Honors Thesis

LCC 4700: Undergraduate Thesis Writing

"The Evolutionary Impact of Functional RNA Secondary Structures within Protein-Coding Regions in Yeast"

Charles Warden

Advisor: Dr. Soojin Yi

Reviewers: Dr. Soojin Yi and Dr. King Jordan

Abstract

In recent years, the importance of functional RNAs in biological processes has been repeatedly demonstrated. Furthermore, a number of studies have indicated that the number and biological significance of functional RNAs has been underestimated. Here we focus on the impact of functional RNAs on the evolutionary rate of protein-coding genes. We used computational methods to predict a set of novel functional RNA secondary structures in the yeast genome. We found that the presence of predicted functional RNAs within coding regions causes a decrease in both synonymous and nonsynonymous evolutionary rates of the corresponding protein. We studied yeast genomes because abundant functional data is available for S*accharomyces cerevisiae*, a common model genetic organism. Furthermore, the set of yeast duplicate genes produced by an ancient whole genome duplication provides a useful model for studying the evolution of gene function in yeast, and results from duplicate gene analysis also demonstrate that the duplicate gene with a greater number of coding functional RNAs also tends to have a relatively slower rate of protein evolution. Moreover, we discovered that genes with a significant portion of functional RNAs within their coding regions tend to be associated with either protein biosynthesis or transport.

Key Words: functional RNA, yeast comparative genomics, comparative genomics, evolutionary rate, gene ontology

Introduction

The central dogma posits that genetic information is encoded by DNA, which is transcribed into RNA, which in turn is translated into protein. Since the conception of the central dogma, there have been numerous exceptions to the flow of genetic information. For example, some genes code for "non-coding" RNA genes that are never transcribed into proteins. In many studies, the term non-coding RNA is often referred to as functional RNA (fRNA). In this study, the term functional RNA more broadly refers to any RNA molecule whose structure can serve some function aside from coding for proteins. Recently, several studies have discovered many ways in which such non-coding RNAs play an important role in biological processes (Storz et al. 2005, Ouellet et al. 2006, Pedersen et al. 2006). Furthermore, a recent study has predicted a surprisingly large proportion of fRNAs *within* protein-coding regions in vertebrates, which requires the DNA primary sequence to *simultaneously* code for fRNA structures and amino acid codons (Pedersen et al. 2006). However, there has been relatively little research on the existence of fRNA secondary structures in yeast. Therefore, studying the distribution of predicted fRNA structures in yeast may provide additional evidence to encourage experimental biologists to investigate the potentially underestimated role for fRNA secondary structures within protein coding regions.

In order to study the distribution of functional RNA secondary structures, we produced a set of predicted folds (based on stringent criteria) that was combined with known folds. This combined dataset was used to analyze 1) the impact of functional RNAs on the evolutionary rates of protein-coding genes in the yeast genome and 2) the influence of gene duplication on the distribution of functional RNAs located in and around duplicate genes produced by the whole genome duplication in yeast. We found

that functional RNAs constrain the evolutionary rate of protein coding regions; this result is observed for both the dataset of all genes in the yeast genome as well as the more selective set of duplicate genes produced by the whole genome duplication in yeast. In particular, there was a negative correlation between percent fRNA coverage and synonymous site evolutionary rates, and this correlation was stronger than for nonsynonymous sites. We performed functional analysis to investigate functional asymmetry in gene duplicates as well as to determine which types of genes are enriched with fRNA secondary structures. We also found that genes associated with protein biosynthesis and transport have an abundance of functional RNA secondary structures within their coding regions. In short, this study provides exciting evidence that functional RNAs *within* protein-coding regions play a more important role than previously thought.

Literature Review

The EvoFold and RNAz programs utilize comparative genomic techniques to produce predictions of functional RNA secondary structures. Most previous fRNA studies have focused on non-coding RNA. Therefore, it will be useful to briefly review previous studies of functional RNAs within non-coding regions in order to contrast coding and non-coding fRNA folds. Because non-coding RNAs often interact with mRNAs, it will be useful to review the possible impact that non-coding RNAs can have on mRNA sequences. Finally, this study focuses on the evolutionary impact of functional RNA secondary structures within protein-coding regions. Therefore, it will obviously be useful to review previous studies that discuss the existence and/or

prevalence of conserved functional RNA secondary structures on messenger RNA strands.

Comparative Genomic Analysis

This project utilizes the EvoFold program to screen the *Saccharomyces cerevisae* genome for functional RNA secondary structures. Evofold is a program that uses comparative genomic analysis to identify RNA sequences that show conserved secondary structure based upon pairs of nucleotides that demonstrate conserved base pairing even if the primary sequence of nucleotides is not conserved among species. Briefly, comparative genomics is a field of computational biology that focuses on the comparison of DNA sequences of various organisms in order to answer the following fundamental question: if certain regions in multiple species' genomes are nearly identical and have thus remained conserved for millions of years, then what vital role must such sequences play in the successful development of each respective organism?

Previous studies have proven that comparative genomic analysis can provide powerful tools to predict the existence of novel genomic structures whose existence can be validated by experimental scientists (Bejerano et al. 2006, Pollard et al. 2006). Thus, if care is taken to produce a sufficiently accurate dataset (including independent verification using a different program), then this project will result in the production of potentially novel fRNA secondary structures whose existence can be verified experimentally in future experiments. Furthermore, *Saccharomyces cerevisae* is an important model organism for genetic studies, so the discovery of novel fRNA secondary structures can provide invaluable information for topics ranging from evolutionary investigations to medical research.

Role of non-coding functional RNAs

A number of RNA genes do not code for proteins and are known as non-coding RNA (ncRNA); non-coding RNA includes ribosomal RNA (rRNA), transfer RNA (tRNA), small nuclear RNA (snRNA), and many categories of other relatively newly discovered ncRNA. There are also a large number of proteins devoted to processing of non-coding RNA (Peng et al. 2003). Thus, the discovery of novel ncRNAs in yeast is a very exciting possibility. In fact, this investigation will not be the first computational screening of non-coding RNAs in yeast (and, more specifically, S*accharomyces cerevisiae*). For example, a study conducted a few years ago used the QRNA program to produce a list of a little less than 100 ncRNA candidate genes, and the manual for the RNAz program described a screening for noncoding RNAs that has not yet been published (McCutcheon and Eddy 2003, Washietl et al. 2005). However, there has not been a whole genome screening of yeast with the EvoFold program (which, based upon preliminary results, is likely to produce a much larger set of candidate genes) and there has been no study to conduct a whole genome screening *simultaneously* in both coding and noncoding regions of the yeast genome.

Role of functional RNA secondary structures in messenger RNAs

Unlike non-coding RNAs, messenger RNAs (mRNAs) are unique in that they are transcribed in order to contain the genetic information for translation of the mRNA to the corresponding protein sequence. However, fRNA secondary structures can play very different roles in gene expression depending if they are found within protein coding exons, non-coding introns, or untranslated regions (UTRs) at the beginning and end of the messenger RNA). As mentioned in the introduction, a recent study has predicted a

surprisingly large proportion of fRNAs *within* protein-coding regions, which requires the DNA primary sequence to *simultaneously* code for fRNA structures and amino acid codons (Pedersen et al. 2006). Thus, it will be especially interesting to observe the impact of fRNA folds predicted within protein coding regions on asymmetric evolution of WGD paralogs.

Synonymous sites are regions of the nucleotide sequence for a protein-coding gene that do not directly affect the protein encoded by the gene. Although synonymous sites were initially thought to evolve neutrally, there has been an increasing amount of evidence that there is selection on such sites though a variety of mechanisms, such as mRNA stability and codon usage bias. In mammals, many studies have indicated the role of mRNA stability and alternative splicing on synonymous sites (Chamary et al. 2006, Chamary and Hurst 2005). There are a number of specific models (such as functional loss, translation efficiency, and translational robustness) that describe an association between gene expression and evolutionary rates, and some of these models may be useful when interpreting selection at synonymous sites (Rocha and Danchin 2004, Akashi 2003, Drummond et al. 2005). In conclusion, my investigation of conserved fRNA secondary structures in coding regions will hopefully help provide evidence to explain the broader phenomenon of selection on synonymous sites in yeast.

Results

Whole Genome Screening for Functional RNAs

Before functional analysis could be conducted, an accurate dataset of predicted fRNA folds had to be produced by independent verification from the EvoFold and RNAz RNA secondary structure prediction programs (see Methods and Supplementary

Material). This optimal set contained 921 predicted folds, and this dataset was combined with known annotations to produce a combined dataset of 1252 folds (see Figures 1-4). Although the majority of folds were predicted in intergenic regions, there was still a considerable amount of folds (36.5 %) found within coding regions. As was observed in a previous screening of vertebrate genomes with the EvoFold program, some categories of known non-coding RNA gene annotations were recovered by this prediction process at a higher rate than other genes. For example, the genome screening did a relatively good job recovering snRNAs and miscellaneous RNAs and a relatively poor job recovering tRNAs and snoRNAs (see Table 2). So, it is likely that there may be unknown fRNAs not included in our dataset of predicted folds. We also erred on the side of accuracy rather than risking false positives. Nevertheless, when conservative definitions of protein-coding genes and prediction fRNAs are used (see Methods), 272 genes in the yeast genome are found to contain at least one coding functional RNA secondary structure. Because there are a little less than 6,000 genes in the yeast genome, the 272 genes that are predicted to contain at least one coding fRNA amount to approximately 5% of the genes in the yeast genome. This study clearly indicates a substantial number of fRNA secondary structures can be found within protein-coding regions of genes.

Functional RNAs within Coding Regions

There is relatively little information available on the impact of functional RNAs *within* protein-coding regions, so it is necessary to first demonstrate that fRNAs have an influence on these regions before evolutionary asymmetry of gene duplicates can be studied. To accomplish this, we tested for a correlation between d_N (nonsynonomous divergence per site) and the percent of the coding region covered by functional RNAs

(either predicted or known). Out of the 272 protein-coding genes containing at least one coding fRNA secondary structure, 126 genes had clearly defined values of d_N, as calculated by Wall et al. 2005. A significant negative correlation between d_N and percent fRNA coverage was observed (ρ= -0.211, p-value = 0.018). Because of the non-linear association between d_N and percent fRNA coverage, this trend can be seen even more clearly when these values are log normalized (ρ= -0.349, p-value = 6.23×10^{-5}; see Figure 5). This observation shows that genes with a higher proportion of fRNA secondary structures within their coding regions evolve more slowly, which implies that the conserved secondary structure imposes constraints on the sequence evolution of nonsynonymous sites. This conclusion is corroborated by calculations of asymmetric evolution of WGD pairs (see next section). There was also a significant correlation between d_S (synonomous divergence per site) and percent fRNA coverage (ρ= -0.506, p-value = 9.41×10^{-9}), which can be seen more clearly when the variables are log normalized (ρ= -0.547, p-value = 3.14×10^{-10}; see Figure 6). It is important to note that the correlation is stronger for synonymous than nonsynonymous sites because synonymous sites have traditionally been considered to evolve neutrally. This observation implies that coding fRNAs have a greater evolutionary impact on synonymous sites than nonsynonymous sites, so synonymous sites may evolve under even greater constraint than nonsynonymous sites in order to maintain functionality of coding fRNAs.

Asymmetric Evolution of Gene Duplicates

We observe a negative correlation between SRR (Signed Relative RNA secondary structure count) and SRN (Signed Relative Nonsynonomous Divergence), and this

correlation between SRR and SRN was significant at the 10% level (see Table 1 and Figures S1-S5). This observation means that the duplicate gene with the greater enrichment of functional RNAs evolves more slowly, and this observation may be due to functional constraint from the relatively greater number of conserved secondary structures within one of the duplicate genes. It should also be noted that there is a negative correlation between SRR and SRN for every set of EvoFold or RNAz predictions when considered separately; in fact, this correlation is observed for coding folds in the RNAz datasets with 4, 5, and 6 species with p-values less than 0.001 (see Figure S6). This is especially important because RNAz predicts secondary structures based mostly on thermodynamic analysis whereas EvoFold's evolutionary model may be biased towards predicting folds within more conserved genes (because it requires a moderately high level of primary sequence conservation). Thus, this observation provides evidence that the correlation between SRR and SRN in the WGD dataset (as well as the correlation between d_N and percent fRNA coverage for the entire set of coding folds) is not likely to be the result of any potential bias towards conserved genes.

Functional Enrichment of Genes Containing fRNAs

Of all the protein-coding regions in the yeast genome, only 25 genes are enriched with 2 or more functional RNA secondary structures (see Table 3). The top 25 genes with the highest coverage of fRNAs within coding regions are also presented as a comparison of the method of counting the number of functional RNA secondary structures versus looking at the percent coverage of RNA secondary structures (see Table 4). In both cases, the top 25 scores are enriched with genes that are associated with either protein biosynthesis or transport, where functionality of genes was defined using the

"Biological Process" annotations available from the Gene Ontology database (Gene Ontology Consortium 2000). There are five genes shared by both of these tables, so the two methods provide somewhat similar results. However, the functional enrichment of genes associated with protein biosynthesis is much more clearly illustrated with the top 25 genes in terms of percent fRNA coverage (13-15 versus 6 genes; see tables 3 and 4). Thus, there is an exciting possibility that there is a biologically significant connection between genes associated with protein biosynthesis and mRNAs containing conserved fRNA secondary structures. However, protein biosynthesis is a vital biological function, so this observation also stresses the need to ensure that the dataset is not biased towards conserved categories of genes (which could potentially be more likely to yield EvoFold predictions).

Conclusions

Thus far, we have found a number of indications that conserved functional RNA secondary structures play an important role in the evolution of coding regions in protein-coding genes. Namely, we show that an increase in the coverage of functional RNAs within protein coding regions is correlated with a decrease in the evolutionary rate of protein evolution (see Figure 5 and Table 1). There is also a negative correlation between percent fRNA coverage and synonymous site evolution (see Figure 6), and this correlation is even stronger than at nonsynonymous sites. Analysis of synonymous site evolution will be useful because synonymous sites have traditionally been considered to evolve neutrally, and the presence of a significant number of conserved fRNA secondary structures within coding regions would provide evidence that a considerable number of synonymous sites undergo non-neutral evolution. We also discovered that genes

enriched with functional RNAs tend to be associated with protein biosynthesis and transport (see Tables 3 and 4). However, more sophisticated functional analysis will be required to determine if these observations are truly caused by conserved coding fRNA secondary structures or a bias towards conserved genes in the dataset. Such a bias may be due to the fact that the predictions were made from the set of phastCons conserved blocks in yeast. However, there is also evidence that the functional constraint on coding fRNAs is the cause of the negative correlation between the frequency of coding fRNAs and divergence at both synonymous and nonsynonymous sites. For example, the EvoFold predictions were filtered through windows of RNAz predictions, and RNAz depends mostly on thermodynamic analysis that will not be biased towards conserved genes.

There are not any easily accessible annotations for known coding fRNAs in yeast. However, it is known that programmed frameshifting occurs in eEF1α in yeast (Farabaugh and Vimaladithan 1998). Both duplicate pairs of this gene contained more than one predicted coding fRNA (and one duplicate has the highest number of predicted coding fRNAs in any yeast gene; see table 3) and one copy of this gene is also in the list of top 25 genes with the greatest percent fRNA coverage (see table 4). Thus, this evidence from previous studies shows that some of the predicted coding fRNAs are likely to be genuine, and this evidence also demonstrates that the numerous genes associated with protein biosynthesis are capable of coding for conserved RNA secondary structures within their protein-coding regions.

This study is based largely on a new set of predicted functional RNA secondary structures, which may be useful to experimental scientists in that it may illuminate novel

roles for novel functional RNAs (within both coding and non-coding regions) in *Saccharomyces cerevisiae* (a common model organism for genetic studies).

This study leads to several interesting future research topics, some of which can be pursued during summer in the Yi laboratory. For example, the work so far does not distinguish between conserved functional RNA secondary structures on the messenger RNA itself versus conserved secondary structures of non-coding RNA genes that overlap with coding regions of protein-coding genes but are on the opposite strand of DNA. To help with this analysis, the "folding potential" for mRNAs in the yeast genome can be calculated, as described by Katz and Burge 2003. This will also provide yet another method to see if mRNAs with favorable folding potentials also show a decrease in the rate of protein evolution. Furthermore, if selection for folding potential within yeast genes containing coding fRNA secondary structures is observed, then this may help control for the effects of a potential bias towards conserved genes in the dataset. A better way to analyze asymmetric evolution of gene duplicates in yeast also needs to be developed. Finally, an exhaustive literature search can be conducted in order to search for more genes that are already known to contain coding fRNAs. After completing this additional analysis, we should be able to provide in-depth analysis of the impact of functional RNAs within protein coding regions.

Methods

Functional RNA Predictions

Screening for functional RNAs was conducted using EvoFold and RNAz programs to provide independent predictions of fRNA folds (Pedersen et al. 2006, Washietl et al. 2005). These two programs should predict folds independently because

EvoFold utilizes a functional RNA model based on stochastic context-free grammars while RNAz primarily utilizes thermodynamic information to predict RNA secondary structures (while also considering covariance between secondary structures in a multi-species alignment). For more information about the multi-species alignments used for these fRNA prediction programs, see the supplementary material. The optimal multi-species alignments from the SGD database (Cherry et al. 1997) for each program was determined by comparing the proportion of known annotations recovered to the proportion of known fRNA in the overall dataset of predicted folds (see supplementary material). For the combined dataset of known and predicted folds, any predicted folds overlapping known folds were removed and replaced with the coordinates for the corresponding known fold from the SGD annotations.

Duplicate Gene Analysis

SRR (Signed Relative RNA secondary structure) was used to quantify the distribution of fRNA secondary structure predictions. Methods to quantify functional and sequence divergence include SRK, SRA, SRF, and SRN. SRK refers to the signed relative connectivity, as determined by yeast protein-protein interaction networks, for a duplicate pair (Breitkreutz et al. 2003). SRA refers to signed relative abundance and is a measure of protein abundance (Ghaemmaghami et al. 2003). SRF refers to signed relative fitness, where genes whose removal from a network is *more detrimental* to survivability are represented by a *lower* value (Gu et al. 2005). SRN refers to signed relative nonsynonomous divergence. A slightly more detailed description is contained in the "SR Calculations" section (below). Unless otherwise stated, all statistical analysis (e.g.

correlation values) was performed using the R software package (R Development Core Team 2006).

SR Calculations

SRR, SRK, SRA, and SRF are a modification of the Canberra distance (in one dimension) and can be expressed in the general form, $SRX = (X_1 – X_2)/(X_1 + X_2)$ where X_1 equals the given quantity (e.g., number of fRNA folds, etc.) for the first duplicate gene and X_2 equals the given quantity for the second duplicate gene (Johnson and Wichern 2002). SRN is also a modification of the Canberra distance and is calculated as follows: $SRN = (d_{N12}\text{-}d_{N13})/d_{N23}$, where d_{N12} equals the divergence in nucleotide sequence between the first duplicate gene and the ortholog in *Kluyveromyces waltii*, d_{N13} equals the divergence in nucleotide sequence between the second duplicate gene and ortholog in *Kluyveromyces waltii*, and d_{N23} equals the divergence in nucleotide sequence between the two duplicate genes. Orthologs were based upon results from Kellis et al. 2004. In general, SRK, SRA, SRF, and SRN were calculated and applied as used in Kim and Yi 2006.

Calculation of Nonsynonymous and Synonymous Divergence

Analysis of nonsynonymous and synonymous divergence was conducting using previously calculated values available from the supplementary material for Drummond et al. 2006. This table presented data that was originally published by Wall et al. 2005. However, it may be valuable to recalculate these values in order to 1) increase the size of the dataset and 2) consider synonymous divergence between more closely related yeast species. In order to recalculate nonsynonymous and synonymous divergence for post-WGD species of yeast genomes, the Multiz alignment has been downloaded for all the

conservatively defined protein-coding genes containing at least one coding functional RNA secondary structure (Blanchette et al. 2004). The Multiz alignment for these genes was obtained using the Galaxy server on the UCSC Genome Browser (Giardine et al. 2005, Karolchik et al. 2003). All coding fRNAs were first defined using annotations for protein-coding genes from the SGD database (Cherry et al. 1997). Questionable annotations for protein-coding genes were filtered to produce a more conservative definition of protein-coding genes. Any genes that were removed from the initial dataset are listed in Table S11. PAML will be used to calculate d_N and d_S for the genes with a valid Multiz alignment (Yang 1997). Possible reasons that the Multiz alignment for a gene would not be valid include gaps in the Multiz alignment that produce a frameshift resulting in the creation of premature stop codons in the genomic sequence. Some preliminary results from these recalculations for synonymous sites can be observed in figure S12.

Acknowledgements

I would like to thank Jakob Pedersen for explaining the technical details of the EvoFold program. I would also like to thank the Office of Information Technology at Georgia Tech for allowing me to conduct research on the PACE Linux cluster. I would also like to thank Navin Elango and Marco Lin for computational assistance. Finally, I would like to thank Navin Elango for reviewing this manuscript for my honors thesis.

References

Akashi, H. 2003. "Translational Selection and Yeast Proteome Evolution" *Genetics* **164**: 1291-1303

Bejerano, G., Lowe, C.B., Ahituv, N., King, B., Siepel, A., Salama, S.R., Rubin, E.M., Kent, W.J., Haussler, D. 2006 "A distal enhancer and an ultraconserved exon are derived from a novel retroposon" *Nature* **441**(7089): 87-90.

Blanchette, M. Kent, M.J., Riemer, C., Elnitski, L., Smit A.F.A., Roskin, K.M., Baertsch, R. Rosenbloom, K., Clawson, H., Green, E.D., Haussler, D., and Miller, W. 2004. "Aligning Multiple Genomic Sequences with the Threaded Blockset Aligner" *Genome Research* **14** (4): 708-715

Breitkreutz, B.J., Stark, C., and Tyers, M. 2003. "The GRID: the general repository for interaction datasets" *Genome Biology* **4**:R23

Chamary, J.V., and Hurst, L.D. 2005. "Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals" *Genome Biology* **6**: R75

Chamary, J.V., Parmley, J.L., and Hurst, L.D. 2006. "Hearing silence: non-neutral evolution at synonymous sites in mammals" *Nature Reviews. Genetics.* **7**: 98-108

Cherry, J.M., Ball C, Weng S, Juvik G, Schmidt R, Adler C, Dunn B, Dwight S, Riles L, Mortimer RK, and Botstein D. 1997. "Genetic and physical maps of *Saccharomyces cerevisiae*" *Nature* **387**: 67073

Drummond, D.A., Bloom, J.D., Adami, C., Wilke, C.O., and Arnold, F.H. 2005. "Why highly expressed protein evolve slowly" *Proceedings of the National Academy of Sciences, U.S.A.* **102**: 14338-14343

Drummond, D.A., Raval, A., and Wilke, C.O. 2006. "A Single Determinant Dominates the Rate of Yeast Protein Evolution" *Molecular Biology and Evolution.* **23**(2): 327-337

Farabaugh, P.J., and Vimaladithan, A. 1998. "Effect of frameshift-inducing mutants of elongation factor 1alpha on programmed +1 frameshifting in yeast" *RNA* **4**: 38-46

The Gene Ontology Consortium. 2000. "Gene Ontology: tool for the unification of biology." *Nature Genetics* **25**: 25-29

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

Giardine, B., Riemer, C., Hardison, R.C., Burhans, R., Elnitski, L., Shah, P., Zhang, Y., Blankenberg, D., Albert, I., Taylor, J., Miller, W., Kent, W.J., and Nekrutenko, A. "Galaxy: A platform for interactive large-scale genome analysis" *Genome Research* **15**: 1451-1455

Gu, X., Zhang, Z., and Huang, W. 2005. "Rapid evolution of expression and regulatory divergences after yeast gene duplication" *Proceedings of the National Academy of Science, USA* **102**: 707-712

Guldener U., Munsterkotter, M., Kastenmuller, G. Strack, N. van Helden, J., Lemer, C., Richelles, J., Wodak, S.J., Garcia-Martinez, J., Perez-Ortin, J.E., Michael, H. Kaps, A., Talla, E., Dujon, B., Andre, B., Souciet, J.L., De Montigny, J., Bon E., Gaillardin, C., and Mewes, H.W. 2005. "CYGD: the Comprehensive Yeast Genome Database" *Nucleic Acids Research* **33**: D364-8

Hofacker, I.L. 2003. "Vienna RNA secondary structure server" *Nucleic Acids Research* **31** (13): 3429-3431

Johnson, R.A., and Wichern, D.W. 2002. *Applied Multivariate Statistical Analysis*. 5th ed. Prentice Hall, NJ.

Kellis, M., Birren, B., and Lander, E. 2004. "Proof and evolutionary analysis of ancient genome duplication in yeast *Saccharomyces cerevisiae*" *Nature* **428**:617-24

Karolchik, D., Baertsch, R., Diekhans, M., Furey, T.S., Hinrichs, A., Lu, Y.T., Roskin, K.M., Schwartz, M., Sugnet, C.W., Thomas, D.J., Weber, R.J., Haussler, D. and Kent, W.J. 2003. "The UCSC Genome Browser Database" *Nucleic Acids Research* **31**: 51-54

Katz, L. and Burge, C. B. 2003. "Widespread selection for local RNA structure in coding regions of bacterial genes" *Genome Res.* **13:** 2042-2051.

Kim, S-H and Yi, S.V. 2006. "Correlated asymmetry between sequence and functional divergence of duplicate proteins in *Saccharomyces cerevisiae*" *Molecular Biology. and Evolution* **23**: 1068-1075

McCutcheon, J.P., and Eddy, S.R. 2003. "Computational identification of non-coding RNAs in *Saccharomyces cerevisiae* by comparative genomics" *Nucleic Acids Research* **31** (14): 4119-4128

Mewes, H.W., Amid, C., Arnold, R., Rishman, D., Guldener, U., Mannhaupt, G., Munsterkotter, M., Pagel, P., Strack, N., Stumpflen, V., Warfsmann, J., Ruepp, A. 2004. "MIPS: analysis and annotation of proteins from whole genomes" *Nucleic Acids Research* **32** :D41-44

Ouellet, D.L., Perron, M.P., Gobell, L-A, Plante, P., and Provost, P. 2006. "MicroRNAs in Gene Regulation: When the Smallest Governs It All" Journal of Biomedicine and Biotechnology. Article ID 69616: 20 pages

Pedersen, J.S., Bejerano, G., Siepel, A., Rosenbloom, K., Lindblad-Toh, K., Lander, E.S., Kent, J., Miller, W., and Haussler, D. 2006. "Identification and Classification of

Conserved RNA Secondary Structures in the Human Genome" *PLoS Computational Biology* **2**: e33

Peng, W-T., Robinson, M.D., Mnaimneh, S., Krogan, N.J., Cagney, G., Morris, Q., Davierwala, A.P., Grigull, J., Yang, X., Zhang, W., Mitsakakis, N., Ryan, O.W., Datta, N., Jojic, V., Pal, C., Canadien, V., Richards, D., Beattle, B., Wu, L.F., Altschuler, S.J., Roweis, S., Frey, B.J., Emili, A., Greenblatt, J.F., and Hughes, T.R. 2003. "A Panoramic View of Yeast Noncoding RNA Processing" *Cell* **113**: 919-933

Pollard, K.S., Salama, S.R., Lambert, N., Lambot, M.A., Coppens, S., Pedersen, J.S., Katzman, S., King, B., Onodera, C., Siepel, A., Kern, A.D., Dehay, C., Igel, H., Ares, M. Jr., Vanderhaeghen, P., and Haussler, D. 2006. "An RNA gene expressed during cortical development evolved rapidly in humans" *Nature* **443** (7108): 167-172

R Development Core Team. 2006. *R: A language and environment for statistical computing* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org

Rocha, E.P.C., and Danchin, A. 2004. "An Analyisis of Determinants of Amino Acids Substitution Rates in Bacterial Proteins" *Molecular Biology and Evolution* **21** (1): 108-116

Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou M., Rosenbloom, K., Clawson, H., Spieth, J., Hiller, L.W., Richards, S., Weinstock, G.M., Wilson, R.K., Gibbs, R.A., Kent, W.J., Miller, W., and Haussler, D. 2005. "Evolutionarily

conserved elements in vertebrate, insect, worm, and yeast genomes" *Genome Research* **15**:1034-1050

Storz, G., Altuvia, S., Wassarman, K.M. 2005. "An abundance of RNA regulators" *Annual Review of Biochemistry* **74**: 199-217

Wall, D.P., Hirsh, A.E., Fraser, H.B., Kumm, J., Giaever, G., Eisen, M.B., and Feldman, M.W. 2005. "Functional genomic analysis of the rates of protein evolution" *Proceedings of the National Academy of Sciences, U.S.A.* **102**: 5483-5488

Washietl, S., Hofacker, I.L, and Stadler, P.F. 2005. "Fast and reliable prediction of noncoding RNAs" *Proceedings of the National Academy of Sciences, U.S.A*. **102** (7): 2454-2459

Yang, Ziheng. 1997. "PAML: a program package for phylogenetic analysis by maximum likelihood" *Computer applications in the biosciences*. **13** (5): 555-556

Table S10: Pearson Correlation Coefficients (ρ) for All Folds in Combined Dataset

	SRR~SRA	SRR~SRK	SRR~SRF	SRR~SRN
All	0.007	-0.002	-0.011	-0.051
Coding	0.090	-0.133	0.026	-0.203*
5` Flank	0.181	0.096	-0.273*	0.135
3` Flank	-0.075	0.130	0.076	-0.084
Intron	0.392	0.183	-0.344	-0.531

Significance levels: * = $p<0.10$

Table 2: Annotated folds recovered with Stringent Dataset (without known fRNAs)

fRNA Category	# Recovered / # Annotations
tRNA	39 / 275
snoRNA	18 / 66
snRNA	5 / 6
rRNA	6 / 11
Miscellaneous fRNA	10 / 14

Table 3: Genes with Multiple Folds within Coding Regions

Gene Name	GO Biological Process	# fRNA folds
YPR080W (TEF1/eEF1)	**protein biosynthesis**; translational elongation	5
YLR342W (FKS1)	beta-1,3 glucan biosynthesis; endocytosis; cell wall organization and biogenesis	4
YOR086C	[Unknown]	4
YOR096W (RPS7A)	**protein biosynthesis**	3
YLR089C	Biosynthesis (putative alanine transaminase)	3
YBR118W (TEF2/eEF1)	**protein biosynthesis**; translational elongation	2
YIL148W (RPL40A)	DNA repair; protein modification; response to DNA damage stimulus	2
YOR109W (INP53)	Endocytosis; cell wall organization and biogenesis; inositol lipid-mediated signaling	2
YOR310C (NOP58)	rRNA modification; rRNA processing; 35S primary transcript processing; snRNP protein import into nucleus; ribosome biogenesis; processing of 20S pre-rRNA	2
YOR133W (EFT1/eEF2)	**protein biosynthesis**; translational elongation	2
YGR214W (RPS0A)	ribosomal small subunit assembly and maintenance; **protein biosynthesis**	2
YML100W (TSL1)	trehalose biosynthesis; response to stress	2
YKL157W (APE2)	Proteolysis; peptide metabolism	2
YDR457W (TOM1)	protein polyubiquitination; transcription; regulation of transcription, DNA-dependent; rRNA processing; mRNA processing; mRNA export from nucleus; protein modification; ubiquitin cycle; protein monoubiquitination; **transport**; nucleocytoplasmic transport; nuclear organization and biogenesis; cell cycle; mitosis; mRNA transport	2
YER155C (BEM2)	microtubule cytoskeleton organization and biogenesis; cell wall organization and biogenesis; budding cell bud growth; intracellular signaling cascade; small GTPase mediated signal transduction; actin cytoskeleton organization and biogenesis; establishment of cell polarity	2
YDR385W (EFT2/eEF2)	**protein biosynthesis**; translational elongation	2
YPL036W (PMA2)	**Transport**; ion transport; cation transport; regulation of pH; metabolism; proton transport	2
YDL145C	**Transport**; intracellular protein transport; ER to Golgi	2

(COP1)	vesicle-mediated transport; retrograde vesicle-mediated transport, Golgi to ER; protein transport	
YER091C (MET6)	amino acid biosynthesis; methionine biosynthesis	2
YDR456W (NHX1)	**Transport**; ion transport; sodium ion transport; regulation of pH; vacuolar acidification; monovalent inorganic cation transport; endosome transport; monovalent inorganic cation homeostasis	2
YDR406W (PDR15)	**transport**	2
YDR093W (DNF2)	cation transport; intracellular protein transport; endocytosis; phospholipid translocation	2
YDL056W (MBP1)	regulation of progression through cell cycle; DNA replication; transcription; regulation of transcription, DNA-dependent	2
YDL140C (RPO21)	Transcription; transcription from RNA polymerase II promoter	2
YPL058C (PDR12)	**Transport**; organic acid transport; propionate metabolism	2

Table 4: Top 25 Genes within Coding Regions with the greatest % coverage of fRNAs

Gene Name	GO Biological Process	% fRNA coverage
YKL157W	Proteolysis; peptide metabolism	100
YBR152W	formation of catalytic spliceosome for first transesterification step; nuclear mRNA splicing, via spliceosome	85.6
YMR079W	**Transport**; Golgi to plasma membrane transport; etc.	79.6
YIL148W	DNA repair; protein modification; response to DNA damage stimulus	74.3
YOR096W	**protein biosynthesis**	57.4
YER044C-A	Meiosis; sporulation	41.4
YML073C	ribosomal large subunit assembly and maintenance; **protein biosynthesis**	38.1
YKR094C	[Homology to rat L40]	33.8
YPL081W	**protein biosynthesis**; regulation of translational fidelity	30.6
YKL006W	**protein biosynthesis**	30.0
YHL001W	**protein biosynthesis**	30.0
YPL079W	**protein biosynthesis**	30.0
YGR214W	ribosomal small subunit assembly and maintenance; **protein biosynthesis**	26.3
YDR524W-A	[Identified by SAGE]	22.2
YLR130C	**Transport**; ion transport; zinc ion transport; low-affinity zinc ion transport; metal ion transport	21.4
YPR132W	[Homology to rat S23 and E. coli S12]	19.7
YIL052C	**protein biosynthesis**; filamentous growth	19.0
YLR448W	ribosomal large subunit assembly and maintenance; **protein biosynthesis**	18.7
YPR080W	**protein biosynthesis**; translational elongation	15.3
YGL076C	**protein biosynthesis**	15.1
YDL136W	**protein biosynthesis**	14.2
YDL191W	[Homology to rat L35]	12.8
YOL127W	ribosomal large subunit assembly and maintenance; **protein biosynthesis**	11.2
YBR084C-A	[Homology to rat L19]	10.4
YNL070W	**Transport**; protein transport; protein import into mitochondrial matrix; protein import into mitochondrial outer membrane	10.4

Figure 1: Confirmed Predictions Based Upon Category

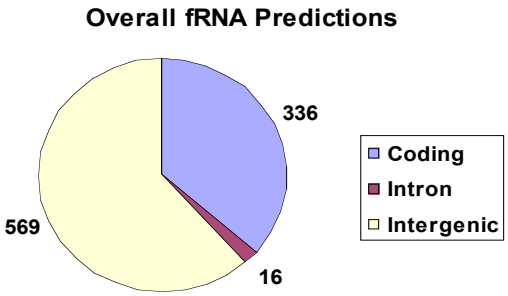


Figure 2: Combined Folds Based Upon Category

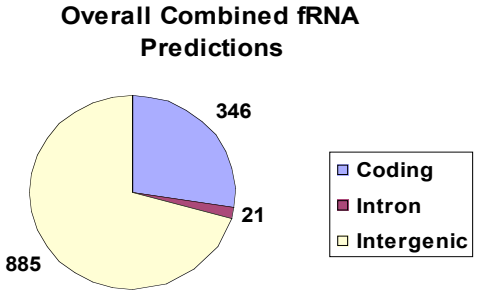


Figure 3: Confirmed Predictions Found Around Duplicate Genes

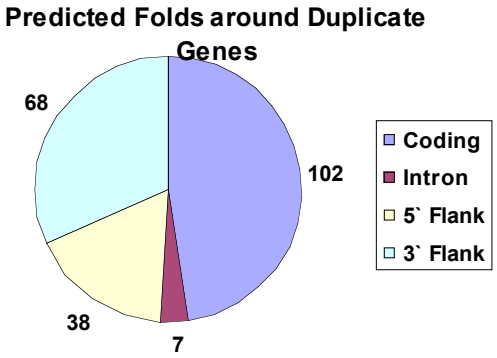


Figure 4: Combined Folds Found Around Duplicate Genes

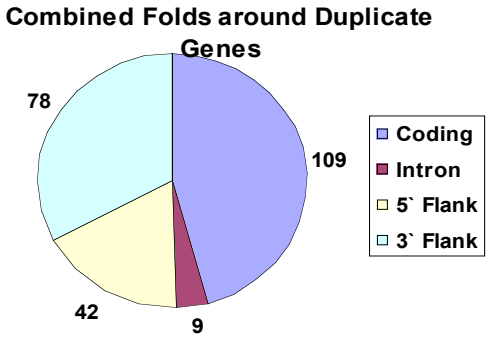
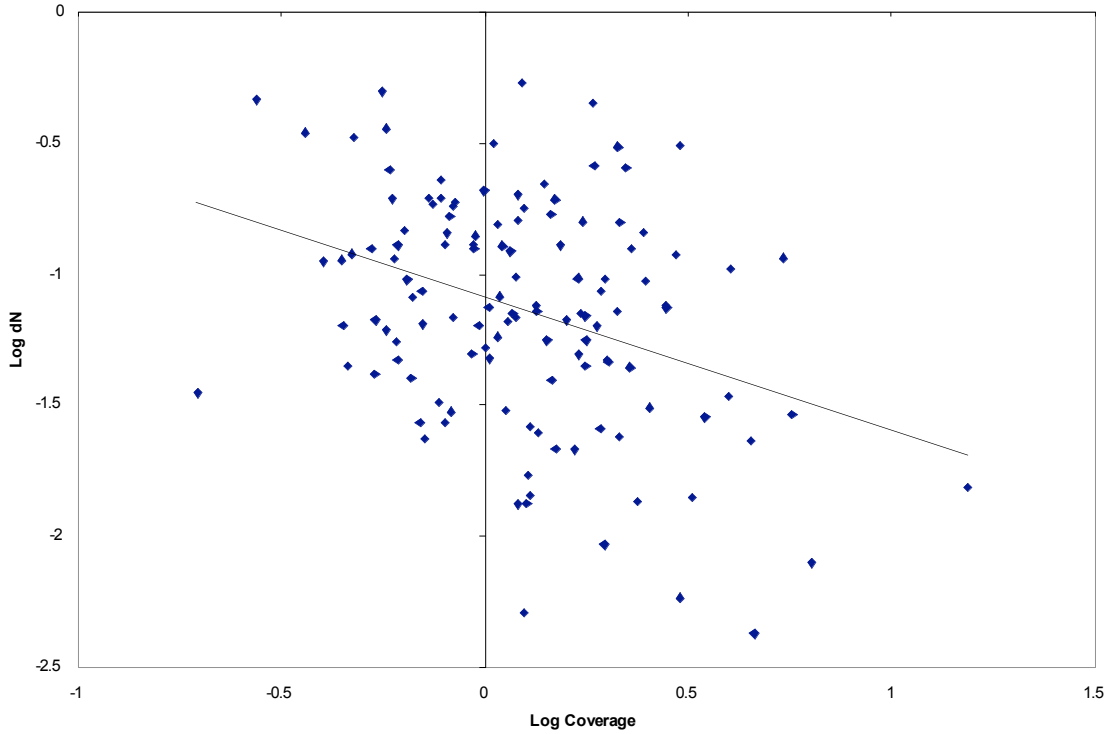
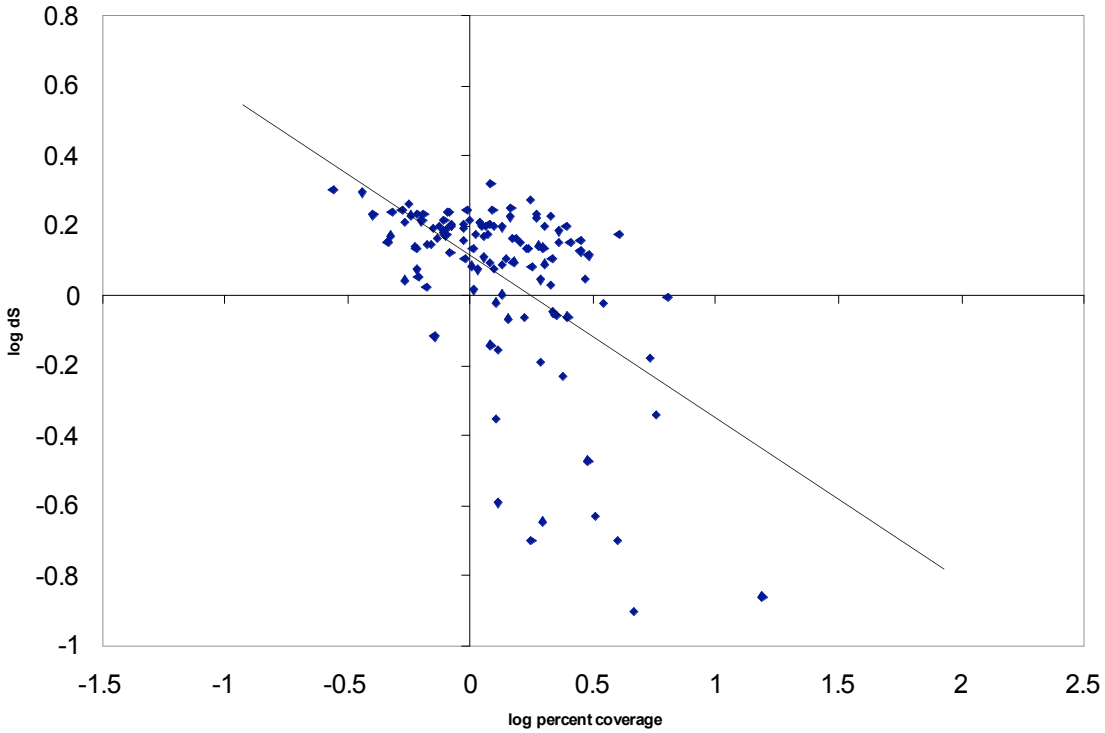


Figure 5: Constraint of Protein Evolution due to Presence of Coding fRNAs



log dN ~ log fRNA coverage: ρ= -0.349, p-value = 6.23×10^{-5}

Figure 6: Constraint of Synonymous Site Evolution Due to the Presence of Coding fRNAs



log dS ~ log fRNA coverage: $\rho = -0.547$, p-value = 3.14×10^{-10}