BMC Genomics

BioMed Central

Research article

Open Access

Weak correlation between sequence conservation in promoter regions and in protein-coding regions of human-mouse orthologous gene pairs

Hirokazu Chiba[1], Riu Yamashita[1], Kengo Kinoshita[1] and Kenta Nakai*[1,2]

Address: [1]Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan and [2]Institute for Bioinformatics Research and Development (BIRD), Japan Science and Technology Agency (JST), Science Plaza, 5-3 Yonban-cho, Chiyoda-ku, Tokyo 102-8666, Japan

Email: Hirokazu Chiba - hchiba@hgc.jp; Riu Yamashita - ryamasi@hgc.jp; Kengo Kinoshita - kinosita@hgc.jp; Kenta Nakai* - knakai@ims.u-tokyo.ac.jp

* Corresponding author

Published: 2 April 2008

BMC Genomics 2008, **9**:152 doi:10.1186/1471-2164-9-152

Received: 21 August 2007
Accepted: 2 April 2008

This article is available from: http://www.biomedcentral.com/1471-2164/9/152

Abstract

Background: Interspecies sequence comparison is a powerful tool to extract functional or evolutionary information from the genomes of organisms. A number of studies have compared protein sequences or promoter sequences between mammals, which provided many insights into genomics. However, the correlation between protein conservation and promoter conservation remains controversial.

Results: We examined promoter conservation as well as protein conservation for 6,901 human and mouse orthologous genes, and observed a very weak correlation between them. We further investigated their relationship by decomposing it based on functional categories, and identified categories with significant tendencies. Remarkably, the 'ribosome' category showed significantly low promoter conservation, despite its high protein conservation, and the 'extracellular matrix' category showed significantly high promoter conservation, in spite of its low protein conservation.

Conclusion: Our results show the relation of gene function to protein conservation and promoter conservation, and revealed that there seem to be nonparallel components between protein and promoter sequence evolution.

Background

Comparative analysis is a powerful approach to extract functional or evolutionary information from biological sequences (reviewed in [1-3]). There were many pioneering works on the molecular evolution of mammalian protein sequences [4], which were followed by large scale comparative analyses between species. Wolfe and Sharp [5] analyzed a collection of 363 mouse and rat orthologous gene pairs, and Murphy [6] examined 615 pairs of orthologous genes between human and rodents. Makalowski et al. [7] performed a comparative analysis for 1,196 cDNA pairs between human and rodents. These studies revealed that the evolutionary rates of protein sequences depend on the protein functions. For example, ribosomal proteins and Ras-like GTPases are highly conserved [7], while proteins for antimicrobial host defenses are highly divergent [6].

On the other hand, comparisons of upstream non-coding sequences have been conducted to investigate the regula-

tory sequences. The complete sequences of mammalian genomes [8-10] facilitated large scale comparisons of non-coding sequences, which provided insights about regulatory sequences. Iwama and Gojobori [11] compared the upstream sequences of 3,750 human-mouse orthologous gene pairs and found that transcription factor genes, particularly those related to developmental processes, show high upstream sequence conservation. Lee et al. [12] also reported that genes involved in adaptive processes tend to have highly conserved upstream regions in mammalian genomes. Choi et al. [13] investigated the levels of non-coding conservation, focusing on tissue-specific genes.

While many efforts have been made to examine protein sequence conservation or regulatory sequence conservation, the relationships between them are poorly understood. Although several researchers have addressed a similar issue, where the relationship between protein evolution and regulatory evolution was examined based on microarray expression data [14-19], there is a discrepancy among their conclusions. Some of the researchers concluded that these two kinds of evolution are decoupled [14,17], while others claimed that there was indeed a correlation between them [15,16,18,19]. Since a substantial amount of the regulatory information is embedded in the promoter region, which is located proximal to the transcriptional start site, examining the protein sequence evolution in relation to the promoter sequence is an alternative approach to address this problem. Recently, Castillo-Davis et al. [20] made the first investigation of the relationship between protein and cis-regulatory sequence evolution using nematode genomes, and observed a weak correlation. As a step to broaden our understanding of genome evolution and function, it seems important to examine these sequences in mammalian genomes, and to analyze them in detail to dissect the relationship. However, such a sequence level analysis has not been carried out for mammals. One of the main problems is the precise determination of the TSS, which is indispensable for identifying reliable promoter regions.

Experimentally validated TSS information can provide a basis for a reliable promoter analysis. Based on large-scale collections of full-length cDNAs [21-24], our group constructed DBTSS, database of transcriptional start sites [25,26], which enabled the reliable identification, annotation and analysis of promoter regions [27-29]. Since abundant TSS data for human and mouse were integrated into DBTSS, large scale cross-species comparisons of promoter regions became possible [30,31]. Recently, our group reported an updated version of DBTSS [32], in which the amount of data was significantly increased.

In this study, we compared promoter sequences as well as protein sequences for 6,901 human and mouse orthologous genes, aiming at two points. First, we carried out a comprehensive comparison of human and mouse promoter sequences, to examine the relationship between promoter conservation and gene function. Second, we tried to elucidate what kinds of relationships exist between promoter conservation and protein conservation in mammals. In the second part, we not only examined the extent of correlation between them, but also investigated the relationship in further detail, by decomposing it based on the functional categories of genes. The results revealed that there seem to be nonparallel components between protein and promoter sequence evolution.

Results

Promoter sequence comparison between human and mouse

We began the analysis with 8,429 promoter pairs of one-to-one orthologous genes between human and mouse. These pairs were compared by using the local alignment program **water** from the EMBOSS package [33]. The resulting distributions of the alignment scores are shown in Figure 1. The distribution has two peaks: a major peak around 1000, and a minor peak a little lower than 100. The minor peak corresponds to the negative control distribution created from randomly shuffled promoter pairs (depicted with a dashed line), indicating the presence of non-orthologous promoters that are not evolutionally related to each other (for an explanation of this phenomenon, see Discussion). The apparent separation of the major and minor peaks indicates that we can discriminate orthologous promoters from non-orthologous ones by examining the local alignment scores. For the following analyses, we used the 6,901 promoter pairs with alignment scores ≥ 200 (82% of the initial data set) to eliminate non-orthologous pairs. The threshold of 200 was chosen so that the proportion of non-orthologous pairs with scores over the threshold was low enough: 200 is the 1.5 percentile of the negative control distribution, and the height of the minor peak is 0.16 times that of the negative control, and thus the proportion of non-orthologous pairs with scores ≥ 200 is estimated to be 0.24% (see Additional file 1). It was possible that the offset of representative TSSs between human and mouse could bias the alignment scores. We evaluated this effect by estimating the offset from the differences in the local alignment end positions and shifting the mouse promoter as much as the offset. As a result of the promoter alignment with the offset correction, we confirmed that the bias was very small (data not shown). Therefore, we retained the original approach.

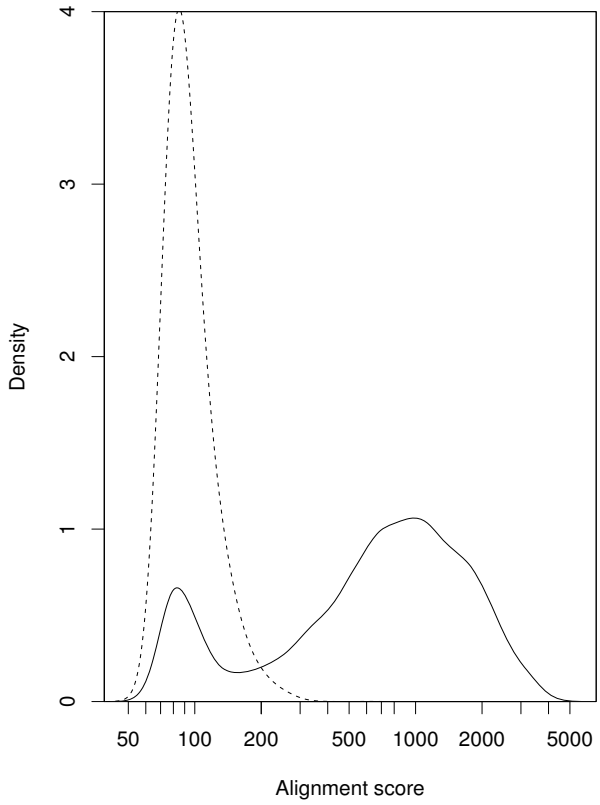

Figure 1
Distribution of alignment scores of human and mouse promoters. The distribution for the orthologous gene pairs is depicted by the solid line, and the distribution for the negative control pairs is shown by the dashed line. The x-axis is shown in a logarithmic scale.

Relationship between gene function and promoter conservation

Based on the promoter sequence comparison between human and mouse for the 6,901 genes, we investigated the relationship between gene function and promoter conservation. Annotations of genes were made by associating human genes with GO terms. To this end, we developed a slimmed-down version of the GO vocabulary, containing 52 terms for biological process (P), 22 for cellular component (C) and 26 for molecular function (F) (see Materials and Methods and Additional file 2 for details). We tested whether the alignment scores for a set of genes associated with a GO term are significantly high or low by a Wilcoxon rank sum test. The resulting GO terms with high promoter conservation are listed in Table 1, and those with low conservation are in Table 2 (only terms with a P-value < 0.01 are in the tables; for the complete list of results, see Additional file 3). Figure 2 shows the distributions of the alignment scores for several GO terms with significant tendencies (all of the distributions for the GO terms listed in Table 1 and 2 are shown in Additional file 4). When we tried the global alignment score, we obtained quite similar tendencies (data not shown). We also confirmed that eliminating the coding sequences from the promoter dataset does not significantly influence the observed tendencies (data not shown, see Materials and Methods for details).

In Table 1, we confirmed that the most significant terms are P:development and P:regulation of transcription [11,12]. Furthermore, an overall observation of the table revealed that the terms with high promoter conservation are related to signaling events inside as well as outside of the cell (P:cell-cell signaling, P:cell surface receptor linked signal transduction, P:ion transport, and P:intracellular signaling cascade). On the other hand, Table 2 covers a wide range of metabolism (P:lipid metabolism, P:carbohydrate metabolism, P:protein biosynthesis, P:proteolysis, P:electron transport, F:oxidoreductase activity, F:nuclease activity). Table 2 also contains cellular components, such as C:mitochondrion, C:lysosome, C:ribosome and C:peroxisome, which correspond to the metabolism-related terms.

Relationship between gene function and protein conservation

The protein conservation tendencies were examined in a similar manner to those of the promoter conservation, using protein sequences obtained from the RefSeq database. Since the alignment score largely depends on the protein length, we used the percentage identity for protein sequences, instead of the alignment scores. GO terms showing high protein conservation are listed in Table 3, and those with low conservation are in Table 4 (only terms with a P-value < 0.01; for the complete list of results, see Additional file 5). Figure 3 shows the distributions of conservation levels for several GO terms with significant tendencies (all of the distributions for the GO terms in Table 3 and 4 are shown in Additional file 6). When we tried global alignment, we obtained quite similar tendencies (data not shown), which is reasonable, given that the coverages of the local alignments were mostly over 95% (data not shown).

Table 3 includes well-known categories for high protein conservation: actins [4], ribosomal proteins, Ras-like GTPases [7] and RNA processing [34], and for low protein conservation, P:immune response [6]. By looking over Table 3, we realized that the categories are composed of a series of processes required for gene expression; from intracellular signaling cascade and regulation of transcription, to RNA processing, protein biosynthesis and intracellular transport. We also find C:cytosol and C:nucleoplasm, where the above-mentioned processes take place, and C:actin cytoskeleton, which is known to be

Table 1: GO categories with high promoter conservation. Terms of biological process are labeled as P, cellular component as C, molecular function as F.

GO term	Number of genes	P-value
P:development	649	0
P:regulation of transcription	602	1.67E-15
F:transcription factor activity	263	3.44E-15
P:transcription	640	4.11E-14
P:nervous system development	154	1.99E-10
P:organ development	213	2.30E-10
P:signal transduction	994	5.19E-10
F:DNA binding	628	3.19E-08
P:morphogenesis	212	9.78E-08
P:cell surface receptor linked signal transduction	363	2.23E-06
P:negative regulation of metabolism	107	1.02E-05
F:receptor binding	221	1.90E-05
P:cell-cell signaling	176	2.27E-05
F:cytoskeletal protein binding	137	4.97E-05
P:negative regulation of biological process	327	6.87E-05
F:ion channel activity	98	9.87E-05
C:extracellular matrix	111	0.000119
C:actin cytoskeleton	85	0.000164
P:cell differentiation	173	0.000179
P:cell adhesion	242	0.000182
P:cellular morphogenesis	111	0.000607
F:ion transporter activity	237	0.001493
P:protein amino acid phosphorylation	213	0.001593
P:ion transport	239	0.001825
F:protein kinase activity	220	0.002033
P:intracellular signaling cascade	431	0.006872
P:chromosome organization and biogenesis	105	0.007832
C:plasma membrane	608	0.008026

Table 2: GO categories with low promoter conservation. Terms of biological process are labeled as P, cellular component as C, molecular function as F.

GO term	Number of genes	P-value
C:mitochondrion	398	5.31E-09
F:oxidoreductase activity	309	2.07E-08
C:lysosome	77	9.94E-08
C:ribosome	114	7.54E-07
P:lipid metabolism	260	1.04E-06
P:carboxylic acid metabolism	225	4.43E-06
F:structural constituent of ribosome	130	5.76E-06
P:amino acid metabolism	112	0.000102
P:electron transport	151	0.000236
P:catabolism	260	0.000251
P:carbohydrate metabolism	220	0.000278
C:peroxisome	49	0.000623
P:protein biosynthesis	283	0.00063
F:nuclease activity	60	0.000772
P:response to biotic stimulus	318	0.000893
C:nucleolus	63	0.004455
P:immune response	270	0.005437
F:iron ion binding	111	0.0055
F:peptidase activity	227	0.005592
P:proteolysis	259	0.006844

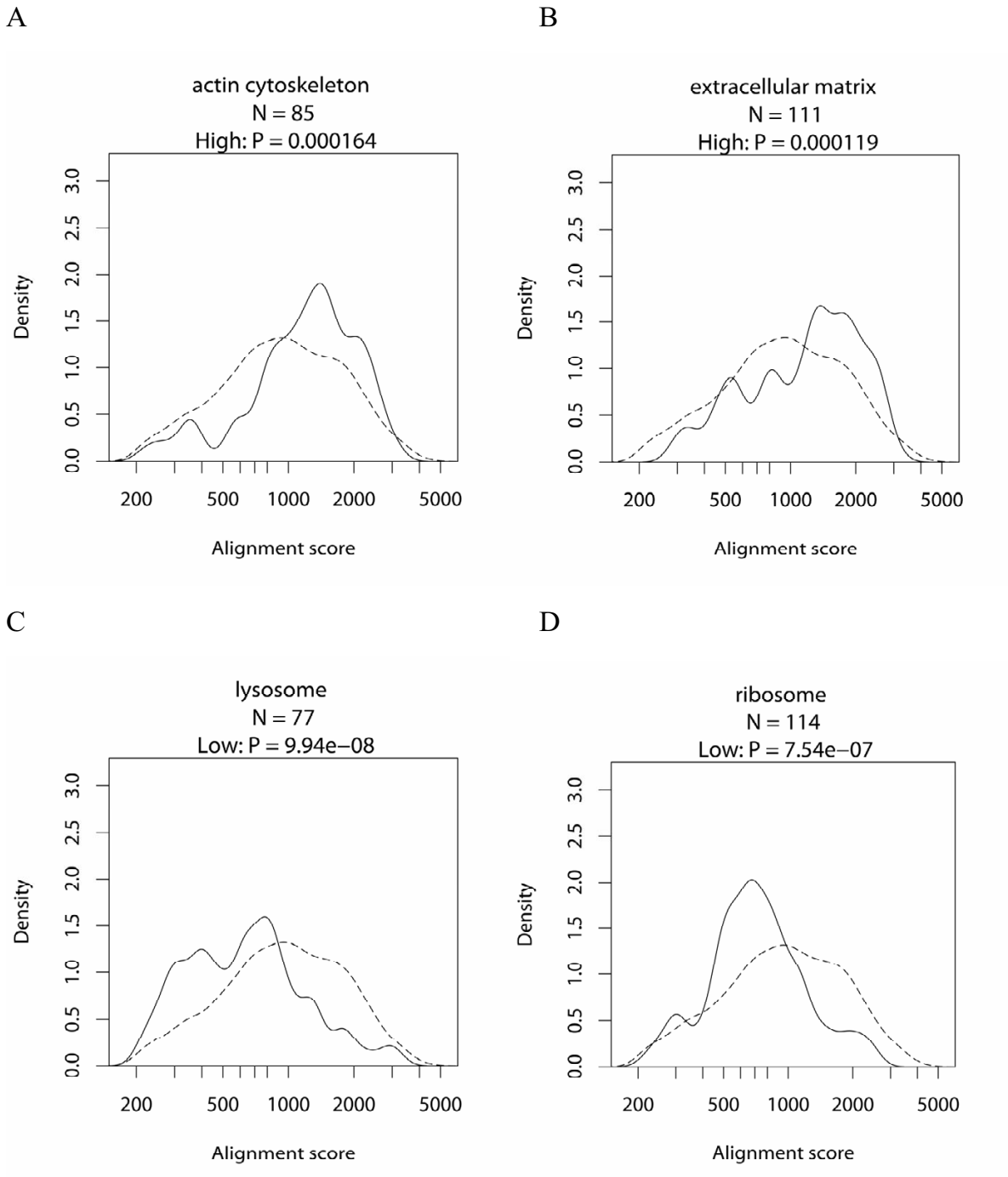

Figure 2

Distribution of alignment scores of human and mouse promoters for several categories with significant tendencies. For the high conservation tendency, actin cytoskeleton (A) and extracellular matrix (B), for the low conservation tendency, lysosome (C) and ribosome (D). For each of A-D, the solid line shows the distribution of the alignment scores for genes with the specific GO term, and the dashed line shows the distribution for the control gene set (see Materials and Methods for details).

Table 3: GO categories with high protein conservation. Terms of biological process are labeled as P, cellular component as C, molecular function as F.

GO term	Number of genes	P-value
F:GTPase activity	88	0
F:GTP binding	160	0
P:intracellular transport	350	0
P:small GTPase mediated signal transduction	126	1.11E-16
F:RNA binding	290	1.33E-15
C:cytosol	171	3.70E-11
P:RNA processing	198	3.25E-10
C:Golgi apparatus	216	5.63E-10
P:intracellular signaling cascade	431	2.57E-09
C:spliceosome complex	37	6.46E-09
P:transcription	640	1.76E-08
P:regulation of transcription	602	2.02E-08
F:ATP binding	520	2.85E-08
C:actin cytoskeleton	85	5.37E-08
P:vesicle-mediated transport	190	7.02E-08
P:cytoskeleton organization and biogenesis	155	9.26E-08
F:cytoskeletal protein binding	137	1.44E-07
P:secretory pathway	102	7.91E-07
C:nucleoplasm	107	1.22E-06
C:ribosome	114	1.36E-06
P:protein biosynthesis	283	1.56E-06
P:ubiquitin cycle	235	2.86E-06
F:ion channel activity	98	7.08E-05
P:protein amino acid phosphorylation	213	0.000101
F:ATPase activity	130	0.000143
C:endomembrane system	163	0.000154
F:protein kinase activity	220	0.000178
P:nervous system development	154	0.000293
F:transcription factor activity	263	0.000465
C:microtubule cytoskeleton	115	0.000595
C:vesicle	86	0.000732
F:structural molecule activity	307	0.000801
F:structural constituent of ribosome	130	0.000843
F:ubiquitin-protein ligase activity	144	0.001969
C:organelle membrane	242	0.004122
P:cell cycle	340	0.006445

involved in transcription [35]. On the other hand, in Table 4, the terms with low conservation are related to extracellular regions or cell surface (C:extracellular space, C:extracellular matrix, C:plasma membrane, F:receptor activity, F:receptor binding, P:cell-cell signaling or P:cell adhesion) or to membrane-bounded organelles (C:lysosome, C:mitochondrion or C:peroxisome). Other terms, such as F:oxidoreductase activity, F:peptidase activity, F:nuclease activity, P:electron tansport and P:proteolysis, correspond to the functions of these cellular components.

Relationship between promoter conservation and protein conservation

To examine the relationship between promoter conservation and protein conservation, we calculated the correlation coefficient of promoter conservation (raw alignment score obtained by **water**) and protein conservation (percentage identity obtained by **water**). This correlation was very weak (the Kendall's rank correlation is 0.193, see Additional file 7 for the scatter plot), suggesting that the promoter and protein sequences are under different types of selective pressure. We further investigated the relationship between protein and promoter conservation in detail, by decomposing it based on GO categories. From Tables 1, 2, 3 and 4, the terms that have significant conservation tendencies for both protein sequences and promoter sequences were extracted and compiled as a 2 by 2 cross table (Table 5). Although this table was basically made by the GO annotations of human genes, the results of the same analysis based on mouse annotations are superimposed, as both analyses were consistent. P:cell-cell signaling was the only exceptional case, showing low protein conservation based on human annotation and high protein conservation on mouse annotation. An

Table 4: GO categories with low protein conservation. Terms of biological process are labeled as P, cellular component as C, molecular function as F.

GO term	Number of genes	P-value
P:response to biotic stimulus	318	4.08E-49
P:immune response	270	1.16E-44
C:extracellular space	179	3.49E-37
P:response to stress	446	5.05E-26
F:oxidoreductase activity	309	2.35E-12
F:receptor activity	391	1.11E-11
F:receptor binding	221	2.15E-11
P:lipid metabolism	260	5.95E-11
P:electron transport	151	7.64E-10
C:lysosome	77	6.38E-08
F:peptidase activity	227	6.15E-07
P:cell proliferation	258	1.65E-06
P:cell adhesion	242	2.16E-06
C:mitochondrion	398	3.00E-05
P:proteolysis	259	4.53E-05
C:extracellular matrix	111	5.52E-05
C:peroxisome	49	8.02E-05
F:nuclease activity	60	8.50E-05
C:plasma membrane	608	0.000291
P:apoptosis	244	0.001373
P:carboxylic acid metabolism	225	0.002527
P:response to abiotic stimulus	148	0.004444
P:positive regulation of biological process	275	0.004599
P:response to chemical stimulus	129	0.004626
P:lipid biosynthesis	101	0.005576
P:cell-cell signaling	176	0.006021
P:sensory perception	111	0.008042

examination of the contents of the two gene sets revealed that the observed difference seems to be derived from the different GO annotation status between human and mouse. Specifically, 151 genes out of 176 are annotated as P:cell-cell signaling only in human, and these genes seems to contribute to the low protein conservation tendency (see Additional file 8). Since human annotations are more abundant, we made the tables with human annotations, and added marks for mouse annotations.

Table 5 illustrates the relationship between protein conservation and promoter conservation, on the functional category basis. GO terms in the upper right cell, which have high conservation for both protein and promoter sequences, are related to transcription regulation or intracellular signaling. In contrast, the membrane-bounded organelles engaged in metabolism are in the lower left cell, showing low conservation for both protein and promoter. Interestingly, several terms are in the upper left and lower right cell, indicating opposite characteristics for protein and promoter conservation. For example, although genes related to signaling events showed high promoter conservation, they do not always have high protein conservation, but can even have low protein conservation; P:cell-cell signaling shows low protein conservation, while F:regulation of transcription shows high protein conservation. An analogous situation can be seen in the case of genes with low promoter conservation; among metabolism-related terms, C:ribosome shows high protein conservation, while C:mitochondrion shows low protein conservation. These results illustrate that there seems to be a nonparallel component in protein and promoter sequence evolution.

Protein and promoter conservation of ribosomal proteins

Unlike other categories, C:ribosome shows a bimodal distribution of protein conservation (Figure 3B); one is around 100% identity, and the other ranges from 70% to 90%. Consistently, several categories related to C:ribosome (P:protein biosynthesis and F:structural constituent of ribosome) also show bimodal distributions (Additional file 6). This result could be due to different evolutionary rates between cytoplasmic and mitochondrial ribosomal protein [36]. Therefore, we checked the annotations for the genes in the C:ribosome category, using the NCBI RefSeq database [37]. In fact, the peak with high protein conservation is substantially composed of cytoplasmic ribosomal proteins, while the peak with lower protein conservation mainly comprises nuclear-encoded mitochondrial ribosomal proteins (Additional file 9).

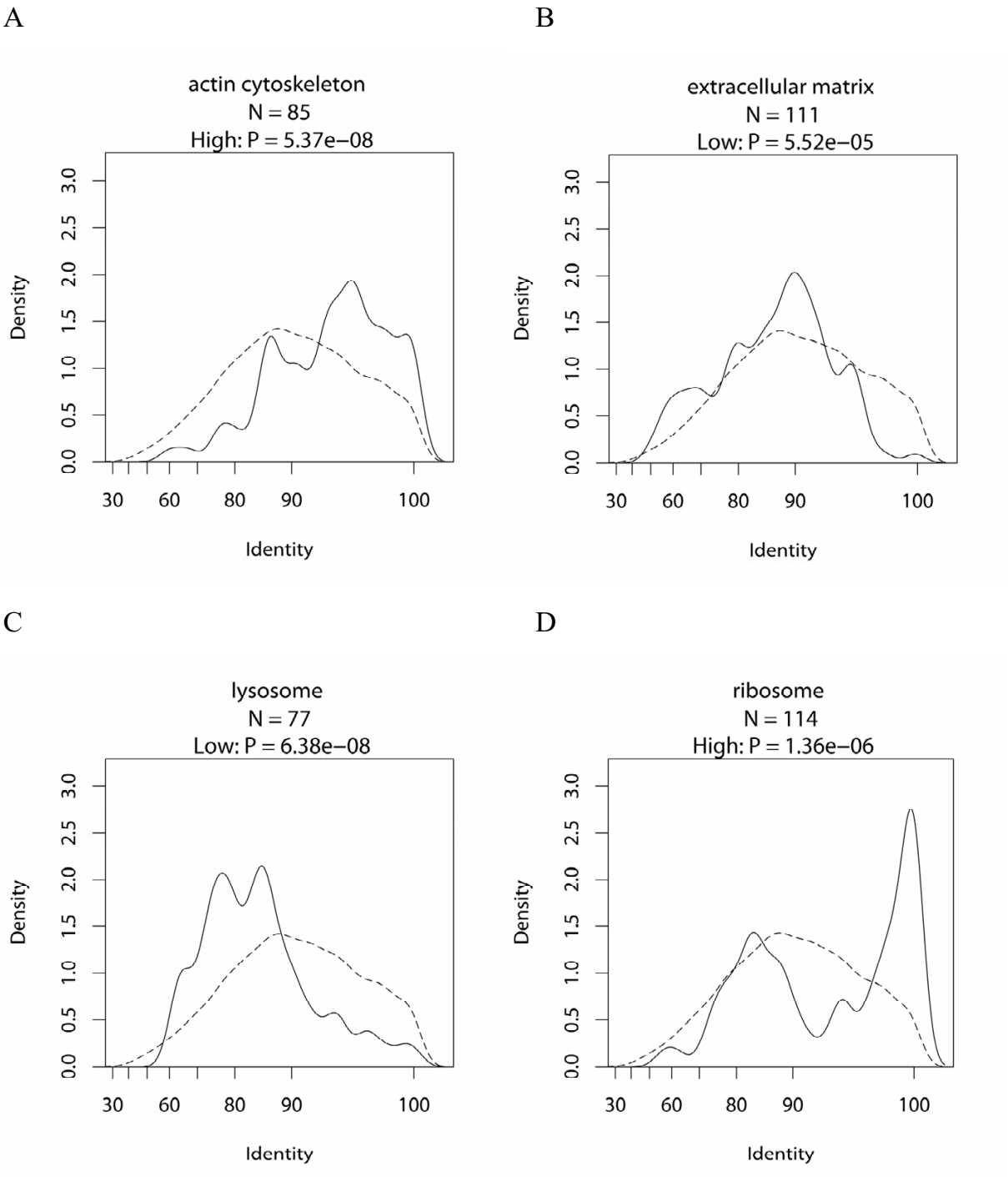

Figure 3

Distribution of percentage identities of human and mouse protein sequences. For the high conservation tendency, actin cytoskeleton (A) and ribosome (D), for the low conservation tendency, extracellular matrix (B) and lysosome (C). For each of A-D, the solid line shows the distribution of the identities for genes with the specific GO term, and the dashed line shows the distribution for the control gene set (see Materials and Methods for details).

Table 5: Summary of GO categories that show significant conservation tendencies for both protein and promoter sequences.

Promoter conservation	Protein conservation: Low	Protein conservation: High
High	F:receptor binding (221) * P:cell-cell signaling (176) * C:extracellular matrix (111) * P:cell adhesion (242) C:plasma membrane (608)	P:regulation of transcription (602) * F:transcription factor activity (263) * P:transcription (640) * P:nervous system development (154) * F:cytoskeletal protein binding (137) * F:ion channel activity (98) * C:actin cytoskeleton (85) * P:protein amino acid phosphorylation (213) * F:protein kinase activity (220) * P:intracellular signaling cascade (431) *
Low	P:proteolysis (259) * F:peptidase activity (227) * P:immune response (270) P:response to biotic stimulus (318) F:nuclease activity (60) * C:peroxisome (49) * P:electron transport (151) * P:carboxylic acid metabolism (225) P:lipid metabolism (260) * C:lysosome (77) * F:oxidoreductase activity (309) * C:mitochondrion (398) *	P:protein biosynthesis (283) * F:structural constituent of ribosome (130) * C:ribosome (114) *

In each cell, the GO categories are ordered by promoter conservation. The number of genes for each term is shown in parentheses. GO annotations associated with human genes were used to make this table. '*' represent GO terms that show a significant tendency not only for the human annotation but also for the mouse annotation.

Notably, the general protein conservation tendency described in previous sections holds here: proteins in the cytosol show high protein conservation, while proteins in membrane-bounded organelles, such as mitochondria, have low protein conservation.

Besides the protein conservation, we examined the promoter conservation tendency for the two subsets of the C:ribosome category, cytoplasmic and mitochondrial ribosomal proteins. In contrast to the protein conservation, we could not observe a significant difference in the conservation levels between these two subgroups (P-value = 0.34 by Wilcoxon rank sum test; see Additional file 10 for details of the distributions). The plot of promoter conservation levels against protein conservation is shown in Figure 4. Apparently, the protein conservation is drastically different between cytoplasmic and mitochondrial ribosomal proteins, whereas the distribution of promoter conservation is quite similar. This result underscores the decoupled property of protein and promoter sequence evolution.

Discussion

When we conducted a comprehensive comparison of promoter sequences for human and mouse orthologous genes, we noted that the promoter pairs of orthologous genes contained non-orthologous promoters. The source of these non-orthologous promoters could be the potential false pairings in the orthologous table. Another possible reason is the presence of alternative promoters [38,39], which can result in the failure to select the corresponding TSSs between human and mouse. The other possible cause is the existence of species-specific promoters; for example, our group recently reported that there are human promoter sequences whose counterparts are completely missing in the mouse genomic sequences [40]. Nevertheless, despite these problems that may cause mispairing of non-orthologous promoters, as much as 82% of the promoter pairs were shown to be evolutionally related in the data set. Although the dynamic aspects of TSSs, such as TSS diversification ad TSS turn over, have been highlighted recently [38,39,41,42], our results show that the representative TSS for each gene has been generally sustained during the evolution of the human and mouse lineages.

We focused on gene pairs with promoters that appeared to be truly evolutionally related, and examined the relationship between promoter conservation and gene function. We found that the terms with high promoter conservation

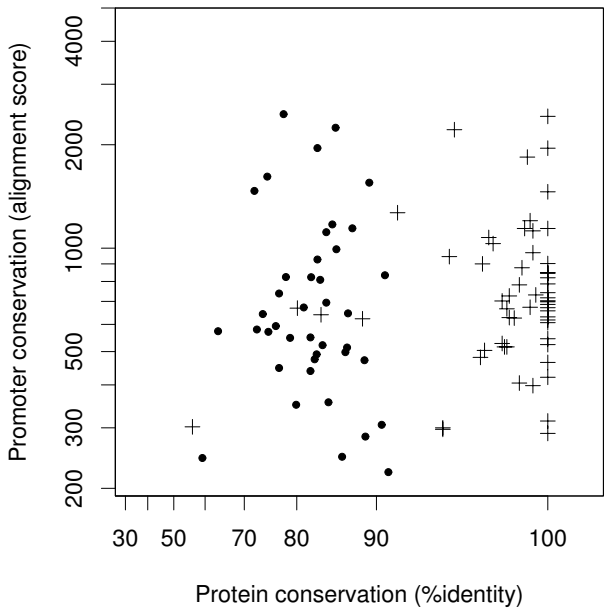

Figure 4
Scatter plot of protein conservation and promoter conservation for two subsets of ribosomal proteins. Crosses represent cytoplasmic ribosomal proteins (58 genes). Dots represent mitochondrial ribosomal proteins (41 genes). The conspicuous outlier corresponding to (56, 302) does not seem to be an actual ribosomal protein, and might have been erroneously annotated by an electronic procedure (see Additional file 9).

are related to signaling events inside as well as outside of the cell. Considering that the promoter conservation levels reflect the regulatory information contained in the sequence, the results suggest that these genes require more regulatory information embedded in the promoter. It is reasonable to suppose that more regulatory information enables more sophisticated changes of expression levels, thereby allowing these proteins to work effectively as signaling molecules. On the other hand, genes involved in metabolism, which showed low promoter conservation, may require relatively less regulatory information in their promoter sequences. Consistently, a recent study revealed that housekeeping genes tend to show reduced upstream sequence conservation [43]. Specifically, in relation to ribosomal proteins, Perry et al. [44] pointed out that most of their promoters contain transposable elements, resulting in a low conservation. The reduced regulatory information in the promoters of ribosomal proteins might be compensated by the translational regulation mechanism directed by the 5' terminal oligopyrimidine sequence in their mRNAs [45].

Related discussions on regulatory sequence conservation have been made for specific categories of genes. Iwama and Gojobori [11] found that transcription factor genes, particularly those related to developmental processes, show high upstream sequence conservation, suggesting that these genes form highly connected regulatory networks. Lee et al. [12] reported that genes involved in adaptive processes tend to have highly conserved upstream regions in mammalian genomes, and also suggested the complex combinatorial circuitry of their transcriptional regulation. There have been other approaches based on whole genome comparisons, where highly conserved non-coding regions were found to be associated with developmental genes [34,46,47]. However, as Lee et al. suggested [12], most of these regions are far from genes and have little overlap with promoter regions. Thus, it seems that these regions are conserved independently from the promoter regions.

The conserved elements in the promoter may be either very short, or spread over a much longer region than the 1,200 bases. In both cases, our measures will report poor conservation when there is just a right amount of it. The local alignment score we used to measure promoter conservation can be roughly considered as a combination of identity and alignment length. Identity reflects the rates of substitutions and indels, and length reflects larger insertions, such as transposon insertions. When we examined the promoter conservation tendency for each GO term, by using alignment length or percentage identity as a measure of conservation, the tendencies were consistent with each other (Additional file 11). Thus, the evolutionary pressures of each functional category on alignment length and identity work in the same direction.

When we investigated the relationship between protein conservation and promoter conservation in mammals, we observed a very weak correlation between them. This suggests that substantial portions of the evolutionary changes of promoter and protein sequences are under different types of selective pressures. This observation is consistent with the nematode [20] and yeast [48] cases, and thus the very weak correlation between protein and promoter conservation might be universal from unicellular organisms to higher vertebrates.

In order to understand the relationship of protein and promoter sequence conservation in terms of gene functions, we examined it by a decomposition based on GO categories. When we dissected not only promoter conservation but also protein conservation, different trends were observed for proteins and promoters. As for proteins, high conservations were observed for terms related to a wide range of gene expression processes that occur in the cytosol and the nucleoplasm, while low conservations

were observed for terms related to extracellular regions, cell surface and membrane-bounded organelles (such as mitochondrion, peroxisome and lysosome). Although the results for the membrane-bounded organelles seem surprising, considering that they often carry out basic, conserved metabolic process, they can also be considered as being topologically "outside" of the cell, given that they are on the opposite side of the membrane from the cytosol. The problem of the determinant of the protein evolutionary rate [49,50] needs to be solved to fully clarify the phenomenon. Nevertheless, our observation provides the trends of the protein sequence evolution in terms of functional categories. Comparing these trends with those of promoters, we found that these two kinds of trends are nonparallel: protein conservation depends on whether they are on the cytosolic side or not, while promoter conservation seems to depend on whether the gene is related to signaling or metabolism. Specifically, cytoplasmic ribosomal proteins, which exist in the cytosol and are engaged in metabolism, have high protein conservation in spite of low promoter conservation. On the other hand, cell-cell signaling genes, which act outside or at the surface of the cell to convey signals, show low protein conservation in spite of high promoter conservation. These terms may provide evidence that decoupled properties exist between protein and promoter sequence evolution.

Conclusion

In this study, we examined the relationship between protein conservation and promoter conservation in detail, by decomposing it based on functional categories. Our results show the relation of gene function to protein conservation and promoter conservation, and revealed that there seem to be nonparallel components between protein and promoter sequence evolution. We believe that this study will provide a basis to understand the evolution of mammalian genes and their regulation. Further efforts are now being made to construct reliable promoter sequences based on full-length cDNAs. Future analyses of multiple species will clarify the evolutionary mechanisms of the coding and regulatory sequences more precisely.

Methods

Sequence comparison

From DBTSS, we obtained human and mouse orthologous gene pairs with experimentally validated TSS information. The definition of an orthologous relationship is based on HomoloGene [51]. One-to-multi orthologous relationships were removed, resulting in 8,429 one-to-one orthologous gene pairs. Since the TSSs for a given gene are not fixed but vary on the chromosome, a representative TSS was defined for each gene, as described in Yamashita et al. [27]. Based on the positions of representative TSSs, sequences from -1000 to +200 were defined as putative promoter sequences. Promoters of orthologous gene pairs were aligned by the local alignment program **water** from the EMBOSS package [33]. In addition, promoter pairs to be used as a negative control were created by shuffling the original pairings, and were aligned similarly. The protein sequences of orthologous gene pairs were obtained from the NCBI reference sequence (RefSeq) database [37]. They were also aligned with **water**. For additional analyses by global alignments, **needle** from the EMBOSS package was used. Furthermore, we confirmed the results after eliminating coding sequences contained in promoter sequences, as follows. The coding sequences downstream of the TSSs were removed by restricting the promoter sequences from -1000 to -1 of the TSSs. In addition, since 16% of the shortened sequences (1,101 out of 6,901) still contained coding sequences, we used the other 5,800 sequences for the additional analyses.

To display the distributions of the alignment scores, they were transformed by common logarithmic transformation, and then the densities were estimated by R with the Gaussian kernel and a band width of 0.5. For protein sequences, protein diversity, instead of identity, was subjected to the logarithmic transformation. In addition, to avoid zero before the logarithmic transformation, a small number was added. Thus, 105 - *identity* was subjected to the logarithmic transformation. This transformation is similar to that described in a previous study on protein evolutionary rates [49].

Annotations of genes

Annotations of genes were based on the gene ontology (GO) [52]. The GO annotations for the human and mouse genes were obtained from the gene2go file at NCBI [53]. In this study, to summarize the attributes of the genes, we developed a slimmed-down version of the GO vocabulary (GO slim), as follows. A set of high level terms was selected to cover most aspects of each of the three ontologies (52 terms for biological process, 22 terms for cellular component and 26 terms for molecular function; for the complete list of selected GO terms, see Additional file 2). Basically, GO terms containing over 100 genes were selected, although well-known cellular components with smaller number of genes, such as C:lysosome and C:peroxisome, were also included. Overly general terms, such as C:cell, P:physiological process and F:binding, were removed, because their biological interpretation seems uninformative. Each GO term was mapped to the GO slim terms using map2slim.pl from the go-perl package [54]. Note that several GO slim terms can be assigned to a single gene; that is, the GO slim terms are not mutually exclusive. In the other sections of the paper, the GO slim terms are referred to as "GO term" for short.

Significance test for the extent of conservation

We tested whether the alignment scores (or percentage identities) of a set of genes associated with a given GO term are significantly high or low by a Wilcoxon rank sum test. It should be noted that the genes used as a control group of a term are those that are not associated with the term, but with other terms. For example, in the case of 'transcription' of biological process, 640 genes are associated with the term among 6,901 genes. Of the 6,261 genes that are not associated with 'transcription', 2,116 genes are missing terms of biological processes. Since these "uncharacterized" genes had low sequence conservation tendencies, we eliminated them from the control gene set. The resulting control set in the case of 'transcription' is thus composed of 4,145 genes.

Authors' contributions

HC carried out the analysis and drafted the manuscript. RY, KK and KN were involved in the study design. KN coordinated the research. All authors read and approved the final manuscript.

Additional material

Additional file 1

Estimated distributions of orthologous and non-orthologous promoter pairs contained in the dataset.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-9-152-S1.pdf]

Additional file 2

Complete list of 100 GO terms selected for this analysis.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-9-152-S2.pdf]

Additional file 3

Complete list of the results of significance tests for promoter conservation of each category.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-9-152-S3.pdf]

Additional file 4

Distribution of alignment scores of human and mouse promoters for GO categories with significant tendencies*. A. GO categories with high protein conservation, B. GO categories with low protein conservation.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-9-152-S4.pdf]

Additional file 5

Complete list of the results of significance tests for protein conservation of each category.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-9-152-S5.pdf]

Additional file 6

Distribution of percentage identities of human and mouse protein sequences for GO categories with significant tendencies*. A. GO categories with high protein conservation, B. GO categories with low protein conservation.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-9-152-S6.pdf]

Additional file 7

Scatter plot of protein conservation and promoter conservation for human and mouse orthologous genes.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-9-152-S7.pdf]

Additional file 8

Protein conservation of human and mouse 'cell-cell signaling' genes.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-9-152-S8.pdf]

Additional file 9

Protein conservations and RefSeq annotations for 'ribosome' category*. Genes are sorted by the percentage identity.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-9-152-S9.pdf]

Additional file 10

Promoter conservations and RefSeq annotations for 'ribosome' category*. Genes are sorted by the alignment score.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-9-152-S10.pdf]

Additional file 11

Promoter conservation tendency for each GO category based on alignment length and percentage identity.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-9-152-S11.pdf]

Acknowledgements

We thank Nicolas Sierro for careful reading of the manuscript and valuable comments. Computation time was provided by the Super Computer System, Human Genome Center, Institute of Medical Science, University of Tokyo. This work was supported by BIRD of Japan Science and Technology Agency (JST).

References

1. O'Brien SJ, Fraser CM: **Genomes and evolution: the power of comparative genomics.** *Curr Opin Genet Dev* 2005, **15(6):**569-571.
2. Ureta-Vidal A, Ettwiller L, Birney E: **Comparative genomics: genome-wide analysis in metazoan eukaryotes.** *Nat Rev Genet* 2003, **4(4):**251-262.
3. Batzoglou S, Pachter L, Mesirov JP, Berger B, Lander ES: **Human and mouse gene structure: comparative analysis and application to exon prediction.** *Genome Res* 2000, **10(7):**950-958.
4. Nei M: **.** In *Molecular Evolutionary Genetics* New York , Columbia University Press; 1987.

5. Wolfe KH, Sharp PM: **Mammalian gene evolution: nucleotide sequence divergence between mouse and rat.** *J Mol Evol* 1993, **37(4):**441-456.
6. Murphy PM: **Molecular mimicry and the generation of host defense protein diversity.** *Cell* 1993, **72(6):**823-826.
7. Makalowski W, Zhang J, Boguski MS: **Comparative analysis of 1196 orthologous mouse and human full-length mRNA and protein sequences.** *Genome Res* 1996, **6(9):**846-857.
8. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Smith DR, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee HM, Dubois J, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis RW, Federspiel NA, Abola AP, Proctor MJ, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson MV, Kaul R, Raymond C, Shimizu N, Kawasaki K, Minoshima S, Evans GA, Athanasiou M, Schultz R, Roe BA, Chen F, Pan H, Ramser J, Lehrach H, Reinhardt R, McCombie WR, de la Bastide M, Dedhia N, Blocker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey JA, Bateman A, Batzoglou S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen HC, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey TS, Galagan J, Gilbert JG, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang W, Johnson LS, Jones TA, Kasif S, Kaspryzk A, Kennedy S, Kent WJ, Kitts P, Koonin EV, Korf I, Kulp D, Lancet D, Lowe TM, McLysaght A, Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz J, Slater G, Smit AF, Stupka E, Szustakowski J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang SP, Yeh RF, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Patrinos A, Morgan MJ, de Jong P, Catanese JJ, Osoegawa K, Shizuya H, Choi S, Chen YJ: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409(6822):**860-921.
9. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, Levine AJ, Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng Z, Di Francesco V, Dunn P, Eilbeck K, Evangelista C, Gabrielian AE, Gan W, Ge W, Gong F, Gu Z, Guan P, Heiman TJ, Higgins ME, Ji RR, Ke Z, Ketchum KA, Lai Z, Lei Y, Li Z, Li J, Liang Y, Lin X, Lu F, Merkulov GV, Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S, Shao W, Shue B, Sun J, Wang Z, Wang A, Wang X, Wang J, Wei M, Wides R, Xiao C, Yan C, Yao A, Ye J, Zhan M, Zhang W, Zhang H, Zhao Q, Zheng L, Zhong F, Zhong W, Zhu S, Zhao S, Gilbert D, Baumhueter S, Spier G, Carter C, Cravchik A, Woodage T, Ali F, An H, Awe A, Baldwin D, Baden H, Barnstead M, Barrow I, Beeson K, Busam D, Carver A, Center A, Cheng ML, Curry L, Danaher S, Davenport L, Desilets R, Dietz S, Dodson K, Doup L, Ferriera S, Garg N, Gluecksmann A, Hart B, Haynes J, Haynes C, Heiner C, Hladun S, Hostin D, Houck J, Howland T, Ibegwam C, Johnson J, Kalush F, Kline L, Koduru S, Love A, Mann F, May D, McCawley S, McIntosh T, McMullen I, Moy M, Moy L, Murphy B, Nelson K, Pfannkoch C, Pratts E, Puri V, Qureshi H, Reardon M, Rodriguez R, Rogers YH, Romblad D, Ruhfel B, Scott R, Sitter C, Smallwood M, Stewart E, Strong R, Suh E, Thomas R, Tint NN, Tse S, Vech C, Wang G, Wetter J, Williams S, Williams M, Windsor S, Winn-Deen E, Wolfe K, Zaveri J, Zaveri K, Abril JF, Guigo R, Campbell MJ, Sjolander KV, Karlak B, Kejariwal A, Mi H, Lazareva B, Hatton T, Narechania A, Diemer K, Muruganujan A, Guo N, Sato S, Bafna V, Istrail S, Lippert R, Schwartz R, Walenz B, Yooseph S, Allen D, Basu A, Baxendale J, Blick L, Caminha M, Carnes-Stine J, Caulk P, Chiang YH, Coyne M, Dahlke C, Mays A, Dombroski M, Donnelly M, Ely D, Esparham S, Fosler C, Gire H, Glanowski S, Glasser K, Glodek A, Gorokhov M, Graham K, Gropman B, Harris M, Heil J, Henderson S, Hoover J, Jennings D, Jordan C, Jordan J, Kasha J, Kagan L, Kraft C, Levitsky A, Lewis M, Liu X, Lopez J, Ma D, Majoros W, McDaniel J, Murphy S, Newman M, Nguyen T, Nguyen N, Nodell M, Pan S, Peck J, Peterson M, Rowe W, Sanders R, Scott J, Simpson M, Smith T, Sprague A, Stockwell T, Turner R, Venter E, Wang M, Wen M, Wu D, Wu M, Xia A, Zandieh A, Zhu X: **The sequence of the human genome.** *Science* 2001, **291(5507):**1304-1351.
10. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, Antonarakis SE, Attwood J, Baertsch R, Bailey J, Barlow K, Beck S, Berry E, Birren B, Bloom T, Bork P, Botcherby M, Bray N, Brent MR, Brown DG, Brown SD, Bult C, Burton J, Butler J, Campbell RD, Carninci P, Cawley S, Chiaromonte F, Chinwalla AT, Church DM, Clamp M, Clee C, Collins FS, Cook LL, Copley RR, Coulson A, Couronne O, Cuff J, Curwen V, Cutts T, Daly M, David R, Davies J, Delehaunty KD, Deri J, Dermitzakis ET, Dewey C, Dickens NJ, Diekhans M, Dodge S, Dubchak I, Dunn DM, Eddy SR, Elnitski L, Emes RD, Eswara P, Eyras E, Felsenfeld A, Fewell GA, Flicek P, Foley K, Frankel WN, Fulton LA, Fulton RS, Furey TS, Gage D, Gibbs RA, Glusman G, Gnerre S, Goldman N, Goodstadt L, Grafham D, Graves TA, Green ED, Gregory S, Guigo R, Guyer M, Hardison RC, Haussler D, Hayashizaki Y, Hillier LW, Hinrichs A, Hlavina W, Holzer T, Hsu F, Hua A, Hubbard T, Hunt A, Jackson I, Jaffe DB, Johnson LS, Jones M, Jones TA, Joy A, Kamal M, Karlsson EK, Karolchik D, Kasprzyk A, Kawai J, Keibler E, Kells C, Kent WJ, Kirby A, Kolbe DL, Korf I, Kucherlapati RS, Kulbokas EJ, Kulp D, Landers T, Leger JP, Leonard S, Letunic I, Levine R, Li J, Li M, Lloyd C, Lucas S, Ma B, Maglott DR, Mardis ER, Matthews L, Mauceli E, Mayer JH, McCarthy M, McCombie WR, McLaren S, McLay K, McPherson JD, Meldrim J, Meredith B, Mesirov JP, Miller W, Miner TL, Mongin E, Montgomery KT, Morgan M, Mott R, Mullikin JC, Muzny DM, Nash WE, Nelson JO, Nhan MN, Nicol R, Ning Z, Nusbaum C, O'Connor MJ, Okazaki Y, Oliver K, Overton-Larty E, Pachter L, Parra G, Pepin KH, Peterson J, Pevzner P, Plumb R, Pohl CS, Poliakov A, Ponce TC, Ponting CP, Potter S, Quail M, Reymond A, Roe BA, Roskin KM, Rubin EM, Rust AG, Santos R, Sapojnikov V, Schultz B, Schultz J, Schwartz MS, Schwartz S, Scott C, Seaman S, Searle S, Sharpe T, Sheridan A, Shownkeen R, Sims S, Singer JB, Slater G, Smit A, Smith DR, Spencer B, Stabenau A, Stange-Thomann N, Sugnet C, Suyama M, Tesler G, Thompson J, Torrents D, Trevaskis E, Tromp J, Ucla C, Ureta-Vidal A, Vinson JP, Von Niederhausern AC, Wade CM, Wall M, Weber RJ, Weiss RB, Wendl MC, West AP, Wetterstrand K, Wheeler R, Whelan S, Wierzbowski J, Willey D, Williams S, Wilson RK, Winter E, Worley KC, Wyman D, Yang S, Yang SP, Zdobnov EM, Zody MC, Lander ES: **Initial sequencing and comparative analysis of the mouse genome.** *Nature* 2002, **420(6915):**520-562.
11. Iwama H, Gojobori T: **Highly conserved upstream sequences for transcription factor genes and implications for the regulatory network.** *Proc Natl Acad Sci U S A* 2004, **101(49):**17156-17161.
12. Lee S, Kohane I, Kasif S: **Genes involved in complex adaptive processes tend to have highly conserved upstream regions in mammalian genomes.** *BMC Genomics* 2005, **6:**168.
13. Choi SS, Bush EC, Lahn BT: **Different classes of tissue-specific genes show different levels of noncoding conservation.** *Genomics* 2006, **87(3):**433-436.
14. Wagner A: **Decoupled evolution of coding region and mRNA expression patterns after gene duplication: implications for the neutralist-selectionist debate.** *Proc Natl Acad Sci U S A* 2000, **97(12):**6579-6584.
15. Gu Z, Nicolae D, Lu HH, Li WH: **Rapid divergence in expression between duplicate genes inferred from microarray data.** *Trends Genet* 2002, **18(12):**609-613.

16. Makova KD, Li WH: **Divergence in the spatial pattern of gene expression between human duplicate genes.** *Genome Res* 2003, **13(7):**1638-1645.
17. Jordan IK, Marino-Ramirez L, Wolf YI, Koonin EV: **Conservation and coevolution in the scale-free human gene coexpression network.** *Mol Biol Evol* 2004, **21(11):**2058-2070.
18. Nuzhdin SV, Wayne ML, Harmon KL, McIntyre LM: **Common pattern of evolution of gene expression level and protein sequence in Drosophila.** *Mol Biol Evol* 2004, **21(7):**1308-1317.
19. Khaitovich P, Hellmann I, Enard W, Nowick K, Leinweber M, Franz H, Weiss G, Lachmann M, Paabo S: **Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees.** *Science* 2005, **309(5742):**1850-1854.
20. Castillo-Davis CI, Hartl DL, Achaz G: **cis-Regulatory and protein evolution in orthologous and duplicate genes.** *Genome Res* 2004, **14(8):**1530-1536.
21. Suzuki Y, Sugano S: **Construction of a full-length enriched and a 5'-end enriched cDNA library using the oligo-capping method.** *Methods Mol Biol* 2003, **221:**73-91.
22. Ota T, Suzuki Y, Nishikawa T, Otsuki T, Sugiyama T, Irie R, Wakamatsu A, Hayashi K, Sato H, Nagai K, Kimura K, Makita H, Sekine M, Obayashi M, Nishi T, Shibahara T, Tanaka T, Ishii S, Yamamoto J, Saito K, Kawai Y, Isono Y, Nakamura Y, Nagahari K, Murakami K, Yasuda T, Iwayanagi T, Wagatsuma M, Shiratori A, Sudo H, Hosoiri T, Kaku Y, Kodaira H, Kondo H, Sugawara M, Takahashi M, Kanda K, Yokoi T, Furuya T, Kikkawa E, Omura Y, Abe K, Kamihara K, Katsuta N, Sato K, Tanikawa M, Yamazaki M, Ninomiya K, Ishibashi T, Yamashita H, Murakawa K, Fujimori K, Tanai H, Kimata M, Watanabe M, Hiraoka S, Chiba Y, Ishida S, Ono Y, Takiguchi S, Watanabe S, Yosida M, Hotuta T, Kusano J, Kanehori K, Takahashi-Fujii A, Hara H, Tanase TO, Nomura Y, Togiya S, Komai F, Hara R, Takeuchi K, Arita M, Imose N, Musashino K, Yuuki H, Oshima A, Sasaki N, Aotsuka S, Yoshikawa Y, Matsunawa H, Ichihara T, Shiohata N, Sano S, Moriya S, Momiyama H, Satoh N, Takami S, Terashima Y, Suzuki O, Nakagawa S, Senoh A, Mizoguchi H, Goto Y, Shimizu F, Wakebe H, Hishigaki H, Watanabe T, Sugiyama A, Takemoto M, Kawakami B, Yamazaki M, Watanabe K, Kumagai A, Itakura S, Fukuzumi Y, Fujimori Y, Komiyama M, Tashiro H, Tanigami A, Fujiwara T, Ono T, Yamada K, Fujii Y, Ozaki K, Hirao M, Ohmori Y, Kawabata A, Hikiji T, Kobatake N, Inagaki H, Ikema Y, Okamoto S, Okitani R, Kawakami T, Noguchi S, Itoh T, Shigeta K, Senba T, Matsumura K, Nakajima Y, Mizuno T, Morinaga M, Sasaki M, Togashi T, Oyama M, Hata H, Watanabe M, Komatsu T, Mizushima-Sugano J, Satoh T, Shirai Y, Takahashi Y, Nakagawa K, Okumura K, Nagase T, Nomura N, Kikuchi H, Masuho Y, Yamashita R, Nakai K, Yada T, Nakamura Y, Ohara O, Isogai T, Sugano S: **Complete sequencing and characterization of 21,243 full-length human cDNAs.** *Nat Genet* 2004, **36(1):**40-45.
23. Carninci P, Hayashizaki Y: **High-efficiency full-length cDNA cloning.** *Methods Enzymol* 1999, **303:**19-44.
24. Okazaki Y, Furuno M, Kasukawa T, Adachi J, Bono H, Kondo S, Nikaido I, Osato N, Saito R, Suzuki H, Yamanaka I, Kiyosawa H, Yagi K, Tomaru Y, Hasegawa Y, Nogami A, Schonbach C, Gojobori T, Baldarelli R, Hill DP, Bult C, Hume DA, Quackenbush J, Schriml LM, Kanapin A, Matsuda H, Batalov S, Beisel KW, Blake JA, Bradt D, Brusic V, Chothia C, Corbani LE, Cousins S, Dalla E, Dragani TA, Fletcher CF, Forrest A, Frazer KS, Gaasterland T, Gariboldi M, Gissi C, Godzik A, Gough J, Grimmond S, Gustincich S, Hirokawa N, Jackson IJ, Jarvis ED, Kanai A, Kawaji H, Kawasawa Y, Kedzierski RM, King BL, Konagaya A, Kurochkin IV, Lee Y, Lenhard B, Lyons PA, Maglott DR, Maltais L, Marchionni L, McKenzie L, Miki H, Nagashima T, Numata K, Okido T, Pavan WJ, Pertea G, Pesole G, Petrovsky N, Pillai R, Pontius JU, Qi D, Ramachandran S, Ravasi T, Reed JC, Reed DJ, Reid J, Ring BZ, Ringwald M, Sandelin A, Schneider C, Semple CA, Setou M, Shimada K, Sultana R, Takenaka Y, Taylor MS, Teasdale RD, Tomita M, Verardo R, Wagner L, Wahlestedt C, Wang Y, Watanabe Y, Wells C, Wilming LG, Wynshaw-Boris A, Yanagisawa M, Yang I, Yang L, Yuan Z, Zavolan M, Zhu Y, Zimmer A, Carninci P, Hayatsu N, Hirozane-Kishikawa T, Konno H, Nakamura M, Sakazume N, Sato K, Shiraki T, Waki K, Kawai J, Aizawa K, Arakawa T, Fukuda S, Hara A, Hashizume W, Imotani K, Ishii Y, Itoh M, Kagawa I, Miyazaki A, Sakai K, Sasaki D, Shibata K, Shinagawa A, Yasunishi A, Yoshino M, Waterston R, Lander ES, Rogers J, Birney E, Hayashizaki Y: **Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs.** *Nature* 2002, **420(6915):**563-573.
25. Suzuki Y, Yamashita R, Nakai K, Sugano S: **DBTSS: DataBase of human Transcriptional Start Sites and full-length cDNAs.** *Nucleic Acids Res* 2002, **30(1):**328-331.
26. Suzuki Y, Yamashita R, Sugano S, Nakai K: **DBTSS, DataBase of Transcriptional Start Sites: progress report 2004.** *Nucleic Acids Res* 2004, **32(Database issue):**D78-81.
27. Yamashita R, Suzuki Y, Sugano S, Nakai K: **Genome-wide analysis reveals strong correlation between CpG islands with nearby transcription start sites of genes and their tissue specificity.** *Gene* 2005, **350(2):**129-136.
28. Sun H, Palaniswamy SK, Pohar TT, Jin VX, Huang TH, Davuluri RV: **MPromDb: an integrated resource for annotation and visualization of mammalian gene promoters and ChIP-chip experimental data.** *Nucleic Acids Res* 2006, **34(Database issue):**D98-103.
29. Jin VX, Singer GA, Agosto-Perez FJ, Liyanarachchi S, Davuluri RV: **Genome-wide analysis of core promoter elements from conserved human and mouse orthologous pairs.** *BMC Bioinformatics* 2006, **7:**114.
30. Suzuki Y, Yamashita R, Shirota M, Sakakibara Y, Chiba J, Mizushima-Sugano J, Nakai K, Sugano S: **Sequence comparison of human and mouse genes reveals a homologous block structure in the promoter regions.** *Genome Res* 2004, **14(9):**1711-1718.
31. Palaniswamy SK, Jin VX, Sun H, Davuluri RV: **OMGProm: a database of orthologous mammalian gene promoters.** *Bioinformatics* 2005, **21(6):**835-836.
32. Yamashita R, Suzuki Y, Wakaguri H, Tsuritani K, Nakai K, Sugano S: **DBTSS: DataBase of Human Transcription Start Sites, progress report 2006.** *Nucleic Acids Res* 2006, **34(Database issue):**D86-9.
33. Rice P, Longden I, Bleasby A: **EMBOSS: the European Molecular Biology Open Software Suite.** *Trends Genet* 2000, **16(6):**276-277.
34. Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, Haussler D: **Ultraconserved elements in the human genome.** *Science* 2004, **304(5675):**1321-1325.
35. Obrdlik A, Kukalev A, Percipalle P: **The function of actin in gene transcription.** *Histol Histopathol* 2007, **22(9):**1051-1055.
36. Pietromonaco SF, Hessler RA, O'Brien TW: **Evolution of proteins in mammalian cytoplasmic and mitochondrial ribosomes.** *J Mol Evol* 1986, **24(1-2):**110-117.
37. Pruitt KD, Tatusova T, Maglott DR: **NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins.** *Nucleic Acids Res* 2005, **33(Database issue):**D501-4.
38. Landry JR, Mager DL, Wilhelm BT: **Complex controls: the role of alternative promoters in mammalian genomes.** *Trends Genet* 2003, **19(11):**640-648.
39. Kimura K, Wakamatsu A, Suzuki Y, Ota T, Nishikawa T, Yamashita R, Yamamoto J, Sekine M, Tsuritani K, Wakaguri H, Ishii S, Sugiyama T, Saito K, Isono Y, Irie R, Kushida N, Yoneyama T, Otsuka R, Kanda K, Yokoi T, Kondo H, Wagatsuma M, Murakawa K, Ishida S, Ishibashi T, Takahashi-Fujii A, Tanase T, Nagai K, Kikuchi H, Nakai K, Isogai T, Sugano S: **Diversification of transcriptional modulation: large-scale identification and characterization of putative alternative promoters of human genes.** *Genome Res* 2006, **16(1):**55-65.
40. Tsuritani K, Irie T, Yamashita R, Sakakibara Y, Wakaguri H, Kanai A, Mizushima-Sugano J, Sugano S, Nakai K, Suzuki Y: **Distinct class of putative "non-conserved" promoters in humans: Comparative studies of alternative promoters of human and mouse genes.** *Genome Res* 2007, **17(7):**1005-1014.
41. Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, Oyama R, Ravasi T, Lenhard B, Wells C, Kodzius R, Shimokawa K, Bajic VB, Brenner SE, Batalov S, Forrest AR, Zavolan M, Davis MJ, Wilming LG, Aidinis V, Allen JE, Ambesi-Impiombato A, Apweiler R, Aturaliya RN, Bailey TL, Bansal M, Baxter L, Beisel KW, Bersano T, Bono H, Chalk AM, Chiu KP, Choudhary V, Christoffels A, Clutterbuck DR, Crowe ML, Dalla E, Dalrymple BP, de Bono B, Della Gatta G, di Bernardo D, Down T, Engstrom P, Fagiolini M, Faulkner G, Fletcher CF, Fukushima T, Furuno M, Futaki S, Gariboldi M, Georgii-Hemming P, Gingeras TR, Gojobori T, Green RE, Gustincich S, Harbers M, Hayashi Y, Hensch TK, Hirokawa N, Hill D, Huminiecki L, Iacono M, Ikeo K, Iwama A, Ishikawa T, Jakt M, Kanapin A, Katoh M, Kawasawa Y, Kelso J, Kitamura H, Kitano H, Kollias G, Krishnan SP, Kruger A, Kummerfeld SK, Kurochkin IV, Lareau LF, Lazarevic D, Lipovich L, Liu J, Liuni S, McWilliam S, Madan Babu M, Madera M, Mar-

chionni L, Matsuda H, Matsuzawa S, Miki H, Mignone F, Miyake S, Morris K, Mottagui-Tabar S, Mulder N, Nakano N, Nakauchi H, Ng P, Nilsson R, Nishiguchi S, Nishikawa S, Nori F, Ohara O, Okazaki Y, Orlando V, Pang KC, Pavan WJ, Pavesi G, Pesole G, Petrovsky N, Piazza S, Reed J, Reid JF, Ring BZ, Ringwald M, Rost B, Ruan Y, Salzberg SL, Sandelin A, Schneider C, Schonbach C, Sekiguchi K, Semple CA, Seno S, Sessa L, Sheng Y, Shibata Y, Shimada H, Shimada K, Silva D, Sinclair B, Sperling S, Stupka E, Sugiura K, Sultana R, Takenaka Y, Taki K, Tammoja K, Tan SL, Tang S, Taylor MS, Tegner J, Teichmann SA, Ueda HR, van Nimwegen E, Verardo R, Wei CL, Yagi K, Yamanishi H, Zabarovsky E, Zhu S, Zimmer A, Hide W, Bult C, Grimmond SM, Teasdale RD, Liu ET, Brusic V, Quackenbush J, Wahlestedt C, Mattick JS, Hume DA, Kai C, Sasaki D, Tomaru Y, Fukuda S, Kanamori-Katayama M, Suzuki M, Aoki J, Arakawa T, Iida J, Imamura K, Itoh M, Kato T, Kawaji H, Kawagashira N, Kawashima T, Kojima M, Kondo S, Konno H, Nakano K, Ninomiya N, Nishio T, Okada M, Plessy C, Shibata K, Shiraki T, Suzuki S, Tagami M, Waki K, Watahiki A, Okamura-Oho Y, Suzuki H, Kawai J, Hayashizaki Y: **The transcriptional landscape of the mammalian genome.** *Science* 2005, **309(5740):**1559-1563.

42. Frith MC, Ponjavic J, Fredman D, Kai C, Kawai J, Carninci P, Hayashizaki Y, Sandelin A: **Evolutionary turnover of mammalian transcription start sites.** *Genome Res* 2006, **16(6):**713-722.
43. Farre D, Bellora N, Mularoni L, Messeguer X, Alba MM: **Housekeeping genes tend to show reduced upstream sequence conservation.** *Genome Biol* 2007, **8(7):**R140.
44. Perry RP: **The architecture of mammalian ribosomal protein promoters.** *BMC Evol Biol* 2005, **5(1):**15.
45. Yoshihama M, Uechi T, Asakawa S, Kawasaki K, Kato S, Higa S, Maeda N, Minoshima S, Tanaka T, Shimizu N, Kenmochi N: **The human ribosomal protein genes: sequencing and comparative analysis of 73 genes.** *Genome Res* 2002, **12(3):**379-390.
46. Sandelin A, Bailey P, Bruce S, Engstrom PG, Klos JM, Wasserman WW, Ericson J, Lenhard B: **Arrays of ultraconserved non-coding regions span the loci of key developmental genes in vertebrate genomes.** *BMC Genomics* 2004, **5(1):**99.
47. Woolfe A, Goodson M, Goode DK, Snell P, McEwen GK, Vavouri T, Smith SF, North P, Callaway H, Kelly K, Walter K, Abnizova I, Gilks W, Edwards YJ, Cooke JE, Elgar G: **Highly conserved non-coding sequences are associated with vertebrate development.** *PLoS Biol* 2005, **3(1):**e7.
48. Chin CS, Chuang JH, Li H: **Genome-wide regulatory complexity in yeast promoters: separation of functionally conserved and neutral sequence.** *Genome Res* 2005, **15(2):**205-213.
49. Wall DP, Hirsh AE, Fraser HB, Kumm J, Giaever G, Eisen MB, Feldman MW: **Functional genomic analysis of the rates of protein evolution.** *Proc Natl Acad Sci U S A* 2005, **102(15):**5483-5488.
50. Pal C, Papp B, Lercher MJ: **An integrated view of protein evolution.** *Nat Rev Genet* 2006, **7(5):**337-348.
51. Maglott D, Ostell J, Pruitt KD, Tatusova T: **Entrez Gene: gene-centered information at NCBI.** *Nucleic Acids Res* 2005, **33(Database issue):**D54-8.
52. Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B, Mungall C, Richter J, Rubin GM, Blake JA, Bult C, Dolan M, Drabkin H, Eppig JT, Hill DP, Ni L, Ringwald M, Balakrishnan R, Cherry JM, Christie KR, Costanzo MC, Dwight SS, Engel S, Fisk DG, Hirschman JE, Hong EL, Nash RS, Sethuraman A, Theesfeld CL, Botstein D, Dolinski K, Feierbach B, Berardini T, Mundodi S, Rhee SY, Apweiler R, Barrell D, Camon E, Dimmer E, Lee V, Chisholm R, Gaudet P, Kibbe W, Kishore R, Schwarz EM, Sternberg P, Gwinn M, Hannick L, Wortman J, Berriman M, Wood V, de la Cruz N, Tonellato P, Jaiswal P, Seigfried T, White R: **The Gene Ontology (GO) database and informatics resource.** *Nucleic Acids Res* 2004, **32(Database issue):**D258-61.
53. **gene2go** [ftp://ftp.ncbi.nih.gov/gene/DATA/]
54. Mungall C: **go-perl.** [http://search.cpan.org/~cmungall/go-perl/].