Published online 8 May 2013 *Nucleic Acids Research, 2013, Vol. 41, No. 13* **6371–6380**
doi:10.1093/nar/gkt349

The Impact of *Trans*-Regulation on the Evolutionary Rates of Metazoan Proteins

Yi-Ching Chen[1], Jen-Hao Cheng[1], Zing Tsung-Yeh Tsai[1,2], Huai-Kuang Tsai[1,*] and Trees-Juen Chuang[3,*]

[1]Institute of Information Science, Academia Sinica, Taipei 115, Taiwan, [2]Bioinformatics Program, Taiwan International Graduate Program, Academia Sinica, Taipei 115, Taiwan and [3]Genomic Research Center, Academia Sinica, Taipei 115, Taiwan

Received February 7, 2013; Revised April 10, 2013; Accepted April 14, 2013

ABSTRACT

Transcription factor (TF) and microRNA (miRNA) are two crucial *trans*-regulatory factors that coordinately control gene expression. Understanding the impacts of these two factors on the rate of protein sequence evolution is of great importance in evolutionary biology. While many biological factors associated with evolutionary rate variations have been studied, evolutionary analysis of simultaneously accounting for TF and miRNA regulations across metazoans is still uninvestigated. Here, we provide a series of statistical analyses to assess the influences of TF and miRNA regulations on evolutionary rates across metazoans (human, mouse and fruit fly). Our results reveal that the negative correlations between *trans*-regulation and evolutionary rates hold well across metazoans, but the strength of TF regulation as a rate indicator becomes weak when the other confounding factors that may affect evolutionary rates are controlled. We show that miRNA regulation tends to be a more essential indicator of evolutionary rates than TF regulation, and the combination of TF and miRNA regulations has a significant dependent effect on protein evolutionary rates. We also show that *trans*-regulation (especially miRNA regulation) is much more important in human/mouse than in fruit fly in determining protein evolutionary rates, suggesting a considerable variation in rate determinants between vertebrates and invertebrates.

INTRODUCTION

Gene expression is largely controlled by actions of various *trans*-regulatory factors. Undoubtedly, transcription factor (TF) and microRNA (miRNA) are the most conspicuous classes of *trans*-regulatory factors and are regarded as primary gene regulators in metazoans. TFs are proteins that facilitate or repress the transcription of their target genes through binding to specific DNA sequences, the so-called TF-binding sites (TFBSs), in the gene promoter regions (1). On the other hand, miRNAs are ~22 nucleotide noncoding RNAs, which target mRNAs and reduce stability and/or translation activity of mRNA to regulate gene expression at the posttranscriptional level (2). TFs and miRNAs may work together and form a complex regulatory network that generally consists of intricate feedback and feed-forward loops (3–5). The coordinated regulation of TFs and miRNAs may play important roles in a wider diversity of biological processes (6). A previous study reported that genes with more TFBSs tend to be targeted by miRNAs and have more miRNA-binding sites, suggesting a positive correlation between these two *trans*-regulatory factors (7). Although the mechanism of how miRNAs cooperate with TFs in the regulatory network remains largely unknown (8), accumulating evidence indicates its biological significance. Thus, it is of interest to investigate the relationship between these two *trans*-regulatory factors.

In terms of molecular evolution, it was shown that genes regulated by more different TFs (N_{TF}) tend to evolve more slowly in yeast (9,10). Similarly, genes targeted by more distinct miRNAs (N_{miR}) were suggested to experience more functional constraints and thereby evolve more slowly in human and mouse (11,12). These observations revealed that *trans*-regulation complexity is an important indicator of evolutionary rates, regardless of TF regulation at the transcriptional level or miRNA regulation at the posttranscriptional level. However, comparative studies of *trans*-regulatory factors have been hampered by the paucity or incompleteness of TF and miRNA information. To our knowledge, there is currently no systematic evolutionary analysis available that simultaneously accounts for these two *trans*-regulatory factors across metazoans. Whether the negative correlation between the number of *trans*-regulators that regulate a

*To whom correspondence should be addressed. Tel: +886 2 27871244; Fax: +886 2 27898757; Email: trees@gate.sinica.edu.tw
Correspondence may also be addressed to Huai-Kuang Tsai. Tel: +886 2 2788 3799 (Ext. 1718); Fax: +886 2 2782 4814; Email: hktsai@iis.sinica.edu.tw

gene (i.e. N_{TF} and N_{miR}) and evolutionary rates is maintained across metazoans, whether N_{TF} and N_{miR} have a dependent effect on evolutionary rates, and which of these two factors has a greater effect on metazoan protein evolution still await investigation.

In addition to *trans*-regulation, many other biological factors associated with and potentially underlying evolutionary rates of proteins have been reported. These factors include protein connectivity in protein–protein interaction (PPI) networks (9,10,13–19), expression level (or expression abundance) (9,11,13,14,17–30), tissue specificity (or expression breadth) (13,21,23,25,26,31–34), length of untranslated regions (UTRs) (12,21,26), intron length (13,21,26,35), intron number (23,26), solvent accessibility (36–39) and disorder content (11,40–42). Some of these factors were also shown to be correlated with N_{TF} or N_{miR} (9,11,12,43,44). We classified these 10 factors into five categories: *trans*-regulation (N_{TF} and N_{miR}), protein connectivity, gene expression (expression level and tissue specificity), gene compactness (UTR length, intron length and intron number) and protein structure (solvent accessibility and disorder content). It is worth exploring whether the last four categories of confounding factors contribute to the strength of N_{TF} and N_{miR} as indicators of evolutionary rates.

To address these issues, we widely collect TF- and miRNA-binding data from human (*Homo sapiens*), mouse (*Mus musculus*) and fruit fly (*Drosophila melanogaster*) and then systematically examine the correlations between these two *trans*-regulatory factors (N_{TF} and N_{miR}) and evolutionary rates: nonsynonymous substitution rate (d_N), synonymous substitution rate (d_S) and d_N/d_S ratio. We show that genes regulated by more different TFs/miRNAs evolve more slowly is generally maintained in human, mouse and fruit fly. By controlling for the other confounding factors (i.e. protein connectivity, gene expression, gene compactness and protein structure), the partial correlations between N_{miR} and evolutionary rates still hold well, whereas the strength of N_{TF} as a rate indicator is greatly decreased in human/mouse and even disappears in fruit fly. We further find two trends: miRNA regulation tends to be much stronger than TF regulation in determining the rate of protein sequence evolution, and TF and miRNA regulations have a dependent effect on evolutionary rates, both of which are generally maintained across metazoans (human, mouse and fruit fly). We also observe that *trans*-regulation seems to play a much greater role in human/mouse than in fruit fly in causing variation in protein evolutionary rates. This result reveals that the relative impact of *trans*-regulation on the evolutionary rates appears to be different between vertebrates and invertebrates.

MATERIALS AND METHODS

Data retrieval and extraction

The protein-coding genes in human and mouse, orthology assignments and human–mouse and mouse–human evolutionary rates (d_N, d_S and d_N/d_S) were downloaded from the Ensembl genome browser at http://www.ensembl.org/ (release 69) (45). For an alternatively spliced gene, only its longest isoform was selected. To avoid the confounding factor of gene duplication, only 1:1 orthologs between human and mouse genes were considered. Meanwhile, fruit fly protein-coding genes were downloaded from Flybase (release 4.3) (46). Fruit fly genes with single-copy orthologs across five other *Drosophila* species (i.e. *D.melanogaster*, *Drosophila simulans*, *Drosophila yakuba*, *Drosophila erecta*, *Drosophila sechellia* and *Drosophila ananassae*) and the evolutionary rates were obtained from Larracuente *et al.*'s study (23). Here, d_S values = 0 and d_N/d_S values ≥ 2 were not considered. Additionally, the human and mouse genes on chromosome Y were excluded for reducing the possibility of irregular evolutionary rates in the short single-copy sex chromosome. Chromatin immunoprecipitation (ChIP) data including 162 human TF ChIP-seq datasets and 59 mouse TF ChIP-chip and ChIP-seq datasets were downloaded from ENCODE project (47) and hmChIP (48), respectively. The promoter of each human gene was defined as the intergenic region of 8 kb upstream to 2 kb downstream of the gene start position (4 kb upstream to 1 kb downstream for each mouse gene). Also, we defined that a TF regulates a gene if at least one ChIP-seq peak of the TF lies within the promoter region of the gene. The nonredundant associations for 149 fruit fly TFs and their target genes were obtained from DroID (May 2011) (49), which integrates TF-gene associations from modENCODE (50) and REDfly (51). For extraction of predicted human TFBS data, 843 human position frequency matrices were downloaded from TRANSFAC® free trial (December 2011) (52). The position weight matrix (PWM) of each position frequency matrix and the cutoffs were obtained by using PATSER (53). This study considered the potential binding motifs whose *P*-values were smaller than or equal to the minimum of the default cutoff and 10^{-3}. We further considered the general-binding preference (GBP) score (54) to obtain more reliable predictions. Only the binding sites with GBP scores >0.2, as Ernst *et al.* (54) suggested, were retained. A TF was defined to regulate a gene if at least one potential binding motif of the TF locates within the promoter region of the gene. Human, mouse and fruit fly miRNA target prediction data were downloaded from TargetScan release 6.2 (including TargetScanHuman, TargetScanMouse and TargetScanFly) (55,56). For accuracy, this study considered all human, mouse and fruit fly miRNA families whose target sites were conserved. The site conservation is defined by conserved branch length as determined in TargetScan (56). The used human, mouse and fruit fly genes and the related information are available at http://bits.iis.sinica.edu.tw/TransRegEvoRate/index.html.

Protein connectivity

The connectivity of a protein was defined by the total number of distinct proteins interacting with the protein. The PPI datasets of human, mouse and fruit fly were downloaded from STRING 9.0 (57), which retrieved known and predicted PPIs from literature.

Gene expression

Normalized expression datasets of 78 nonpathogenic human tissues and 77 nonpathogenic mouse tissues were downloaded from BioGPS (58), and a normalized expression dataset of 27 fruit fly nonpathogenic tissues was downloaded from FlyAtlas (59). If multiple probe sets refer to the same gene, the signals from different probe sets of the same gene were averaged. Here, expression was analyzed in terms of expression level and tissue specificity (τ). The expression level of a gene was defined as the average signal intensity across all examined tissues. The tissue specificity of a gene is defined by

$$\tau = \frac{\sum_{j=1}^{n}\left(1 - \frac{\log S(j)}{\log S_{\max}}\right)}{n-1}$$

in which n denotes the number of the examined tissues, $S(j)$ denotes the signal intensity and $S_{\max}$ denotes the highest signal across all examined tissues (60). A large τ value represents high tissue specificity. Of note, to minimize potential noise that might be caused by low signal intensities, we set the signal to 100 if it is <100 (21,61,62).

Gene compactness and protein structural features

Gene compactness was measured by the intron number and the average lengths of UTRs and introns of a gene. Regarding protein structure, it was analyzed in terms of solvent accessibility and disorder content. The solvent accessibility of a protein was calculated by the maximum number of exposed residues that interact with solvent molecules over the length of the protein, in which the exposed residues were predicted by ACCPro release 4.1 with the default threshold of 25% (63). We only considered the proteins of lengths <8000 amino acids owing to the limitation of ACCPro. The disorder content of a protein, defined by the percentage of intrinsically disordered region, was estimated by the number of disordered residues over the length of the protein. The disordered residues were predicted by DISOPRED2 version 2.4 with the default 5% false-positive threshold (64). To ensure a lower standard error, we only considered the proteins of length longer than 100 amino acids.

Calculation of the relative contribution to variability explained

The relative contribution to variability explained (RCVE) is used to measure the relative importance of each tested factor, which is calculated as follows:

$$RCVE = \frac{R^2_{full} - R^2_{reduced}}{R^2_{full}}$$

where R^2_{full} and $R^2_{reduced}$ denotes the R^2 value of the full model (including all of the factors examined) and that of the reduced model (excluding the factor of interest), respectively. A larger RCVE indicates a more important contribution of the factor of interest to the regression model (65).

RESULTS AND DISCUSSION

miRNA regulation is much more important than TF regulation in determining the evolutionary rates of metazoans

Previous studies have shown that the number of regulatory TFs that regulate a gene (N_{TF}) is negatively correlated with $d_{\mathrm{N}}/d_{\mathrm{S}}$ in a yeast transcriptional regulatory network (9,10), leading to that genes with more regulatory TFs tend to evolve more slowly. We are then interested to know whether the trend is maintained in multicellular organisms. We first extract experimentally determined TFBS data (i.e. TF ChIP-binding datasets) from human, mouse and fruit fly ('Materials and Methods' section; Table 1) and estimate the Spearman's rank correlation (ρ) between N_{TF} and evolutionary rates (i.e. d_N, d_{S} and $d_{\mathrm{N}}/d_{\mathrm{S}}$). In general, we find that evolutionary rates are negatively correlated with N_{TF} in the three species examined (Table 2). In terms of miRNA regulation, we extract human, mouse and fruit fly miRNA target data (Table 1) and also show negative correlations between N_{miR} and evolutionary rates in the three species examined (Table 2). These observations reveal a common trend that genes regulated by more TFs or miRNAs evolve more slowly at both the protein and RNA levels in metazoans.

The above results, however, should be treated carefully because many confounding factors that may affect evolutionary rates of protein-coding genes have not been controlled. As stated above (see 'Introduction' section), the confounding factors include protein connectivity, gene expression [expression level and tissue specificity (or expression breadth)], gene compactness (UTR length, intron length and intron number), protein structure (solvent accessibility and disorder content) and so on. Some of these confounding factors have also been reported to be correlated with N_{TF} or N_{miR}. For example, N_{TF} was reported to be positively correlated with mRNA expression (9) and UTR length (44). Meanwhile, N_{miR} was shown to be positively correlated with protein connectivity (11,43), expression breadth (43), 3′UTR length (12) and disorder content (11). Thus, we reevaluate the correlations between *trans*-regulation (N_{TF} and N_{miR}) and evolutionary rates by using partial correlation analyses (66) to simultaneously control for these confounding factors. As shown in Table 2, N_{TF} is still negatively correlated with evolutionary rates in human and mouse after controlling for N_{miR} and the other eight potential confounding factors. However, the partial correlations between N_{TF} and evolutionary rates are substantially reduced in human/mouse and even disappear in fruit fly (Table 2). This result suggests that the evolutionary effect of N_{TF} is considerably affected by these confounding factors. On the other hand, the negative correlations between N_{miR} and evolutionary rates remains strong in all three species examined after controlling for N_{TF} and the other confounding factors (Table 2). These observations reveal that N_{TF} and N_{miR} tend to have different effects on evolutionary rates. We find that the partial correlation between N_{miR} and evolutionary rates is remarkably stronger than that between N_{TF} and evolutionary rates

Table 1. The numbers of TFs and miRNAs used in this study

Species	TF		miRNA	
	Data source	Number of TFs	Data source	Number of miRNAs
Human	ENCODE (hg19)	162	TargetScanHuman (release 6.2)	1267
Mouse	hmChIP (mm8)	59	TargetScanMouse (release 6.2)	663
Fruit fly	DroID (May 2011)	149	TargetScanFly (release 6.2)	121

Table 2. Spearman's rank coefficient of correlation (ρ) between evolutionary rates (d_N, d_S and d_N/d_S) and experimentally determined N_{TF} (or N_{miR}) before and after controlling for N_{miR} (or experimentally determined N_{TF}) and the other eight confounding factors: protein connectivity, expression level, tissue specificity (τ), UTR length, intron length, intron number, solvent accessibility and disorder content

Indicator and species	Before control			After control		
	d_N	d_S	d_N/d_S	d_N	d_S	d_N/d_S
Evolutionary rate versus N_{TF}						
Human[a]	−0.1852***	−0.1641***	−0.1388***	−0.1099***	−0.1076***	−0.0763***
Mouse[b]	−0.1499***	−0.1087***	−0.1213***	−0.0832***	−0.0501***	−0.0683***
Fruit fly[c]	−0.0759**	−0.1081***	−0.0460	0.0006	0.0092	0.0049
Evolutionary rate versus N_{miR}						
Human[a]	−0.3769***	−0.2941***	−0.2916***	−0.3343***	−0.2068***	−0.2686***
Mouse[b]	−0.3165***	−0.2710***	−0.2369***	−0.3004***	−0.1990***	−0.2405***
Fruit fly[c]	−0.1172***	−0.2027***	−0.0406	−0.0931***	−0.0522*	−0.0745**

[a]The analysis was based on 6870 human genes and their mouse orthologs.
[b]The analysis was based on 4903 mouse genes and their human orthologs.
[c]The analysis was based on 1768 fruit fly genes. The d_N, d_S and d_N/d_S values were estimated in single-copy orthologs within the six *Drosophila* group species (23).
Significance: $*P < 0.05$, $**P < 0.01$, and $***P < 0.001$.

when the other confounding factors are controlled, suggesting that N_{miR} is much more important than N_{TF} in affecting d_N, d_S and d_N/d_S (Table 2). This trend is maintained across metazoans.

Dependent effects of TF and miRNA regulations on protein evolutionary rates in metazoans

It is known that TF and miRNA would cooperate with each other in gene regulation (3–5). In addition, genes with more TFBSs have a higher probability to be targeted by miRNAs and tend to have more miRNA-binding sites in human (7). Also, highly connected TFs in human regulatory network tend to regulate more miRNAs and to be more regulated by miRNAs (67). Accordingly, we speculate that there is a positive correlation between N_{TF} and N_{miR}. To address this, we examine the Pearson's coefficient of correlation (r) between these two *trans*-regulatory factors for human, mouse and fruit fly. Figure 1 shows that N_{TF} is indeed positively correlated with N_{miR} and such a trend holds in these three species examined (all $P < 0.001$).

To further investigate the relationship between TF and miRNA regulations in evolution, we then ask whether these two *trans*-regulatory factors have an interaction impact on evolutionary rate. To address this question, we respectively divide the human, mouse and fly protein-coding genes into three groups: (i) genes regulated by TFs but not by any miRNAs collected in this study (denoted as 'G_{TF}'); (ii) genes regulated by miRNAs but not by any TFs examined in this study (denoted as 'G_{miR}'); and (iii) genes regulated by both two *trans*-regulatory factors (denoted as 'G_{Both}'). In general, we observe that the median d_N/d_S values are significantly lower in G_{Both} than in G_{TF}/G_{miR}, regardless of examinations in human, mouse and fruit fly (all $P < 0.001$ by the two-tailed Wilcoxon rank sum test; Figure 2). Our result suggests that genes simultaneously regulated by these two types of *trans*-regulatory factors tend to evolve more slowly than those regulated by only one type of *trans*-regulatory factors, suggesting that combination of TF and miRNA regulations has a dependent effect on protein evolutionary rates in metazoans. We further conduct a stepwise multiple regression analysis including N_{TF}, N_{miR} and the other eight confounding factors to explore the interaction effects on d_N/d_S between any two of these 10 factors. According to the stepwise model selection, the trend that the coefficients of the N_{TF}–N_{miR} interaction term ($\beta_{1,2}$) significantly deviate from zero holds in all three species examined (Supplementary Table S1), further supporting the dependence between N_{TF} and N_{miR} in affecting d_N/d_S.

Trans-regulation is much more important in mammals than in fruit fly in determining protein evolutionary rates

We have shown that *trans*-regulation (N_{TF} and N_{miR}) is an important indicator of evolutionary rates in metazoans (Table 2). Considering the other biological factors

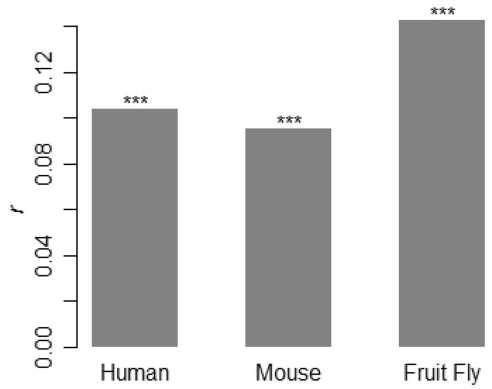

Figure 1. Pearson's coefficient of correlations (r) between N_{TF} and N_{miR} in human, mouse and fruit fly. The analyses were based on genes with TFBSs, miRNA targets and the other eight confounding factors (6870 human genes, 4903 mouse genes and 1768 fruit fly genes). Significance: $***P < 0.001$.

associated with evolutionary rates of proteins [protein connectivity, gene expression (expression level and tissue specificity), gene compactness (UTR length, intron length and intron number) and protein structure (solvent accessibility and disorder content)], we then ask which biological factor(s) is/are the dominant determinant(s) of evolutionary rates. To this end, we measure the relative effect of each individual factor in determining the evolutionary rates by calculating the RCVE (see 'Materials and Methods' section). As shown in Figure 3 and Supplementary Figure S1A, the most dominant determinants of d_N and d_N/d_S common to human and mouse are *trans*-regulation (N_{TF} and N_{miR}) and protein structure (solvent accessibility and disorder content), whereas only protein structure is shown as a dominant determinant in fruit fly. Regarding d_S, *trans*-regulation also exhibits influential determinants in human and mouse; however, the trend is not observed in fruit fly (Supplementary Figure S1B). Our results suggest that the effect of *trans*-regulation (especially miRNA regulation) on protein evolutionary rates is much stronger in mammals than in insects, in consistent with our above finding that the correlations between *trans*-regulation and evolutionary rates are relatively less significant in fruit fly than in human/mouse (Table 2). The results reveal that *trans*-regulation seems to be much more important in human/mouse than in fly in determining the rate of protein sequence evolution.

The above results thus suggest that the relative impacts of *trans*-regulation on evolutionary rates are different between vertebrates and invertebrates. In view of the relationship between regulatory complexity and organismal complexity, there are two possible reasons. First, for TF regulations, previous studies have indicated that organismal complexity might arise from progressively more elaborate gene regulation and the number of TFs per gene is positively correlated with the size of the genome (68–70). Second, in terms of miRNAs, a recent study showed an exponential correlation between the 3′UTR length and morphological complexity (71). The median 3′UTR length is much longer in human than in fruit fly, leading to the conclusion that human genes generally have longer potential miRNA-targeted regions and more complex miRNA regulations (71). Several studies also demonstrated that miRNAs regulate 20−30% of vertebrate genes (72–75) but only 15% of *Drosophila* genes (75). These notions imply that regulatory complexity might increase with the increase of organismal complexity and *trans*-regulations tend to play a much greater role in mammals than in insects, leading to a higher correlation between *trans*-regulation and evolutionary rates in mammals.

Potential caveats

Although the trends that miRNA regulation is much stronger than TF regulation in determining the rate of protein sequence evolution, and TF and miRNA regulations have a dependent effect on evolutionary rates generally hold in metazoans (human, mouse and fruit fly), the limited experimental data (i.e. ChIP-supported TFs and TFBSs) probably cause bias in our results. To address this possibility, we retrieve 843 TRANSFAC human TFs with known PWMs, filter out potentially false-positive TFBSs using the GBP scores (see 'Materials and Methods' section) and then conduct the same analyses. Obviously, the number of the TRANSFAC human TFs is much larger than that of the ChIP-supported TFs used above (843 versus 162; Table 1). On the basis of TRANSFAC-based N_{TF} (or predicted N_{TF}), we find that the abovementioned trends still hold well (Table 3, Supplementary Figures S2A and S3A and Supplementary Table S1). Although highly accurate TFBS predictions (which are currently more comprehensive than experimental data) remain challenging (76–78), the predicted TFBS data used here were generated by integrating multiple evidence sources (including sequence conservation, *cis*-feature, transcriptional information, epigenetic information, and so on) with motif information (54), which were shown to be highly predictive of true locations of TF binding (54,79–81). It is worthwhile to apply our evolutionary analyses to other species (or newly generated data) as the dramatic increase of publicly available *trans*-regulation data. Because the probability of observing the same trends from two biased datasets appears to be small, our results are likely unbiased.

Moreover, because rodents have a faster molecular clock than primates (82,83), it is possible to yield different tendencies between comparison of human–mouse orthologs and that of two species with similar molecular clocks. We therefore ask whether our results may be biased toward different molecular clocks. To address this question, we conduct the same statistical analyses for mouse–rat orthologs, which have similar molecular clocks, and show the same tendencies as above (Table 4, Supplementary Figures S2B and S3B and Supplementary Table S1). These results indicate that these observed trends are not affected by species selection or different molecular clocks. Therefore, our results can be regarded, in a broad view, as exploring the impacts of *trans*-regulation on evolutionary rates.

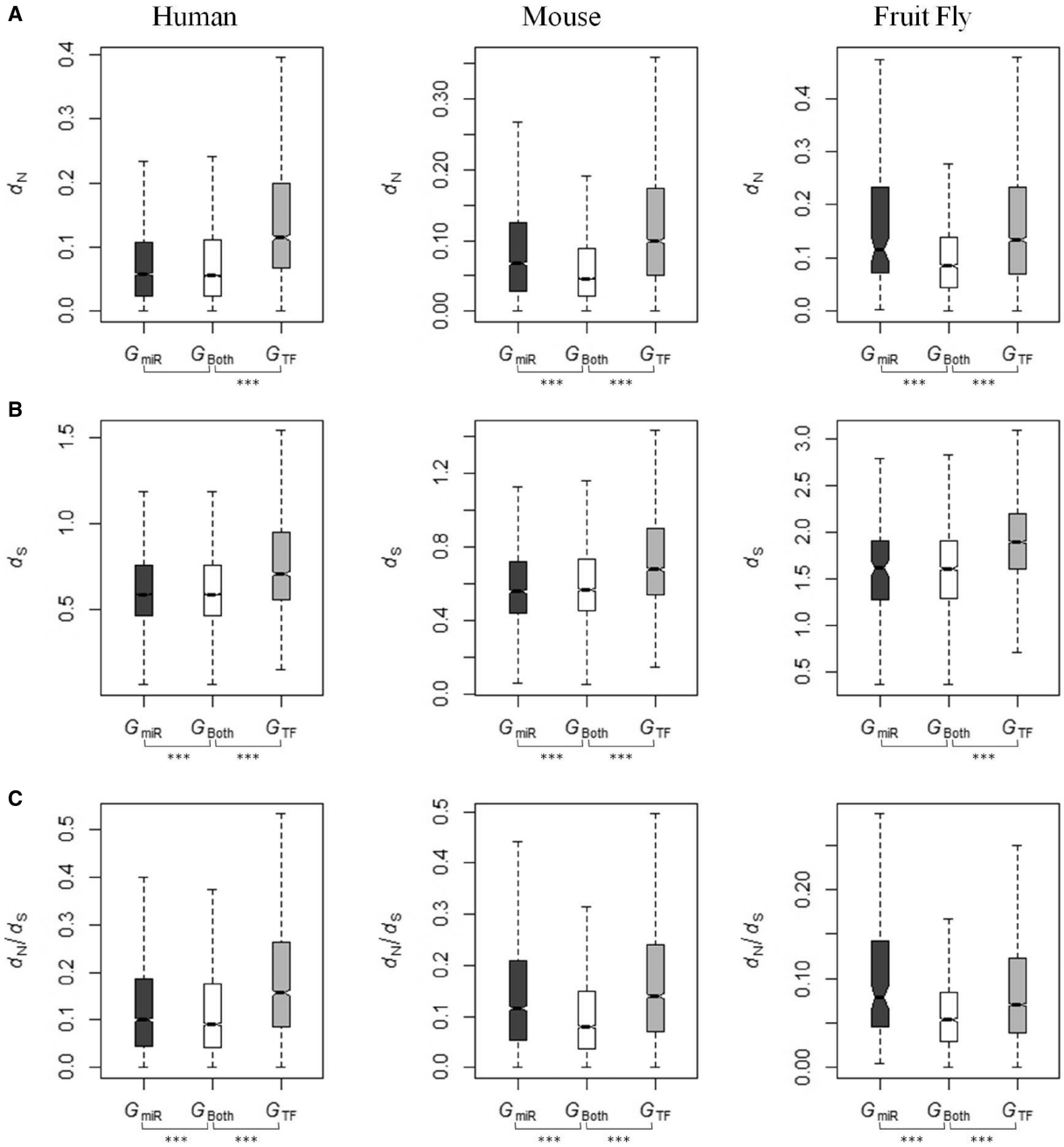

Figure 2. The distributions of relative impact of single-factor and dual-factor regulations on evolutionary rates: (**A**) d_N, (**B**) d_S and (**C**) d_N/d_S. The three vertical bars separately represent a set of genes regulated by miRNAs alone (denoted as 'G_{miR}'), by both of TFs and miRNAs (denoted as 'G_{Both}') and by TFs alone (denoted as 'G_{TF}'). Statistical significance is estimated using the two-tailed Wilcoxon rank sum test: ***$P < 0.001$.

Concluding remarks

This study analyzes the impacts of two *trans*-regulatory factors (N_{TF} and N_{miR}) on the evolutionary rates in the metazoan protein-coding genes. Our results indicate that (i) both N_{TF} and N_{miR} are negatively correlated with evolutionary rates (d_N, d_S and d_N/d_S) in metazoans, but the strength of N_{TF} becomes weak in human/mouse and even disappears in fruit fly if the other confounding factors are controlled for; (ii) evolutionary rates tend to more strongly correlated with N_{miR} than with N_{TF}; (iii) genes simultaneously regulated by TFs and miRNAs are subject to stronger selection pressure than those regulated by only TFs or miRNAs, and the stepwise multiple regression analysis also reveals that the coefficients of the N_{TF}–N_{miR} interaction term ($\beta_{1,2}$) significantly deviate from zero, both of which suggest the dependence between N_{TF} and N_{miR} in affecting d_N/d_S; and (iv) compared with other biological factors, *trans*-regulation exhibits an influential

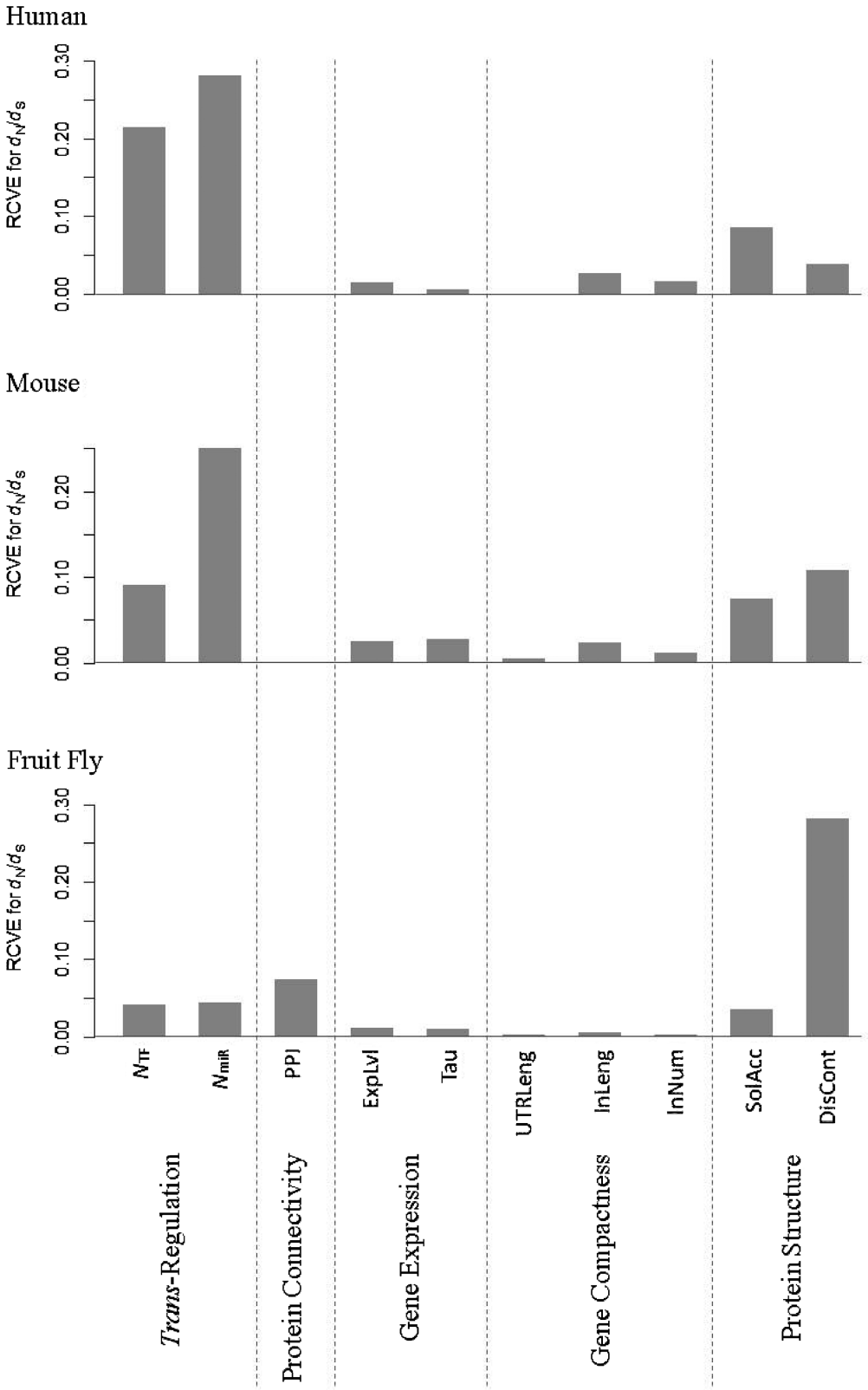

Figure 3. The RCVE of the ten factors: N_{TF}, N_{miR}, protein connectivity (PPI), expression level (ExpLvl), tissue specificity (τ), UTR length (UTRLeng), intron length (InLeng), intron number (InNum), solvent accessibility (SolAcc) and disorder content (DisCont) on d_N/d_S in human, mouse and fruit fly. The analyses were based on 6870 human genes, 4903 mouse genes and 1768 fruit fly genes.

Table 3. Spearman's rank coefficient of correlation (ρ) between human–mouse evolutionary rates (d_N, d_S and d_N/d_S) and predicted N_{TF} (or N_{miR}) before and after controlling for N_{miR} (or predicted N_{TF}) and the other eight confounding factors: protein connectivity, expression level, tissue specificity (τ), UTR length, intron length, intron number, solvent accessibility and disorder content in human

Indicator	Before control			After control		
	d_N	d_S	d_N/d_S	d_N	d_S	d_N/d_S
Evolutionary rate versus N_{TF}	−0.1494***	−0.2075***	−0.0831***	−0.0792***	−0.1520***	−0.0258*
Evolutionary rate versus N_{miR}	−0.3424***	−0.2912***	−0.2567***	−0.3221***	−0.2027***	−0.2610***

The analysis was based on 6653 human genes and their mouse orthologs.
Significance: $*P < 0.05$ and $***P < 0.001$.

Table 4. Spearman's rank coefficient of correlation (ρ) between mouse–rat evolutionary rates (d_N, d_S and d_N/d_S) and experimentally-determined N_{TF} (or N_{miR}) before and after controlling for N_{miR} (or experimentally-determined N_{TF}) and the other eight confounding factors: protein connectivity, expression level, tissue specificity (τ), UTR length, intron length, intron number, solvent accessibility and disorder content in mouse

Indicator	Before control			After control		
	d_N	d_S	d_N/d_S	d_N	d_S	d_N/d_S
Evolutionary rate versus N_{TF}	−0.1116***	−0.0640***	−0.1018***	−0.0629***	−0.0223	−0.0629***
Evolutionary rate versus N_{miR}	−0.2246***	−0.1339***	−0.2000***	−0.2219***	−0.1359***	−0.1986***

The analysis was based on 4620 mouse genes and their rat orthologs.
Significance: $***P < 0.001$.

determinants in determining d_N and d_N/d_S in vertebrates, whereas the effect of *trans*-regulation on protein evolutionary rates is relatively weaker in invertebrates. The first and fourth trends show a great variation in rate determinants between vertebrates and invertebrates, also echoing the previous notion that the rules governing evolutionary rates may not be the same for all species (21). Because the currently available *trans*-regulatory data may only partially represent the reality, we compare the impacts of TFs and miRNAs across species and evaluated the impacts of them by controlling for potential confounding factors. It is found that the second and third trends hold well in diverse species including vertebrates and invertebrate. We therefore suggest that these two observations should be generally maintained in metazoans, although the roles of various rate determinants might be different between species (21) (also see the first and fourth trends). In addition, our result shows remarkable dependent effects of TF and miRNA regulations on protein evolutionary rates. We thus demonstrate the intricate relationships between gene regulations and the actions of natural selection in metazoan protein evolution.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Table 1 and Supplementary Figures 1–3.

ACKNOWLEDGEMENTS

We especially thank Ben-Yang Liao and Chia-Ying Chen for their valuable suggestions, and Ting-Wei Hsu for a part of the data preprocessing.

FUNDING

Funding for open access charge: Genomics Research Center and Institute of Information Science of Academia Sinica; National Science Council of Taiwan [NSC99-2628-B-001-008-MY3 to T.-J.C., and NSC100-2628-E-001-006-MY3 to H.-K.T.].

Conflict of interest statement. None declared.

REFERENCES

1. Latchman,D.S. (1997) Transcription factors: an overview. *Int. J. Biochem. Cell Biol.*, **29**, 1305–1312.
2. Bartel,D.P. (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, **116**, 281–297.
3. Tsang,J., Zhu,J. and van Oudenaarden,A. (2007) MicroRNA-mediated feedback and feedforward loops are recurrent network motifs in mammals. *Mol. Cell*, **26**, 753–767.
4. Shalgi,R., Lieber,D., Oren,M. and Pilpel,Y. (2007) Global and local architecture of the mammalian microRNA-transcription factor regulatory network. *PLoS Computat. Biol.*, **3**, e131.
5. Su,N., Wang,Y., Qian,M. and Deng,M. (2010) Combinatorial regulation of transcription factors and microRNAs. *BMC Syst. Biol.*, **4**, 150.
6. Dahan,O., Gingold,H. and Pilpel,Y. (2011) Regulatory mechanisms and networks couple the different phases of gene expression. *Trends Genet.*, **27**, 316–322.
7. Cui,Q., Yu,Z., Pan,Y., Purisima,E.O. and Wang,E. (2007) MicroRNAs preferentially target the genes with high transcriptional regulation complexity. *Biochem. Biophys. Res. Commun.*, **352**, 733–738.
8. Chen,C.Y., Chen,S.T., Fuh,C.S., Juan,H.F. and Huang,H.C. (2011) Coregulation of transcription factors and microRNAs in human transcriptional regulatory network. *BMC Bioinformatics*, **12(Suppl.1)**, S41.

9. Xia,Y., Franzosa,E.A. and Gerstein,M.B. (2009) Integrated assessment of genomic correlates of protein evolutionary rate. *PLoS Computat. Biol.*, **5**, e1000413.
10. Wang,Y., Franzosa,E.A., Zhang,X.S. and Xia,Y. (2010) Protein evolution in yeast transcription factor subnetworks. *Nucleic Acids Res.*, **38**, 5959–5969.
11. Chen,S.C., Chuang,T.J. and Li,W.H. (2011) The relationships among microRNA regulation, intrinsically disordered regions, and other indicators of protein evolutionary rate. *Mol. Biol. Evol.*, **28**, 2513–2520.
12. Cheng,C., Bhardwaj,N. and Gerstein,M. (2009) The relationship between the evolution of microRNA targets and the length of their UTRs. *BMC Genomics*, **10**, 431.
13. Liao,B.Y., Weng,M.P. and Zhang,J. (2010) Impact of extracellularity on the evolutionary rate of mammalian proteins. *Genome Biol. Evol.*, **2**, 39–43.
14. Lemos,B., Bettencourt,B.R., Meiklejohn,C.D. and Hartl,D.L. (2005) Evolution of proteins and gene expression levels are coupled in *Drosophila* and are independently associated with mRNA abundance, protein length, and number of protein-protein interactions. *Mol. Biol. Evol.*, **22**, 1345–1354.
15. Fraser,H.B., Hirsh,A.E., Steinmetz,L.M., Scharfe,C. and Feldman,M.W. (2002) Evolutionary rate in the protein interaction network. *Science*, **296**, 750–752.
16. Fraser,H.B., Wall,D.P. and Hirsh,A.E. (2003) A simple dependence between protein evolution rate and the number of protein-protein interactions. *BMC Evol. Biol.*, **3**, 11.
17. Drummond,D.A., Raval,A. and Wilke,C.O. (2006) A single determinant dominates the rate of yeast protein evolution. *Mol. Biol. Evol.*, **23**, 327–337.
18. Krylov,D.M., Wolf,Y.I., Rogozin,I.B. and Koonin,E.V. (2003) Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution. *Genome Res.*, **13**, 2229–2235.
19. Plotkin,J.B. and Fraser,H.B. (2007) Assessing the determinants of evolutionary rates in the presence of noise. *Mol. Biol. Evol.*, **24**, 1113–1121.
20. Drummond,D.A. and Wilke,C.O. (2008) Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell*, **134**, 341–352.
21. Liao,B.Y., Scott,N.M. and Zhang,J. (2006) Impacts of gene essentiality, expression pattern, and gene compactness on the evolutionary rate of mammalian proteins. *Mol. Biol. Evol.*, **23**, 2072–2080.
22. Subramanian,S. and Kumar,S. (2004) Gene expression intensity shapes evolutionary rates of the proteins encoded by the vertebrate genome. *Genetics*, **168**, 373–381.
23. Larracuente,A.M., Sackton,T.B., Greenberg,A.J., Wong,A., Singh,N.D., Sturgill,D., Zhang,Y., Oliver,B. and Clark,A.G. (2008) Evolution of protein-coding genes in *Drosophila*. *Trends Genet.*, **24**, 114–123.
24. Marais,G., Domazet-Loso,T., Tautz,D. and Charlesworth,B. (2004) Correlated evolution of synonymous and nonsynonymous sites in *Drosophila*. *J. Mol. Evol.*, **59**, 771–779.
25. Wright,S.I., Yau,C.B., Looseley,M. and Meyers,B.C. (2004) Effects of gene expression on molecular evolution in Arabidopsis thaliana and Arabidopsis lyrata. *Mol. Biol. Evol.*, **21**, 1719–1726.
26. Yang,L. and Gaut,B.S. (2011) Factors that contribute to variation in evolutionary rate among Arabidopsis genes. *Mol. Biol. Evol.*, **28**, 2359–2369.
27. Pal,C., Papp,B. and Hurst,L.D. (2001) Highly expressed genes in yeast evolve slowly. *Genetics*, **158**, 927–931.
28. Drummond,D.A., Bloom,J.D., Adami,C., Wilke,C.O. and Arnold,F.H. (2005) Why highly expressed proteins evolve slowly. *Proc. Natl Acad. Sci. USA*, **102**, 14338–14343.
29. Jovelin,R. and Phillips,P.C. (2009) Evolutionary rates and centrality in the yeast gene regulatory network. *Genome Biol.*, **10**, R35.
30. Bloom,J.D. and Adami,C. (2003) Apparent dependence of protein evolutionary rate on number of interactions is linked to biases in protein-protein interactions data sets. *BMC Evol. Biol.*, **3**, 21.
31. Winter,E.E., Goodstadt,L. and Ponting,C.P. (2004) Elevated rates of protein secretion, evolution, and disease among tissue-specific genes. *Genome Res.*, **14**, 54–61.
32. Duret,L. and Mouchiroud,D. (2000) Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. *Mol. Biol. Evol.*, **17**, 68–74.
33. Park,S.G. and Choi,S.S. (2010) Expression breadth and expression abundance behave differently in correlations with evolutionary rates. *BMC Evol. Biol.*, **10**, 241.
34. Ingvarsson,P.K. (2007) Gene expression and protein length influence codon usage and rates of sequence evolution in Populus tremula. *Mol. Biol. Evol.*, **24**, 836–844.
35. Marais,G., Nouvellet,P., Keightley,P.D. and Charlesworth,B. (2005) Intron size and exon evolution in *Drosophila*. *Genetics*, **170**, 481–485.
36. Lin,Y.S., Hsu,W.L., Hwang,J.K. and Li,W.H. (2007) Proportion of solvent-exposed amino acids in a protein and rate of protein evolution. *Mol. Biol. Evol.*, **24**, 1005–1011.
37. Franzosa,E.A. and Xia,Y. (2009) Structural determinants of protein evolution are context-sensitive at the residue level. *Mol. Biol. Evol.*, **26**, 2387–2395.
38. Bloom,J.D., Drummond,D.A., Arnold,F.H. and Wilke,C.O. (2006) Structural determinants of the rate of protein evolution in yeast. *Mol. Biol. Evol.*, **23**, 1751–1761.
39. Zhou,T., Drummond,D.A. and Wilke,C.O. (2008) Contact density affects protein evolutionary rate from bacteria to animals. *J. Mol. Biol.*, **66**, 395–404.
40. Kim,P.M., Sboner,A., Xia,Y. and Gerstein,M. (2008) The role of disorder in interaction networks: a structural analysis. *Mol. Syst. Biol.*, **4**, 179.
41. Brown,C.J., Johnson,A.K. and Daughdrill,G.W. (2010) Comparing models of evolution for ordered and disordered proteins. *Mol. Biol. Evol.*, **27**, 609–621.
42. Brown,C.J., Takayama,S., Campen,A.M., Vise,P., Marshall,T.W., Oldfield,C.J., Williams,C.J. and Dunker,A.K. (2002) Evolutionary rate heterogeneity in proteins with long disordered regions. *J. Mol. Biol.*, **55**, 104–110.
43. Liang,H. and Li,W.H. (2007) MicroRNA regulation of human protein protein interaction network. *RNA*, **13**, 1402–1408.
44. Tuller,T., Ruppin,E. and Kupiec,M. (2009) Properties of untranslated regions of the S. cerevisiae genome. *BMC Genomics*, **10**, 391.
45. Flicek,P., Amode,M.R., Barrell,D., Beal,K., Brent,S., Chen,Y., Clapham,P., Coates,G., Fairley,S., Fitzgerald,S. *et al.* (2011) Ensembl 2011. *Nucleic Acids Res.*, **39**, D800–D806.
46. McQuilton,P., St Pierre,S.E. and Thurmond,J. (2012) FlyBase 101—the basics of navigating FlyBase. *Nucleic Acids Res.*, **40**, D706–D714.
47. Dunham,I., Kundaje,A., Aldred,S.F., Collins,P.J., Davis,C.A., Doyle,F., Epstein,C.B., Frietze,S., Harrow,J., Kaul,R. *et al.* (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
48. Chen,L., Wu,G. and Ji,H. (2011) hmChIP: a database and web server for exploring publicly available human and mouse ChIP-seq and ChIP-chip data. *Bioinformatics*, **27**, 1447–1448.
49. Murali,T., Pacifico,S., Yu,J., Guest,S., Roberts,G.G. III and Finley,R.L. Jr (2011) DroID 2011: a comprehensive, integrated resource for protein, transcription factor, RNA and gene interactions for *Drosophila*. *Nucleic Acids Res.*, **39**, D736–D743.
50. Roy,S., Ernst,J., Kharchenko,P.V., Kheradpour,P., Negre,N., Eaton,M.L., Landolin,J.M., Bristow,C.A., Ma,L., Lin,M.F. *et al.* (2010) Identification of functional elements and regulatory circuits by *Drosophila* modENCODE. *Science*, **330**, 1787–1797.
51. Gallo,S.M., Gerrard,D.T., Miner,D., Simich,M., Des Soye,B., Bergman,C.M. and Halfon,M.S. (2010) REDfly v3.0: toward a comprehensive database of transcriptional regulatory elements in *Drosophila*. *Nucleic Acids Res.*, **39**, D118–D123.
52. Matys,V., Kel-Margoulis,O.V., Fricke,E., Liebich,I., Land,S., Barre-Dirrie,A., Reuter,I., Chekmenev,D., Krull,M., Hornischer,K. *et al.* (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.*, **34**, D108–D110.
53. Hertz,G.Z. and Stormo,G.D. (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, **15**, 563–577.

54. Ernst,J., Plasterer,H.L., Simon,I. and Bar-Joseph,Z. (2010) Integrating multiple evidence sources to predict transcription factor binding in the human genome. *Genome Res.*, **20**, 526–536.
55. Ruby,J.G., Stark,A., Johnston,W.K., Kellis,M., Bartel,D.P. and Lai,E.C. (2007) Evolution, biogenesis, expression, and target predictions of a substantially expanded set of *Drosophila* microRNAs. *Genome Res.*, **17**, 1850–1864.
56. Friedman,R.C., Farh,K.K., Burge,C.B. and Bartel,D.P. (2009) Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res.*, **19**, 92–105.
57. Szklarczyk,D., Franceschini,A., Kuhn,M., Simonovic,M., Roth,A., Minguez,P., Doerks,T., Stark,M., Muller,J., Bork,P. *et al.* (2011) The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res.*, **39**, D561–D568.
58. Wu,C., Orozco,C., Boyer,J., Leglise,M., Goodale,J., Batalov,S., Hodge,C.L., Haase,J., Janes,J., Huss,J.W. 3rd *et al.* (2009) BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. *Genome Biol.*, **10**, R130.
59. Chintapalli,V.R., Wang,J. and Dow,J.A. (2007) Using FlyAtlas to identify better *Drosophila* melanogaster models of human disease. *Nat. Genet.*, **39**, 715–720.
60. Yanai,I., Benjamin,H., Shmoish,M., Chalifa-Caspi,V., Shklar,M., Ophir,R., Bar-Even,A., Horn-Saban,S., Safran,M., Domany,E. *et al.* (2005) Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics*, **21**, 650–659.
61. Liao,B.Y. and Zhang,J. (2006) Low rates of expression profile divergence in highly expressed genes and tissue-specific genes during mammalian evolution. *Mol. Biol. Evol.*, **23**, 1119–1128.
62. Chen,F.C., Chen,C.J., Li,W.H. and Chuang,T.J. (2010) Gene family size conservation is a good indicator of evolutionary rates. *Mol. Biol. Evol.*, **27**, 1750–1758.
63. Cheng,J., Randall,A.Z., Sweredoski,M.J. and Baldi,P. (2005) SCRATCH: a protein structure and structural feature prediction server. *Nucleic Acids Res.*, **33**, W72–W76.
64. Akgul,C., Moulding,D.A. and Edwards,S.W. (2004) Alternative splicing of Bcl-2-related genes: functional consequences and potential therapeutic applications. *Cell. Mol. Life Sci.*, **61**, 2189–2199.
65. Kvikstad,E.M., Tyekucheva,S., Chiaromonte,F. and Makova,K.D. (2007) A macaque's-eye view of human insertions and deletions: differences in mechanisms. *PLoS Computat. Biol.*, **3**, 1772–1782.
66. Kim,S.H. and Yi,S.V. (2007) Understanding relationship between sequence and functional evolution in yeast proteins. *Genetica*, **131**, 151–156.
67. Gerstein,M.B., Kundaje,A., Hariharan,M., Landt,S.G., Yan,K.K., Cheng,C., Mu,X.J., Khurana,E., Rozowsky,J., Alexander,R. *et al.* (2012) Architecture of the human regulatory network derived from ENCODE data. *Nature*, **489**, 91–100.
68. van Nimwegen,E. (2003) Scaling laws in the functional content of genomes. *Trends Genet.*, **19**, 479–484.
69. Koonin,E.V. and Wolf,Y.I. (2010) Constraints and plasticity in genome and molecular-phenome evolution. *Nat. Rev. Genet.*, **11**, 487–498.
70. Levine,M. and Tjian,R. (2003) Transcription regulation and animal diversity. *Nature*, **424**, 147–151.
71. Chen,C.Y., Chen,S.T., Juan,H.F. and Huang,H.C. (2012) Lengthening of 3'UTR increases with morphological complexity in animal evolution. *Bioinformatics*, **28**, 3178–3181.
72. John,B., Enright,A.J., Aravin,A., Tuschl,T., Sander,C. and Marks,D.S. (2004) Human MicroRNA targets. *PLoS Biol.*, **2**, e363.
73. Lewis,B.P., Burge,C.B. and Bartel,D.P. (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, **120**, 15–20.
74. Xi,R. and Xie,T. (2005) Stem cell self-renewal controlled by chromatin remodeling factors. *Science*, **310**, 1487–1489.
75. Grun,D., Wang,Y.L., Langenberger,D., Gunsalus,K.C. and Rajewsky,N. (2005) microRNA target predictions across seven *Drosophila* species and comparison to mammalian targets. *PLoS Comput. Biol.*, **1**, e13.
76. Furey,T.S. (2012) ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. *Nat. Rev. Genet.*, **13**, 840–852.
77. Orenstein,Y., Linhart,C. and Shamir,R. (2012) Assessment of algorithms for inferring positional weight matrix motifs of transcription factor binding sites using protein binding microarray data. *PLoS One*, **7**, e46145.
78. Hardison,R.C. and Taylor,J. (2012) Genomic approaches towards finding cis-regulatory modules in animals. *Nat. Rev. Genet.*, **13**, 469–483.
79. Joseph,R., Orlov,Y.L., Huss,M., Sun,W., Kong,S.L., Ukil,L., Pan,Y.F., Li,G., Lim,M., Thomsen,J.S. *et al.* (2010) Integrative model of genomic factors for determining binding site selection by estrogen receptor-alpha. *Mol. Syst. Biol.*, **6**, 456.
80. Sandve,G.K., Gundersen,S., Rydbeck,H., Glad,I.K., Holden,L., Holden,M., Liestol,K., Clancy,T., Drablos,F., Ferkingstad,E. *et al.* (2011) The differential disease regulome. *BMC Genomics*, **12**, 353.
81. Oh,Y.M., Kim,J.K., Choi,S. and Yoo,J.Y. (2012) Identification of co-occurring transcription factor binding sites from DNA sequence using clustered position weight matrices. *Nucleic Acids Res.*, **40**, e38.
82. Li,W.H. (1997) *Rates and Patterns of Nucleotide Substitutions*. Sinauer Associates, Sunderland, MA.
83. Nekrutenko,A., Chung,W.Y. and Li,W.H. (2003) An evolutionary approach reveals a high protein-coding capacity of the human genome. *Trends Genet.*, **19**, 306–310.