then calculated from the distribution of correlations obtained from the randomizations. All calculations were done using the R software for statistical computing (http://www.r-project.org).

1.9. Construction of the multi-species matrix of translation efficiencies

1.9.1. Generation of a table of orthologous groups

Using the inparanoid algorithm (Remm et al. 2001), I constructed two-species ortholog lists for every pair of species in my sample (excluding *A. gossypii*, which was discarded previously, see Results), using *C. neoformans*, a basidiomycotic fungus, as an outgroup. There is a discrepancy between the inparanoid algorithm, as reported by Remm et al. (Remm et al. 2001), and the programs supplied by the authors at http://inparanoid.cgb.ki.se/: while the paper specifies that the matched segment between two sequences must cover at least 50% of the longer sequence for the sequences to be considered homologous, the program applies this cutoff to the shorter sequence. In order to avoid domain-level matches, I modified the inparanoid program to reflect the algorithm as presented in the paper. I used the modified version of the inparanoid program without bootstrapping to generate the two-species groups of orthologs. The MultiParanoid program (Alexeyenko et al. 2006) was used to merge these two-species ortholog lists into one matrix. The order of species in the input to the program was as follows: *S. pombe*, *S. cerevisiae*, *A. nidulans*, *Y. lipolytica*, *C.*

albicans, *K. lactis*, *C. glabrata*, *D. hansenii*, and *S. bayanus*. With this order, two-species ortholog lists with large evolutionary distances between the relevant species, such as *S. pombe* and *A. nidulans*, were processed before ortholog lists of close species, such as *S. cerevisiae* and *C. glabrata*. The output of the MultiParanoid program was converted into a matrix of orthologs where each row corresponds to a gene and each column to a species. Note that if duplication had occurred after the divergence of *S. pombe* and *A. nidulans* from the remaining species, there would be more than one gene representing the same species in the same orthologous group (row). Since I assume that all genes in a single orthologous group have the same function, I will henceforth refer to them as representing a single gene, even when there is more than one representative per species. Finally, I retained only those orthologous groups (rows in the table) that had representatives from both *S. cerevisiae* and *S. pombe* (2883 out of 6226 rows).

1.9.2. Generation of a matrix of translational efficiencies across species

I combined the orthologous groups table with the tAI values computed for all ORFs of the nine species to create a matrix of translational efficiencies across species. In cases where the orthologous group contained several paralogs I used the maximal tAI among the representatives of the species, reasoning that this tAI would be the most adequate surrogate measure for the maximal levels of the activity expected from members of the orthologous group. Cells for which there were no representative of the corresponding species were left empty. Each row in the table will henceforth be referred to as a profile. This matrix of translational efficiencies was then submitted for preprocessing at the GEPAS (Montaner et al. 2006) server v3.0 (http://www.gepas.org): 73 profiles with more than 30% missing values (i.e. more than two missing values) were removed, missing values in the remaining 2810 profiles were imputed using the KNNimpute algorithm with k=15 (this was necessary for 839 profiles). Each column was then standardized so that its mean and standard deviation were 0 and 1, respectively. The same standardization was then applied to the rows of the matrix, emphasizing the efficiency of genes relative to their orthologous counterparts, rather than efficiency relative to genes in the same species.

1.10. Analysis of physically interacting pairs of proteins

Pairs of physically interacting proteins were obtained from the work of von Mering et al. (von Mering et al. 2002) by filtering out those protein pairs that were marked as "previously annotated: no". Thus, my set of physically interacting protein pairs was based upon a manually curated catalogue of proteins (Munich Information Center for Protein Sequences, MIPS (Mewes et al. 2002)) - the constituents of each pair within this set were both members in the same MIPS complex. The control set of non-interacting pairs was constructed by calculating all possible pairs using the proteins in the interacting pairs set, and then subtracting those pairs that are known to interact.

1.11. Gene Ontology (GO) data

The Gene Ontology (GO) database (Harris et al. 2004) was downloaded from http://www.geneontology.org on 23 August 2006. A file relating each *S. cerevisiae* ORF to GO terms (gene association file) was downloaded from SGD (Balakrishnan et al.) on the same date. The terms in the gene association file are the most specific descriptors of the ORF, and were expanded using the directed acyclic graph (DAG) structure of the GO database to generate for each ORF a comprehensive list of terms, containing both specific and general terms, that describe the ORF. Each GO term was considered to annotate any orthologous group (and corresponding translational efficiency profile) containing a *S. cerevisiae* gene that is associated with this term. If an orthologous group contained more than one *S. cerevisiae* gene then it was sufficient for one of these genes to be associated with the term in order to associate the whole orthologous group and the corresponding profile with the term. For the purpose of avoiding redundant statistical tests, I reduced the list of GO terms annotating my data to a non-redundant list, by arbitrarily choosing one representative term from each group of terms that annotate the exact same profiles in my data. Files listing all GO terms and their synonyms in my data (either when considering individual genes or orthologous groups) can be found under the Downloads section at http://longitude.weizmann.ac.il/pub/papers/Man2007_tai/suppl.

1.12. Intron data

Data regarding introns in the genes of *S. cerevisiae* was obtained from SGD (Balakrishnan et al.) on April 9 2007. Note that SGD contains only introns that are found between coding exons (as opposed to 5' UTR introns). For *S. pombe* and *A. nidulans* I computed the number of introns for a gene as the number of exons minus one. For *S. pombe* a file relating genes to exon numbers was downloaded from the *S. pombe* Genome Project site at the Sanger Institute (http://www.sanger.ac.uk/Projects/S_pombe/); for *A. nidulans* exon numbers were inferred from the gene structure file, which was downloaded from the *Aspergillus nidulans* Database at the Broad Institute (http://www.broad.mit.edu/annotation/genome/aspergillus_nidulans/Home.html).

1.13. Statistical analyses

1.13.1. Cluster analysis

Hierarchical clustering of the translational efficiency profiles was performed using the MATLAB/Math Works Inc. package. I used the Euclidean distance between normalized profiles as a distance measure and the average linkage algorithm for the construction of the hierarchical tree. The granularity of the clustering was chosen by eye.

1.13.2. Calculation of functional enrichment for clusters

In each cluster I checked for the enrichment of each of the non-redundant GO-terms, conditional on there being at least one gene within the cluster that was annotated with this term, and that the term annotates at least three genes in the whole dataset. Enrichment was assessed using the one-sided hypergeometric test. I corrected for multiple testing using the False Discovery Rate (FDR) method (Benjamini and Hochberg 1995) with an FDR of 5%, pooling together the results from all clusters. Due to the hierarchical structure of the gene ontology the tests in this analysis are nested, making it difficult to account for the multiple testing. I therefore complemented the analysis with an empirical evaluation of the significance of enrichment. For each cluster I noted the indices of the genes that constitute the cluster. I then permuted the genes in the whole dataset 10,000 times, using for each permutation the gene indices of the clusters to create new clusters with the same size

distribution and a random dispersion of the annotation among the clusters that conserves the hierarchy between the GO terms. For each permutation I counted the number of genes each of the tested terms annotates in each of the clusters. The empirical significance for the enrichment of a term T observed to annotate N genes in cluster C, was calculated as the number of permutations in which term T annotated at least N genes in cluster C, divided by 10,000. The Pearson correlation between the theoretical and empirical p-values, excluding cases where the empirical p-value was 0, was greater than 0.99.

1.13.3. Analysis of the species-effect on translation efficiency

I used the Friedman test (Rice 1995), a non-parametric analog of the two-way analysis of variance (ANOVA) without any replicates, to test for a difference of the median translation efficiency between species in a selected subset of the orthologous groups (Fig. 1). This test was applied both to the clusters obtained through hierarchical clustering, and to all sets of genes defined by a GO term that annotates all genes in the set, conditional on the set containing at least three genes. I corrected for multiple testing using the FDR method (Benjamini and Hochberg 1995) with an FDR of 5%, obtaining a significant species-effect on translational efficiency for 571 GO terms. As a control I repeated the Friedman tests for GO terms after randomizing the genes relative to the lists of GO terms associated with each gene. In contrast to the original assignment, in which there were 571 significant terms with an FDR of 5%, in the randomized case I found only 17 terms using the same FDR.

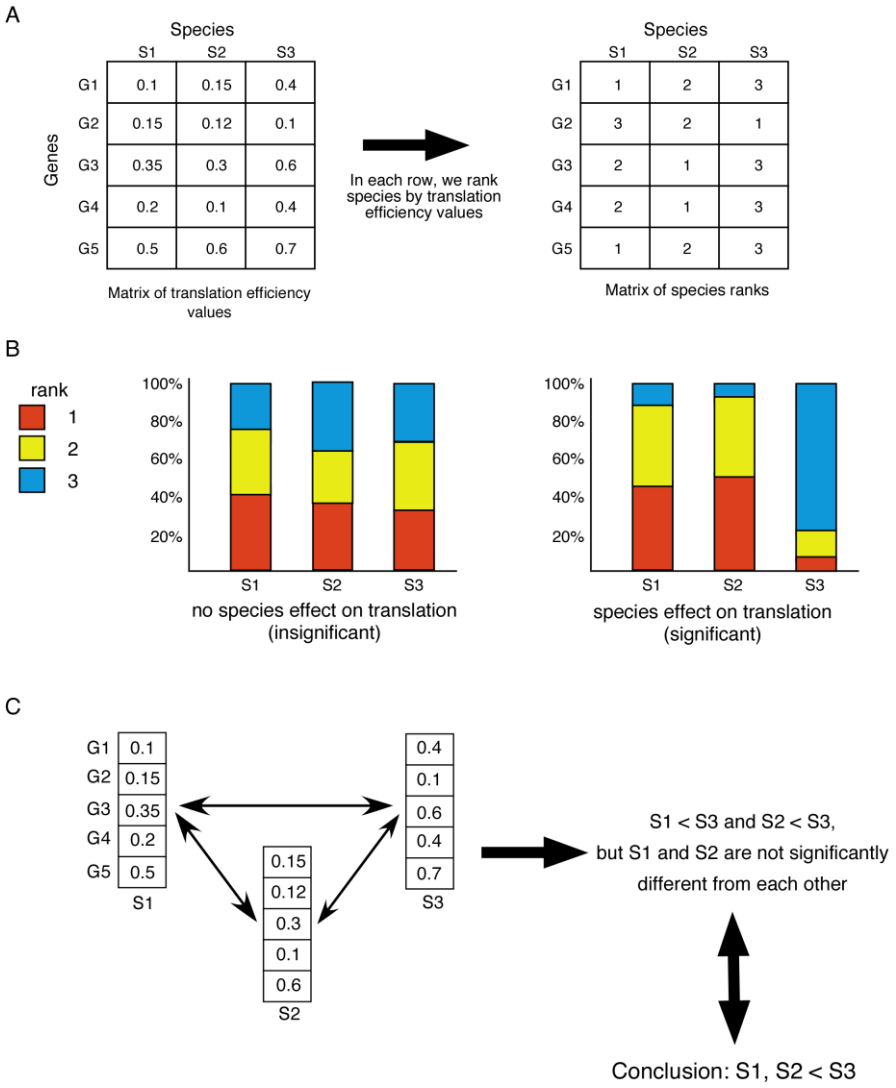


Fig. 1 Scheme for testing for a species effect on translation efficiency for a group of genes. The presented scheme can be applied to any group of genes (in my study I applied it to clusters of genes derived from the hierarchical clustering procedure, as well as groups of genes associated with various Gene Ontology (GO) (Harris et al. 2004) categories. A. The presence of a species effect on translation efficiency is tested using the Friedman test (Rice 1995). This test is based on the ranking of the values in each row; that is, for each gene, the species are ranked according to how efficiently they translate it (approximated by tAI). The matrix of ranks is then summarized and its significance is assessed. B. Stacked histograms of ranks representing two hypothetical gene sets: one in which there is no species effect on translation efficiency (the ranks are almost equally divided among the species; left) and one (corresponding to the matrix in A) where there is such an effect (the third species has an excess of high ranks; right). C. Once a species effect on translation has been established, an attempt is made to discover the source of the signal of differential efficiency using *post hoc* tests. For this purpose, the translation efficiency values for the group of genes in question are compared for all possible pairs of species. The conclusion that is drawn from these pairwise comparisons, after correction for multiple hypotheses testing, is presented as a species stratification.

1.13.4. Post-hoc tests

Each set of genes that was found to be statistically significant using the Friedman test (with an FDR of 5%) was further tested to find the source of difference in medians (Fig. 1). For this purpose I used the Wilcoxon signed-rank test (Rice 1995), applied to all pairwise comparisons among species (columns), using an FDR (Benjamini and Hochberg 1995) of 20% in the correction of multiple tests. Note that

for very small sets of genes the minimal p-value possible for the Wilcoxon signed-rank test exceeds the threshold set by the multiple testing procedure. Therefore, for such small sets, the comparisons that obtained the minimal p-value possible for the test, considering the size of the set, were considered to be statistically significant. For those pairwise comparisons that turned out to be statistically significant, I used the median values for the species in the relevant set of genes to determine the direction of the relationship. These directional pairwise relationships were then utilized to order the species according to their relative translational efficiency for the set of genes considered.

2. Methods used for the comparison of the human and chimpanzee olfactory receptor repertoire

2.1. Identification of chimpanzee OR genes

We used Gene-IT's Biofacet software (Gene-IT) to compare the chimpanzee genome draft (PCAP1026, NCBI Build 1.1, November 2003, http://www.ncbi.nlm.nih.gov/; R. Waterson, personal communication) to all human nucleotide sequences in HORDE v. 40 (http://bioportal.weizmann.ac.il/HORDE/), a database of OR gene sequences, with an expectation value cutoff of 0.00001. We selected all resulting alignments with a Smith-Waterman score >50 and over 70% identity and coalesced overlapping results, thus obtaining 1091 genomic segments. The most frequent length of these genomic segments was ~930 bp, corresponding to a complete OR gene (Pilpel and Lancet 1999). We generated a library of potential chimpanzee OR genes by extracting these genomic ranges, padding them with 200 bp in each direction (where possible) and masked repeats using RepeatMasker (http://repeatmasker.systemsbiology.net/). Using FASTX (Pearson et al. 1997), we compared each potential chimp OR gene to the intact protein sequences in HORDE v.40, with an expectation value cutoff of 0.01, and kept up to 10 results. We then used the protein match with highest identity to the query to reconstruct a conceptual translation for each chimpanzee OR gene. All chimpanzee OR sequences were submitted to the HORDE (http://bioportal.weizmann.ac.il/HORDE/) database.

2.2. Phylogenetic analysis

We selected those human OR genes that had a nucleotide sequence of at least 800 bp. Since the chimpanzee collection of ORs was more likely to contain fragments, we used an alternative criterion to select chimpanzee OR genes; if the conceptually translated nucleotide sequence was flanked on both side by untranslated sequence, then the conceptually translated region had to span at least 800 bp. Since the protein sequences of genes are better conserved than the nucleotide sequences, we chose a protein multiple-sequence alignment as a starting point for the phylogenetic analysis. We used ClustalX v1.83 (Chenna et al. 2003) in "Profile alignment" mode to align the conceptual translations of the selected OR genes against a template alignment – a previously published, manually curated, OR multiple sequence alignment that

contained representatives from all OR families (Man et al. 2004). An overlap of at least 70 amino acids in the alignment was selected as a criterion to determine whether two genes could be compared. We scanned the resultant alignment for pairs of sequences that did not meet this criterion. We then excluded a minimal number of sequences from our set of human and chimpanzee genes, so that all pairs of sequences had an overlap that is longer than the cutoff. The remaining sequences, 694 from chimpanzee and 762 from human, were aligned against the template alignment. We used seaview (Galtier et al. 1996) to correct any obvious errors in the alignment. We manually added the protein sequence of bovine rhodopsin to the alignment, according to a previously published alignment (Man et al. 2004). We then back-translated the resultant protein sequence alignment into a nucleotide sequence alignment, from which we computed a distance matrix using only overlap regions for each pair of sequences. We constructed a phylogenetic tree with the neighbor program from the PHYLIP (Phylogeny Inference Package) package v.3.62 (Felsenstein, J. 2005. *Distributed by the author. Department of Genome Sciences, University of Washington, Seattle*), using bovine rhodopsin as an outgroup. Trees were drawn using TreeExplorer (K. Tamura; http://evolgen/biol.metro-u.ac.jp/TE/TE_man.html).

2.3. Identification of human-chimpanzee orthologs

We generated the human-chimpanzee OR gene ortholog list using a novel statistical approach for comparing and ranking sequence alignments. This method (1) generates all pairwise alignments between human and chimpanzee OR protein sequences, (2) sorts the alignments by ranking higher those with statistically significant higher identity levels (statistical significance is determined using a chi-square test with categories being number of identical positions and number of non-identical positions), or in the case of statistical equivalence, preferring longer alignments, and (3) generates the list of potential orthologs by scanning the ranked alignments, accepting best-matching human-chimpanzee pairs, and discarding pairs involving previously assigned sequences.

2.4. Identification of shared and human-specific pseudogenes

Using conceptual translation, we identified all of the coding region disruptions in human OR pseudogenes that are present in the human-chimpanzee OR ortholog list. If an uninterrupted ORF was found, the gene was annotated as intact. If no ORF was identified, the gene was annotated as a pseudogene. We then performed a pairwise alignment of the conceptual protein sequence of each human pseudogene with its conceptually translated chimpanzee ortholog. A coding-region disruption was considered to be "shared" between the two species if the same codon carried the mutation (a stop codon, or a single base pair insertion/deletion within a codon). In all cases, we noted how many coding-region disruptions are shared versus human specific. If no shared disruptions were found, the locus was inferred to be a human-specific pseudogene.

2.5. Estimation of the time since human rapid accumulation of OR pseudogenes began

We assumed that the number of coding-region disruptions per locus is Poisson distributed (i.e., that disrupting mutations occur at a constant rate, are independent and infrequent). Let *n* be the number of genes with disruptions and *T* be the total number of observed disruptions in human-specific pseudogenes. We cannot directly observe the number of human OR genes that could have been disrupted (i.e. are under no constraint) but by chance were not. Instead, we observe all intact genes, a subset of which were not disrupted by chance and a subset of which are probably intact due to evolutionary constraint. Thus, we are missing information about the number of unconstrained loci with zero disruptions, X. Conditional on X loci with 0 disruptions, $\lambda = T/(X+n)$. In order to estimate X and λ jointly, we solved for the λ that minimized the sum of χ^2 deviations (across classes, for zero to infinity observations), setting $X = n \cdot e^{[-\lambda]}/(1-e^{[-\lambda]})$. To assess the error associated with our estimate of the mean, we performed the following bootstrapping procedure: we drew repeatedly from a Poisson distribution with mean $\hat{\lambda} = 0.451$ (the mean number of disruptions observed for the human-specific pseudogenes) until there were n (or more) non-zero observations, then estimated the sample mean. As our ~95% confidence interval, we took the central 95 percentile of the distribution of sample means across 10,000 replicates.

2.6. PAML analysis

We used the PAML package (Yang 1997), with substitution model (4; HKY85), in order to infer the sequence ancestral to human and chimpanzee for each OR gene in the ortholog trio list. We also used PAML to assess the likelihood of two models of protein evolution given our data. The null model (H0), allows one Dn/Ds (i.e. the ratio of nonsynonymous to synonymous divergence) parameter for the entire tree, while the alternative model (H1) permits a separate Dn/Ds ratio for each lineage. We use a likelihood ratio test (LRT; see below) (Rice 1995) to test the null model and a χ^2 distribution with two degrees of freedom to obtain p-values. Since sequence divergence between human and chimpanzee is only ~1% ("The Chimpanzee Sequencing and Analysis Consortium" 2005), in some cases there were no sequence differences in one or more substitution categories (synonymous or nonsynonymous substitutions) in one or more lineages. In these cases, we could not estimate meaningful Dn/Ds ratios for all the lineages, and we therefore excluded these loci from the analysis.

2.7. Likelihood ratio test (LRT)

The likelihood ratio test (LRT) (Rice 1995) is used to compare the goodness-of-fit between two models: a null model (H0) and an alternative model (H1). The LRT is calculated by first computing the likelihood scores of the two models, L0 and L1 for the null and alternative model, respectively. The LRT statistic is then calculated as: $\delta = 2(\ln L1 - \ln L0)$. When δ is larger than some predefined cutoff value (determined from the null distribution for δ) the null model H0 is rejected in favor of the alternative model H1. When the two models compared are nested within each other, i.e. H0 is a special case of H1, then the χ^2 distribution is a good approximation of the null distribution of δ, with the number of degrees of freedom being the difference in number of free parameters between the two models.

3. Results

1. Selection for translation efficiency and its relation to species divergence in yeasts

In comparing the translation efficiency among the genes of different ascomycotic species, I focused on one aspect of the translation process, namely elongation of the nascent peptide, with the efficiency of this process for a certain gene being gauged by its codon usage. For the purpose of the study, I selected ten ascomycotic fungal species whose genomes have been fully sequenced. My sample of species spans a wide range of evolutionary distances, with dates of divergence ranging from ~20 MYA between *S. cerevisiae* and *S. bayanus* to 350-1,000 MYA between *S. pombe* and the hemiascomycotic species (Berbee and Taylor 2001). As most of the species I used are yeasts, I will henceforth refer to them as yeast species.

The ideal choice of species for the analyses described in this work is far from trivial. On the one hand, too remote species may have too few shared orthologs. On the other hand, in very close species orthologs may present translational efficiencies that are close merely due to phylogenetic relatedness. This means that not all species contribute equally and non-independently. These two considerations were taken into account when choosing species for analysis. Thus, I excluded from analysis some *Saccharomyces* sensu stricto species that are closer to *S. cerevisiae* than is *S. bayanus* (c.f. *S. paradoxus*), and included in the analysis only ascomycotic fungi.

1.1. Slow evolution of tRNA repertoires in ascomycotic yeasts

The translation efficiency of a coding sequence is commonly gauged by the extent of its adaptation to the cellular tRNA pools (Sharp and Li 1987; dos Reis et al. 2004), which serves as a surrogate measure for the speed of elongation during translation. An investigation of the evolution of cellular tRNA repertoires is therefore pertinent to the study of translational selection among related species. Experimental data regarding the concentrations of the various tRNA types in the cell is available for only one species in my sample (*S. cerevisiae* (Ikemura 1982)). However, relying on the observation that the *in vivo* concentration of a tRNA type is highly proportional to the number of gene copies coding for it, facilitates the investigation of the tRNA pools of any species that has been fully sequenced, and in particular the species in my sample.

Using a HMM-based approach (Lowe and Eddy 1997) I was able to reliably obtain the tRNA gene copy numbers for nine of the ten species (Appendix 1). The total size of the repertoire (counting only tRNAs decoding sense codons) ranges from 133 (*C. albicans*) to 510 (*Y. lipolytica*) genes. Despite this wide range of repertoire sizes the distribution of gene copy numbers among the different anticodons tends to be highly correlated between pairs of species (Table 1). This extremely slow evolution of the tRNA repertoire among the species analyzed is expected since changes in dominance of the different codons encoding for the same amino acid may be highly pleiotropic, affecting the translation efficiency of many genes in the genome. It can be concluded that the codon usage of orthologous genes in these yeast species evolved against a very slowly evolving, or essentially invariant, tRNA pool. Interestingly, although all species pairwise correlations are statistically significant, the correlations between the tRNA repertoires of *A. nidulans* and *Y. lipolytica* and those of the hemiascomycotic species (all species in our sample except *A. nidulans* and *S. pombe* are hemiascomycotic) stand out as much lower than among the rest of the species pairs (Table 1). However, closer examination (Fig. 2) reveals that these low correlations are a result of a minority of outlying anticodons that represent cases of dominance shifts between synonymous anticodons. For example, while in *Y. lipolytica* (and in *A. nidulans*) the codon CAG for glutamine corresponds to the highly abundant tRNA, in *S. cerevisiae* (and in the rest of the analyzed species, except *A. gossypii*, which has equal numbers of tRNA genes for both glutamine anticodons), another codon, CAA, for this amino acid corresponds to the high copy number tRNA. Due to the pleiotropic effects of potential changes in the tRNA pool on cellular protein concentrations I expect that the unusual decoding seen in *Y. lipolytica* and *A. nidulans* arose very gradually, perhaps through an intermediate stage of redundancy, where the old and new major tRNAs co-existed in similar proportions during the time in which the coding sequences changed their codon preference. The identification of such intermediate redundant stages may require genome sequences of additional species, which are optimally distant from the currently available species. In that respect this process may bear similarity to the dominance switch process that was recently inferred to have occurred in the promoters of the ribosomal protein genes in yeast, that through a redundant intermediate stage have switched between two different transcription factor regimes (Tanay et al. 2005).

	Sc	Cg	Kl	Ag	Dh	Ca	Yl	An	Sp
Sc	-	0.97	0.98	0.87	0.95	0.9	0.58	0.66	0.8
Cg	7.78E-33	-	0.97	0.9	0.91	0.84	0.62	0.69	0.81
Kl	4.44E-37	2.31E-33	-	0.87	0.93	0.89	0.55	0.62	0.77
Ag	1.43E-17	3.43E-20	7.25E-18	-	0.79	0.73	0.68	0.77	0.81
Dh	2.29E-28	1.98E-21	1.40E-23	1.25E-12	-	0.88	0.56	0.64	0.77
Ca	6.65E-21	1.57E-15	2.38E-19	3.90E-10	3.61E-18	-	0.47	0.56	0.7
Yl	3.48E-06	5.91E-07	1.36E-05	1.21E-08	1.19E-05	3.33E-04	-	0.84	0.82
An	4.72E-08	1.09E-08	6.34E-07	9.84E-12	1.90E-07	1.26E-05	2.42E-15	-	0.93
Sp	5.51E-13	1.68E-13	1.27E-11	1.15E-13	1.33E-11	4.15E-09	5.43E-14	1.14E-23	-

Table 1. Correlations among the tRNA repertoires of the various species. The correlation coefficients were calculated using the MATLAB/Math Works Inc package, utilizing 54 tRNA species: all tRNAs decoding sense codons, excluding those tRNAs that are assumed to be absent in all living species. The upper triangle contains the Pearson correlation coefficients; the lower triangle contains the corresponding p-values. Species abbreviations: Sc – *S. cerevisiae*; Cg – *C. glabrata*; Kl – *K. lactis*; Ag – *A. gossypii*; Dh – *D. hansenii*; Ca – *C. albicans*; Yl – *Y. lipolytica*; An – *A. nidulans*; Sp – *S. pombe*. The correlations of *Y. lipolytica and A. nidulans* with the hemiascomycotic species (all species in our sample except *A. nidulans* and *S. pombe* are hemiascomycotic), which stand out as being much lower than the rest of the observed correlations, are highlighted in red.

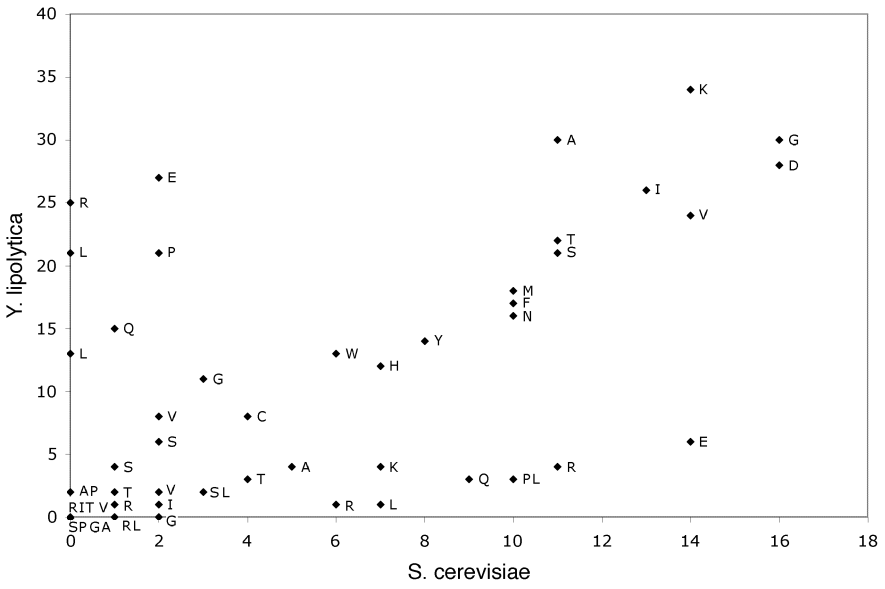


Fig. 2 Comparison of the tRNA gene repertoires of *S. cerevisiae and Y. lipolytica*. The gene copy numbers for each tRNA and each species were determined from the whole genome sequence using an HMM-based approach (Lowe and Eddy 1997) (see Methods). For each anticodon the gene copy number in *S. cerevisiae* (x-axis) and *Y. lipolytica* (y-axis) is displayed. The points are annotated with the one-letter symbol of the amino acid the anticodon translates. The Pearson correlation between the two tRNA gene repertoires is 0.58 (p=3.48e-06). The balance-swaps among anticodons translating glutamic acid (E), proline (P), glutamine (Q), arginine (R), and leucine (L) can be clearly seen. These switches in dominance of tRNA species are accompanied by corresponding changes in the usage of the codons they translate. For example, in *S. cerevisiae* there are nine genes encoding a tRNA bearing the anticodon of CAA (encoding glutamine – Q), and only one tRNA gene for the anticodon of CAG, the second codon for Q. Accordingly, *S. cerevisiae* uses CAA to encode Q 79,139 times (69% of the time) and CAG only 36,234 times. In *Y. lipolytica*, on the other hand, there are only three genes for tRNAs bearing the anticodon of CAA and 15 genes for the tRNAs bearing the anticodon of CAG. *Y. lipolytica* uses CAA to encode Q only 30,507 times (23% of the time), whereas CAG is used 100,228 times.

1.2. The relationship between translation efficiency and experimentally-determined protein and mRNA levels

While the tRNA pool largely evolves very slowly, it is possible that individual genes and gene modules evolved different extents of adaptation to that pool in different species. As a preliminary step towards the comparison of translational efficiency of genes among the species, I assessed the ability of translation elongation efficiency to predict protein levels, since such a comparison is pertinent to phenotypic divergence only if translation efficiency bears some relevance to protein levels. I chose the tRNA adaptation index (tAI) (dos Reis et al. 2004), an index which scores coding sequences based on the optimality of the codons they use, as a measure of translation efficiency of a gene. The individual codon scores are based on the availability of each of the tRNAs, as approximated by its gene copy number (available in Appendix 1), in a procedure that also incorporates codon-anticodon wobble interactions.

The relationship between translation efficiency and gene expression levels has been examined previously (Coghlan and Wolfe 2000; Jansen et al. 2003; Friberg et al. 2004; Goetz and Fuglsang 2005; Supek and Vlahovicek 2005). However, tAI was developed relatively recently for the purpose of inferring the presence of selection for translation efficiency in a genome (dos Reis et al. 2004), and has not been examined to date for its capacity to predict gene expression levels. Since my study largely relied on tAI as a surrogate measure for protein levels, I decided to re-examine the correlation between translation efficiency and protein expression levels using tAI. In the following, all quantities were log-transformed in order to achieve a distribution of values that is more normal in nature, thus allowing standard linear model analyses.

1.2.1. tAI as a predictor of *S. cerevisiae* protein and transcript levels measured in rich-medium conditions

Using experimentally-determined protein levels of almost 4000 open reading frames (ORFs) (Ghaemmaghami et al. 2003) I obtained a statistically significant positive correlation (Pearson $r=0.63$; $p<1e-363$) between tAI values and the corresponding protein levels (Fig. 3A). The same analysis using a different data set constituting 150 proteins (Greenbaum et al. 2002), yielded similar results (Fig. 3B). The correlation between the values of CAI (Sharp and Li 1987), a closely related

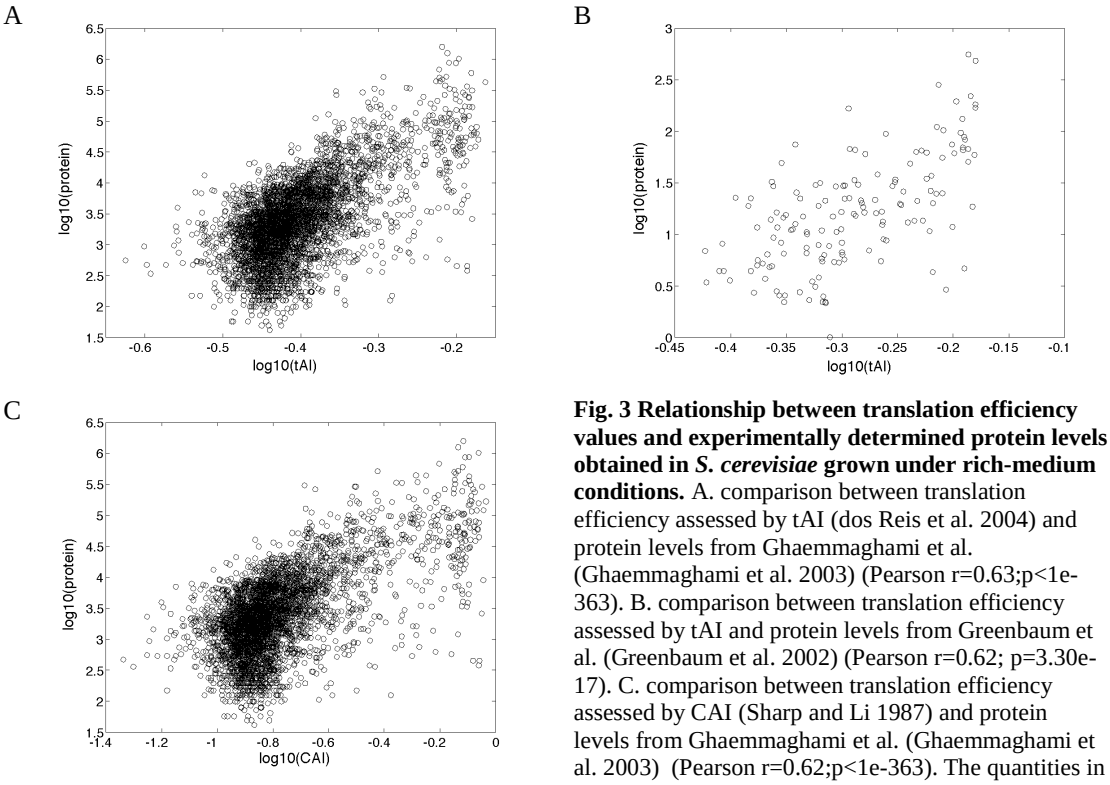


Fig. 3 Relationship between translation efficiency values and experimentally determined protein levels obtained in *S. cerevisiae* grown under rich-medium conditions. A. comparison between translation efficiency assessed by tAI (dos Reis et al. 2004) and protein levels from Ghaemmaghami et al. (Ghaemmaghami et al. 2003) (Pearson r=0.63;p<1e-363). B. comparison between translation efficiency assessed by tAI and protein levels from Greenbaum et al. (Greenbaum et al. 2002) (Pearson r=0.62; p=3.30e-17). C. comparison between translation efficiency assessed by CAI (Sharp and Li 1987) and protein levels from Ghaemmaghami et al. (Ghaemmaghami et al. 2003) (Pearson r=0.62;p<1e-363). The quantities in all panels have been log-transformed in order to achieve a distribution of values that is more normal in nature, thus allowing standard linear model analyses.

index of translation efficiency, and protein levels is comparable to the correlation between tAI values and protein levels (r=0.62;p<1e-363; Fig. 3C).

Comparison of genome-wide mRNA (Holstege et al. 1998) and protein (Ghaemmaghami et al. 2003) levels obtained under similar conditions yielded a statistically significant positive correlation (Pearson r=0.62;p<1e-363; Fig. 4A). However, as seen, despite a considerable correlation, similar mRNA levels still correspond to a wide range (up to 20-fold difference) of protein levels, and the same protein level may be obtained by transcripts that span a broad range of expression values (Gygi et al. 1999). A recent study analyzed experimentally the half-lives of thousands of *S. cerevisiae* proteins (Belle et al. 2006). The differential stability of the different proteins, demonstrated by the above study, may account for the only limited correlation, yet it is conceivable that different levels of translational efficiency may explain some of the discrepancy between mRNA and protein levels. Therefore, the tAI could potentially provide complementary information to mRNA levels when predicting protein levels. To examine this hypothesis, I computed a multiple linear

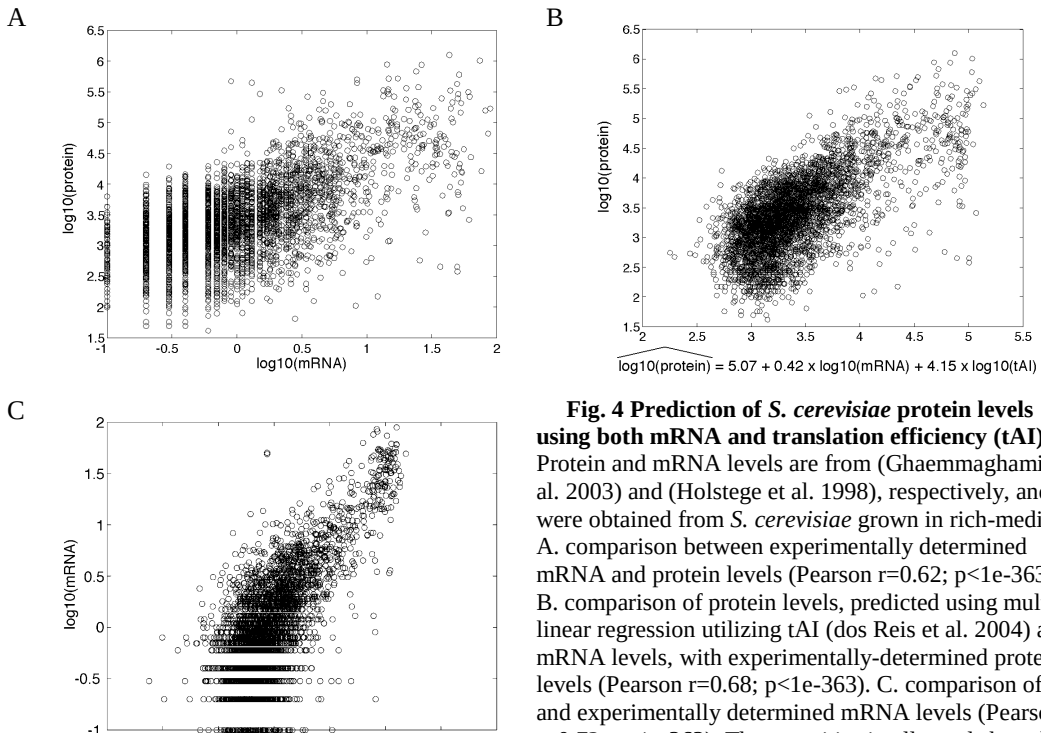


Fig. 4 Prediction of *S. cerevisiae* protein levels using both mRNA and translation efficiency (tAI). Protein and mRNA levels are from (Ghaemmaghami et al. 2003) and (Holstege et al. 1998), respectively, and were obtained from *S. cerevisiae* grown in rich-medium. A. comparison between experimentally determined mRNA and protein levels (Pearson r=0.62; p<1e-363). B. comparison of protein levels, predicted using multiple linear regression utilizing tAI (dos Reis et al. 2004) and mRNA levels, with experimentally-determined protein levels (Pearson r=0.68; p<1e-363). C. comparison of tAI and experimentally determined mRNA levels (Pearson r=0.72; p<1e-363). The quantities in all panels have been log-transformed in order to achieve a distribution of values that is more normal in nature, thus allowing standard linear model analyses.

regression model utilizing both tAI and mRNA levels (Holstege et al. 1998) to predict the protein levels (Ghaemmaghami et al. 2003) (Fig. 4B). The model's improvement over the individual predictors seems quite modest, with the Pearson correlation coefficient of the fitted values with the protein levels being 0.68 (p<1e-363). This is in line with the statistically significant positive correlation between tAI and mRNA levels (Fig. 4C; r=0.72; p<1e-363), a correlation that is higher than the correlation coefficients of each of these two separate factors with protein levels. However, computation of the partial correlations indicates that each of the individual variables makes a significant contribution: the partial correlation of tAI with the protein levels, at fixed mRNA levels is r=0.35 (p=4.14e-105); and a partial correlation of r=0.30 (p=1.21e-76) is seen for the mRNA and protein levels, given the tAI. I further demonstrated the significant relationship of tAI to protein levels by computing the correlation between tAI and protein levels for groups of genes having the same mRNA levels (Fig. 5A). Using only mRNA populations that correspond to at least 20 genes, I obtained 29 groups of genes, accounting for 3063 out of 3463 genes that have a value for all three measures (mRNA, protein and tAI). In all cases the correlation coefficients between tAI and protein levels are positive, and in 26 out of 29 groups

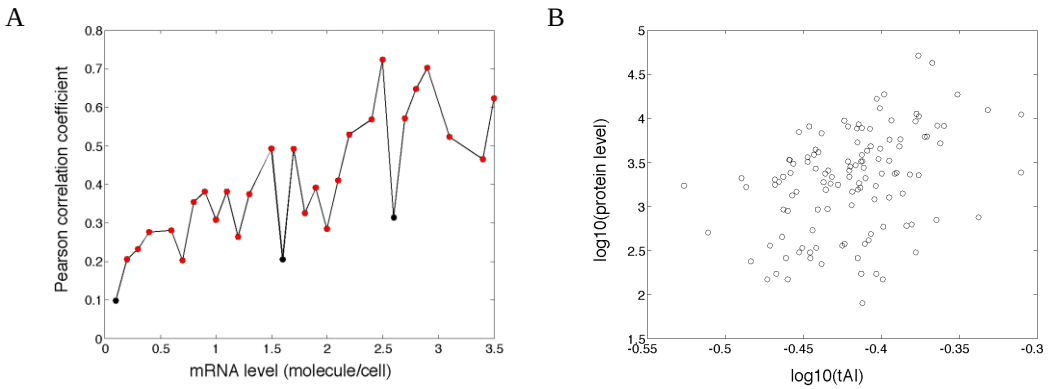


Fig. 5 Correlation of translation efficiency (tAI) with protein levels when mRNA levels are held constant. A. Pearson correlations between tAI (dos Reis et al. 2004) and *S. cerevisiae* protein levels obtained under rich medium conditions (Ghaemmaghami et al. 2003), using sets of genes where in each set all genes have the same mRNA level (Holstege et al. 1998). In the calculation of the correlations both tAI and protein levels were log-transformed in order to achieve a distribution of values that is more normal in nature, thus allowing a standard linear model analysis. Red points denote statistically significant correlations (p<=0.05). Only mRNA levels corresponding to at least 20 proteins were used. The points shown account for 3063 out of 3463 genes for which we have both protein and mRNA data. The set sizes are (mRNA levels are given in parentheses): 85 (0.1), 269 (0.2), 312 (0.3), 311 (0.4), 288 (0.6), 233 (0.7), 211 (0.8), 147 (0.9), 134 (1.0), 121 (1.1), 118 (1.2), 108 (1.3), 74 (1.5), 72 (1.6), 65 (1.7), 66 (1.8), 59 (1.9), 50 (2.0), 37 (2.1), 39 (2.2), 41 (2.2), 37 (2.5), 33 (2.6), 27 (2.7), 27 (2.8), 22 (2.9), 31 (3.1), 25 (3.4), 21 (3.5). B. Comparison of tAI and protein levels (Ghaemmaghami et al. 2003) for 121 genes in which mRNA levels (Holstege et al. 1998) equal 1.1 molecule/cell (Pearson r=0.38; p=1.57e-05).

they are statistically significant (p<=0.05). Interestingly this correlation tends to increase with mRNA levels. Fig. 5B shows a typical scatter-plot of tAI vs. protein levels for the population of mRNAs that are present on average in 1.1 molecules per cell. The fact that even genes with similar levels of mRNA display a positive correlation between tAI and protein levels means that the adaptation to the tRNA pool is not merely meant to facilitate the translation of a highly abundant mRNA that needs to produce a large amount of protein. Rather, this adaptation is a crucial component that given the mRNA level can produce significant additional modulation in determining protein levels.

1.2.2. tAI as a predictor of maximal potential protein levels

Examination of the scatter-plot of tAI vs. protein levels for *S. cerevisiae* in rich medium conditions (Fig. 3) revealed that although the correlation between these two variables is statistically significant, similar to the observation for mRNA levels (Gygi et al. 1999), tAI is not an accurate quantitative predictor for protein levels, e.g. proteins of the same tAI values may range in their protein levels by as much as 2,267 fold. A possible explanation for the inaccuracies in the predictions of the tAI is that, whereas protein and transcript levels vary across different conditions, the tAI, as a

measure derived from sequence information alone, is independent of conditions. Therefore, it is possible that tAI is an indicator of the maximal potential protein levels, rather than the protein levels at a specific condition. This hypothesis was also presented by Carbone and Madden (Carbone and Madden 2005), who supported it with two examples, the seripauperin gene family and the hem13 gene. In both examples transcript levels, measured under rich-medium conditions (Holstege et al. 1998), are lower than would be expected from the translation efficiency of the respective coding sequences, but the literature indicates that the proteins are highly translated under certain conditions.

To further examine the validity of the above hypothesis I capitalized on the many microarray experiments of recent years, as compared to few proteomic studies. The fact that high mRNA levels generally correspond to high protein levels, allows us to make a comparison of tAI with mRNA levels rather than protein levels. I examined 24 outlier genes from the scatter plot of the correlation between the tAI and mRNA levels, i.e. genes that exhibited relatively high tAI (tAI>0.5; log10(tAI)>-0.301), but transcript levels that are lower than would be expected, and looked for experiments where these outliers were induced. This analysis is obviously limited to the conditions covered by experiments published to date, and therefore doesn't necessarily cover all the conditions yeast cells may experience. However, despite this limitation, in the majority of cases (21 out of 24 cases) I could find a condition under which the ORF was at least two-fold induced, with the lowest maximal induction being 3.4-fold for YLR461W (PAU4), a member of the seripauperin family, during the unfolded protein response (Travers et al. 2000). In many cases I could find an experiment for which the product of the mRNA at log-phase in rich medium conditions (Holstege et al. 1998) and the fold-induction value was in line with the expected mRNA level. For example, YPL240C (HSP82), a cytoplasmic chaperone of the HSP90 family with a relatively high tAI value of 0.60, but very modest transcript levels under rich-medium conditions (Holstege et al. 1998), is induced 11.7-fold during a heat shock experiment from 21°C to 37°C (Gasch et al. 2000). Thus, available data support the hypothesis of tAI as an indicator of maximal protein levels under all possible conditions encountered by the cell.

1.3. tAI as a predictor of translational selection in a genome

The application of the tAI to the sequences of a genome is useful only if translational selection has played a significant part in shaping the codon usage of the genome. Thus, before selecting species for a multi-species analysis I checked whether translational selection can be detected in their genome.

If translational selection were the main force shaping codon usage, sequences showing high bias in their codon usage would typically be those that were selected for optimal translation efficiency. However, codon usage is largely affected by the silent GC content (Xg), i.e. the percentage of codons that have guanine or cytosine at their third nucleotide position. dos Reis et al. (dos Reis et al. 2004) have suggested testing for the presence of translational selection in a genome by assessing the correlation between the extent of adaptation of a coding sequence to the tRNA pool and the bias in codon usage that is unaccounted for by silent GC content. More specifically, they suggested testing for the correlation between tAI and f(Xg)-Nc, where Nc is the observed effective number of codons (Wright 1990) , and f(Xg) is a function predicting the effective number of codons based solely on Xg. Since f(Xg) represents a lower-bound on the amount of bias a sequence can display, and the amount of bias is negatively correlated with Nc (and f(Xg)), it is expected that for most sequences f(Xg)-Nc would be non- negative, and that greater differences would be observed when f(Xg) accounts poorly for the observed bias. And, since higher values of tAI are predictive of greater extent of adaptation to the tRNA pool (as approximated by the tRNA gene copy numbers), a strong positive correlation between f(Xg)-Nc and tAI would indicate co-adaptation between codon usage and the tRNA pool. I applied this test to the ten yeast species in my sample and found statistically significant correlations between f(Xg)-Nc and tAI, thus concluding translational selection to be present in all of them (Table 2; Fig. 6 and Appendix 2). In spite of this, it may be that while translational selection shaped the residual codon bias in coding sequences left after accounting for the effect of silent GC, mutation pressure has been so strong that the effect of translational selection on the overall codon bias in the sequence might be minute. In such a case, changes in protein levels may be achieved in different ways, for example by raising the levels of transcript. This suggests that the test by do Reis et al. may not be appropriate for testing the extent of the effect of translational selection

species	correlation of tAI with f_1(Xg)-Nc	significance	correlation of tAI with Nc	significance
A. gossypii	0.60	<0.001	-0.38	0.384
A. nidulans	0.58	<0.001	-0.68	<0.001
C. albicans	0.63	0.003	-0.65	0.005
C. glabrata	0.86	<0.001	-0.79	<0.001
D. hansenii	0.78	<0.001	-0.75	<0.001
K. lactis	0.85	<0.001	-0.83	<0.001
S. bayanus	0.81	<0.001	-0.73	<0.001
S. cerevisiae	0.81	<0.001	-0.79	<0.001
S. pombe	0.83	<0.001	-0.66	<0.001
Y. lipolytica	0.83	<0.001	-0.84	<0.001

Table 2. Correlation of translation efficiency (tAI) with overall codon bias for the ten yeast species analyzed. Pearson correlations and their significance, computed over all nuclear-encoded coding sequences, are shown for the comparison of tAI with the overall codon bias after accounting for the effect of silent GC content (f_1(Xg)-Nc) and for the comparison of tAI with overall codon bias (Nc). The significance of the correlations was computed by comparing them to 1000 correlations obtained using tAI values that were computed from randomized codon weights (see Methods).

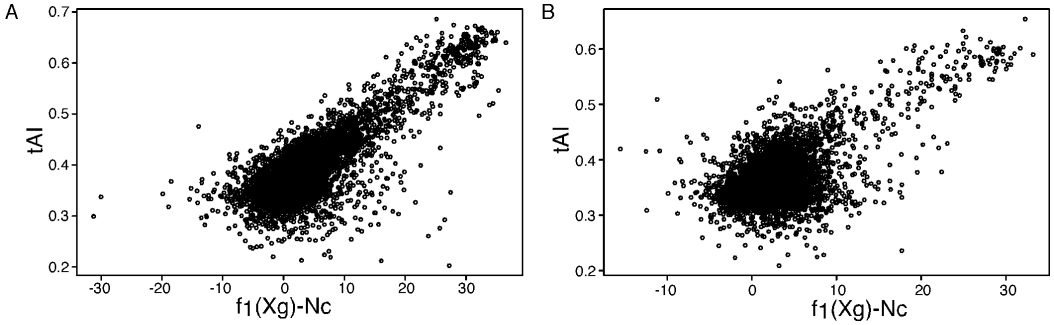


Fig. 6 tAI vs. f1(Xg)-Nc for *S. cerevisiae* (A) and *A. gossypii* (B)

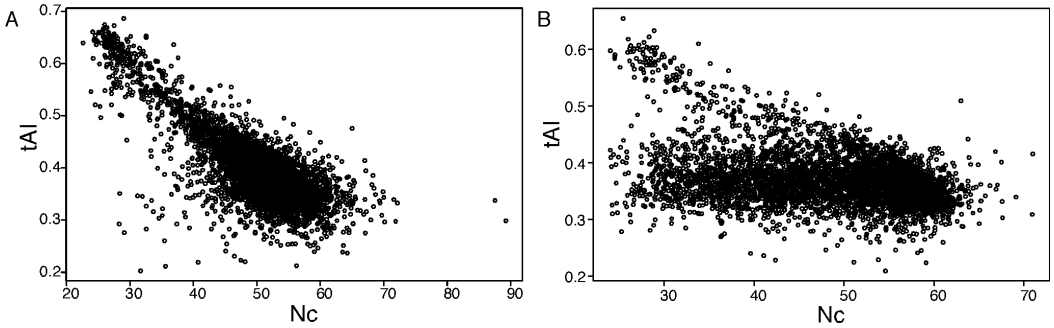


Fig. 7 tAI vs. Nc for *S. cerevisiae* (A) and *A. gossypii* (B)

in shaping codon usage, and as a consequence on expression. Therefore, to test the contribution of translational selection to overall codon usage, I tested the correlation between tAI and Nc. This time I expected strong negative correlation if codon usage is highly adapted to the cellular tRNA pools. I found that for nine of the species this correlation was statistically significant (Table 2; Fig. 7A and Appendix 3). However, for *A. gossypii* the magnitude of correlation was low and insignificant (Table 2; Fig. 7B),

suggesting that for this species tAI would not be a good predictor of expression levels. I therefore excluded *A. gossypii* from all subsequent analysis that involved multiple species.

1.4. Relationship between the translation efficiencies of physically interacting proteins

To validate the use of the tAI (dos Reis et al. 2004) for functional inferences I examined the relationship between the translational efficiencies of physically interacting pairs of proteins, as designated by tAI. First, I compared the translational efficiencies of the physically interacting proteins within a single species. For this purpose I used a manually-curated set of about 2,300 experimentally-determined physically interacting protein pairs obtained from the work of von Mering et al. (von Mering et al. 2002). I expected that physically interacting pairs of proteins would be found in similar quantities *in vivo*. This expectation is supported by the fact that for protein levels measured in *S. cerevisiae* in rich-medium (Ghaemmaghami et al. 2003), ~85% of physically interacting pairs show protein levels of the same order of magnitude, and overall the fold-differences within physically interacting pairs are significantly smaller than those found within a control set of non-interacting pairs ($p=2.16e-42$; Wilcoxon-Cox rank sum test). I then compared the distribution of squared tAI differences of the interacting pairs with the differences obtained from a control set of non-interacting pairs. I found the squared differences among the physically interacting pairs to be significantly smaller than those found in the set of non-interacting pairs ($p<1e-174$; Wilcoxon-Cox rank sum test; Fig. 8A). Thus, *S. cerevisiae* physically interacting pairs of proteins show both similar protein levels *in vivo* and similar translation efficiencies.

I next examined the behavior of translation efficiencies of physically interacting protein pairs across the nine species that were found to be under translational selection. For this purpose I constructed a matrix of over 2,800 orthologous groups that are inferred to have been present in their last common ancestor. In this matrix the i,j^{th} element contains the inferred translation efficiency of gene i in species j, as calculated by the tAI (dos Reis et al. 2004) (subject to further normalization, see Methods). It is expected that the levels of physically interacting pairs of proteins would vary across species and conditions in a coordinated manner. In this respect the

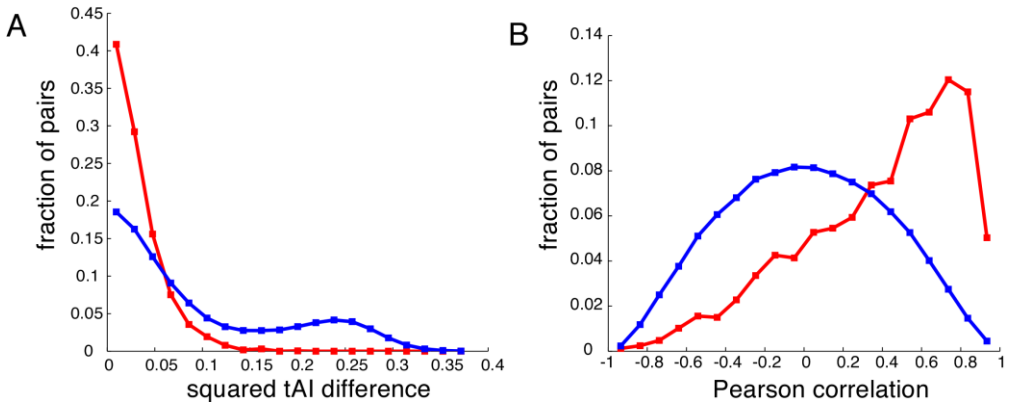


Fig. 8 Physically interacting proteins tend to have similar translational efficiencies across species. A. Histogram of squared tAI differences among physically interacting protein pairs (von Mering et al. 2002) (red) and among non-interacting pairs of proteins (blue; see Methods). B. Histogram of Pearson correlations among the across-species translation-efficiency profiles of physically interacting protein pairs (von Mering et al. 2002) (red) and among non-interacting pairs of proteins (blue; see Methods).

prediction of protein levels from coding sequences seems problematic, since these predictions are condition-independent. Yet, recent studies (Fraser et al. 2004; Lithwick and Margalit 2005), using CAI (Sharp and Li 1987) as a predictor of protein expression levels, showed that profiles of predicted expression levels across species tended to be correlated for functionally interacting protein pairs. I filtered out from the data of Von Mering et al. (von Mering et al. 2002) pairs of paralogs that belong to the same orthologous group (i.e. are a result of a duplication that occurred after the divergence of the analyzed species from each other), leaving me with about 1,700 pairs of physically interacting proteins. I then compared the Pearson correlation among the profiles (rows in the matrix of translation efficiencies) corresponding to the interacting pairs with correlations obtained for a control set of non-interacting pairs. The correlations among the profiles of physically interacting pairs were significantly higher than those found for non-interacting pairs ($p<1e-100$; Wilcoxon-Cox rank sum test; Fig. 8B). Protein pairs that interact in *S. cerevisiae* and that have similar tAI in this species but not in some of the more remote species may represent cases of physical interactions that evolved after the divergence of the species.

1.5. Translation efficiency across species correlates with the glucose repression phenotype

Given the comparative translation efficiency matrix I could examine relationships between gene functions and lifestyle properties of the different yeasts. A recent analysis established among yeast species a connection between the presence vs.

absence of the glucose repression phenotype and transcription regulation of the cytosolic vs. the mitochondrial ribosomal proteins (CRPs and MRPs, respectively) (Ihmels et al. 2005). Glucose repression is the preference of metabolizing glucose through fermentation rather than respiration even under aerobic conditions (Barnett and Entian 2005). One of the groups of genes repressed by glucose under this phenotype is the MRPs (Barnett and Entian 2005). Ihmels et al. (Ihmels et al. 2005) found that yeasts that do not display the glucose repression phenotype maintained the capacity to co-regulate the two types of ribosomes, while the yeasts that prefer fermentation over respiration have lost this capacity. This motivated me to examine the translation efficiency of the MRPs and CRPs in the species in my dataset, in which four species display the glucose repression phenotype. I found that each of these two groups of genes shows a strikingly coherent, yet markedly different pattern of relative translational efficiencies across the species (Fig. 9A and 9B). Interestingly, the MRPs show the lowest translational efficiency in the four species that display glucose repression (Fig. 9A). This likely reflects the reduced need of these species for MRP genes, which function to synthesize components utilized in oxidative energy metabolism. In addition to the ribosomal components, I checked whether translation efficiency of metabolic enzymes that are needed either for fermentation or respiration segregate according to the glucose repression phenotype. Specifically, I examined the tricarboxylic acid (TCA) cycle and the glycolytic genes (Fig. 9C and 9D, respectively). As expected, I found the glycolytic enzymes to have maximal translation efficiency in the four species that exhibit the glucose repression phenotype, while the TCA cycle genes show the highest translation efficiency in the five other species. These results cannot be simply explained by the species phylogeny, since in the species tree neither group of species is monophyletic (Fig. 9E). Specifically, *S. pombe* and the other three species that display glucose repression seem to have converged upon similar translation efficiency profiles of the above genes, despite the fact that *S. pombe* (together with *A. nidulans*) is farthest away from these species (Fig. 9A-D).

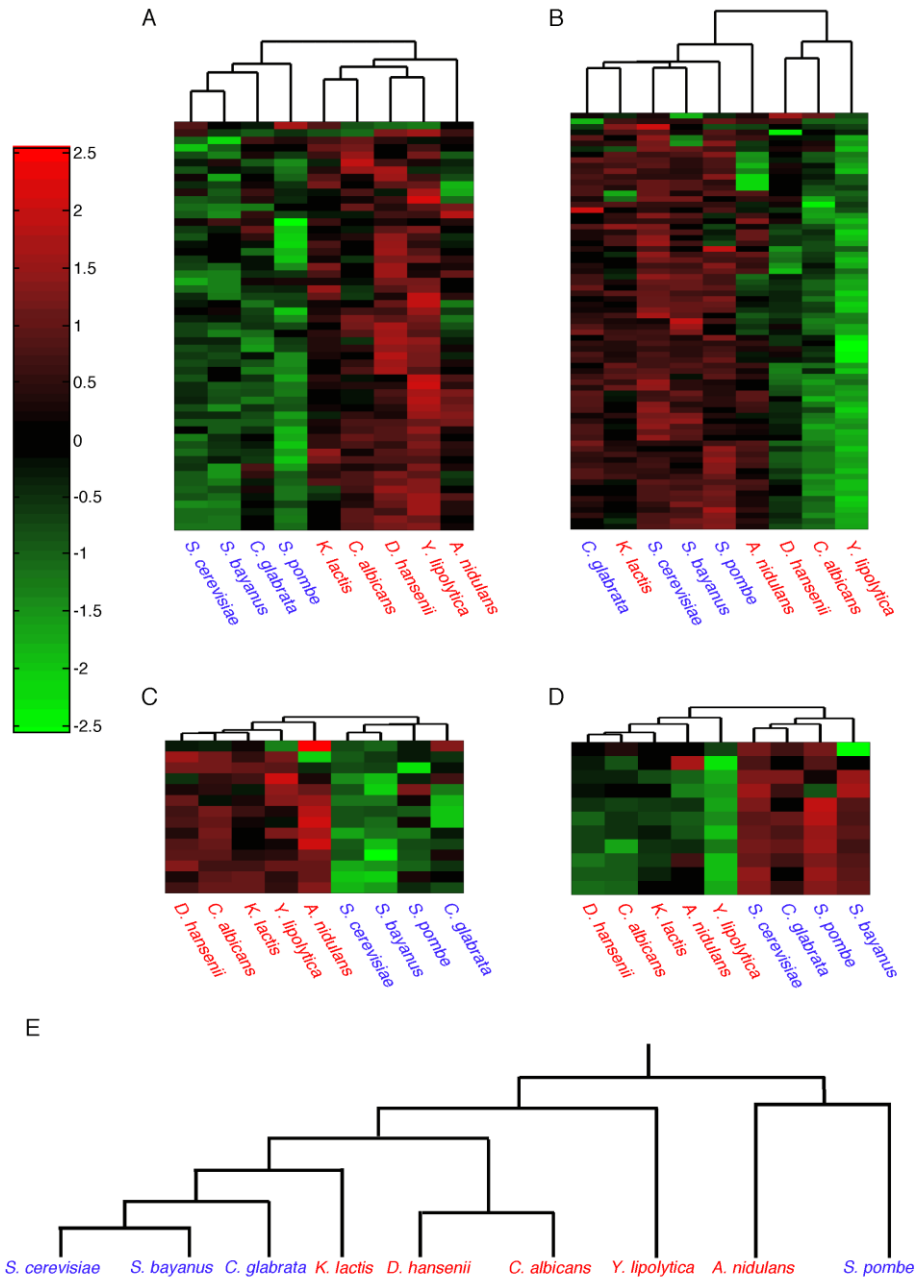


Fig. 9 The translation efficiency profiles of mitochondrial and cytosolic ribosomal proteins, glycolysis, and tricarboxylic acid cycle display coherent patterns. The relative translational efficiencies across species (normalized tAI; see Methods) of the mitochondrial ribosomal proteins (MRPs, A), cytosolic ribosomal proteins (CRPs, B), tricarboxylic acid (TCA) cycle (C), and glycolysis genes (D) are shown on heat maps. A colorbar specifies the values indicated by the colors. Note that for panels B to D the range of values is only from -2 to 2. Both rows (genes) and columns (species) have been sorted by average linkage hierarchical clustering, with Euclidean distance as a distance measure. The clustering of the species is indicated by dendrograms. E. Topology of the phylogenetic tree of the species analyzed based on 18S rRNA (Prillinger et al. 2002).; the branch lengths are not proportional to time. In all panels species displaying the glucose repression phenotype are colored in blue, whereas the other five species are colored in red. It can be seen that the translation efficiency profiles of MRPs, TCA cycle and glycolysis genes segregate according to the glucose repression phenotype.

1.6. Cluster analysis of translation efficiency across species

I next clustered all the genes in the multi-species translation efficiency matrix. Clustering is commonly used in the analysis of microarray data to identify co-expressed genes and generate hypotheses as to the involvement of these genes in the conditions or species examined (Boutros and Okey 2005). Analogously, I used hierarchical clustering of the genes and species in the translation efficiency matrix and partitioned them into 40 clusters (Fig. 10; see Methods). It is of note that the species dendrogram resulting from the clustering procedure bears only limited resemblance to the underlying species tree (Fig. 10A vs. Fig. 9E), suggesting that the evolution of the tAI values reflects more than mere evolutionary drift.

I next turned to analyze the 40 clusters of genes that were obtained. I used the Friedman test (Rice 1995) (Fig. 1), a non-parametric analog of the two-way analysis of variance (ANOVA) without replicates, to test, in each cluster, the null hypothesis that there are no differences in the translational efficiency (tAI) of orthologous genes (rows) across species (columns), and found that in all 40 clusters there is a species-effect on relative translational efficiencies (the worst p-value among the 40 clusters was 7e-05; Appendix 4). I then performed post-hoc tests (Fig. 1), utilizing the Wilcoxon signed rank test (Rice 1995), to find all pairs of species for which the genes of the cluster differ significantly in their translational efficiencies. In all clusters I was able to stratify the species into at least two groups that differ in translational efficiency, based on the results of these tests and the median values for the genes in the cluster (Appendix 4).

In an effort to shed light on phenotypic differences that might be implied by the species stratification in each cluster, I looked for enrichments of functional terms from the Gene Ontology (GO) database (Harris et al. 2004) in each of the clusters. I found such enrichments in 22 of the clusters, including the pathways and modules shown in Fig. 9 (a complete list of the functional enrichments can be found at http://longitude.weizmann.ac.il/pub/papers/Man2007_tai/suppl under the "Supplementary Tables" section). Fig. 11 shows representative clusters, along with dendrograms depicting similarity between species using the tAI of the genes in each cluster. Here, too, it is apparent that using particular gene sets for species clustering results in significant distortions relative to the phylogenetic species tree.

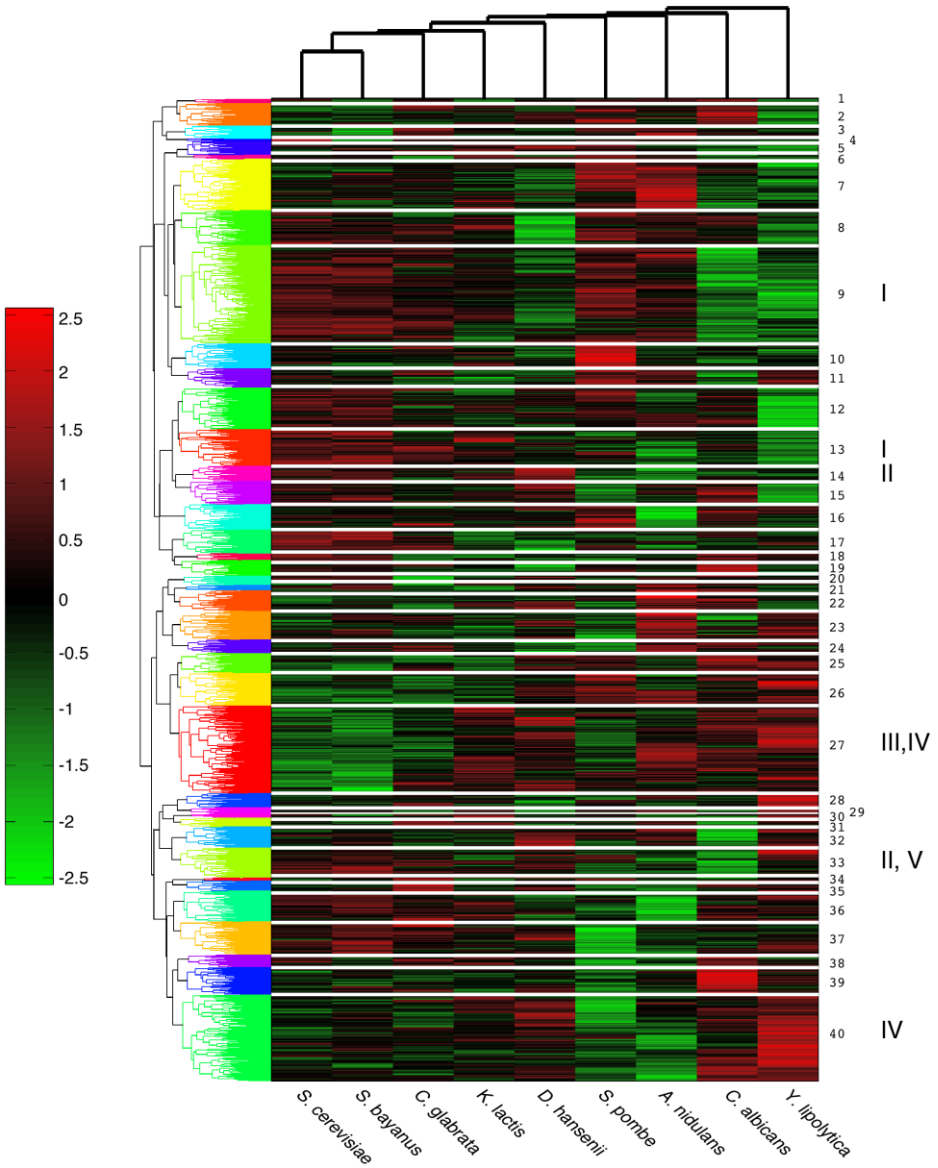


Fig. 10 Two-way hierarchical clustering of the multi-species translation efficiency profiles into 40 clusters. The 2810 translation efficiency profiles are shown with genes and species ordered according to the clustering. The clustering of the species, ordered by their translation efficiency profiles across genes, differs from the accepted phylogeny for these species (Prillinger et al. 2002) (Fig. 9E). The numbering of the clusters is indicated. I note some highly significant functional enrichments (p-values result from one-sided hypergeometric tests). I. cytosolic ribosome (sensu Eukaryota): 47/96 (p=2.19e-23) of the genes annotated with this term are contained in cluster #9 and 18/96 (p=5.86e-09) genes are contained in cluster #13. II. ribosome biogenesis: 14/186 (p=2.73e-07) genes are contained in cluster #14 and 21/186 (p=9.17e-08) are contained in cluster #33. III. mitochondrial part: 97/286 (p=8.99e-38) genes are contained in cluster #27; oxidative phosphorylation: 20/24 (p=3.93e-18) genes are contained in cluster #27; aerobic respiration: 23/50 (p=3.42e-12) genes are contained in cluster #27; tricarboxylic acid cycle: 10/14 (p=1.98e-08) genes are contained in cluster #27. IV. mitochondrial ribosome: 29/55 (p=3.57e17) genes are contained in cluster #27 and 18/55 (p=3.71e-07) genes are contained in cluster #40. V. rRNA processing: 20/146 (p=6.87e-09) genes are contained in cluster #33 and 22/146 (p=1.18e-09) genes are contained in cluster #37; nucleolus: 21/181 (p=5.64e-08) genes are contained in cluster #33 and 24/181 (p=2.67e-09) genes are contained in cluster #37; RNA metabolism: 29/345 (p=1.51e-07) genes are contained in cluster #33 and 37/345 (p=1.95e-11) are contained in cluster #37.

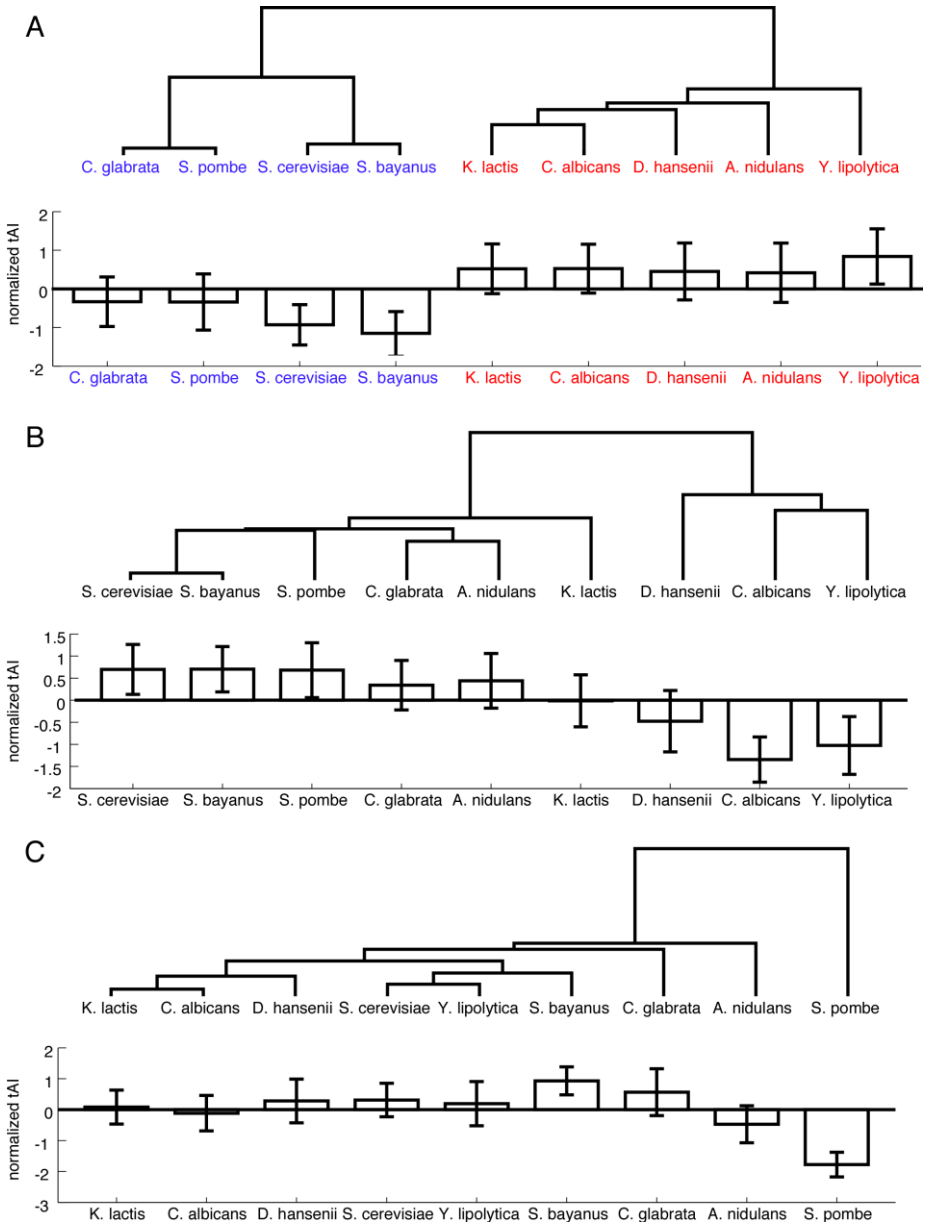


Fig. 11 Clustering of the yeast species according to their translation efficiency in a number of clusters. Dendrograms and bar plots of the translation efficiency profiles of three of the clusters of Fig. 10 are shown: A. #27 B. #9 and C. #37. The dendrograms represent the clustering of the species according to the translation efficiency values of the genes in the various clusters. The bar plot shows the mean translation efficiency value for the genes in the cluster for each species, with error bars indicating the standard deviation. The order of the species in the bar plots is according to the relevant dendrogram. In A (cluster #27) species displaying the glucose repression phenotype are colored in blue, whereas those that do not display this phenotype are colored in red.

1.7. Supervised analysis of translation efficiency across species

Despite experimenting with various numbers of clusters, I did not identify a value that gives rise to a simple one-to-one correspondence between GO terms and cluster. This resulted from either splitting of genes with the same GO terms among multiple clusters (in the case of a high number of clusters), or the generation of incoherent clusters (where a small number of clusters was attempted). I therefore turned to a

supervised approach. For each non-redundant GO term (see Methods) I examined the profiles of the group of genes associated with it and, using the Friedman test (Rice 1995) (Fig. 1), checked for a species effect on translational efficiency. I found 571 terms that display a significant species effect with a false discovery rate (FDR) (Benjamini and Hochberg 1995) of 5% (see Methods for a comparison with results obtained with randomized data). For 273 out of 571 of these terms I was able to stratify the species based on the translational efficiency values of the genes into two or more groups (Fig. 1), utilizing *post-hoc* Wilcoxon signed rank tests (Rice 1995) (Table 3; for a complete listing of the results see the "Additional Tables" section of http://longitude.weizmann.ac.il/pub/papers/Man2007_tai/suppl). Consistent with my previous analysis I observed that for terms related to respiration the species stratification indicated a tendency for lower translational efficiencies in the species displaying the glucose repression phenotype, whereas for the term glycolysis the analysis pointed at higher translational efficiencies for these species (Table 3). Interestingly, according to this analysis the cytosolic ribosome shows higher levels of translational efficiency in the four species exhibiting glucose repression compared to the other five species (Table 3).

Species stratification	GO category	number of profiles	Friedman test p-value
S. cerevisiae, S. bayanus, S. pombe < C. glabrata < K. lactis, D. hansenii, C. albicans, A. nidulans < Y. lipolytica	aerobic respiration*	50	<1.11e-16
	mitochondrion*	607	<1.11e-16
S. pombe < S. cerevisiae, S. bayanus < C. glabrata < A. nidulans < K. lactis, C. albicans < D. hansenii < Y. lipolytica	mitochondrial ribosome*	55	<1.11e-16
S. bayanus < S. cerevisiae, S. pombe < C. glabrata < D. hansenii, Y. lipolytica < C. albicans < K. lactis, A. nidulans	oxidative phosphorylation*	24	<1.11e-16
S. cerevisiae, S. bayanus, S. pombe < C. glabrata < D. hansenii, C. albicans, Y. lipolytica, A. nidulans < K. lactis	mitochondrial electron transport chain*	17	6.22e-15
	respiratory chain complex III (sensu Eukaryota)*	7	1.41e-06
S. cerevisiae, S. bayanus < C. glabrata, S. pombe < K. lactis, D. hansenii, C. albicans, Y. lipolytica, A. nidulans	tricarboxylic acid cycle*	14	5.87e-11
Y. lipolytica < C. albicans < D. hansenii < K. lactis, A. nidulans < S. bayanus, C. glabrata, S. pombe < S. cerevisiae	cytosolic ribosome (sensu Eukaryota)	96	<1.11e-16
Y. lipolytica < K. lactis, D. hansenii, C. albicans, A. nidulans < S. cerevisiae, S. bayanus, C. glabrata, S. pombe	glycolysis*	11	9.28e-11
S. cerevisiae, S. bayanus, C. glabrata, K. lactis, D. hansenii, C. albicans, Y. lipolytica, S. pombe < A. nidulans	spliceosome complex*	41	4.41e-08
S. cerevisiae, S. bayanus, C. glabrata, K. lactis, D. hansenii, C. albicans, Y. lipolytica < S. pombe < A. nidulans	nuclear mRNA splicing, via spliceosome*	52	7.83e-08
	snRNP U1*	12	1.76e-05
S. cerevisiae, S. bayanus, C. glabrata, K. lactis, D. hansenii, C. albicans, Y. lipolytica < A. nidulans, S. pombe	mRNA processing*	79	4.32e-07
S. cerevisiae, S. bayanus, C. glabrata, K. lactis, D. hansenii, C. albicans, A. nidulans, S. pombe < Y. lipolytica	organic acid metabolism*	200	0.0024
Y. lipolytica < K. lactis, D. hansenii, C. albicans < C. glabrata < S. cerevisiae, S. bayanus, A. nidulans, S. pombe	M phase	126	<1.11e-16
	cell cycle	216	<1.11e-16
Y. lipolytica < S. cerevisiae, C. glabrata, K. lactis, D. hansenii, C. albicans, A. nidulans, S. pombe < S. bayanus	DNA metabolism	285	<1.11e-16
	transcription	270	<1.11e-16
	nucleoplasm	224	<1.11e-16
	nuclear part	608	<1.11e-16
	nuclear chromosome	104	6.66e-16
	transcription from RNA polymerase II promoter	165	4.77e-15
	DNA replication	78	1.67e-07
A. nidulans, S. pombe < C. albicans < C. glabrata, K. lactis < S. cerevisiae < S. bayanus, D. hansenii, Y. lipolytica	ribosome biogenesis	186	<1.11e-16
D. hansenii, Y. lipolytica < S. cerevisiae, S. bayanus, C. glabrata, K. lactis, C. albicans, A. nidulans, S. pombe	response to stimulus	323	3.00e-14
	DNA repair	110	8.15e-12
	meiosis	62	1.50e-11

Species stratification	GO category	number of profiles	Friedman test p-value
	response to stress	258	8.55e-10
	chromosome, pericentric region	34	7.45e-06
	autophagy	21	1.66e-05
D. hansenii, C. albicans, Y. lipolytica < S. cerevisiae, S. bayanus, C. glabrata, K. lactis, A. nidulans, S. pombe	telomere maintenance	168	7.17e-11
	meiotic recombination	26	5.07e-07
A. nidulans < S. cerevisiae, S. bayanus, C. glabrata, K. lactis, D. hansenii, C. albicans, Y. lipolytica, S. pombe	vesicle-mediated transport	198	1.26e-09
	secretion	138	1.15e-08
	Golgi vesicle transport	97	5.44e-07
	ER to Golgi vesicle-mediated transport	55	7.20e-06
	actin cytoskeleton	53	4.05e-05
	cortical cytoskeleton	38	1.64e-04
	threonine metabolism	6	3.97e-04
S. cerevisiae, S. bayanus, C. glabrata, K. lactis, Y. lipolytica, A. nidulans, S. pombe < D. hansenii, C. albicans	proteasome complex (sensu Eukaryota)	43	1.94e-09
C. glabrata, K. lactis, C. albicans, Y. lipolytica, A. nidulans, S. pombe < S. cerevisiae, S. bayanus, D. hansenii	helicase activity	56	7.07e-09
	DNA-directed RNA polymerase activity	31	1.27e-08
C. albicans, A. nidulans, S. pombe < S. cerevisiae, S. bayanus, C. glabrata, K. lactis, D. hansenii, Y. lipolytica	RNA modification	49	1.75e-07
	RNA methyltransferase activity	21	5.40e-06
C. glabrata < S. cerevisiae, S. bayanus, K. lactis, D. hansenii, C. albicans, Y. lipolytica, A. nidulans, S. pombe	pyrophosphatase activity	176	1.34e-05
	regulation of pH	17	0.0001

Table 3. Species stratification patterns according to translational efficiencies for various GO categories. The genes associated with each of the GO terms appearing in the table display a species effect on translation efficiency. Pairwise significant relationships between species have been summarized as a species stratification. Note that the same stratification may be implied by several GO categories. GO categories for which the implied species stratification is related to a known phenotype discussed in the text, are marked by an asterisk ('*'). For the remaining GO categories no phenotypic explanation is currently available.

The above analysis revealed additional instances, beyond the glucose repression phenotype, in which the patterns of translational efficiencies of genes annotated with a certain term could be linked to a known phenotypic difference between the organisms. For example, the orthologous groups annotated with the terms "mRNA processing",

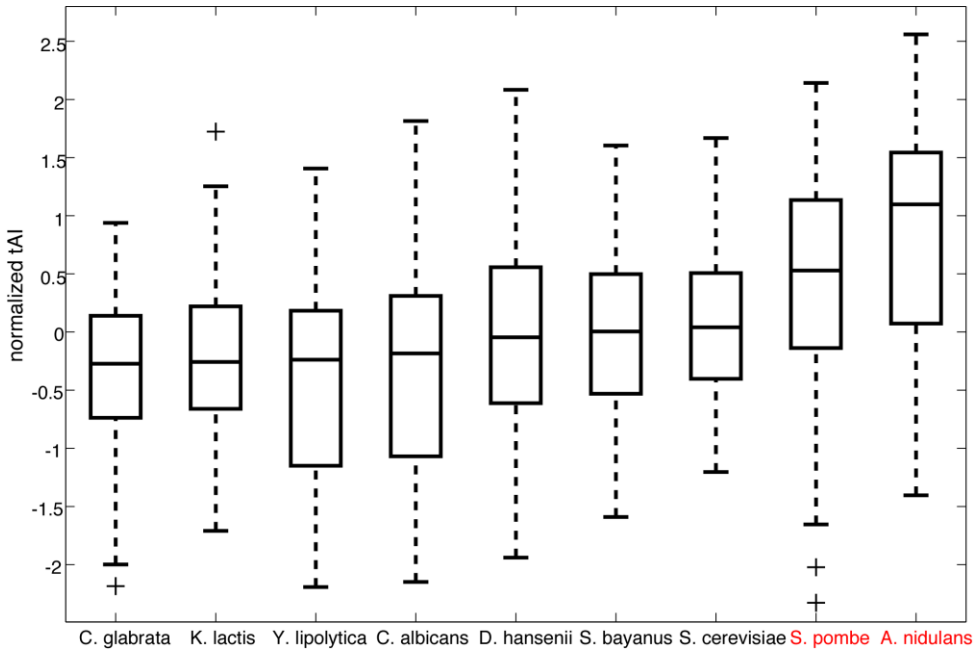


Fig. 12 The translation efficiency profiles of genes related to mRNA splicing is consistent with the number of introns in the genomes of species. A boxplot of the relative translational efficiencies (normalized tAI; see methods) of the genes annotated with the GO term "nuclear mRNA splicing, via spliceosome" (52 genes; p=7.83e-08; Friedman test) is shown. Species are sorted in ascending order of median of translation efficiencies for this groups of genes. The genes belonging to this group show higher translation efficiency in *A. nidulans* and *S. pombe* (highlighted in red), relative to the other seven species, in concordance with the high proportion of introns in these two species (see text).

"nuclear mRNA splicing, via spliceosome", "spliceosome complex", as well as other splicing related terms, displayed the highest translational efficiencies in *A. nidulans*, followed by *S. pombe*, while the rest of the species show significantly lower values (Fig. 12). Strikingly, this order perfectly corresponds to the number of intron containing genes in the various genomes. In *A. nidulans* 9,227 genes (~85% of the genes in the genome) contain altogether 24,824 introns; in *S. pombe* 2,173 genes (~50% of the genes in the genome) contain a total of 4,548 introns. In contrast, in *S. cerevisiae* only 261 genes (~5% of the genes) contain a total of 270 introns, and in *C. albicans* the proportion of intron-containing genes is predicted to be only 3% (Nantel 2006). Thus, I can predict that the fraction of intron-containing genes in the genomes of the remaining five species, which translate the splicing-related genes relatively inefficiently, is low, perhaps similar to those of *S. cerevisiae* and *C. albicans*.

	S. cerevisiae genome	*S. cerevisiae* mRNA measurements	*S. pombe* genome	*S. pombe* mRNA measurements
total number of genes	5869	5048	4766	1342
number of CRPs	129	123	118	104
average number of introns per gene	0.046	0.23	0.93	0.59
average number of introns per CRP gene	0.67	0.67	0.61	0.62
average number of introns per non-CRP gene	0.032	0.028	0.96	0.95

Table 4. Summary of gene data for *S. cerevisiae* and *S. pombe*. Summarized are the gene numbers and average number of introns per gene for the whole genome and for datasets of mRNA measurements for the two species. Only genes for which we have information regarding the number of introns are counted. The mRNA measurements are from (Holstege et al. 1998) and (Schmidt et al. 2007) for *S. cerevisiae* and *S. pombe*, respectively. Special attention is given to the CRPs, as despite their small percentage within the genome (~2% in both species) they account for a substantial share of the transcripts within the cell (~30% in both species).

In *S. cerevisiae* the cytosolic ribosomal protein (CRP) genes account for 30% of the mRNA molecules in the cell under rich medium conditions (Holstege et al. 1998), and because ~75% of the CRP genes contain introns, and these genes are short lived, about 40% of *S. cerevisiae* transcripts are spliced (Warner 1999). This contrasts with the fact that only about 5% of the S. cerevisiae genes contain introns. I was therefore concerned that, while the number of genes with introns is significantly smaller in *S. cerevisiae* compared to *S. pombe* and *A. nidulans*, this result may not hold for the number of splicing events taking place in the cell. To address this concern I compared the intron-content of transcripts in *S. cerevisiae* to that of *S. pombe*, using the datasets of Holstege et al, (Holstege et al. 1998) and Schmidt et al. (Schmidt et al. 2007), for *S. cerevisiae* and *S. pombe*, respectively (both datasets were obtained under rich medium conditions). Despite the fact that the dataset published by Schmidt et al. (Schmidt et al. 2007) contains mRNA levels for only about 30% of *S. pombe* genes, it seems representative of the entire genome in terms of its intron numbers (Table 4), and like in *S. cerevisiae*, the CRP genes account for 32% of the transcripts. I found that *S. cerevisiae* transcripts contain an average of 0.23 introns per transcript. In contrast, in *S. pombe*, even though the average number of introns per CRP gene is lower than in *S. cerevisiae* (Table 4), the average number of introns per transcript is 0.59, more than twice than the amount found for *S. cerevisiae*. For *A. nidulans* since CRP genes contain an average of 2.6 introns per gene, and the remaining genes contain an

average of 2.3 introns per gene, it can safely be assumed that the average number of introns per transcript would significantly exceed the corresponding numbers for both *S. cerevisiae* and *S. pombe*. It is worth noting that in *S. cerevisiae* ribosomal proteins are repressed as part of the stress response (Gasch et al. 2000). Therefore, since CRP genes in *S. cerevisiae* contain a much higher proportion of introns than the rest of the genes (Table 4), the average number of introns per transcript calculated here, based on measurements taken under rich medium conditions, is an upper bound for this quantity across conditions. In contrast, in *S. pombe* and *A. nidulans* the average number of introns in CRPs is lower than (*S. pombe*) or similar to (*A. nidulans*) the average number of introns in the rest of the genes, so that in these species the average number of introns per transcript would be similar or even higher in other conditions. In conclusion, as indicated by the translation efficiencies of the genes participating in mRNA splicing, *A. nidulans* requires the most splicing events, followed by *S. pombe*, and then *S. cerevisiae*. It is predicted that the remaining species in the dataset will require a proportion of splicing events that is similar to that of *S. cerevisiae*.

As an additional example for a relationship between a known species characteristic and the translational efficiency of the relevant gene set, I found the genes annotated with “organic acid metabolism” to be of higher translational efficiency in *Y. lipolytica* compared to the rest of the species (Fig. 13, Table 3). This is in line with the use of this yeast for the industrial production of organic acids, such as 2-ketoglutaric acid and citric acid (Barth and Gaillardin 1997).

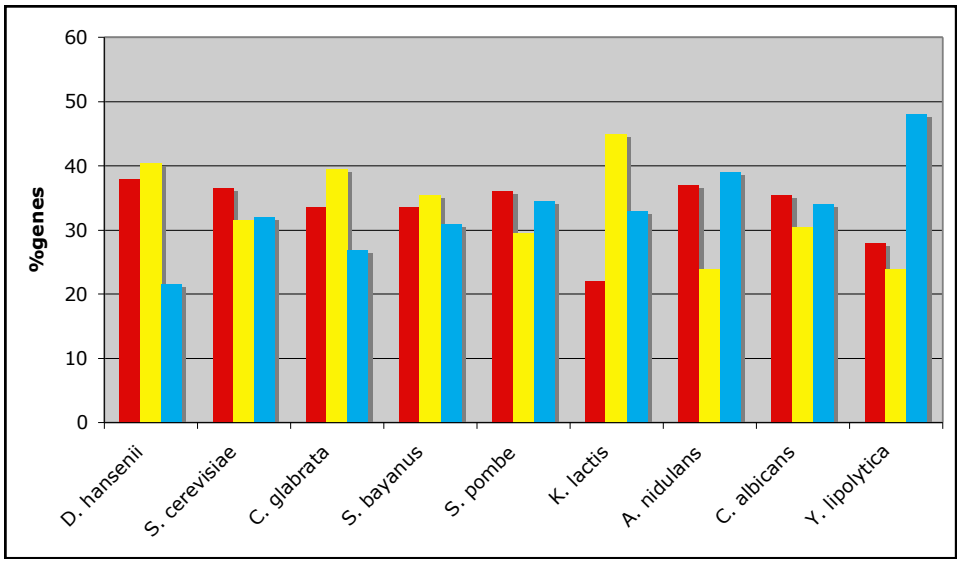


Fig. 13 Organic acid metabolism genes are translated more efficiently in *Y. lipolytica*, a yeast used for the industrial production of organic acids. For each gene associated with the GO category "organic acid metabolism" the species were given ranks according to how efficiently they translate the gene (the species translating the gene the least efficiently was given a rank of 1; the species translating the gene the most efficiently was given a rank of 9). The ranks of the species for all 200 genes associated with this category were summarized as a histogram. For clarity the ranks were grouped into three groups: 1-3 (red); 4-6 (yellow); and 7-9 (blue). Species are ordered in ascending order of the median of their translation efficiency for this group of genes. It can be seen that *Y. lipolytica*, a species used for the industrial production of organic acids (Barth and Gaillardin 1997), tends to have higher ranks than other species, and thus translates these genes more efficiently (p=0.0024; Friedman test).

My data also contains some highly significant patterns that I cannot presently explain by known phenotypic differences among the species (Table 3, as well as the complete listing of the supervised analysis results found under the "Additional Tables" section in http://longitude.weizmann.ac.il/pub/papers/Man2007_tai/suppl). As an example, the genes of the M phase of the cell cycle were found to exhibit the following order of translational efficiencies: *Y. lipolytica* < *K. lactis*, *D. hansenii*, *C. albicans* < *C. glabrata* < *S. cerevisiae*, *S. bayanus*, *A. nidulans*, *S. pombe* (Table 3, Fig. 14A). In search for particular M-phase related genes that gave rise to this signal, I clustered (using the kmeans algorithm (MacQueen 1967)) the 126 translational efficiency profiles of genes associated with this GO category into five clusters. One of the clusters, which contains 33 profiles, resembles the above-mentioned pattern (Fig. 14B, Appendix 5). Interestingly, this cluster is composed of genes that belong to meiosis as well as genes that participate in mitosis. Although not much is known about the physiology of most of the species in my sample, the species stratification found for these genes, as well as other significant patterns, could be used to raise predictions that would direct future research of these species.

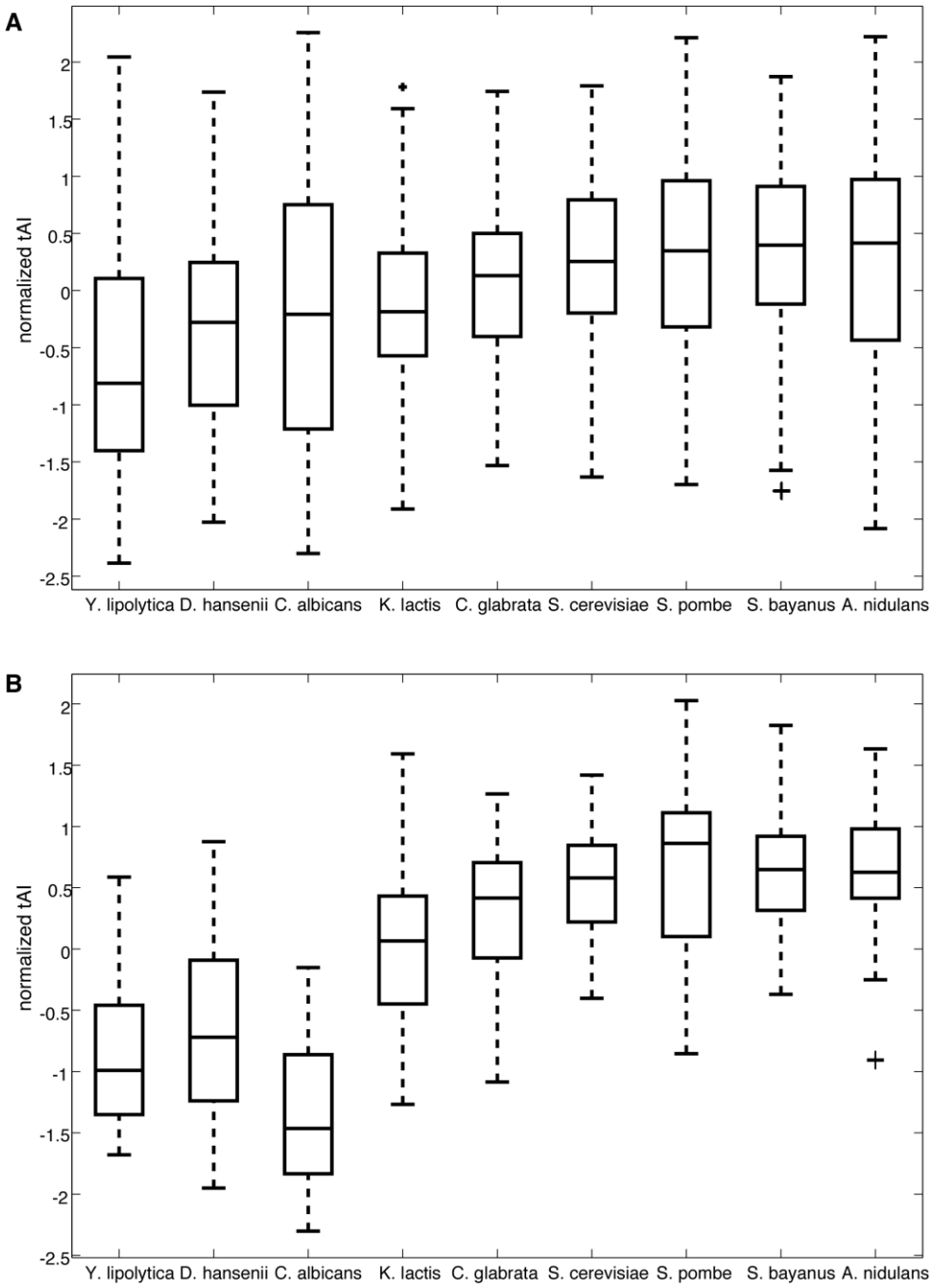


Fig. 14 The translation efficiency profiles of genes related to the M-phase of the cell cycle. Shown are boxplots of the relative translational efficiencies (normalized tAI; see Methods) of two groups of genes: A. genes annotated with "M phase" (126 profiles). Species are ordered in ascending order of the median translation efficiency for this group of genes. B. a subset of 33 of these profiles that resembles the species stratification implied statistically by the M phase genes (Table 3). Species are ordered as in A.

2. A comparison of the human and chimpanzee olfactory receptor gene repertoires

2.1. The chimpanzee OR repertoire

We identified 1,091 putative OR genes in the draft of the chimpanzee genome. Of these, 192 sequences were shorter than 300 bp (corresponding to 1/3 of the entire OR protein length), and were therefore excluded from subsequent analyses. Of 899 chimpanzee OR genes, 353 (39%) have an uninterrupted (intact) open reading frame, and hence, may be considered functional. This fraction of intact chimpanzee OR genes may be an underestimate due to sequencing errors that have been incorporated into the chimpanzee genome assembly and that appear to disrupt coding regions. In order to test this possibility, we used the previously published sequences of 30 chimpanzee intact OR genes (Gilad et al. 2003b) and compared them with the corresponding sequences in the chimpanzee genome draft. We found an average of 0.71% sequence differences between the sequences obtained by Gilad et al. (Gilad et al. 2003b) and those of the chimpanzee assembly. Seven (23.3%) of the OR genes annotated as "intact" by Gilad et al. (Gilad et al. 2003b) contain either nonsense mutations or single base-pair insertions/deletions in the chimpanzee assembly that lead to one or more in-frame premature stop codons. If these disruptions are in fact sequencing errors, then, extrapolating to the whole repertoire, the corrected fraction of intact genes in the chimpanzee OR gene repertoire is ~50%.

Next, we used the full-length (>800 bp) OR sequences from human and chimpanzee to build a distance-based phylogenetic tree of both OR gene repertoires (Fig. 15). Following the family-subfamily classification of OR genes (Glusman et al. 2000; Glusman et al. 2001), the overlap of the represented OR subfamilies in the repertoires of human and chimpanzee is nearly complete (Fig. 15), and in particular, in most OR subfamilies, there is a human ortholog for almost every chimpanzee OR gene. However, there are also some species-specific expansions. We note here the largest expansions. A chimpanzee expansion within OR subfamily 4C (Fig. 16A), and three human expansions in subfamilies 2A, 4F (also noted by Linardopoulou et al. (Linardopoulou et al. 2001)), and 6C (Fig. 16B,C,D). The sequence similarity between the human-specific OR genes in subfamily 6C is only 70%. This suggests

that these genes existed in the common ancestor of human and chimpanzee, and that their orthologs were either deleted from the chimpanzee genome, or were not found by us (possibly due to properties of the assembly). In addition, the chimpanzee has roughly ~60% more loci from the 7E subfamily compared with human (84 and 132 7E OR genes in human and chimpanzee, respectively). The 7E OR subfamily in human consists almost entirely of pseudogenes (Newman and Trask 2003); similarly, there is only one intact OR gene among the chimpanzee 7E OR subfamily sequences.

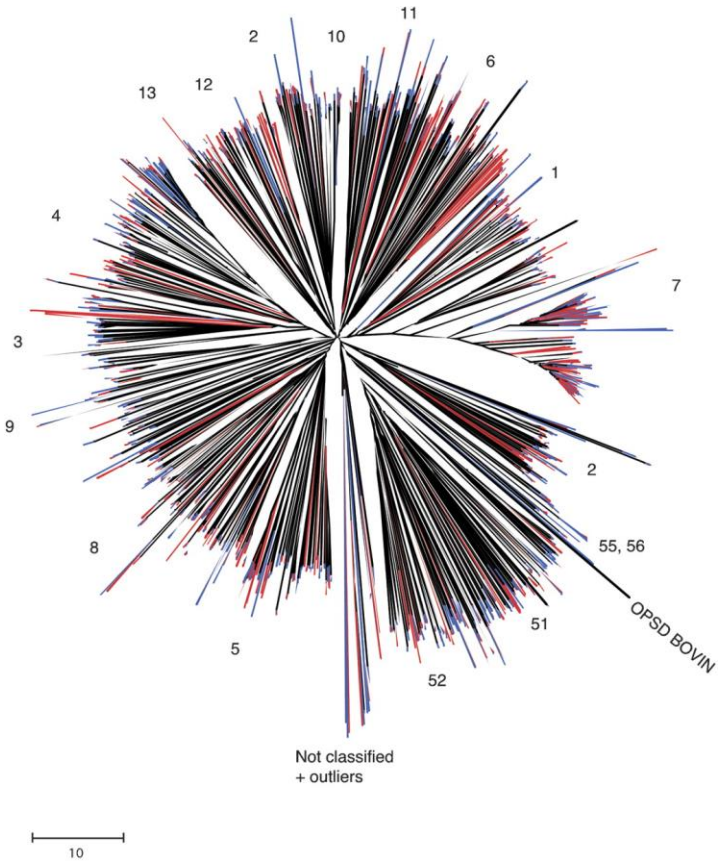


Fig. 15 A neighbor-joining tree of the olfactory receptor repertoires of human and chimpanzee. The sequence of the bovine rhodopsin protein was used as outgroup (indicated as OPSD_BOVIN). Numbers indicate the different OR gene families. Human external branches are red, chimpanzee ones are blue.

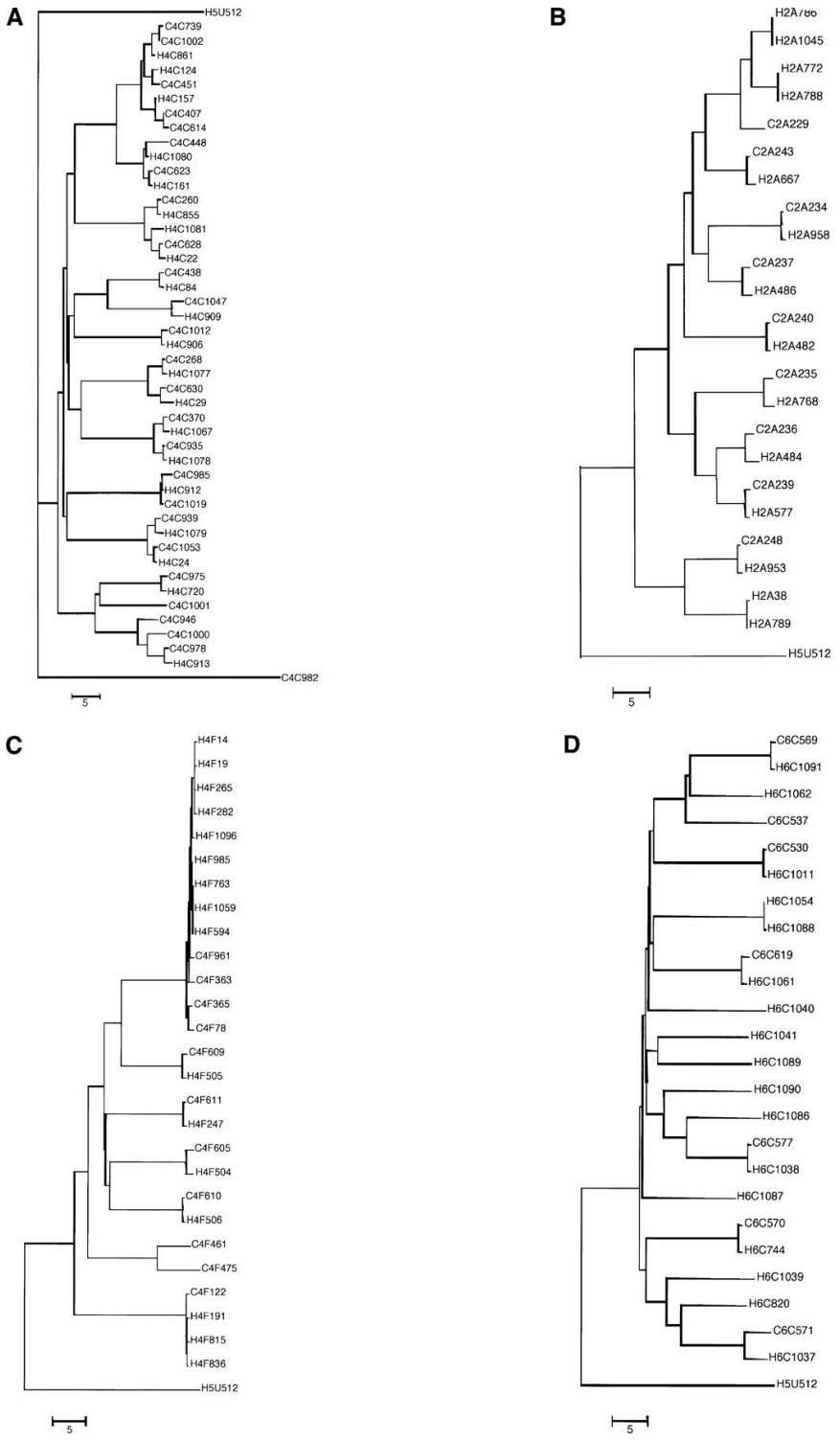


Fig. 16 Distance matrix trees for specific OR subfamilies in human and chimpanzee. The first letter of the OR name indicates the species name (H and C for human and chimpanzee, respectively). Human OR sequence H5U512 was used as an outgroup in all cases. A. subfamily 4C: 20 human sequences (10 intact) and 27 chimpanzee sequences (12 intact) B. Subfamily 2A: 14 human sequences (nine intact) and nine chimpanzee sequences (seven intact) C. Subfamily 4F: 16 human sequences (nine intact) and 11 chimpanzee sequences (four intact) D. Subfamily 6C: 17 human sequences (10 intact) and nine chimpanzee sequences (seven intact).

2.2. OR genes under selection

Previous studies of human and chimpanzee OR genes indicated that this gene superfamily evolves under different selection pressures in each species (Gilad et al. 2003a). The availability of the complete chimpanzee OR gene repertoire enabled us to ask whether we can identify specific OR genes that may have been a target of natural selection in one of the species. Specifically, we used an analysis that is sensitive to differences between the species in the type of selection pressures acting on a given locus.

As a starting point for our analysis, we used the previously published set of human-mouse OR gene orthologs (Man et al. 2004). We then used the "reciprocal best hit" human-chimpanzee OR gene list in order to identify the corresponding chimpanzee orthologs. Thus, we obtained clear human-chimpanzee-mouse ortholog trios for 201 OR genes. By using the mouse ortholog as the outgroup, we were able to estimate the OR gene sequences of the human-chimpanzee common ancestor, and thereby infer lineage-specific substitutions for each OR gene.

In order to test for differences in selection pressures among the species, we compared the rate of synonymous and nonsynonymous divergence on each lineage. Under the null model, there is a single ratio of nonsynonymous to synonymous divergence (Dn/Ds) for the trio of species. Under the alternative, each lineage is allowed a separate Dn/Ds ratio. For each OR gene, we maximized the likelihood of the parameters given the data. We then used a likelihood ratio test (LRT) (Rice 1995) to test the null model. In this way, we could reject the null model for 52 OR genes ($p < 0.05$; 5 genes are expected to be significant by chance, after excluding genes with zero counts in any class of substitutions).

Since our main goal in this section was to identify specific genes that are most likely to evolve under positive selection, we concentrated on 18 OR genes that were significant at a false discovery rate (FDR) (Benjamini and Hochberg 1995) of 1% (Table 5). A significant LRT result could reflect differences in selective constraint between orthologs, or might result from positive selection acting on an OR gene in only one of the lineages. We inspected the Dn/Ds values for each of these OR genes on individual lineages to help interpret the rejection of the null model. In six cases, the Dn/Ds value for substitutions on the chimpanzee lineage was below one, while the Dn/Ds value for the human lineage was

Locus (human name)	human [a] Dn/Ds	Chimpanzee [a] Dn/Ds	LRT	P value [b]
OR52H1	**2.80**	0.97	34.24	0.000001
OR5M3	0.78	0.48	25.98	0.000002
OR5M8	0.47	1.13	21.34	0.000023
OR11L1	1.00	0.78	20.23	0.000040
OR1L8	0.66	**1.20**	19.01	0.000074
OR52B2	0.79	0.33	18.26	0.000108
OR4F29	0.48	**3.20**	17.30	0.000175
OR6K2	**2.03**	0.71	17.13	0.000190
OR51G1	**1.35**	0.69	15.76	0.000378
OR4D11	**1.73**	0.68	15.76	0.000379
OR10G7	0.60	1.14	15.34	0.000468
OR4C11	0.67	0.84	15.11	0.000525
OR5AP2	1.10	0.92	14.19	0.000829
OR4D10	0.65	0.94	14.16	0.000841
OR51Q1	0.78	0.94	13.70	0.001059
OR56B2P	0.62	**2.49**	13.29	0.001297
OR1P1P	**1.60**	0.33	13.08	0.001444
OR4F13P	**2.20**	0.43	12.91	0.001576

Table 5. LRT results and Dn/Ds value for human and chimpanzee OR genes

[a]Dn/Ds values that seem indicative of positive selection are in bold.

[b]P values of the Likelihood Ratio test (LRT) of the null vs. alternative models.

higher than 1.2 (Table 5). This suggests that the rejection of the null model in these cases is due to positive selection driving the evolution of the human, but not the chimpanzee OR gene. Similarly, we find three cases for which it seems that the chimpanzee, but not the human gene, has evolved under positive selection. In nine cases, the rejection of the null model may be due to strong purifying selection on one lineage (possibly the mouse) and relaxed constraint on at least one of the others.

2.3. Estimating the age of human pseudogenes

We identified 761 clear cases of human-chimpanzee OR gene orthologous pairs (see Methods). Of these, the number of apparent pseudogenes in human and chimpanzee is 403 and 440, respectively. Using the "conceptual translation" of coding sequences to protein (i.e. the inferred protein sequences, allowing for frameshifts and stop codons, based on a library of related protein sequences) we identified all coding-region disruptions in each pseudogene. We then defined two groups of human OR pseudogenes as follows: (1) shared pseudogenes, i.e., those that share at least one coding-region disruption with their chimpanzee ortholog, and hence were most likely pseudogenes in the human-chimpanzee common ancestor, and (2) human-specific pseudogenes, i.e., those that do not share any disruption with their chimpanzee

orthologs, and most likely were intact in the common ancestor of human and chimpanzee.

Species-specific and shared coding-region disruptions in OR genes have been described in the past (Rouquier et al. 1998). Here, we concentrated on human-specific disruption in the shared pseudogenes (by definition, at least one disruption in these loci is shared with chimpanzee, but any additional ones may be human specific). We assume that these disruptions are neutral mutations, as they occurred in pseudogenes. As expected under the hypothesis of a neutral molecular clock, the number of human-specific disruptions per shared pseudogene appears to be approximately Poisson distributed (Fig. 17). If we fit a Poisson distribution to the data, the estimate of the mean is $\hat{\lambda} = 0.701$ disruptions per gene (~1 kb) since the divergence of human and chimpanzee. Assuming six million years (MY) since the common ancestor of a human and chimpanzee sequence, this corresponds to a neutral OR gene disruption rate of $\frac{0.701}{6 \cdot 10^6} = 1.17 \cdot 10^{-6}$ disruptions per gene per year. This calculation provides a general estimate of the rate at which neutral gene disruptions accumulate in OR genes, and possibly in other human genes with similar GC content.

Next, we tabulated the number of coding-region disruptions in the human-specific pseudogenes. Interestingly, we could not reject a Poisson distribution for these disruptions either (by using a χ^2 test, excluding the class corresponding to zero disruptions, p=0.58), meaning that coding sequence disruptions have been accumulating along these sequences independently of each other. This implies that the first coding sequence disruption that appeared in each of these pseudogenes was as selectively neutral as the remaining disruptions within the sequence. We then proceeded by assuming that at a certain point in human evolution, a subset of OR genes became unnecessary and were free to accumulate coding-region disruptions. In order to estimate this time point, we used the mean of the Poisson distribution, which we estimated to be 0.451. Under our model, assuming 6 MY since the common ancestor of a human and a chimpanzee sequence, we estimate that the relaxation of selective constraint started (0.451/0.701)•6 MYA = 3.86 MYA, with 3.28-4.56 MY as rough 95% confidence intervals (obtained by parametrical bootstrapping 10,000 times; see Methods). Thus, we can reject the hypothesis that OR genes have been accumulating disruptions at a neutral rate over the past 6 MY. This hypothesis can

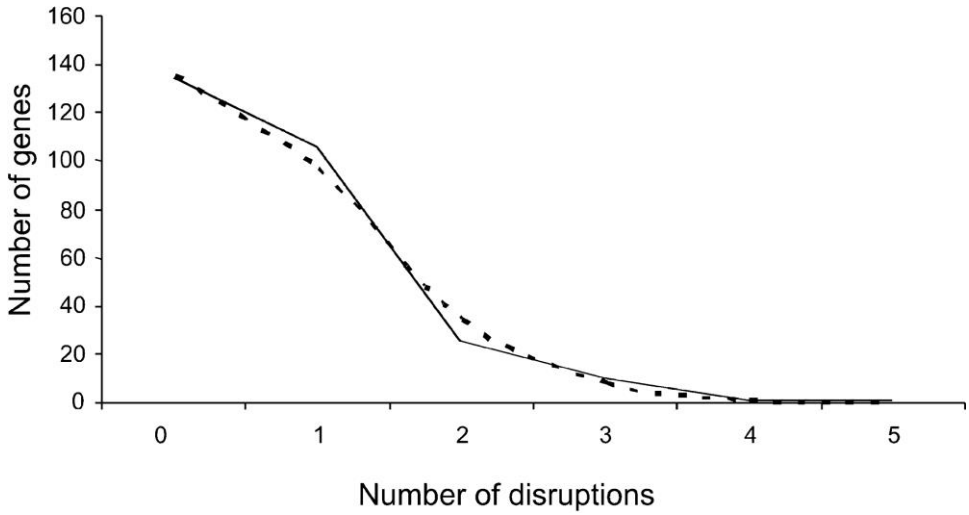


Fig. 17 The distribution of human-specific OR gene disruptions in pseudogenes shared among humans and chimpanzees. The broken line is the Poisson fit for the data (λ=0.701).

also be rejected by testing whether the Poisson distribution obtained for neutral disruptions over 6 MY fits the data for human-specific disruptions; it does not ($p < 0.01$).

3. Discussion

Both studies presented here have revealed differences among the coding sequences of different species that may take part in their phenotypic divergence. For the analysis of translational efficiency among yeast species I have found many groups of functionally related genes that show higher translation efficiency in one group of species relative to the other species. Among these, I found cases where the patterns of translation efficiency for the group of genes were in concordance with a phenotype. However, in many cases the patterns I found, though very significant statistically, could not be explained by known phenotypic differences and may open avenues to further investigation into the physiology of the species analyzed. Altogether I found extensive selection on synonymous codon usage that modulates translation according to gene function and phenotype. I conclude that, like factors such as transcription regulation, translation efficiency affects and is affected by the process of species divergence.

The comparison of the human and chimpanzee olfactory repertoires also revealed differences between the two species. As previously predicted, the proportion of functional OR genes in the human repertoire is much smaller than the corresponding proportion in the chimpanzee repertoire, implying that the human sense of smell is indeed diminishing at a greater rate than the chimpanzee sense of smell. We were also able to rule out the possibility that humans have been accumulating OR pseudogenes at a constant neutral rate since their divergence from chimpanzees. A possible interpretation of this finding is that the changes in human lifestyle that allowed such a deterioration of the OR repertoire did not occur immediately with the human-chimpanzee divergence. The comparison of the two repertoires also revealed two chimpanzee-specific OR subfamily expansions and three expansions specific to humans were found, as well as several examples in each of the species where an OR is evolving under positive selection in one species but not the other. Thus, we conclude that, despite the overall relaxation of constraint on human olfaction, the evolution of the OR gene repertoires of both species was shaped by species-specific sensory requirements.

The two studies also revealed conserved aspects among the species. The analysis of the yeast species revealed that the proportions of tRNA genes, which serve as a

surrogate measure for the cellular tRNA pools, are largely conserved among the analyzed species. This means that the evolution of translation efficiency occurs mainly in a distributive fashion, whereby each gene adapts its coding sequence to an essentially invariable tRNA pool. In the comparison of the OR gene repertoires of human and chimpanzee, we found that the overwhelming majority of the chimpanzee OR genes have one human ortholog, so that the overall structure of the repertoire is conserved.

Some differences between the two studies are also of note. One major difference is the type of signals analyzed. The analysis of translation efficiency focused on subtle signals at the level of synonymous codon usage that are many times considered neutral. On the other hand, the comparison of the olfactory repertoire focused, mainly, on very crude signals: open-reading frame disruptions that obviously prevent the formation of a functional protein and the expansion of certain subfamilies in one species or the other. Another major difference between the studies is the degree of interpretation one can give to their results. The extensive annotation of the *S. cerevisiae* genome and the fact that yeast are model organisms that are relatively easy to research has allowed me to make specific associations between the translation efficiency of groups of genes and the phenotype of these yeasts. Moreover, some of the statistically significant results, which could not be explained at present, may be explained by future experiments. In contrast, the olfactory system is considered a difficult system to study experimentally with the result being that ligands have been found for only a very small fraction of known OR sequences (c.f. (Kajiya et al. 2001; Hatt 2004)). This problem is complicated by the fact that the olfactory system is a combinatorial one where the same receptor recognizes multiple odorants and the same odorant is recognized by multiple receptors, albeit with different affinities (Malnic et al. 1999; Kajiya et al. 2001). Therefore, for the comparison of the olfactory repertoires of human and chimpanzee, although we can detect differences that may be attributed to divergences of sensory requirements, not only is it impossible at present to interpret the exact phenotypic meaning of these differences, but it is also unlikely that we will be able to interpret them in the near future.

Finally, a pertinent question is whether the tools used in the first project could be applied to the species in the second project, and *vice versa*. The comparison of translation efficiency between the genes of human and chimpanzee is problematic, for two reasons. First, it is not clear that translational selection has had a significant

impact on the codon usage of genes in vertebrate genomes (see the subsection entitled "Comparative analysis of translation efficiency in other species" below). Second, since human and chimpanzee have diverged only recently in evolutionary terms the tAI of their genes is probably insufficiently diverged, which greatly reduces the chance that a comparison between the translation efficiency of orthologs from these two species would be meaningful. On the other hand, the tools used in the comparison between the human and chimpanzee olfactory receptor repertoires can be employed, albeit to a limited extent, for the comparison of the yeast species analyzed in the first project. Since yeast species do not have olfactory receptors, such analyses would be employed to yeast gene families in general. A survey of pseudogenes in *S. cerevisiae* (Harrison et al. 2002) indicated most pseudogenization events have occurred in evolutionarily young ORFs, as opposed to ancient ORFs that are conserved between species. So, based on this early report, it would appear that very little phenotypic variation among yeast species is due to pseudogenization. However, since this pseudogene survey was conducted only in one yeast species, it may be worthwhile to repeat it in the other yeast species analyzed here to examine whether pseudogenization has played a part in phenotypic divergence in these species. A more promising avenue for the investigation of phenotypic divergence in yeasts, which is in line with the human-chimpanzee comparison, is to look for expansions/contractions of protein families in these species. For example, in a comparison of five of the species in our sample, Dujon et al. found hundreds of cases of families that were present in all species albeit with a different number of representatives, and in fourteen cases the expansion was found to be statistically significant (Dujon et al. 2004). Such expansions may shed light on phenotypic differences among the yeast species I analyzed. Finally, the maximum likelihood framework, which was utilized to suggest OR coding sequences that have evolved under positive selection in human or chimpanzee, can be applied to yeast orthologs, as long as the analyzed genes are not too diverged in sequence. This would probably mean that the analysis could be useful in comparing close species pairs such as *S. cerevisiae* and *C. glabrata*, but would be meaningless in the comparison of more distant species pairs such as *S. cerevisiae* and *S. pombe*.

Aspects specific to each of the two parts of the research are discussed below.

1. Selection for translation efficiency and its relation to species divergence in yeast

1.1. tAI as a predictor of protein levels

In this study I used tAI (dos Reis et al. 2004) as a surrogate measure for protein levels. The comparison of tAI values with observed protein levels was statistically significant, indicating that this theoretical index captures an aspect of coding sequences that is relevant to protein levels. Moreover, the statistical significance of the partial correlation between tAI and protein levels, whilst controlling for mRNA levels, indicates that tAI adds information that cannot be obtained from mRNA levels. Nevertheless, genes with vastly different protein levels obtain the same tAI, and vice versa. Setting aside the contribution of errors in measurements to the discordance between observations and predictions, there are three possible, non-exclusive, reasons for the lack of accuracy in protein level predictions that are based on tAI. First, tAI focuses on one aspect of the maintenance of protein levels, namely translation. As a result, the effect of other factors that are known to contribute significantly to steady-state protein levels, namely transcript levels and protein half-lives (Belle et al. 2006), are neglected. Second, tAI attempts to quantify the efficiency of a specific aspect of translation, i.e. the effect of codon usage on the elongation of the nascent peptide. However, there are other facets of the translation process that can be utilized in the regulation of protein levels. These include the initiation of translation, which is considered a rate-limiting step in the translation process (Preiss and Hentze 2003), as well as the density of translating ribosomes on transcripts, which is different for different genes (Arava et al. 2003). Finally, tAI may not model accurately the effects of codon usage on the efficiency of peptide elongation. For instance, the relative strength of the wobble interaction of tRNAs with non-cognate codons was optimized based on transcript levels, rather than on protein levels (dos Reis et al. 2004), thus focusing on the demand for the various tRNAs rather than the resultant protein levels. It is possible that using the required protein levels instead would yield a more accurate model of anticodon-codon interactions. In addition, tAI assumes that only the overall codon composition of the coding sequence is relevant for the purposes of translation efficiency, so that any two equal-length sequences containing the exact same codon composition, would obtain the same tAI score. However, it has been observed that codon bias is not uniform along sequences (Qin et al. 2004), so it is possible that an

index that would take into account not only the overall codon composition, but also the dispersion of the codon bias along the sequence, would show an improved performance over tAI. On the other hand, the pattern of dispersion of codon bias along the sequence does not seem to be universal, with *D. melanogaster* showing a different pattern than *S. cerevisiae* and *E. coli*, and the effect of expression level on overall codon usage bias is more pronounced than its effect on the shape of the spatial distribution (Qin et al. 2004). Thus, it seems that the incorporation of spatial patterns of codon bias into translation efficiency calculations would not be trivial, and it is not clear that such an effort would lead to great improvement in predictions. All this being said, my work has shown that tAI, despite being a very crude predictor of protein levels, contains enough information to make predictions regarding the phenotypes of species.

1.2. Comparison to other studies that investigated the relationship between translation efficiency and the lifestyles of species

In addition to my work, only one study has investigated the relationship between the translation efficiency of individual genes and the lifestyles of the species in whose genome they reside (Carbone and Madden 2005). Using conformance to the dominating codon bias in the genome as a measure of translation efficiency (Carbone et al. 2003), Carbone and Madden investigated the relationship between translation efficiency of genes and the lifestyles of thirteen prokaryotes and one eukaryote (*S. cerevisiae*). Contrary to my study, they chose to limit themselves to the metabolism of the investigated species, analyzing metabolic pathways that were found in them. In addition, rather than looking at the values of individual genes they chose to summarize the translation efficiency of a pathway by its mean value of codon bias. Similar to my study, the pathways of each organism were then given a score that reflects its rank among the other pathways. Pathways with a high relative score were then inferred to be central to the metabolism of the relevant species. These relative scores allowed the authors to examine pathways that display similar behavior among the analyzed organisms, as well as pathways that behave differentially. Since each pathway was summarized by a single number, which was further simplified into a color code, the comparisons made among the organisms were, in contrast to my study, qualitative in nature. Such a qualitative comparison probably limits the power of

analyses. On the other hand, the fact that a single number summarized the pathways freed the authors of the need to perform an extensive ortholog analysis that would have most likely been complicated due to the high divergence between the species they analyzed. Another difference between my study and the study of Carbone and Madden was the fact that in my study species that did not show a significant trend of translational selection were excluded from the comparative analysis, whereas Carbone and Madden included some species which, according to their criteria, do not show translational codon bias. Their claim was that, even in the genomes of such organisms, genes that are central to the metabolism of the relevant species show codon bias relative to the rest of the genome. Since in my comparative analyses I normalized translation efficiency values also across species, I was concerned that noise inserted by species that do not show statistically significant translational selection in their genome would reduce the power of the analyses that were more quantitative in nature. Indeed, in earlier analyses, where *A. gossypii* was included, some of the results that were shown here were also obtained, albeit with less significant p-values (results not shown). A final difference between the two studies is my use of tAI, rather than conformance to the dominant codon bias in the genome (Carbone et al. 2003), to measure translation efficiency of coding sequences. which alleviated the need to speculate on the identity of the highly expressed genes in each species. This methodological choice was crucial since it turns out that the classical genes that are assumed to be high in all species (CRPs, glycolysis etc.) showed a lot of variance among species. As a bonus, the use of tAI, which relies on the composition of the tRNA pool, led me to investigate the evolution of the tRNA pools of the species analyzed, exposing the fact that this pool remains relatively constant.

In summary, my study is distinguished from that of Carbone and Madden (Carbone et al. 2003) by a narrower and more conservative choice of species on my part. On the other hand, in my analyses I used a broader range of categories of genes, allowing me to observe, as well as make predictions, regarding differences that are related to non-metabolic processes. Also, by using more closely related species, in which orthology relationships are more easily inferred, as well as by utilizing the values of individual genes, rather than summary values, to compare gene groups among species, I was able to obtain results that are more quantitative in nature.

1.3. Comparative analysis of translation efficiency in other species

The methodology developed in my study opens another avenue to investigate the genes underlying phenotypic divergence. I anticipate that the kind of analyses performed in the current study could in future be extended to other groups of species. However, the analysis may not be applicable to any group of species - a number of criteria would have to be considered when selecting species for analysis. First, only species with a complete genome sequence can be considered for analysis. This requirement is necessary both for the inference of the tRNA gene copy numbers (although if experimentally-determined abundances of tRNAs are available, these can be used in the calculation of tAI instead of tRNA gene copy numbers) and for the inference of orthologs with other species in the sample. Second, a number of phylogenetically close (although not too close) species should be available for analysis. A time of divergence among species that is too large probably entails a very small proportion of shared genes, which makes it likely that most differences among the species are due to different gene repertoires, rather than differences in the expression of shared genes. On the other hand, the comparison of translation efficiencies between species that are too close to each other in evolutionary terms may not be meaningful, as they may not be sufficiently diverged in tAI. Third, in order to be able to interpret the results of the comparison of translation efficiencies among species, at least one species should have an extensively annotated genome sequence. Finally, it should be ascertained that translational selection is a significant contributor to the codon usage patterns of each of the species in the sample.

While the first three criteria may present a problem in analyzing a species of particular interest, this would probably be a transient problem that may be resolved as more genomes are sequenced and annotated. However, the final criterion is one that examines a property of a genome that is permanent, and therefore if the species of interest does not fulfill it, it is permanently barred from being analyzed by the current methodology. The issue of how to determine whether a species passes this last criterion is therefore of special importance. On the one hand, translational selection has been inferred in a large number of species from all kingdoms of life, so that potentially the number of species to which the comparative translation efficiency analysis can be applied is very large. On the other hand, different authors inferred translational selection using different methods and their results are sometimes

contradictory (for example, Pan et al. (Pan et al. 1998) inferred that the genome of *H. influenza* experienced translational selection, but this result was contradicted by the test proposed by dos Reis et al. (dos Reis et al. 2004)). Notably, vertebrates are a group of species in which the existence of translational selection is particularly controversial. In species belonging to this group multivariate analyses have shown that variability in codon usage of a gene is governed mainly by Xg, the GC content at the third codon position (Kanaya et al. 2001; Musto et al. 2001), which is strongly correlated with the GC content of the isochore in which it is located (reviewed in (Bernardi 2000; Eyre-Walker and Hurst 2001)). The fact that Xg (and thus codon usage) as well as introns and intergenic regions seem to be influenced to the same extent by the GC content of the isochore implies that translational selection is unlikely to have a significant impact on codon usage in vertebrates. Indeed, the absence of translational selection in vertebrate genomes was supported by a number of papers (c.f. (Urrutia and Hurst 2001; Semon et al. 2006)). However, other papers found evidence for translational selection in vertebrate genomes (c.f. (Musto et al. 2001; Comeron 2004; Comeron 2006; Kotlar and Lavner 2006)), although some admitted that this selection is weak (Musto et al. 2001; Comeron 2004; Comeron 2006). In this study I compared the extent of adaptation of coding sequences to the tRNA pools to general codon bias, as a test for the significant effects of translational selection on codon usage in the genome. This test doesn't attempt to isolate translational selection from other evolutionary forces, which may influence codon usage in a way that is synergistic or antagonistic to translational selection. Rather, it tests whether the observed codon usage is adapted to the tRNA pools, irrespective of the evolutionary forces that have led to this situation. I suggest that future similar studies would employ a similar strategy in selecting species that are appropriate for analysis.

Finally, since tRNA pools may differ among tissues (Dittmar et al. 2006), developmental stages (White et al. 1973) and perhaps even environmental conditions, the methodology may be modified to accommodate new experimental data. For instance, as experimental data regarding tRNA abundances for different tissues, developmental stages or conditions accumulates, one may use the maximal tAI that can be obtained using any of the available pools, rather than the theoretical tAI that is based on the tRNA gene copy numbers, as a surrogate measure for maximal protein levels. This maximal tAI can then be used both in the test for translational selection and in the comparison among species. Alternatively, one can compare the translation

efficiency in the same tissue or environmental condition across species, by using tAI values computed using the tRNA pools determined experimentally for this specific tissue or condition. A serious disadvantage of such approaches is that that they require experimental data regarding the tRNA pools of all the analyzed species, a requirement that was waived in the current study.

2. A comparison of the human and chimpanzee olfactory gene repertoires

We analyzed the complete chimpanzee OR repertoire and compared it with the repertoire of OR genes in human. Our comparison yielded several findings that could underlie the hypothesized differences in olfactory phenotypes among humans and chimpanzees.

Confirming predictions that were based on small sample sizes (Gilad et al. 2003a; Gilad et al. 2003b), we find that the fraction of OR pseudogenes is significantly greater in humans. This finding is supported by the recent observation that human-specific pseudogenes are enriched for ORs (Wang et al. 2006). Species-specific inactivation of a gene can lead to a loss of a function, e.g. the *Trypanosoma* lytic factor HPR, which has been inactivated in chimpanzee but not in human, likely explains the susceptibility of chimpanzees to *T. brucei* infections (Puente et al. 2005). On the other hand, inactivation of genes can also confer a selective advantage and thus be driven to fixation by positive selection, as has been shown for CASPASE12, whose null allele confers protection from severe sepsis in humans (Wang et al. 2006). In our study we observed that the distribution of number of inactivating mutations per gene in human-specific OR pseudogenes fits a poisson distribution, implying that the first inactivating mutation in each of these pseudogenes was selectively neutral. Thus, it is unlikely that any of the human-specific OR pseudogenes conferred a selective advantage to humans, and that the pseudogenization process only led to loss of olfactory capabilities. It is interesting to note that the relaxation of constraints in the human lineage is not common to all chemosensory functions. For instance, Parry et al. analyzed the T2R family of bitter taste receptors in humans, bonobos and chimpanzees and did not find a human-specific loss of the amount of functional genes (Parry et al. 2004).

We also noted species-specific expansions/contractions among subfamilies of ORs. Such differences in OR family sizes between human and chimpanzee have been also noted by a recent study that examined the evolution of gene family sizes in mammalian species (Demuth et al. 2006). The observed expansions/contractions include both intact OR genes as well as pseudogenes and are probably the product of a neutral process of duplication and deletion (Nei et al. 2000). Alternatively, these

expansions could be the result of species-specific sensory needs, as the number of functional genes within any given OR subfamily may be proportional to the breadth of binding sites within a subfamily (Malnic et al. 1999).

Lastly, we suggest a number of OR genes, both in human and in chimpanzee, as probable candidates for adaptations, as they may have evolved under positive selection. Previously, Gilad et al. (Gilad et al. 2003a) suggested that OR genes in human evolve under positive selection, but no evidence was found for such adaptation in chimpanzee. The authors found that most chimpanzee intact OR genes evolve under strong evolutionary constraint and suggested that this may reduce the power to detect positive selection in small-scale studies. Here, we took advantage of the identification of 201 human-chimpanzee-mouse ortholog trios, to detect rapidly evolving proteins in human and chimpanzee. Our findings are in line with an earlier study that, using an approach similar to ours, looked for evidence for natural selection in either the human or chimpanzee lineage within a very large collection of genes, and found an enrichment for ORs among the genes that were highlighted (Clark et al. 2003). The signature of selection on OR genes can be corroborated by the analysis of polymorphism data (e.g., (Hamblin and Di Rienzo 2000; Hamblin et al. 2002)). Targets of selection identified from the analysis of polymorphism and divergence are promising candidate for human- and chimpanzee-specific chemosensory traits. Indeed, a recent study, which compared human polymorphism data and fixed differences between humans and chimpanzees, also pinpointed a number of olfactory genes as candidates for adaptations in the human lineage (Bustamante et al. 2005). A natural next step is to collect data from additional primates to establish whether selective pressures are truly exclusive to one species. Finally, studies to associate OR genes to their primary odorants will determine whether the genes identified in this study indeed underlie species-specific sensitivity.

2.1. The chimpanzee OR repertoire

On a first pass, the number of chimpanzee genomic segments that our algorithm identified as OR gene candidates is 26% higher than the number of human OR genes. Moreover, we could only find clear human orthologs (see Methods) for 761 (69%) of the chimpanzee candidate OR genes. However, when we only considered the 899 chimpanzee OR sequences that are longer than 300 bp, the size of the chimpanzee OR

repertoire becomes similar to that of human, and the proportion of chimpanzee loci with a human ortholog is 85%. This suggests that many of the short sequences identified as OR genes result from imperfections of the chimpanzee genome assembly. It is not improbable that these sequences should have been collapsed in the assembly, rather than be represented as unique (short) genomic segments. It may also be that some of the short OR candidates are not ORs after all, and would not have been included in our putative chimpanzee OR gene repertoire if we had used a more stringent cutoff in our searches against the chimpanzee genome.

The discrepancy in ortholog matches does not completely disappear when the chimpanzee short sequences are excluded. In most remaining cases, we can explain the lack of an ortholog for ~15% of OR genes by lineage-specific expansions/contractions. We also noted a difference between human and chimpanzee in the size of the 7E OR subfamily. This OR subfamily consists almost exclusively of pseudogenes and was shown to have expanded in the human lineage (Newman and Trask 2003). Our findings suggest a similar, even more pronounced, expansion of family 7E in chimpanzee. The selective advantage of this expansion, if any, is unclear.

2.2. Relaxation of constraint on the human lineage

We used the previously published sequence of 30 chimpanzee intact OR genes (Gilad et al. 2003b) in order to estimate the number of sequencing errors that lead to an apparent coding region disruption in the chimpanzee genome draft. Our corrected estimate of the proportion of pseudogenes in the chimpanzee OR repertoire (~50%) is still significantly higher than the estimate of 32-38% from Gilad et al. (Gilad et al. 2003a; Gilad et al. 2003b). This is probably due to the high number of subfamily 7E OR pseudogenes in chimpanzee. OR genes from this subfamily were excluded from the analysis of Gilad et al. (Gilad et al. 2003a; Gilad et al. 2003b), since a recent expansion has been observed for this subfamily (Newman and Trask 2003), and, except for one sequence, all of the 7E ORs are pseudogenes. If we exclude the 7E subfamily from our analysis, the proportion of pseudogenes in human and chimpanzee are 51% and 41%, respectively. These values are within the 95% CI of the observations of Gilad et al. (Gilad et al. 2003a; Gilad et al. 2003b). We note that if we underestimated the number of sequence errors that result in an apparent disrupted coding region, the correct proportion of OR pseudogenes in chimpanzee may be

lower. Thus, the use of the entire repertoire confirms that a greater proportion of OR genes evolve under no or little constraint in humans relative to chimpanzees.

In our attempt to date the time since humans have started to rapidly accumulate OR pseudogenes, we made the simplistic assumption that all human intact OR genes were under evolutionary constraint until some point in human evolution. Then, a subset of OR genes became unnecessary, and hence neutrally evolving. We know, however, that not all ORs are under constraint in nonhuman primates (Gilad et al. 2003b). Thus, a more realistic model might include a background rate of OR disruptions for all primates, with additional sets of OR genes becoming unnecessary at various time points during human evolution. Unfortunately, without additional information, it is difficult to make inferences about such a model. However, by assuming no background rate of OR gene disruptions in our calculation, our estimate is an upper bound on the time since humans experienced relaxed evolutionary constraint relative to other primates. Hence, we are able to exclude the possibility that humans have been accumulating OR pseudogenes at a neutral rate since the divergence of these two species.

Another assumption made in this analysis was that any OR gene with an uninterrupted ORF is intact. This approach probably results in an underestimate of the proportion of pseudogenes, as not all OR genes with an intact coding region are functional. Mutations in promoter or control regions of OR genes may lead to reduced or no expression. Similarly, radical missense mutations in highly conserved positions of the OR protein may result in dysfunction (Young et al. 2002; Menashe et al. 2003). Although it is known that there are several highly conserved positions among OR genes, it is not always straightforward to ascertain which, if any, of these positions is necessary to retain function. Some changes will alter, rather than completely abolish the function of the receptor (Gaillard et al. 2004). We therefore chose the most straightforward definition of a pseudogene, a gene without a full open reading frame. In the future, this analysis may be repeated using a wider definition of OR pseudogenes, for instance through the use of the tool provided by Menashe et al. (Menashe et al. 2006).

2.3. Positive selection on OR genes

We analyzed, using a maximum-likelihood based methodology, 201 human-chimpanzee-mouse ortholog trios. We find 52 OR genes whose phylogenetic trees are significantly more likely under a model where Dn/Ds varies among evolutionary lineages. A significant LRT result could reflect differences in selective constraint between orthologs, or might result from positive selection acting on an OR gene in only one of the lineages. By inspecting the data further, we highlighted several OR genes, both in human and in chimpanzee, which have probably evolved under positive selection. These OR genes have experienced, on average, seven amino acid substitutions per gene. Interestingly, in some cases (OR4D11 and OR1P1P in human, and OR1L8 in chimpanzee), we find that amino acid substitutions occurred in the putative binding site of the OR protein (Man et al. 2004). These changes may have functional significance. However, in the other OR proteins that are inferred to have evolved under selection, amino acid substitutions are scattered with no clear pattern. In addition, we find several substitutions in positions that are otherwise extremely conserved across OR proteins (such as the DRY motif) (Buck and Axel 1991). Substitutions in these positions may result in a dysfunctional receptor (Young et al. 2002). We are unable to provide a satisfactory explanation for our observation of Dn/Ds ratios well above one for these genes.

3. Statement of independent collaboration

The research in the first project described in this thesis was the product of my independent efforts. The second project described herein is the product of a collaboration between myself and Dr. Yoav Gilad (then at Yale University, New Haven, Connecticut, USA) and Dr. Gustavo Glusman (Institute of Systems Biology, Seattle, Washington, USA). The contribution of each of the participants was as follows. Dr. Yoav Gilad conceived the idea for the analysis and performed the calculation of the time since human rapid accumulation of OR pseudogenes began. Dr. Gustavo Glusman identified the chimpanzee OR genes in the chimpanzee genome draft, as well as the human-chimpanzee ortholog pairs. I performed the phylogenetic analysis, the identification of shared and human-specific pseudogenes, the bootstrap analysis for the estimation of the time since human rapid accumulation of OR pseudogenes began, and the PAML analysis for the identification of ORs evolving under positive selection.

4. Literature

1. Literature Cited

Alexeyenko, A., I. Tamas, G. Liu and E. L. Sonnhammer (2006). "Automatic clustering of orthologs and inparalogs shared by multiple proteomes." Bioinformatics **22**(14): e9-15.

Arava, Y., Y. Wang, J. D. Storey, C. L. Liu, P. O. Brown and D. Herschlag (2003). "Genome-wide analysis of mRNA translation profiles in Saccharomyces cerevisiae." Proc Natl Acad Sci U S A **100**(7): 3889-3894.

Arnaud, M. B., M. C. Costanzo, M. S. Skrzypek, G. Binkley, C. Lane, S. R. Miyasato and G. Sherlock "Candida Genome Database" http://www.candidagenome.org/.

Avidor-Reiss, T., A. M. Maer, E. Koundakjian, A. Polyanovsky, T. Keil, S. Subramaniam and C. S. Zuker (2004). "Decoding cilia function: defining specialized genes required for compartmentalized cilia biogenesis." Cell **117**(4): 527-539.

Bakewell, M. A., P. Shi and J. Zhang (2007). "More genes underwent positive selection in chimpanzee evolution than in human evolution." Proc Natl Acad Sci U S A **104**(18): 7489-7494.

Balakrishnan, R., K. R. Christie, M. C. Costanzo, K. Dolinski, S. S. Dwight, S. R. Engel, D. G. Fisk, J. E. Hirschman, E. L. Hong, R. Nash, R. Oughtred, M. Skrzypek, C. L. Theesfeld, G. Binkley, Q. Dong, C. Lane, A. Sethuraman, S. Weng, D. Botstein and J. M. Cherry "Saccharomyces Genome Database" ftp://ftp.yeastgenome.org/yeast/.

Barnett, J. A. and K. D. Entian (2005). "A history of research on yeasts 9: regulation of sugar metabolism." Yeast **22**(11): 835-894.

Barth, G. and C. Gaillardin (1997). "Physiology and genetics of the dimorphic fungus Yarrowia lipolytica." FEMS Microbiol Rev **19**(4): 219-237.

Belle, A., A. Tanay, L. Bitincka, R. Shamir and E. K. O'Shea (2006). "Quantification of protein half-lives in the budding yeast proteome." Proc Natl Acad Sci U S A **103**(35): 13004-13009.

Benjamini, Y. and Y. Hochberg (1995). "Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing." Journal of the Royal Statistical Society Series B-Methodological **57**(1): 289-300.

Berbee, M. and J. Taylor (2001). Systematics and evolution. The Mycota. D. McLaughlin, E. McLaughlin and P. Lemke. Berlin, Springer. **VIIB:** 229-245.

Bernardi, G. (2000). "Isochores and the evolutionary genomics of vertebrates." Gene **241**(1): 3-17.

Boutros, P. C. and A. B. Okey (2005). "Unsupervised pattern recognition: an introduction to the whys and wherefores of clustering microarray data." Brief Bioinform **6**(4): 331-343.

Braun, B. R., M. van Het Hoog, C. d'Enfert, M. Martchenko, J. Dungan, A. Kuo, D. O. Inglis, M. A. Uhl, H. Hogues, M. Berriman, M. Lorenz, A. Levitin, U. Oberholzer, C. Bachewich, D. Harcus, A. Marcil, D. Dignard, T. Iouk, R. Zito, L. Frangeul, F. Tekaia, K. Rutherford, E. Wang, C. A. Munro, S. Bates, N. A. Gow, L. L. Hoyer, G. Kohler, J. Morschhauser, G. Newport, S. Znaidi, M. Raymond, B. Turcotte, G. Sherlock, M. Costanzo, J. Ihmels, J. Berman, D.

Sanglard, N. Agabian, A. P. Mitchell, A. D. Johnson, M. Whiteway and A. Nantel (2005). "A human-curated annotation of the Candida albicans genome." PLoS Genet **1**(1): 36-57.

Brunet, M., F. Guy, D. Pilbeam, H. T. Mackaye, A. Likius, D. Ahounta, A. Beauvilain, C. Blondel, H. Bocherens, J. R. Boisserie, L. De Bonis, Y. Coppens, J. Dejax, C. Denys, P. Duringer, V. Eisenmann, G. Fanone, P. Fronty, D. Geraads, T. Lehmann, F. Lihoreau, A. Louchart, A. Mahamat, G. Merceron, G. Mouchelin, O. Otero, P. Pelaez Campomanes, M. Ponce De Leon, J. C. Rage, M. Sapanet, M. Schuster, J. Sudre, P. Tassy, X. Valentin, P. Vignaud, L. Viriot, A. Zazzo and C. Zollikofer (2002). "A new hominid from the Upper Miocene of Chad, Central Africa." Nature **418**(6894): 145-151.

Buck, L. and R. Axel (1991). "A novel multigene family may encode odorant receptors: a molecular basis for odor recognition." Cell **65**(1): 175-187.

Bustamante, C. D., A. Fledel-Alon, S. Williamson, R. Nielsen, M. T. Hubisz, S. Glanowski, D. M. Tanenbaum, T. J. White, J. J. Sninsky, R. D. Hernandez, D. Civello, M. D. Adams, M. Cargill and A. G. Clark (2005). "Natural selection on protein-coding genes in the human genome." Nature **437**(7062): 1153-1157.

Carbone, A. and R. Madden (2005). "Insights on the evolution of metabolic networks of unicellular translationally biased organisms from transcriptomic data and sequence analysis." J Mol Evol **61**(4): 456-469.

Carbone, A., A. Zinovyev and F. Kepes (2003). "Codon adaptation index as a measure of dominating codon bias." Bioinformatics **19**(16): 2005-2015.

Chen, F. C. and W. H. Li (2001). "Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees." Am J Hum Genet **68**(2): 444-456.

Chenna, R., H. Sugawara, T. Koike, R. Lopez, T. J. Gibson, D. G. Higgins and J. D. Thompson (2003). "Multiple sequence alignment with the Clustal series of programs." Nucleic Acids Res **31**(13): 3497-3500.

Clark, A. G., S. Glanowski, R. Nielsen, P. D. Thomas, A. Kejariwal, M. A. Todd, D. M. Tanenbaum, D. Civello, F. Lu, B. Murphy, S. Ferriera, G. Wang, X. Zheng, T. J. White, J. J. Sninsky, M. D. Adams and M. Cargill (2003). "Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios." Science **302**(5652): 1960-1963.

Cliften, P., P. Sudarsanam, A. Desikan, L. Fulton, B. Fulton, J. Majors, R. Waterston, B. A. Cohen and M. Johnston (2003). "Finding functional features in Saccharomyces genomes by phylogenetic footprinting." Science **301**(5629): 71-76.

Coghlan, A. and K. H. Wolfe (2000). "Relationship of codon bias to mRNA concentration and protein length in Saccharomyces cerevisiae." Yeast **16**(12): 1131-1145.

Comeron, J. M. (2004). "Selective and mutational patterns associated with gene expression in humans: influences on synonymous composition and intron presence." Genetics **167**(3): 1293-1304.

Comeron, J. M. (2006). "Weak selection and recent mutational changes influence polymorphic synonymous mutations in humans." Proc Natl Acad Sci U S A **103**(18): 6940-6945.

Demuth, J. P., T. D. Bie, J. E. Stajich, N. Cristianini and M. W. Hahn (2006). "The evolution of Mammalian gene families." PLoS ONE **1**: e85.

Dittmar, K. A., J. M. Goodenbour and T. Pan (2006). "Tissue-Specific Differences in Human Transfer RNA Expression." PLoS Genet **2**(12): e221.

dos Reis, M., R. Savva and L. Wernisch (2004). "Solving the riddle of codon usage preferences: a test for translational selection." Nucleic Acids Res **32**(17): 5036-5044.

Dujon, B., D. Sherman, G. Fischer, P. Durrens, S. Casaregola, I. Lafontaine, J. De Montigny, C. Marck, C. Neuveglise, E. Talla, N. Goffard, L. Frangeul, M. Aigle, V. Anthouard, A. Babour, V. Barbe, S. Barnay, S. Blanchin, J. M. Beckerich, E. Beyne, C. Bleykasten, A. Boisrame, J. Boyer, L. Cattolico, F. Confanioleri, A. De Daruvar, L. Despons, E. Fabre, C. Fairhead, H. Ferry-Dumazet, A. Groppi, F. Hantraye, C. Hennequin, N. Jauniaux, P. Joyet, R. Kachouri, A. Kerrest, R. Koszul, M. Lemaire, I. Lesur, L. Ma, H. Muller, J. M. Nicaud, M. Nikolski, S. Oztas, O. Ozier-Kalogeropoulos, S. Pellenz, S. Potier, G. F. Richard, M. L. Straub, A. Suleau, D. Swennen, F. Tekaia, M. Wesolowski-Louvel, E. Westhof, B. Wirth, M. Zeniou-Meyer, I. Zivanovic, M. Bolotin-Fukuhara, A. Thierry, C. Bouchier, B. Caudron, C. Scarpelli, C. Gaillardin, J. Weissenbach, P. Wincker and J. L. Souciet (2004). "Genome evolution in yeasts." Nature **430**(6995): 35-44.

Duret, L. and D. Mouchiroud (1999). "Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, and Arabidopsis." Proc Natl Acad Sci U S A **96**(8): 4482-4487.

Escalante, A. A., E. Barrio and F. J. Ayala (1995). "Evolutionary origin of human and primate malarias: evidence from the circumsporozoite protein gene." Mol Biol Evol **12**(4): 616-626.

Eyre-Walker, A. and L. D. Hurst (2001). "The evolution of isochores." Nat Rev Genet **2**(7): 549-555.

Fraser, H. B., A. E. Hirsh, D. P. Wall and M. B. Eisen (2004). "Coevolution of gene expression among interacting proteins." Proc Natl Acad Sci U S A **101**(24): 9033-9038.

Friberg, M., P. von Rohr and G. Gonnet (2004). "Limitations of codon adaptation index and other coding DNA-based features for prediction of protein expression in Saccharomyces cerevisiae." Yeast **21**(13): 1083-1093.

Gaillard, I., S. Rouquier, A. Chavanieu, P. Mollard and D. Giorgi (2004). "Amino-acid changes acquired during evolution by olfactory receptor 912-93 modify the specificity of odorant recognition." Hum Mol Genet **13**(7): 771-780.

Galagan, J. E., S. E. Calvo, C. Cuomo, L. J. Ma, J. R. Wortman, S. Batzoglou, S. I. Lee, M. Basturkmen, C. C. Spevak, J. Clutterbuck, V. Kapitonov, J. Jurka, C. Scazzocchio, M. Farman, J. Butler, S. Purcell, S. Harris, G. H. Braus, O. Draht, S. Busch, C. D'Enfert, C. Bouchier, G. H. Goldman, D. Bell-Pedersen, S. Griffiths-Jones, J. H. Doonan, J. Yu, K. Vienken, A. Pain, M. Freitag, E. U. Selker, D. B. Archer, M. A. Penalva, B. R. Oakley, M. Momany, T. Tanaka, T. Kumagai, K. Asai, M. Machida, W. C. Nierman, D. W. Denning, M. Caddick, M. Hynes, M. Paoletti, R. Fischer, B. Miller, P. Dyer, M. S. Sachs, S. A. Osmani and B. W. Birren (2005). "Sequencing of Aspergillus nidulans and comparative analysis with A. fumigatus and A. oryzae." Nature **438**(7071): 1105-1115.

Galtier, N., M. Gouy and C. Gautier (1996). "SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny." Comput Appl Biosci **12**(6): 543-548.

Gasch, A. P., P. T. Spellman, C. M. Kao, O. Carmel-Harel, M. B. Eisen, G. Storz, D. Botstein and P. O. Brown (2000). "Genomic expression programs in the response of yeast cells to environmental changes." Mol Biol Cell **11**(12): 4241-4257.

Ghaemmaghami, S., W. K. Huh, K. Bower, R. W. Howson, A. Belle, N. Dephoure, E. K. O'Shea and J. S. Weissman (2003). "Global analysis of protein expression in yeast." Nature **425**(6959): 737-741.

Gilad, Y., C. D. Bustamante, D. Lancet and S. Paabo (2003a). "Natural selection on the olfactory receptor gene family in humans and chimpanzees." Am J Hum Genet **73**(3): 489-501.

Gilad, Y., O. Man, S. Paabo and D. Lancet (2003b). "Human specific loss of olfactory receptor genes." Proc Natl Acad Sci U S A **100**(6): 3324-3327.

Gilad, Y., A. Oshlack, G. K. Smyth, T. P. Speed and K. P. White (2006). "Expression profiling in primates reveals a rapid evolution of human transcription factors." Nature **440**(7081): 242-245.

Glusman, G., A. Bahar, D. Sharon, Y. Pilpel, J. White and D. Lancet (2000). "The olfactory receptor gene superfamily: data mining, classification, and nomenclature." Mamm Genome **11**(11): 1016-1023.

Glusman, G., I. Yanai, I. Rubin and D. Lancet (2001). "The complete human olfactory subgenome." Genome Res **11**(5): 685-702.

Godfrey, P. A., B. Malnic and L. B. Buck (2004). "The mouse olfactory receptor gene family." Proc Natl Acad Sci U S A **101**(7): 2156-2161.

Goetz, R. M. and A. Fuglsang (2005). "Correlation of codon bias measures with mRNA levels: analysis of transcriptome data from Escherichia coli." Biochem Biophys Res Commun **327**(1): 4-7.

Goodall, J. (1964). "Tool-Using And Aimed Throwing In A Community Of Free-Living Chimpanzees." Nature **201**: 1264-1266.

Greenbaum, D., R. Jansen and M. Gerstein (2002). "Analysis of mRNA expression and protein abundance data: an approach for the comparison of the enrichment of features in the cellular population of proteins and transcripts." Bioinformatics **18**(4): 585-596.

Gygi, S. P., Y. Rochon, B. R. Franza and R. Aebersold (1999). "Correlation between protein and mRNA abundance in yeast." Mol Cell Biol **19**(3): 1720-1730.

Hamblin, M. T. and A. Di Rienzo (2000). "Detection of the signature of natural selection in humans: evidence from the Duffy blood group locus." Am J Hum Genet **66**(5): 1669-1679.

Hamblin, M. T., E. E. Thompson and A. Di Rienzo (2002). "Complex signatures of natural selection at the Duffy blood group locus." Am J Hum Genet **70**(2): 369-383.

Harris, M. A., J. Clark, A. Ireland, J. Lomax, M. Ashburner, R. Foulger, K. Eilbeck, S. Lewis, B. Marshall, C. Mungall, J. Richter, G. M. Rubin, J. A. Blake, C. Bult, M. Dolan, H. Drabkin, J. T. Eppig, D. P. Hill, L. Ni, M. Ringwald, R. Balakrishnan, J. M. Cherry, K. R. Christie, M. C. Costanzo, S. S. Dwight, S. Engel, D. G. Fisk, J. E. Hirschman, E. L. Hong, R. S. Nash, A. Sethuraman, C. L. Theesfeld, D. Botstein, K. Dolinski, B. Feierbach, T. Berardini, S. Mundodi, S. Y. Rhee, R. Apweiler, D. Barrell, E. Camon, E. Dimmer, V. Lee, R. Chisholm, P. Gaudet, W. Kibbe, R. Kishore, E. M. Schwarz, P. Sternberg, M. Gwinn, L. Hannick, J. Wortman, M. Berriman, V. Wood, N. de la Cruz, P. Tonellato, P. Jaiswal, T. Seigfried and R. White (2004). "The Gene Ontology

(GO) database and informatics resource." Nucleic Acids Res **32**(Database issue): D258-261.
Harrison, P., A. Kumar, N. Lan, N. Echols, M. Snyder and M. Gerstein (2002). "A small reservoir of disabled ORFs in the yeast genome and its implications for the dynamics of proteome evolution." J Mol Biol **316**(3): 409-419.
Hatt, H. (2004). "Molecular and cellular basis of human olfaction." Chem Biodivers **1**(12): 1857-1869.
Hertz-Fowler, C., C. S. Peacock, V. Wood, M. Aslett, A. Kerhornou, P. Mooney, A. Tivey, M. Berriman, N. Hall, K. Rutherford, J. Parkhill, A. C. Ivens, M. A. Rajandream and B. Barrell (2004). "GeneDB: a resource for prokaryotic and eukaryotic organisms." Nucleic Acids Res **32**(Database issue): D339-343.
Holstege, F. C., E. G. Jennings, J. J. Wyrick, T. I. Lee, C. J. Hengartner, M. R. Green, T. R. Golub, E. S. Lander and R. A. Young (1998). "Dissecting the regulatory circuitry of a eukaryotic genome." Cell **95**(5): 717-728.
Ihmels, J., S. Bergmann, M. Gerami-Nejad, I. Yanai, M. McClellan, J. Berman and N. Barkai (2005). "Rewiring of the yeast transcriptional network through the evolution of motif usage." Science **309**(5736): 938-940.
Ikemura, T. (1981). "Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the E. coli translational system." J Mol Biol **151**(3): 389-409.
Ikemura, T. (1982). "Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes. Differences in synonymous codon choice patterns of yeast and Escherichia coli with reference to the abundance of isoaccepting transfer RNAs." J Mol Biol **158**(4): 573-597.
Jansen, R., H. J. Bussemaker and M. Gerstein (2003). "Revisiting the codon adaptation index from a whole-genome perspective: analyzing the relationship between gene expression and codon occurrence in yeast using a variety of models." Nucleic Acids Res **31**(8): 2242-2251.
Kajiya, K., K. Inaki, M. Tanaka, T. Haga, H. Kataoka and K. Touhara (2001). "Molecular bases of odor discrimination: Reconstitution of olfactory receptors that recognize overlapping sets of odorants." J Neurosci **21**(16): 6018-6025.
Kanaya, S., Y. Yamada, M. Kinouchi, Y. Kudo and T. Ikemura (2001). "Codon usage and tRNA genes in eukaryotes: correlation of codon usage diversity with translation efficiency and with CG-dinucleotide usage as assessed by multivariate analysis." J Mol Evol **53**(4-5): 290-298.
Kanaya, S., Y. Yamada, Y. Kudo and T. Ikemura (1999). "Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of Bacillus subtilis tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis." Gene **238**(1): 143-155.
Kanz, C., P. Aldebert, N. Althorpe, W. Baker, A. Baldwin, K. Bates, P. Browne, A. van den Broek, M. Castro, G. Cochrane, K. Duggan, R. Eberhardt, N. Faruque, J. Gamble, F. G. Diez, N. Harte, T. Kulikova, Q. Lin, V. Lombard, R. Lopez, R. Mancuso, M. McHale, F. Nardone, V. Silventoinen, S. Sobhany, P. Stoehr, M. A. Tuli, K. Tzouvara, R. Vaughan, D. Wu, W. Zhu and R. Apweiler (2005). "The EMBL Nucleotide Sequence Database." Nucleic Acids Res **33**(Database issue): D29-33.

Kellis, M., B. W. Birren and E. S. Lander (2004). "Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae." Nature **428**(6983): 617-624.
King, M. C. and A. C. Wilson (1975). "Evolution at two levels in humans and chimpanzees." Science **188**(4184): 107-116.
Kotlar, D. and Y. Lavner (2006). "The action of selection on codon bias in the human genome is related to frequency, complexity, and chronology of amino acids." BMC Genomics **7**: 67.
Lafay, B., A. T. Lloyd, M. J. McLean, K. M. Devine, P. M. Sharp and K. H. Wolfe (1999). "Proteome composition and codon usage in spirochaetes: species-specific and DNA strand-specific mutational biases." Nucleic Acids Res **27**(7): 1642-1649.
Linardopoulou, E., H. C. Mefford, O. Nguyen, C. Friedman, G. van den Engh, D. G. Farwell, M. Coltrera and B. J. Trask (2001). "Transcriptional activity of multiple copies of a subtelomerically located olfactory receptor gene that is polymorphic in number and location." Hum Mol Genet **10**(21): 2373-2383.
Lithwick, G. and H. Margalit (2005). "Relative predicted protein levels of functionally associated proteins are conserved across organisms." Nucleic Acids Res **33**(3): 1051-1057.
Lowe, T. M. and S. R. Eddy (1997). "tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence." Nucleic Acids Res **25**(5): 955-964.
MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. Fifth Berkeley symposium on mathematical statistics and probability, University of California Press, Berkeley.
Malnic, B., P. A. Godfrey and L. B. Buck (2004). "The human olfactory receptor gene family." Proc Natl Acad Sci U S A **101**(8): 2584-2589.
Malnic, B., J. Hirono, T. Sato and L. B. Buck (1999). "Combinatorial receptor codes for odors." Cell **96**(5): 713-723.
Man, O., Y. Gilad and D. Lancet (2004). "Prediction of the odorant binding site of olfactory receptor proteins by human-mouse comparisons." Protein Sci **13**(1): 240-254.
Marashi, S. A. and H. S. Najafabadi (2004). "How reliable re-adjustment is: correspondence regarding A. Fuglsang, "The 'effective number of codons' revisited"." Biochem Biophys Res Commun **324**(1): 1-2.
Menashe, I., R. Aloni and D. Lancet (2006). "A probabilistic classifier for olfactory receptor pseudogenes." BMC Bioinformatics **7**: 393.
Menashe, I., O. Man, D. Lancet and Y. Gilad (2003). "Different noses for different people." Nat Genet **34**(2): 143-144.
Mewes, H. W., D. Frishman, U. Guldener, G. Mannhaupt, K. Mayer, M. Mokrejs, B. Morgenstern, M. Munsterkotter, S. Rudd and B. Weil (2002). "MIPS: a database for genomes and protein sequences." Nucleic Acids Res **30**(1): 31-34.
Montaner, D., J. Tarraga, J. Huerta-Cepas, J. Burguet, J. M. Vaquerizas, L. Conde, P. Minguez, J. Vera, S. Mukherjee, J. Valls, M. A. Pujana, E. Alloza, J. Herrero, F. Al-Shahrour and J. Dopazo (2006). "Next station in microarray data analysis: GEPAS." Nucleic Acids Res **34**(Web Server issue): W486-491.
Moriyama, E. N. and J. R. Powell (1997). "Codon usage bias and tRNA abundance in Drosophila." J Mol Evol **45**(5): 514-523.

Musto, H., S. Cruveiller, G. D'Onofrio, H. Romero and G. Bernardi (2001). "Translational selection on codon usage in Xenopus laevis." Mol Biol Evol **18**(9): 1703-1707.
Nantel, A. (2006). "The long hard road to a completed Candida albicans genome." Fungal Genet Biol **43**(5): 311-315.
Nei, M., I. B. Rogozin and H. Piontkivska (2000). "Purifying selection and birth-and-death evolution in the ubiquitin gene family." Proc Natl Acad Sci U S A **97**(20): 10866-10871.
Newman, T. and B. J. Trask (2003). "Complex evolution of 7E olfactory receptor genes in segmental duplications." Genome Res **13**(5): 781-793.
Ohama, T., A. Muto and S. Osawa (1990). "Role of GC-biased mutation pressure on synonymous codon choice in Micrococcus luteus, a bacterium with a high genomic GC-content." Nucleic Acids Res **18**(6): 1565-1569.
Olender, T., T. Fuchs, C. Linhart, R. Shamir, M. Adams, F. Kalush, M. Khen and D. Lancet (2004). "The canine olfactory subgenome." Genomics **83**(3): 361-372.
Ollomo, B., S. Karch, P. Bureau, N. Elissa, A. J. Georges and P. Millet (1997). "Lack of malaria parasite transmission between apes and humans in Gabon." Am J Trop Med Hyg **56**(4): 440-445.
Olson, M. V. (1999). "When less is more: gene loss as an engine of evolutionary change." Am J Hum Genet **64**(1): 18-23.
Olson, M. V. and A. Varki (2003). "Sequencing the chimpanzee genome: insights into human evolution and disease." Nat Rev Genet **4**(1): 20-28.
Pan, A., C. Dutta and J. Das (1998). "Codon usage in highly expressed genes of Haemophillus influenzae and Mycobacterium tuberculosis: translational selection versus mutational bias." Gene **215**(2): 405-413.
Parry, C. M., A. Erkner and J. le Coutre (2004). "Divergence of T2R chemosensory receptor families in humans, bonobos, and chimpanzees." Proc Natl Acad Sci U S A **101**(41): 14830-14834.
Pearson, W. R., T. Wood, Z. Zhang and W. Miller (1997). "Comparison of DNA sequences with protein sequences." Genomics **46**(1): 24-36.
Percudani, R., A. Pavesi and S. Ottonello (1997). "Transfer RNA gene redundancy and translational selection in Saccharomyces cerevisiae." J Mol Biol **268**(2): 322-330.
Pilpel, Y. and D. Lancet (1999). "The variable and conserved interfaces of modeled olfactory receptor proteins." Protein Sci **8**(5): 969-977.
Powers, D. A. and P. M. Schulte (1998). "Evolutionary adaptations of gene structure and expression in natural populations in relation to a changing environment: a multidisciplinary approach to address the million-year saga of a small fish." J Exp Zool **282**(1-2): 71-94.
Preiss, T. and M. W. Hentze (2003). "Starting the protein synthesis machine: eukaryotic translation initiation." Bioessays **25**(12): 1201-1211.
Prillinger, H., K. Lopandic, W. Schweigkofler, R. Deak, H. J. Aarts, R. Bauer, K. Sterflinger, G. F. Kraus and A. Maraz (2002). "Phylogeny and systematics of the fungi with special reference to the Ascomycota and Basidiomycota." Chem Immunol **81**: 207-295.
Prud'homme, B., N. Gompel and S. B. Carroll (2007). "Emerging principles of regulatory evolution." Proc Natl Acad Sci U S A **104 Suppl 1**: 8605-8612.
Pruess, M., P. Kersey and R. Apweiler (2005). "The Integr8 project--a resource for genomic and proteomic data." In Silico Biol **5**(2): 179-185.

Puente, X. S., A. Gutierrez-Fernandez, G. R. Ordonez, L. W. Hillier and C. Lopez-Otin (2005). "Comparative genomic analysis of human and chimpanzee proteases." Genomics **86**(6): 638-647.

Qin, H., W. B. Wu, J. M. Comeron, M. Kreitman and W. H. Li (2004). "Intragenic spatial patterns of codon usage bias in prokaryotic and eukaryotic genomes." Genetics **168**(4): 2245-2260.

Quignon, P., E. Kirkness, E. Cadieu, N. Touleimat, R. Guyon, C. Renier, C. Hitte, C. Andre, C. Fraser and F. Galibert (2003). "Comparison of the canine and human olfactory receptor gene repertoires." Genome Biol **4**(12): R80.

Remm, M., C. E. Storm and E. L. Sonnhammer (2001). "Automatic clustering of orthologs and in-paralogs from pairwise species comparisons." J Mol Biol **314**(5): 1041-1052.

Rice, J. (1995). Mathematical statistics and data analysis. Belmont, California, Duxbury Press (Wadsworth Publishing Company).

Rouquier, S., A. Blancher and D. Giorgi (2000). "The olfactory receptor gene repertoire in primates and mouse: evidence for reduction of the functional fraction in primates." Proc Natl Acad Sci U S A **97**(6): 2870-2874.

Rouquier, S., S. Taviaux, B. J. Trask, V. Brand-Arpon, G. van den Engh, J. Demaille and D. Giorgi (1998). "Distribution of olfactory receptor genes in the human genome." Nat Genet **18**(3): 243-250.

Schmidt, M. W., A. Houseman, A. R. Ivanov and D. A. Wolf (2007). "Comparative proteomic and transcriptomic profiling of the fission yeast Schizosaccharomyces pombe." Mol Syst Biol **3**: 79.

Segal, E., Y. Fondufe-Mittendorf, L. Chen, A. Thastrom, Y. Field, I. K. Moore, J. P. Wang and J. Widom (2006). "A genomic code for nucleosome positioning." Nature.

Semon, M., J. R. Lobry and L. Duret (2006). "No evidence for tissue-specific adaptation of synonymous codon usage in humans." Mol Biol Evol **23**(3): 523-529.

Serizawa, S., K. Miyamichi, H. Nakatani, M. Suzuki, M. Saito, Y. Yoshihara and H. Sakano (2003). "Negative feedback regulation ensures the one receptor-one olfactory neuron rule in mouse." Science **302**(5653): 2088-2094.

Sharp, P. M. and W. H. Li (1986). "An evolutionary perspective on synonymous codon usage in unicellular organisms." J Mol Evol **24**(1-2): 28-38.

Sharp, P. M. and W. H. Li (1987). "The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications." Nucleic Acids Res **15**(3): 1281-1295.

Sharp, P. M., T. M. Tuohy and K. R. Mosurski (1986). "Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes." Nucleic Acids Res **14**(13): 5125-5143.

Simpson, P. (2007). "The stars and stripes of animal bodies: evolution of regulatory elements mediating pigment and bristle patterns in Drosophila." Trends Genet **23**(7): 350-358.

Stajich, J. E., D. Block, K. Boulez, S. E. Brenner, S. A. Chervitz, C. Dagdigian, G. Fuellen, J. G. Gilbert, I. Korf, H. Lapp, H. Lehvaslaiho, C. Matsalla, C. J. Mungall, B. I. Osborne, M. R. Pocock, P. Schattner, M. Senger, L. D. Stein, E. Stupka, M. D. Wilkinson and E. Birney (2002). "The Bioperl toolkit: Perl modules for the life sciences." Genome Res **12**(10): 1611-1618.

Sugita, T. and T. Nakase (1999). "Non-universal usage of the leucine CUG codon and the molecular phylogeny of the genus Candida." Syst Appl Microbiol **22**(1): 79-86.

Supek, F. and K. Vlahovicek (2005). "Comparison of codon usage measures and their applicability in prediction of microbial gene expressivity." BMC Bioinformatics **6**: 182.

Tanay, A., A. Regev and R. Shamir (2005). "Conservation and evolvability in regulatory networks: the evolution of ribosomal regulation in yeast." Proc Natl Acad Sci U S A **102**(20): 7203-7208.

Tekaia, F., G. Blandin, A. Malpertuy, B. Llorente, P. Durrens, C. Toffano-Nioche, O. Ozier-Kalogeropoulos, E. Bon, C. Gaillardin, M. Aigle, M. Bolotin-Fukuhara, S. Casaregola, J. de Montigny, A. Lepingle, C. Neuveglise, S. Potier, J. Souciet, M. Wesolowski-Louvel and B. Dujon (2000). "Genomic exploration of the hemiascomycetous yeasts: 3. Methods and strategies used for sequence analysis and annotation." FEBS Lett **487**(1): 17-30.

Terai, Y., N. Morikawa, K. Kawakami and N. Okada (2003). "The complexity of alternative splicing of hagoromo mRNAs is increased in an explosively speciated lineage in East African cichlids." Proc Natl Acad Sci U S A **100**(22): 12798-12803.

Terai, Y., O. Seehausen, T. Sasaki, K. Takahashi, S. Mizoiri, T. Sugawara, T. Sato, M. Watanabe, N. Konijnendijk, H. D. Mrosso, H. Tachida, H. Imai, Y. Shichida and N. Okada (2006). "Divergent selection on opsins drives incipient speciation in Lake Victoria cichlids." PLoS Biol **4**(12): e433.

The Chimpanzee Sequencing and Analysis Consortium. "Initial sequence of the chimpanzee genome and comparison with the human genome." (2005). Nature **437**(7055): 69-87.

Travers, K. J., C. K. Patil, L. Wodicka, D. J. Lockhart, J. S. Weissman and P. Walter (2000). "Functional and genomic analyses reveal an essential coordination between the unfolded protein response and ER-associated degradation." Cell **101**(3): 249-258.

Urrutia, A. O. and L. D. Hurst (2001). "Codon usage bias covaries with expression breadth and the rate of synonymous evolution in humans, but this is not evidence for selection." Genetics **159**(3): 1191-1199.

Varenne, S., J. Buc, R. Lloubes and C. Lazdunski (1984). "Translation is a non-uniform process. Effect of tRNA availability on the rate of elongation of nascent polypeptide chains." J Mol Biol **180**(3): 549-576.

Varki, A. and T. K. Altheide (2005). "Comparing the human and chimpanzee genomes: searching for needles in a haystack." Genome Res **15**(12): 1746-1758.

von Mering, C., R. Krause, B. Snel, M. Cornell, S. G. Oliver, S. Fields and P. Bork (2002). "Comparative assessment of large-scale data sets of protein-protein interactions." Nature **417**(6887): 399-403.

Wang, X., W. E. Grus and J. Zhang (2006). "Gene losses during human origins." PLoS Biol **4**(3): e52.

Warner, J. R. (1999). "The economics of ribosome biosynthesis in yeast." Trends Biochem Sci **24**(11): 437-440.

White, B. N., G. M. Tener, J. Holden and D. T. Suzuki (1973). "Analysis of tRNAs during the development of Drosophila." Dev Biol **33**(1): 185-195.

Whiten, A., J. Goodall, W. C. McGrew, T. Nishida, V. Reynolds, Y. Sugiyama, C. E. Tutin, R. W. Wrangham and C. Boesch (1999). "Cultures in chimpanzees." Nature **399**(6737): 682-685.
Wolfe, K. H. and D. C. Shields (1997). "Molecular evidence for an ancient duplication of the entire yeast genome." Nature **387**(6634): 708-713.
Wright, F. (1990). "The 'effective number of codons' used in a gene." Gene **87**(1): 23-29.
Yang, Z. (1997). "PAML: a program package for phylogenetic analysis by maximum likelihood." Comput Appl Biosci **13**(5): 555-556.
Young, J. M., C. Friedman, E. M. Williams, J. A. Ross, L. Tonnes-Priddy and B. J. Trask (2002). "Different evolutionary processes shaped the mouse and human olfactory receptor gene families." Hum Mol Genet **11**(5): 535-546.
Zhang, X. and S. Firestein (2002). "The olfactory receptor gene superfamily of the mouse." Nat Neurosci **5**(2): 124-133.

2. Publications derived from the doctoral research

Gilad, Y.*, O. Man* and G. Glusman (2005). "A comparison of the human and chimpanzee olfactory receptor gene repertoires." Genome Res **15**(2): 224-230.
Man, O., J.L. Sussman and Y. Pilpel (2007). “Examination of the tRNA Adaptation Index as a predictor of protein expression levels.” RECOMB 2005 Workshop on Regulatory Genomics, LNBI 4023 proceedings, Springer-Verlag: 107-118.
Man, O. and Y. Pilpel (2007). "Differential translation efficiency of orthologous genes is involved in phenotypic divergence of yeast species." Nat Genet **39**(3): 415-421.

*Authors contributed equally

3.

4. Appendices

1. Appendix 1 – The tRNA repertoires of the yeast species analyzed

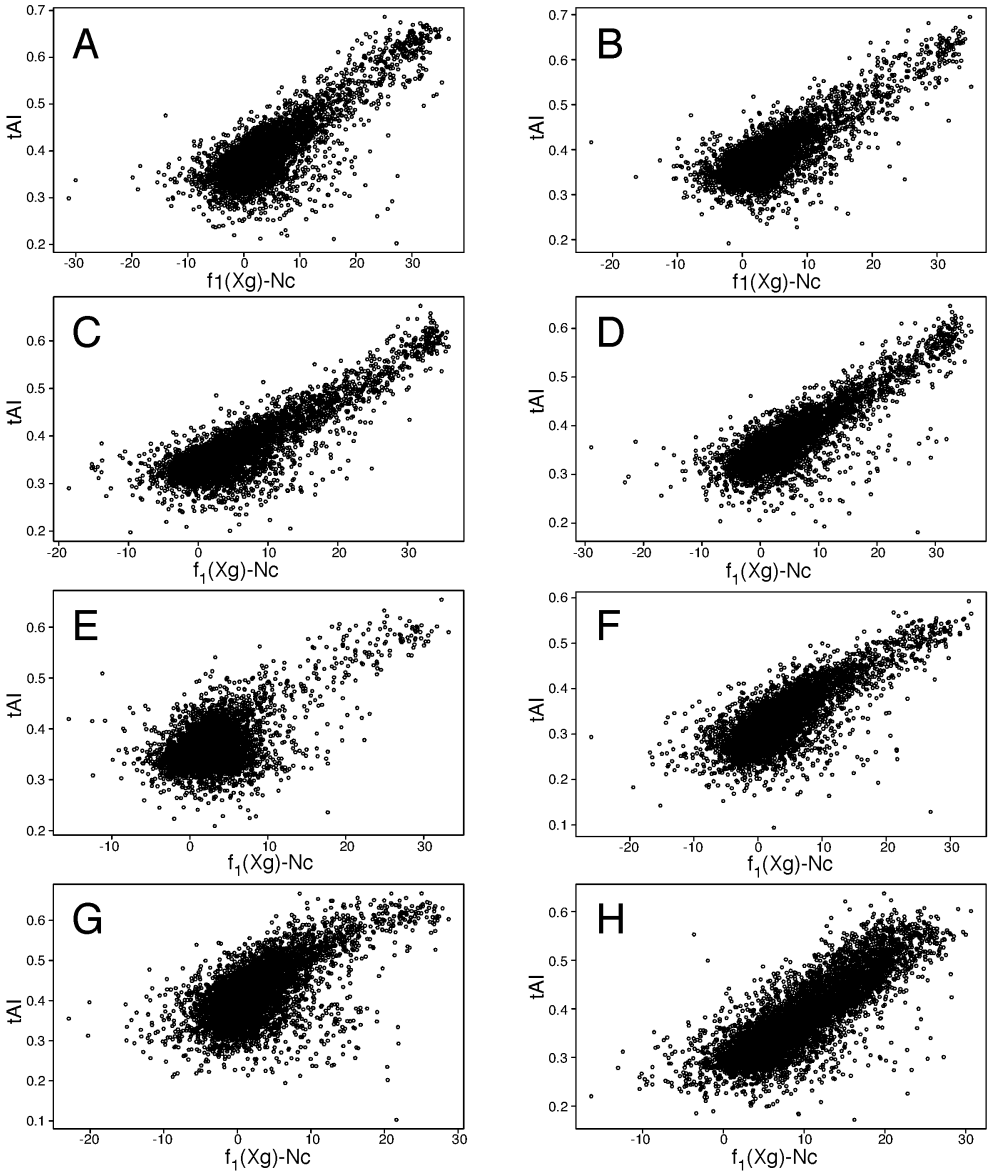
The gene copy numbers of all tRNA species in nine yeasts were determined using an HMM-based approach (Lowe and Eddy 1997) (see Methods). Rows that correspond to the seven tRNAs that are assumed to be absent in all living species (due to the structure of the genetic table and wobble interactions, which imply that their presence may result in mistranslation of some codons) are shown in red. The first three columns show, respectively, the anticodon borne by the tRNA, the codon that is perfectly decoded by the anticodon, and the amino acid that corresponds to the codon. Species abbreviations: Sc – *S. cerevisiae*; Cg – *C. glabrata*; Kl – *K. lactis*; Ag – A. gossypii; Dh – *D. hansenii*; Ca – *C. albicans*; Yl – *Y. lipolytica*; An – *A. nidulans*; Sp – *S. pombe*.

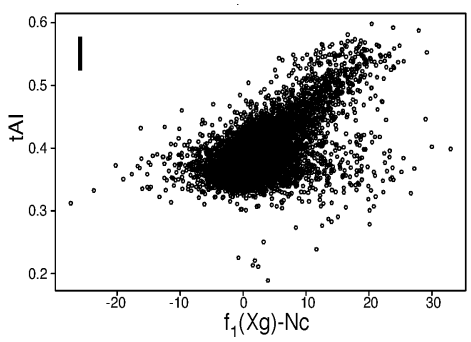
anticodon	codon	amino acid	Sc	Cg	Kl	Ag	Dh	Ca	Yl	An	Sp
AAA	TTT	F	0	0	0	0	0	0	0	0	0
GAA	TTC	F	10	6	5	6	8	5	17	5	5
TAA	TTA	L	7	3	3	2	9	5	1	1	2
CAA	TTG	L	10	8	7	7	4	6	3	2	4
AGA	TCT	S	11	9	6	7	7	4	21	5	7
GGA	TCC	S	0	0	0	0	0	0	0	0	0
TGA	TCA	S	3	2	2	2	3	3	2	1	2
CGA	TCG	S	1	1	1	3	1	1	4	2	1
ATA	TAT	Y	0	0	0	0	0	0	0	0	0
GTA	TAC	Y	8	6	5	4	6	5	14	5	4
ACA	TGT	C	0	0	0	0	0	0	0	0	0
GCA	TGC	C	4	3	3	3	4	2	8	3	3
CCA	TGG	W	6	5	4	5	4	2	13	3	3
AAG	CTT	L	0	0	0	0	2	2	21	6	5
GAG	CTC	L	1	1	1	1	0	0	0	0	0
TAG	CTA	L	3	4	2	4	0	0	2	2	1
CAG	CTG	L	0	0	0	0	1	1	13	3	1
AGG	CCT	P	2	1	1	1	1	1	21	6	6
GGG	CCC	P	0	0	0	0	0	0	0	0	0
TGG	CCA	P	10	7	7	7	7	5	3	2	2
CGG	CCG	P	0	0	0	0	0	0	2	2	1
ATG	CAT	H	0	0	0	0	0	0	0	0	0
GTG	CAC	H	7	6	4	5	5	3	12	5	4
TTG	CAA	Q	9	6	6	4	7	5	3	2	4
CTG	CAG	Q	1	2	1	4	1	1	15	5	2
ACG	CGT	R	6	4	3	4	5	2	1	9	8
GCG	CGC	R	0	0	0	0	0	0	0	0	0
TCG	CGA	R	0	0	0	0	0	0	25	2	1
CCG	CGG	R	1	1	1	2	1	1	0	2	1
AAT	ATT	I	13	9	7	9	9	5	26	7	8
GAT	ATC	I	0	0	0	0	0	0	0	1	0
TAT	ATA	I	2	2	1	1	2	1	1	1	1
CAT	ATG	M	10	7	6	8	7	4	18	7	7
AGT	ACT	T	11	9	6	7	8	6	22	6	7
GGT	ACC	T	0	0	0	0	0	0	0	0	0
TGT	ACA	T	4	3	2	2	2	2	3	2	2
CGT	ACG	T	1	1	1	2	1	1	2	2	1

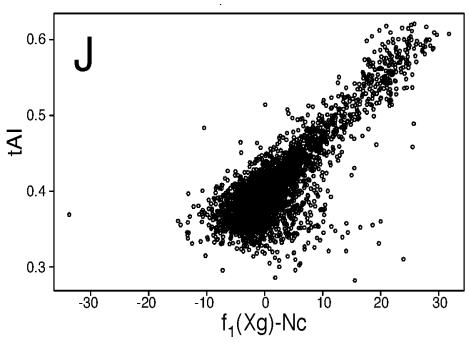
anticodon	codon	amino acid	Sc	Cg	Kl	Ag	Dh	Ca	Yl	An	Sp
ATT	AAT	N	0	0	0	0	0	0	0	0	0
GTT	AAC	N	10	9	6	6	8	4	16	7	6
TTT	AAA	K	7	3	4	3	7	5	4	3	3
CTT	AAG	K	14	12	9	9	12	2	34	8	9
ACT	AGT	S	0	0	0	0	0	0	0	0	0
GCT	AGC	S	2	3	2	3	2	2	6	4	3
TCT	AGA	R	11	9	7	6	10	5	4	2	2
CCT	AGG	R	1	1	1	1	1	1	1	2	1
AAC	GTT	V	14	10	7	8	11	6	24	8	9
GAC	GTC	V	0	0	0	0	0	0	0	0	0
TAC	GTA	V	2	2	1	2	1	1	2	1	2
CAC	GTG	V	2	1	2	4	1	1	8	2	1
AGC	GCT	A	11	10	7	7	7	7	30	8	9
GGC	GCC	A	0	0	0	0	0	0	0	0	0
TGC	GCA	A	5	5	3	4	4	2	4	2	2
CGC	GCG	A	0	0	0	2	1	0	2	3	1
ATC	GAT	D	0	0	0	0	0	0	0	0	0
GTC	GAC	D	16	9	8	10	9	7	28	9	8
TTC	GAA	E	14	9	8	3	9	7	6	3	4
CTC	GAG	E	2	3	2	8	1	1	27	8	6
ACC	GGT	G	0	0	0	0	0	0	0	0	0
GCC	GGC	G	16	12	7	11	11	6	30	11	8
TCC	GGA	G	3	2	2	2	4	2	11	3	3
CCC	GGG	G	2	1	1	2	1	1	0	1	1
Total			273	207	162	191	205	133	510	184	171

2. Appendix 2 – Comparison of the effective number of codons after accounting for silent GC content (f_1(Xg)-Nc) and the tRNA adaptation index (tAI) for the coding sequences of the ten yeast species analyzed

tAI and f_1(Xg)-Nc were computed for all nuclear-encoded coding sequences (see Methods). A. *S. cerevisiae* B. *S. bayanus* C. *C. glabrata* D. *K. lactis* E. *A. gossypii* F. *D. hansenii* G. *C. albicans* H. *Y. lipolytica* I. *A. nidulans* J. *S. pombe*

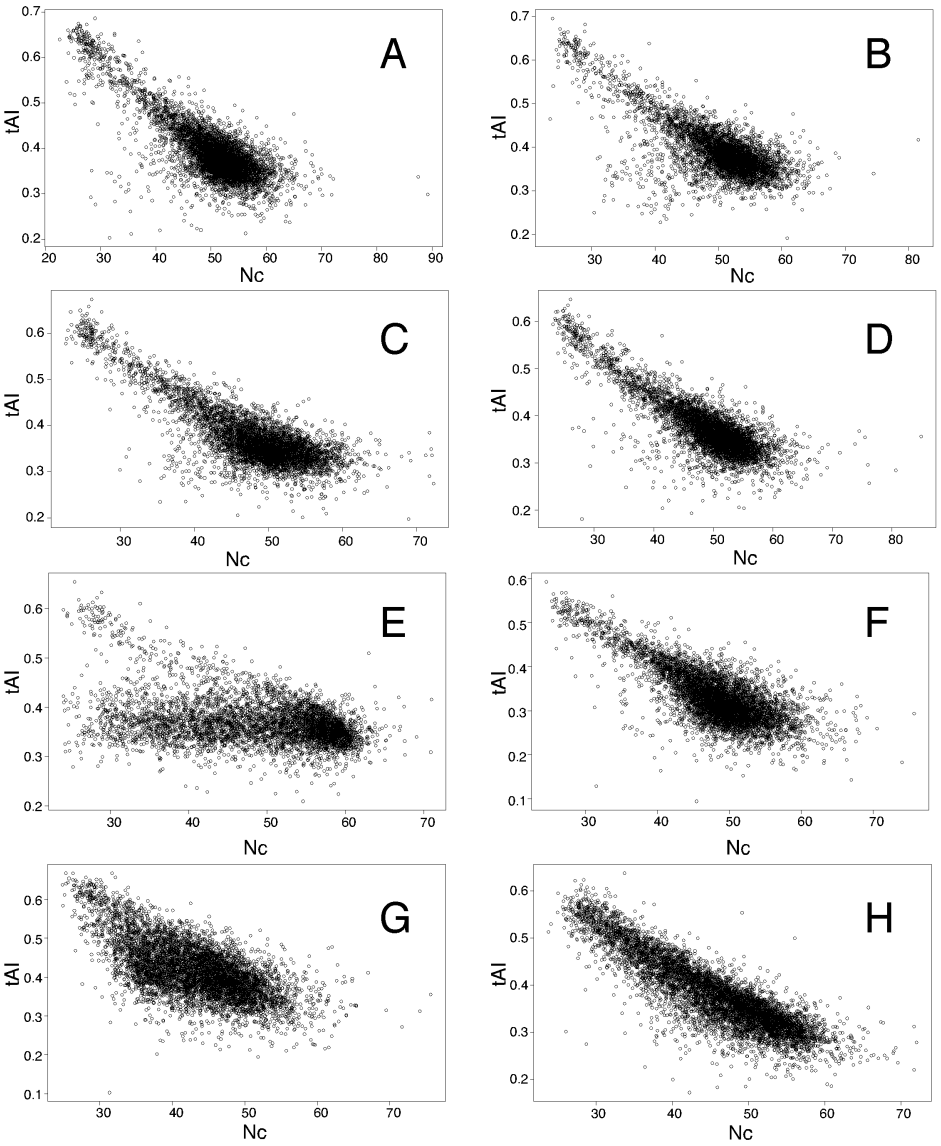


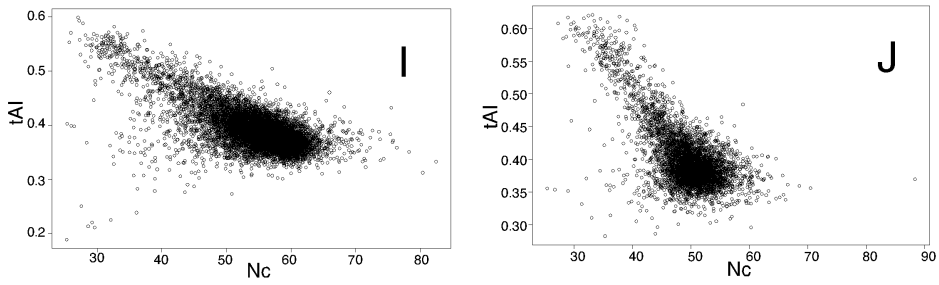

I
tAI
0.6
0.5
0.4
0.3
0.2
-20
-10
0
10
20
30
f1(Xg)-Nc


J
tAI
0.6
0.5
0.4
0.3
-30
-20
-10
0
10
20
30
f1(Xg)-Nc

3. Appendix 3 – comparison of the effective number of codons (Nc) and the tRNA adaptation index (tAI) for the coding sequences of the ten yeast species analyzed

tAI and Nc were computed for all nuclear-encoded coding sequences (see Methods). A. *S. cerevisiae* B. *S. bayanus* C. *C. glabrata* D. *K. lactis* E. *A. gossypii* F. *D. hansenii* G. *C. albicans* H. *Y. lipolytica* I. *A. nidulans* J. *S. pombe*




I
tAI
Nc
0.6
0.5
0.4
0.3
0.2
30
40
50
60
70
80
J
tAI
Nc
0.60
0.55
0.50
0.45
0.40
0.35
0.30
30
40
50
60
70
80
90

4. Appendix 4 – Results of Friedman test and post-hoc analysis for the forty clusters of translation efficiency profiles

For each cluster I performed the Friedman test, and if the p-value was significant (using a FDR (Benjamini and Hochberg 1995) of 5%) I followed with post-hoc tests for all pairwise comparisons. The results of the post-hoc tests are summarized as a species stratification. A '<' sign indicates that each of the species to the left of this sign obtained a significant p-value (using a FDR of 20%) in the comparison of their translation efficiency values with those of each of the species on the right of the sign, and the medians of the values for the species on the left are smaller than those of the species on the right. Cluster numbers are as in Fig. 10.

cluster#	cluster size	Friedman test p-value	conclusion from *post-hoc* tests
1	14	5.04E-12	*K. lactis < Y. lipolytica < S. bayanus < S. cerevisiae, C. glabrata, D. hansenii, A. nidulans, S. pombe < C. albicans*
2	64	<1.11E-16	*Y. lipolytica < S. cerevisiae, S. bayanus < A. nidulans < C. glabrata, K. lactis < D. hansenii < S. pombe < C. albicans*
3	37	<1.11E-16	*S. bayanus < S. cerevisiae, K. lactis, D. hansenii, C. albicans, Y. lipolytica < S. pombe < C. glabrata, A. nidulans*
4	7	7.02E-05	*S. bayanus, C. glabrata, K. lactis, D. hansenii, C. albicans, Y. lipolytica, A. nidulans, S. pombe < S. cerevisiae*
5	40	<1.11E-16	*C. albicans, Y. lipolytica < A. nidulans < S. cerevisiae, S. bayanus < C. glabrata < K. lactis, S. pombe < D. hansenii*
6	11	1.66E-09	*C. glabrata < C. albicans, Y. lipolytica < S. cerevisiae, S. bayanus, K. lactis, D. hansenii, A. nidulans, S. pombe*
7	147	<1.11E-16	*Y. lipolytica < D. hansenii < C. albicans < S. cerevisiae,* S.bayanus, *C. glabrata, K. lactis < S. pombe < A. nidulans*
8	100	<1.11E-16	*D. hansenii < Y. lipolytica < C. glabrata < K. lactis, C. albicans < S. bayanus, A. nidulans < S. cerevisiae,* S.pombe
9	281	<1.11E-16	*C. albicans < Y. lipolytica < D. hansenii < K. lactis < C. glabrata < A. nidulans < S. cerevisiae, S. bayanus, S. pombe*
10	70	<1.11E-16	*D. hansenii < S. bayanus, C. albicans, Y. lipolytica, A. nidulans < S. cerevisiae, C. glabrata, K. lactis < S. pombe*
11	56	<1.11E-16	*C. albicans < K. lactis < D. hansenii < S. cerevisiae, S. bayanus, C. glabrata < Y. lipolytica, A. nidulans < S. pombe*
12	119	<1.11E-16	*Y. lipolytica < K. lactis < C. glabrata, D. hansenii, A. nidulans < C. albicans < S. cerevisiae, S. pombe < S. bayanus*
13	105	<1.11E-16	*Y. lipolytica < A. nidulans < D. hansenii, C. albicans, S. pombe < C. glabrata, K. lactis < S. cerevisiae, S. bayanus*
14	43	<1.11E-16	*A. nidulans < Y. lipolytica, S. pombe < C. glabrata, C. albicans < K. lactis < S. cerevisiae < S. bayanus < D. hansenii*
15	67	<1.11E-16	*Y. lipolytica < S. pombe < C. glabrata, A. nidulans < K. lactis < S. cerevisiae, S. bayanus < D. hansenii < C. albicans*
16	73	<1.11E-16	*A. nidulans < K. lactis, Y. lipolytica < S. cerevisiae, S. bayanus, C. glabrata, D. hansenii, C. albicans < S. pombe*
17	67	<1.11E-16	*K. lactis,* D.hansenii *< A. nidulans < C. albicans, Y. lipolytica, S. pombe < C. glabrata < S. cerevisiae, S. bayanus*
18	20	<1.11E-16	*C. glabrata, K. lactis, D. hansenii, S. pombe < A. nidulans < Y. lipolytica < S. bayanus < S. cerevisiae, C. albicans*
19	45	<1.11E-16	*D. hansenii < K. lactis, Y. lipolytica, A. nidulans < C. glabrata < S. cerevisiae < S. bayanus, S. pombe < C. albicans*
20	26	5.46E-13	*C. glabrata < S. cerevisiae, S. bayanus, K. lactis, D. hansenii, Y. lipolytica, A. nidulans, S. pombe < C. albicans*
21	13	3.60E-11	*K. lactis < S. cerevisiae, C. glabrata, C. albicans, Y. lipolytica, S. pombe < S. bayanus, D. hansenii, A. nidulans*

cluster#	cluster size	Friedman test p-value	conclusion from *post-hoc* tests
22	61	<1.11E-16	*C. glabrata < S. cerevisiae, Y. lipolytica < S. bayanus, K. lactis, S. pombe < D. hansenii < C. albicans < A. nidulans*
23	82	<1.11E-16	*S. pombe < K. lactis < S. cerevisiae, C. albicans < C. glabrata < S. bayanus, D. hansenii < Y. lipolytica < A. nidulans*
24	38	<1.11E-16	*D. hansenii, S. pombe < K. lactis < S. cerevisiae < S. bayanus, C. glabrata, C. albicans, Y. lipolytica < A. nidulans*
25	56	<1.11E-16	*S. bayanus, K. lactis < S. cerevisiae, C. glabrata, A. nidulans < S. pombe < D. hansenii, Y. lipolytica < C. albicans*
26	95	<1.11E-16	*C. glabrata < S. cerevisiae < S. bayanus < K. lactis < D. hansenii, C. albicans, A. nidulans < Y. lipolytica, S. pombe*
27	251	<1.11E-16	*S. bayanus < S. cerevisiae < C. glabrata, S. pombe < K. lactis, D. hansenii, C. albicans, A. nidulans < Y. lipolytica*
28	40	<1.11E-16	*D. hansenii < S. cerevisiae, S. bayanus, C. glabrata, K. lactis, C. albicans, A. nidulans, S. pombe < Y. lipolytica*
29	18	2.22E-16	*C. glabrata, D. hansenii, C. albicans < S. pombe < S. cerevisiae, S. bayanus, K. lactis, A. nidulans < Y. lipolytica*
30	11	2.61E-10	*S. bayanus, D. hansenii, C. albicans, A. nidulans < S. cerevisiae, C. glabrata, S. pombe < K. lactis, Y. lipolytica*
31	26	<1.11E-16	*C. albicans < S. cerevisiae, S. bayanus, S. pombe < D. hansenii, Y. lipolytica, A. nidulans < C. glabrata, K. lactis*
32	60	<1.11E-16	*C. albicans < C. glabrata < S. cerevisiae, S. bayanus, K. lactis, S. pombe < Y. lipolytica, A. nidulans < D. hansenii*
33	87	<1.11E-16	*C. albicans < A. nidulans < S. pombe < K. lactis < C. glabrata < D. hansenii < S. cerevisiae, Y. lipolytica < S. bayanus*
34	7	1.39E-06	*S. cerevisiae, K. lactis < S. bayanus, D. hansenii, C. albicans, Y. lipolytica, A. nidulans < S. pombe < C. glabrata*
35	30	<1.11E-16	*K. lactis, A. nidulans, S. pombe < S. cerevisiae, S. bayanus < D. hansenii, C. albicans, Y. lipolytica < C. glabrata*
36	87	<1.11E-16	*A. nidulans < D. hansenii, S. pombe < S. cerevisiae, S. bayanus, C. glabrata, K. lactis, C. albicans, Y. lipolytica*
37	96	<1.11E-16	*S. pombe < A. nidulans < C. albicans < S. cerevisiae, K. lactis, D. hansenii, Y. lipolytica < C. glabrata < S. bayanus*
38	36	<1.11E-16	*A. nidulans, S. pombe < S. bayanus < Y. lipolytica < S. cerevisiae, D. hansenii < C. glabrata, K. lactis < C. albicans*
39	78	<1.11E-16	*S. pombe < S. cerevisiae, C. glabrata, K. lactis, A. nidulans < S. bayanus < D. hansenii, Y. lipolytica < C. albicans*
40	246	<1.11E-16	*S. pombe < A. nidulans < S. cerevisiae, C. glabrata < S. bayanus < K. lactis < D. hansenii, C. albicans < Y. lipolytica*

5. Appendix 5 – List of M-phase genes most-representative of the ranking of species found by the Friedman test and post-hoc analyses for all the M phase genes

The genes of the M phase of the cell cycle constitute an example of a highly significant pattern that cannot presently be explained by a known pheynotypic difference among the species analyzed. These genes were found to exhibit the following order of translational efficiencies: *Y. lipolytica* < *K. lactis, D. hansenii, C. albicans* < *C. glabrata* < *S. cerevisiae, S. bayanus, A. nidulans, S. pombe*., (Table 3, Fig. 14A). In search for particular M-phase related genes that gave rise to this signal, I clustered (using kmeans (MacQueen 1967)) the 126 translational efficiency profiles of genes associated with this GO category into five clusters. One of the clusters, that contains 33 profiles, resembles the above pattern (Fig. 14B). The members of this cluster are listed below. In most cases each gene represents one profile in the translation efficiency profile. However, in one case two genes (YHR115C and YNL116W), which are highly similar in sequence and redundant in function, were grouped into the same orthologous group, and thus represent the same profile. All data regarding the genes was obtained from the SGD database (Balakrishnan et al.).

Systematic name	Gene	Aliases	Description
YGL183C	MND1		Protein required for recombination and meiotic nuclear division; forms a complex with Hop2p, which is involved in chromosome pairing and repair of meiotic double-strand breaks
YGL194C	HOS2	RTL1	Histone deacetylase required for gene activation via specific deacetylation of lysines in H3 and H4 histone tails; subunit of the Set3 complex, a meiotic-specific repressor of sporulation specific genes that contains deacetylase activity
YNL172W	APC1		Largest subunit of the Anaphase-Promoting Complex/Cyclosome (APC/C), which is a ubiquitin-protein ligase required for degradation of anaphase inhibitors, including mitotic cyclins, during the metaphase/anaphase transition
YDR118W	APC4		Subunit of the Anaphase-Promoting Complex/Cyclosome (APC/C), which is a ubiquitin-protein ligase required for degradation of anaphase inhibitors, including mitotic cyclins, during the metaphase/anaphase transition
YGL003C	CDH1	HCT1	Cell-cycle regulated activator of the anaphase-promoting complex/cyclosome (APC/C), which directs ubiquitination of cyclins resulting in mitotic exit; targets the APC/C to specific substrates including CDC20, ASE1, CIN8 and FIN1
YGR225W	AMA1	SPO70	Activator of meiotic anaphase promoting complex (APC/C); Cdc20p family member; required for initiation of spore wall assembly; required for Clb1p degradation during meiosis
YOR351C	MEK1	MRE4	Meiosis-specific serine/threonine protein kinase, functions in meiotic checkpoint, phosphorylates Red1p, interacts with Hop1p; required for meiotic recombination and normal spore viability

Systematic name	Gene	Aliases	Description
YPR025C	CCL1		Cyclin associated with protein kinase Kin28p, which is the TFIIH-associated carboxy-terminal domain (CTD) kinase involved in transcription initiation at RNA polymerase II promoters
YMR190C	SGS1		Nucleolar DNA helicase of the RecQ family involved in maintenance of genome integrity, regulates chromosome synapsis and meiotic crossing over; has similarity to human BLM and WRN helicases implicated in Bloom and Werner syndromes
YHR039C	MSC7		Protein of unknown function, green fluorescent protein (GFP)-fusion protein localizes to the endoplasmic reticulum; msc7 mutants are defective in directing meiotic recombination events to homologous chromatids
YGR188C	BUB1		Protein kinase that forms a complex with Mad1p and Bub3p that is crucial in the checkpoint mechanism required to prevent cell cycle progression into anaphase in the presence of spindle damage, associates with centromere DNA via Skp1p
YHR115C	DMA1	CHF1	Protein involved in regulating spindle position and orientation, functionally redundant with Dma2p; homolog of S. pombe Dma1 and H. sapiens Chfr
YNL116W	DMA2	CHF2	Protein involved in regulating spindle position and orientation, functionally redundant with Dma1p; homolog of S. pombe Dma1 and H. sapiens Chfr
YLR234W	TOP3	EDR1	DNA Topoisomerase III, conserved protein that functions in a complex with Sgs1p and Rmi1p to relax single-stranded negatively-supercoiled DNA preferentially, involved in telomere stability and regulation of mitotic recombination
YPL022W	RAD1	LPB9	Single-stranded DNA endonuclease (with Rad10p), cleaves single-stranded DNA during nucleotide excision repair and double-strand break repair; subunit of Nucleotide Excision Repair Factor 1 (NEF1); homolog of human XPF protein
YER179W	DMC1	ISC2	Meiosis-specific protein required for repair of double-strand breaks and pairing between homologous chromosomes; homolog of Rad51p and the bacterial RecA protein
YPL122C	TFB2		Subunit of TFIIH and nucleotide excision repair factor 3 complexes, involved in transcription initiation, required for nucleotide excision repair, similar to 52 kDa subunit of human TFIIH
YPL008W	CHL1	CTF1, LPA9	Conserved nuclear protein required to establish sister-chromatid pairing during S-phase, probable DNA helicase with similarity to human BACH1, which associates with tumor suppressor BRCA1; associates with acetyltransferase Ctf7p
YOR058C	ASE1	YOR29-09	Mitotic spindle midzone localized microtubule-associated protein (MAP) family member; required for spindle elongation and stabilization; undergoes cell cycle-regulated degradation by anaphase promoting complex; potential Cdc28p substrate
YDR386W	MUS81	SLX3	Helix-hairpin-helix protein, involved in DNA repair and replication fork stability; functions as an endonuclease in complex with Mms4p; interacts with Rad54p
YML032C	RAD52		Protein that stimulates strand exchange by facilitating Rad51p binding to single-stranded DNA; anneals complementary single-stranded DNA; involved in the repair of double-strand breaks in DNA during vegetative growth and meiosis
YOR368W	RAD17		Checkpoint protein, involved in the activation of the DNA damage and meiotic pachytene checkpoints; with Mec3p and Ddc1p, forms a clamp that is loaded onto partial duplex DNA; homolog of human and S. pombe Rad1 and U. maydis Rec1 proteins
YNL082W	PMS1		ATP-binding protein required for mismatch repair in mitosis and meiosis; functions as a heterodimer with Mlh1p, binds double- and single-stranded DNA via its N-terminal domain, similar to E. coli MutL

Systematic name	Gene	Aliases	Description
YBR160W	CDC28	CDK1, HSL5, SRM5	Catalytic subunit of the main cell cycle cyclin-dependent kinase (CDK); alternately associates with G1 cyclins (CLNs) and G2/M cyclins (CLBs) which direct the CDK to specific substrates
YDR180W	SCC2		Subunit of cohesin loading factor (Scc2p-Scc4p), a complex required for the loading of cohesin complexes onto chromosomes; involved in establishing sister chromatid cohesion during DSB repair via histone H2AX
YPR056W	TFB4		Subunit of TFIIH complex, involved in transcription initiation, similar to 34 kDa subunit of human TFIIH; interacts with Ssl1p
YML095C	RAD10		Single-stranded DNA endonuclease (with Rad1p), cleaves single-stranded DNA during nucleotide excision repair and double-strand break repair; subunit of Nucleotide Excision Repair Factor 1 (NEF1); homolog of human ERCC1 protein
YNL025C	SSN8	GIG3, NUT9, SRB11, UME3, RYE2, CycC	Cyclin-like component of the RNA polymerase II holoenzyme, involved in phosphorylation of the RNA polymerase II C-terminal domain; involved in glucose repression and telomere maintenance.
YER173W	RAD24	RS1	Checkpoint protein, involved in the activation of the DNA damage and meiotic pachytene checkpoints; subunit of a clamp loader that loads Rad17p-Mec3p-Ddc1p onto DNA; homolog of human and S. pombe Rad17 protein
YKL049C	CSE4	CSL2	Centromere protein that resembles histones, required for proper kinetochore function; homolog of human CENP-A
YJR053W	BFA1	IBD1	Component of the GTPase-activating Bfa1p-Bub2p complex involved in multiple cell cycle checkpoint pathways that control exit from mitosis
YOR014W	RTS1	SCS1	B-type regulatory subunit of protein phosphatase 2A (PP2A); homolog of the mammalian B' subunit of PP2A
YBR073W	RDH54	TID1	DNA-dependent ATPase, stimulates strand exchange by modifying the topology of double-stranded DNA; involved in the recombinational repair of double-strand breaks in DNA during mitosis and meiosis; proposed to be involved in crossover interference
YPL194W	DDC1		DNA damage checkpoint protein, part of a PCNA-like complex required for DNA damage response, required for pachytene checkpoint to inhibit cell cycle in response to unrepaired recombination intermediates; potential Cdc28p substrate

6.