RESEARCH ARTICLE

A Preliminary Analysis of Gene Pleiotropy Estimated from Protein Sequences

JOURNAL OF EXPERIMENTAL ZOOLOGY
PART B
MOLECULAR AND DEVELOPMENTAL EVOLUTION

ZHIXI SU[1†], YANWU ZENG[1†], AND XUN GU[1,2*]

[1]Institutes of Biomedical Sciences, School of Life Sciences, Center for Evolutionary Biology, Fudan University, Shanghai, China

[2]Department of Genetics, Developmental and Cell Biology (GDCB), Iowa State University, Ames, Iowa

ABSTRACT

Biologists have long recognized the importance of gene pleiotropy, that is, single genes affect multiple traits, which is one of the most commonly observed attributes of genes. Yet the extent of gene pleiotropy has been seriously under-explored. Theoretically, Fisher's model assumed a universal pleiotropy, that is, a mutation can potentially affect all phenotypic traits. On the other hand, experimental assays of a gene usually showed a few distinct phenotypes. Our recent work provides a new approach by estimating the degree of pleiotropy effectively from the phylogenetic sequence analysis. In this article, we estimated the effective gene pleiotropy for 321 vertebrate genes, and found that a gene typically affects 6–7 molecular phenotypes that correspond to the components of organismal fitness, respectively. The positive correlation of gene pleiotropy with the number of Gene Ontology biological processes, as well as the expression broadness provides a biological basis for the sequence-based estimation of gene pleiotropy. On the other hand, the degree of gene pleiotropy has been restricted to a digital number of molecular phenotypes, indicating that some cautions are needed for theoretical analysis of gene pleiotropy based on the assumption of universal pleiotropy. *J. Exp. Zool. (Mol. Dev. Evol.) 314B:115–122, 2010.* © 2009 Wiley-Liss, Inc.

J. Exp. Zool. (Mol. Dev. Evol. 314B:115–122, 2010

How to cite this article: Su Z, Zeng Y, Gu X. 2010. A preliminary analysis of gene pleiotropy estimated from protein sequences. J. Exp. Zool. (Mol. Dev. Evol.) 314B:115–122.

Gene pleiotropy, broadly defined as the capacity of a gene affecting multiple phenotypic characters (Fisher, '30; Williams, '57; Wright, '68), is a major measure for functional importance of a gene (Wagner, '89; Barton, '90; Waxman and Peck, '98; Orr, 2000; Griswold and Whitlock, 2003; Otto, 2004; Dudley et al., 2005; Albert et al., 2008). Though the concept has been invoked to explain many biological phenomena, the degree of gene pleiotropy remains largely unknown. To study the effect of gene pleiotropy on phenotypic evolution, Fisher ('30) proposed the geometric model with the assumption of universal pleiotropy, i.e., a mutation can potentially affect all phenotypic traits. It should be noticed that Fisher's model, as well as its follow-up variations, becomes a theoretical framework for exploring the evolutionary interplay between the genotype and the phenotype (Wagner, '89; Barton, '90; Waxman and Peck, '98; Orr, 2000; Griswold and Whitlock, 2003; Otto, 2004; Dudley et al., 2005; Albert et al., 2008). Yet the notion of universal pleiotropy has not been well tested. By contrast, experimental assays usually showed that the number of phenotypic traits that may be affected by a gene (or quantitative trait locus (QTL)) is actually limited. For instance, Wagner et al. (2008) showed, by studying pleiotropy in mice with the use of QTLs affecting skeletal characters, that most QTLs affect a relatively small number (a medium of six) of traits. Their result did not support the concept of universal pleiotropy, implying the role of modularity in shaping the degree of gene pleiotropy (modular pleiotropy).

[†]Zhixi Su and Yanwu Zeng contributed equally to this work.

Grant Sponsor: National Science Foundation of China; Grant number: 30700140; Grant Sponsor: Universities of Iowa State and Fudan

Additional Supporting Information may be found in the online version of this article.

*Correspondence to: Xun Gu, Department of Genetics, Development and Cell Biology, Iowa State University, Ames, IA 50011.E-mail: xgu@iastate.edu

Received 1 May 2009; Revised 21 June 2009; Accepted 24 June 2009

Published online 27 July 2009 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/jez.b.21315

As the theoretical definition of pleiotropy may differ from the phenotype-specific empirical approach, it is desirable to estimate the degree of pleiotropy under an evolutionary model that is compatible to the original Fisher's model. Our recent work (Gu, 2007a,b) provided a practically feasible approach to address this issue. It proposed that molecular evolution of a gene occurs in a K-dimensional space corresponding to K orthogonal molecular phenotypes, a theoretical representation for multiple functions of a gene. Random mutations of the gene could affect these molecular phenotypes constrained by the stabilizing selection. Gu (2007a) showed that the dimension of molecular phenotypes (K), a measure for gene pleiotropy, can be effectively estimated from the protein sequence data under the conventional nearly neutral model. In this article, we conduct a preliminary analysis of over 300 vertebrate genes to address two issues: (i) whether pleiotropy is a general feature for most genes and (ii) whether the notion of universal pleiotropy is reasonable.

MATERIALS AND METHODS

Vertebrate Sequence Analysis

Data set. The homology information between human and each of other seven species (mouse, dog, cow, chicken, Xenopus, fugu, and zebrafish) and all protein sequences were obtained from Ensembl EnsMart (March 2006, Ensembl release 47, http://www.ensembl.org/Multi/martview) (Kasprzyk et al., 2004). For those genes having multiple protein sequences, the longest protein sequence was kept for subsequent analysis. There are several annotated homology relationships between human and other species by Ensembl. We only considered those pairs of genes annotated as Unique Best Reciprocal Hit (UBRH, meaning that they were unique reciprocal best hits in all-against-all BlastZ searches) to be orthologous. Only those human genes have orthologous in all other seven species were selected for further analysis. Each human gene and its seven orthologous genes compose a homologous group. The number of synonymous substitutions per synonymous site (d_S) and the number of nonsynonymous substitutions per nonsynonymous site (d_N) between human and mouse orthologs were also retrieved form Ensembl EnsMart. In this database, d_S and d_N were estimated by PAML package (Yang, '97) using the likelihood method.

Analysis of rate variation among sites. Multiple protein sequence alignment for each homologous group was generated by ClustalW at default settings (Thompson et al., '94). The phylogenetic trees were also inferred by ClustalW using Neighbour Joining method. Under the inferred phylogeny, we used the method of Gu and Zhang ('97) to estimate the H-measure for rate variation among sites.

Gene Ontology analysis. The number of biological processes (BPs) according to the Gene Ontology (GO) (The Gene Ontology Consortium, 2002) project for all human genes was retrieved based on the following data sets: GO text file, "process.ontology", which was downloaded from http://www.geneontology.org/; and a file containing the associations between all human genes and GO terms, which were derived from Ensembl EnsMart.

Expression analysis. Mouse Affymetrix microarray data were obtained from http://symatlas.gnf.org (Su et al., 2004), which comprised 122 mouse (GNF1M) microchip experiments, including 61 tissues. The expression level detected by each probe set was obtained as the average difference (AD) value computed from MAS 5.0 algorithm (MAS5) (Hubbell et al., 2002). The AD values were averaged among replicates. We mapped the mouse Ensembl genes onto the Affymatrix probe sets, respectively, using the annotation tables available at http://symatlas.gnf.org and the Ensembl EnsMart tool. There are around 20% cases that multiple probe sets in the microarray were targeted against the single gene. As described in Yang et al. (2005), probe sets containing probes with a higher likelihood of cross-hybridization between genes (indicated by a suffix of "_x_at" or "_s_at" in the Affymetrix IDs) are considered lower confidence reporters of gene expression. So for genes with more than one probe set, we discarded all the low-confidence probe sets if higher confidence ones are available and take the average over the remaining probe sets for the given gene. In this study, we use an AD value of 800 as the threshold to calculate the expression breadth (Loose criteria such as AD values >400 or >200 were also used in the definition of gene expression state, and it did not affect our main results).

Estimation of Gene Pleiotropy

Gu (2007a) analyzed the pleiotropy model of molecular evolution under the following assumptions: (i) K-dimensional molecular phenotypes (y) of the gene are under Gaussian-like stabilizing selection, indicating a single fitness optima for multiple functions. Any deviation from the optima is under the purifying selection. (ii) The fitness optima of y may shift randomly during the course of evolution, according to a multivariate normal distribution. It generates the process of *micro-adaptation* that could be caused by the external (environmental) or internal (physiological) perturbations, or the functional compensation for the previously fixed slightly deleterious mutation. (iii) The distribution of mutational effects, p(y), follows a multivariate normal distribution. Under these assumptions, Gu (2007a) has derived the general formulas for the first and second moment of the evolutionary rate of a gene, respectively. In additional to the mutational rate, the first and second moments of the rate depend on a number (K) of selection intensities corresponding to each molecular phenotype, denoted by $B_1,\ldots,B_K$. Further, Gu (2007a) argued that if $B_1,\ldots,B_K$ are not very small, the two formulas for the first and second moments of the evolutionary rate can be simplified to include only three parameters: mutation rate, K and the mean selection intensity of single molecular phenotype (B_0).

Under this assumption, we call K as the effective gene pleiotropy denoted by K_e and B_0 as the baseline selection intensity.

We developed a computational pipeline to estimate the effective gene pleiotropy (K_e) and the baseline selection intensity (B_0). The estimation procedure can be briefly illustrated (Fig. 1) by the triosephosphate isomerase gene (TPI1, SWISSPROT P60174). (i) Infer the phylogenetic tree from the multiple alignment of vertebrate homologous TPI1 protein sequences (human, mouse, dog, cow, chicken, Xenopus, fugu, and zebrafish) by the software ClustalW (Thompson et al., '94). The inferred phylogeny is consistent with the known vertebrate phylogeny. (ii) Estimate the ratio of nonsynonymous to synonymous distances (d_N/d_S) from closely related coding sequences by the likelihood method using PAML (Yang, '97). It is a conventional estimate for the mean evolutionary rate divided by the mutation rate. For TPI1, we estimated $d_N/d_S = 0.045$ between the human and mouse genes, suggesting a strong sequence conservation. (iii) *Estimation of H-index for the rate variation among sites*, which is related to second moment of the evolutionary rate. It is well known that the evolutionary rate (λ) varies among sites because of different functional constraints (Gu et al., '95). Let $m(\lambda)$ and var(λ) be the mean and variance of the rate over sites, respectively. The index H is then defined by $H = \text{var}(\lambda)/[\text{var}(\lambda)+m^2(\lambda)]$ to describe the relative strength of the rate variation among sites: $H = 0$ when var(λ) = 0, and $H = 1$ when var(λ) = ∞. When the phylogeny is inferred, we used the method of Gu and Zhang ('97) to estimate $H = 0.614$ for TPI1 gene, indicating a mildly strong rate variation among sites. (iv) Estimate the effective gene pleiotropy K_e from d_N/d_S and H, according to the formula

$$\frac{d_N}{d_S} = 2^{-K_e/2}\left[1 + \frac{0.0208K_e(K_e+2)}{1+0.289K_e}\right](1-H) \qquad (1)$$

as well as B_0 (Gu, 2007a). In the case of TPI1 gene, we estimated $K_e = 7.29$ and the mean selection intensity $S = -11.65$. Then, the baseline selection intensity is given by $B_0 = 11.65/7.29 \approx 1.60$.

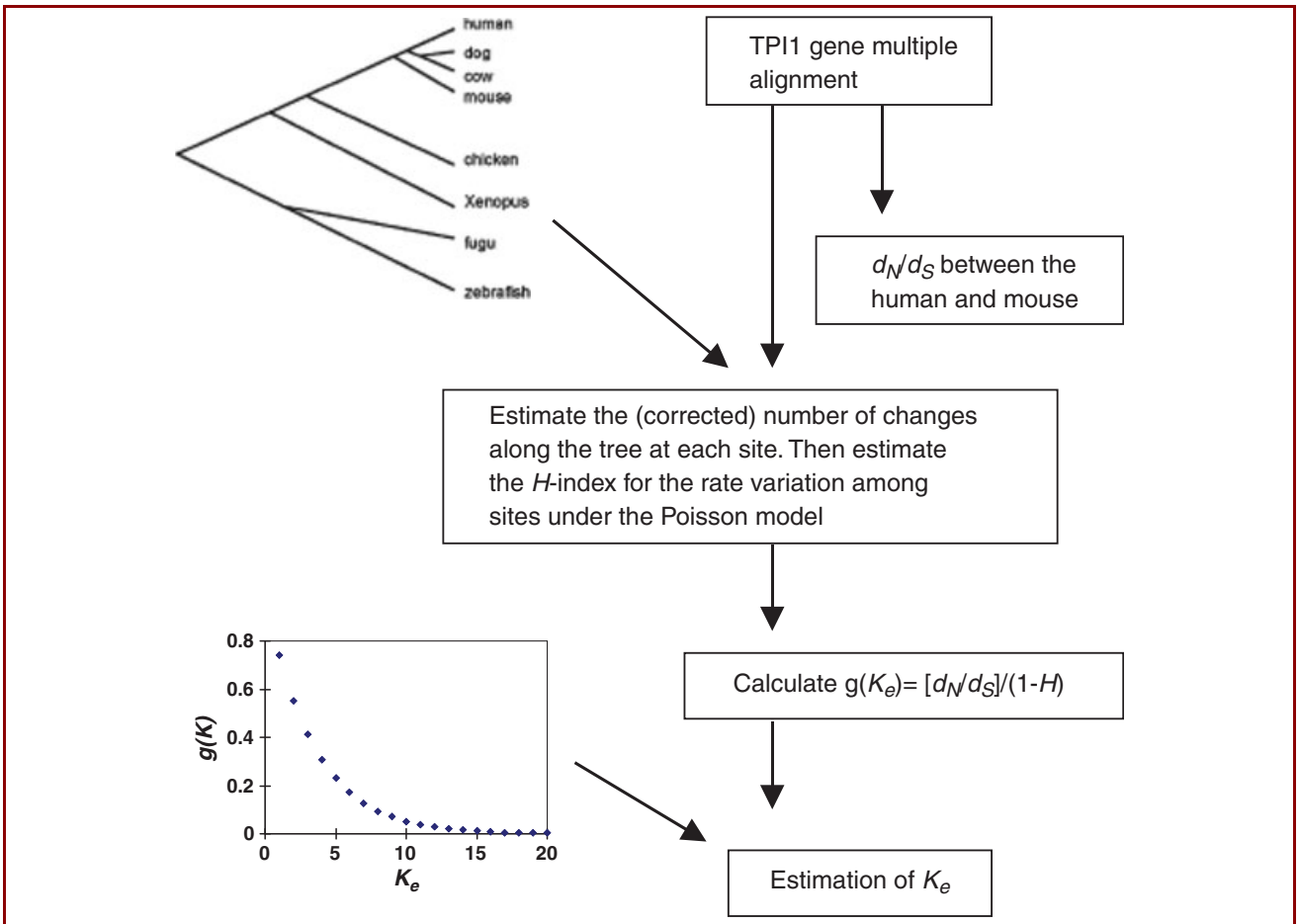

Figure 1. Schematic illustration for the computational procedure of gene pleiotropy estimation.

RESULTS AND DISCUSSION

Extent of Gene Pleiotropy

We analyzed 321 vertebrate genes by the computational procedure as described above. For these vertebrate genes, we observed $d_N/d_S \ll 1$, with the mean 0.098 (between the human and mouse orthologous genes). Based on the known phylogeny of eight vertebrate genomes (Fig. 1), we estimated the *H*-index for each gene. The mean of *H* among vertebrate genes is 0.517, with the 25–75% quantile (0.41–0.61), indicating that most vertebrate genes have a moderate rate variation among sites (Gu et al., '95). For each vertebrate gene with known d_N/d_S and *H*, we estimated the effective gene pleiotropy (K_e) by Eq. (1). As summarized in Table 1, most genes show certain degrees of pleiotropy ($K_e > 3$), supporting the notion that gene pleiotropy is a general feature in gene function. Moreover, the mean of effective gene pleiotropy is $K_e = 6.52$, that is, random mutations of a gene typically affect 6–7 molecular phenotypes, or, correspondingly, fitness components (see Supplementary Table S1 for a list of all studied genes).

To examine whether our analysis is sensitive to the estimation of d_N/d_S, we used all pairwise combinations of four mammals to estimate the d_N/d_S. As summarized in Table 2, the estimated effective gene pleiotropy varies among different treatments but the scale of variation is small. On the other hand, we found that when the number of changes at each site is estimated by the parsimony method without any correction, gene pleiotropy tends to be overestimated (Table 2). At any rate, we conclude that these 5–10% estimation differences should not affect the general pattern about the degree of gene pleiotropy.

In this study, we have selected genes that are one-to-one reciprocal hits across a large number of vertebrate organisms in order to minimize the estimation bias from the heterogeneous sampling of organisms, as well as the effect of duplicate genes. As these genes tend to evolve relatively slowly, our overall view about the degree of gene pleiotropy could be upwardly biased. Though a full evaluation of this issue needs much more genes examined, a small set of mammal-specific genes reveals that K_e, on average, is around 5–6 (data not shown). At any rate, our main conclusion that gene pleiotropy is not universal holds.

Biological Relevance

Moreover, we have conducted some analyses to test whether the effective gene pleiotropy estimated from the protein sequence analysis is biologically interpretable. Biological information relevant to the degree of gene pleiotropy includes a high number of aspects, as revealed by various experimental assays and genomic data types. Not surprisingly, the relationship between the estimated gene pleiotropy (K_e) and any piece of biological evidence of pleiotropy is sophisticated. Consequently, one may expect a positive yet relatively weak correlation between K_e and a single piece of biological evidence.

Table 1. Summary of vertebrate gene pleiotropy analysis.

K_e	Num	d_N/d_S (human–mouse)	d_N/d_S (human–chimpanzee)	d_N (human–mouse)	d_S (human–mouse)	H	S	B_O	Biological processes	Expression broadness
<3	26	0.269 ± 0.022	0.359 ± 0.048	0.161 ± 0.018	0.609 ± 0.033	0.482 ± 0.033	−6.89 ± 0.76	3.38 ± 0.49	1.62 ± 0.28	6.10 ± 2.61
3–4	29	0.184 ± 0.010	0.426 ± 0.084	0.112 ± 0.008	0.607 ± 0.030	0.496 ± 0.025	−7.14 ± 0.24	2.08 ± 0.06	1.76 ± 0.35	5.73 ± 2.76
4–5	47	0.134 ± 0.006	0.347 ± 0.043	0.090 ± 0.006	0.651 ± 0.029	0.492 ± 0.022	−8.30 ± 0.23	1.83 ± 0.05	1.53 ± 0.23	2.26 ± 0.55
5–6	63	0.103 ± 0.003	0.296 ± 0.036	0.072 ± 0.004	0.684 ± 0.031	0.483 ± 0.016	−9.03 ± 0.12	1.65 ± 0.02	1.70 ± 0.27	8.64 ± 2.05
6–7	42	0.065 ± 0.002	0.236 ± 0.037	0.045 ± 0.004	0.678 ± 0.042	0.557 ± 0.016	−10.59 ± 0.14	1.62 ± 0.02	1.74 ± 0.31	8.36 ± 2.41
7–8	37	0.051 ± 0.002	0.164 ± 0.028	0.034 ± 0.003	0.651 ± 0.043	0.537 ± 0.018	−11.42 ± 0.15	1.53 ± 0.02	2.08 ± 0.38	6.82 ± 2.79
8–9	27	0.036 ± 0.002	0.177 ± 0.045	0.022 ± 0.002	0.609 ± 0.046	0.570 ± 0.023	−12.63 ± 0.20	1.50 ± 0.02	2.85 ± 0.80	10.70 ± 3.77
9–10	19	0.026 ± 0.002	0.154 ± 0.052	0.018 ± 0.002	0.668 ± 0.062	0.559 ± 0.031	−13.75 ± 0.25	1.44 ± 0.03	2.00 ± 0.48	10.57 ± 4.03
>10	31	0.010 ± 0.001	0.129 ± 0.047	0.006 ± 0.001	0.579 ± 0.044	0.611 ± 0.027	−18.03 ± 0.61	1.36 ± 0.02	2.26 ± 0.62	13.36 ± 3.24

Note: Effective gene pleiotropy (K_e) was estimated based on the human–mouse orthologous genes for d_N/d_S and Gu-Zhang's (1997) method for *H*. Biological processes was counted from the gene ontology (GO). Expression broadness is the number of mouse tissues in which a gene is expressed, based on the Su et al. (2004).

Intuitively, one may view the concept of molecular phenotypes as canonical BPs that connect between the biochemical-structural features of the encoded protein and the organismal fitness. Hence, the pleiotropic model predicts that more highly pleiotropic genes tend to be involved in more BPs, and vice versa. Though a detailed documentation of BPs that a gene can be involved remains a central matter of functional genomics and systems biology, GO has provided a first-order approximation to addressing this issue. With the help of BP category in GO, we then counted the number of BP GO terms of each gene we study. The mean number of BP is about 1.70. Hence, in spite that BP terms in the BP category of GO are far from the completeness, it indeed indicates a significant portion of genes that may be involved in at least two BPs. Moreover, we found a positive correlation between the number of BP and the effective number of gene pleiotropy (Spearman rank test; $\rho = 0.11$, $n = 321$, $P < 0.001$), as listed in Table 1.

There are some potential sources of bias in the GO analysis. For instance, the number of GO terms might be biased, as the more extensively studied a gene is, the more likely it is to have new GO terms. Consequently, a gene with many GO terms

Table 2. Effects of protein sequence analysis on the estimation of gene pleiotropy.

	d_N/d_S	H	K
(1) Effect of d_N/d_S estimation			
Human–mouse	0.098	0.525	6.52
Human–cow	0.112	0.525	6.25
Human–dog	0.118	0.525	6.14
Mouse–cow	0.096	0.525	6.56
Mouse–dog	0.097	0.525	6.54
Cow–dog	0.120	0.525	6.11
(2) Effect of *H* estimation			
Parsimony	0.098	0.468	6.75

Table 3. A list of highly pleiotropic genes ($K_e > 10$).

Description	d_N/d_S	H	K_e	S	B_0	BP	EB	SWISSPROT /RefSeq(human)
Proteasome subunit α type 3	0.0124	0.738	10.28	−15.80	1.536	1	24	P25788
Tweety 3	0.0231	0.460	10.60	−14.06	1.326	0	5	NP_079526
F-box/LRR-repeat protein 16	0.0183	0.571	10.61	−14.69	1.384	1		Q8N461
Caveolin-3	0.0187	0.521	10.89	−14.68	1.347	2	0	P56539
JNK1-associated membrane protein	0.0181	0.531	10.93	−14.77	1.352	0	8	Q9P055
Syntaxin-binding protein 6	0.0236	0.383	10.97	−14.09	1.284	1		Q8NFX7
Myeloid leukemia factor 2	0.0214	0.429	11.04	−14.36	1.301	2	55	Q15773
Tetraspanin-9	0.0148	0.605	11.05	−15.36	1.390	0	3	O75954
MAM domain containing protein 1	0.0144	0.614	11.05	−15.43	1.396	0		Q7Z553
Ssu72 RNA polymerase II CTD phosphatase homolog	0.0088	0.736	11.42	−16.88	1.479	0	6	NP_054907
Vacuolar ATP synthase subunit E	0.0127	0.610	11.49	−15.84	1.379	4	0	P36543
Thyroid transcription factor 1	0.0120	0.571	11.98	−16.08	1.342	7		P43699
DAZ-associated protein 1	0.0124	0.504	12.35	−16.07	1.301	2	2	Q96EP5
PAF acetylhydrolase 45 kDa subunit	0.0051	0.796	12.37	−18.58	1.502	14	22	P43034
Signal recognition particle 54 kDa protein	0.0076	0.660	12.73	−17.47	1.372	2	0	P61011
Homeobox protein OTX2	0.0142	0.360	12.74	−15.83	1.242	14	2	P32243
Ankyrin repeat and SOCS box protein 13	0.0106	0.490	12.97	−16.63	1.282	1	6	Q8WXK3
Splicing factor 45	0.0066	0.674	13.05	−17.90	1.372	2	0	Q96I25
Cytoplasmic FMR1 interacting protein 1 isoform a	0.0076	0.622	13.08	−17.53	1.340	1	1	NP_055423
Fascin	0.0120	0.353	13.32	−16.40	1.231	3	11	Q16658
Pre-mRNA-processing splicing factor 8	0.0019	0.884	13.78	−21.64	1.570	3	28	Q6P2Q9
Transcriptional adapter 3-like	0.0051	0.599	14.52	−18.83	1.297	3	0	O75528
Cornichon homolog	0.0010	0.886	15.83	−23.63	1.493	2	45	O95406
14-3-3 protein epsilon (14-3-3E)	0.0010	0.777	17.77	−23.61	1.329	1	60	P62258
Defender against cell death 1	0.0010	0.738	18.34	−23.77	1.296	1		P61803
ADP-ribosylation factor-like 10C	0.0008	0.754	18.75	−24.34	1.298	2	30	NP_060654
Breast cancer metastasis-suppressor 1-like	0.0011	0.544	19.98	−24.03	1.203	0	2	NP_115728
F-box/LRR-repeat protein 14	0.0010	0.507	20.40	−24.27	1.189	1	0	Q8N1E6

BP, biological processes and EB, expression broadness.

perhaps is to be well studied. To avoid this problem, we have selected the current data set randomly, without targeting particular genes. Indeed, the distribution of BP number in our gene sampling is virtually the same as that in the human genome, Salathe et al. (2006) suggested to exclude those genes that do not have matched GO for removing the effect of the GO annotation bias to a large extent. Among 321 genes, we observed 98 genes that have received zero BP number. Excluding them we found the pleiotropy-BP correlation is no longer significant, though the trend remains ($P \sim 0.1$ for both spearman rank and linear regression test). Further analysis based on a larger dataset is needed to clarify this issue.

In the same manner, the pleiotropic model predicts that highly pleiotropic genes tend to be expressed in multiple tissues, i.e., expression broadness. It implies that the same biochemical function of a gene (enzyme or binding activity) can be related to several different components of the organismal fitness by having expressions in multiple cell types or tissues. We used the mouse Affimatrix profiles (Su et al., 2004) to test this prediction. To avoid the experimental noise that affects the inference of low-expressed genes severely, we adopted the cut-off of AD value 800. In the literature, a gene that has more than AD value 800 was regarded as the status of "highly expressed". The mean number of tissues in which a mouse gene is expressed is about 8.4. Though the expression in multiple tissues does not necessarily indicate that this gene is functionally pleiotropic, we found that it is significantly correlated with the effective gene pleiotropy (K_e) we have estimated (Spearman rank test; $\rho = 0.13$, $n = 253$, $P = 0.03$). In other words, the increased number of tissues in which a gene is expressed may increase the likelihood to involve distinct phenotypic characters that could affect a number of fitness components.

High and Low Pleiotropic Genes Versus Sequence Conservation

Tables 3 and 4 list the highly pleiotropic genes ($K_e > 10$) and lowly pleiotropic genes ($K_e < 3$), respectively; see Supplementary Table S1 for a complete gene list. It seems that highly pleiotropic genes tend to evolve very slowly at the sequence level (Table 3). For instance, increasing gene pleiotropy from $K_e = 10$ to 20 results in about 20-fold decrease in the rate of protein evolution (d_N/d_S from 0.02 to 0.001). On the other hand, Table 4 summarizes that lowly pleiotropic genes tend to evolve fast. The high correlation between gene pleiotropy (K_e) and sequence conservation (d_N/d_S) estimated by the human and mouse over these hundreds of genes is quite impressive ($r^2 = 0.90$) (Fig. 2).

Table 4. A list of lowly pleiotropic genes ($K_e < 3$).

Description	d_N/d_S	H	K_e	S	B_0	BP	EB	SWISSPROT /RefSeq(human)
MAP3K12-binding inhibitory protein 1 (MBIP)	0.309	0.595	0.911	−10.052	11.03	1	1	Q9NS73
Proliferating-cell nucleolar antigen p120 (NOP2)	0.269	0.626	1.108	−9.808	8.85	2	4	P46087
Polymerase (DNA directed) ν	0.444	0.364	1.202	−3.807	3.17	1	0	NP_861524
Reticulon-1 (Neuroendocrine-specific protein)	0.260	0.573	1.670	−6.712	4.02	2	17	Q16799
ATP synthase coupling factor 6	0.363	0.378	1.810	−4.406	2.43	4	58	P18859
L-threonine dehydrogenase	0.241	0.578	1.883	−6.760	3.59	2		NR_001578
ATP synthase coupling factor B	0.340	0.403	1.901	−4.690	2.47	2	0	Q99766
tRNA-(N1G37) methyltransferase	0.277	0.484	2.103	−5.596	2.66	0	0	NP_065861
Adiponectin receptor protein 2 (ADIPOR2)	0.109	0.785	2.297	−12.324	5.37	3	9	Q86V24
SET and MYND domain containing 4	0.363	0.270	2.364	−4.328	1.83	0	0	NP_443160
Rho GTPase-activating protein 11A isoform 1	0.301	0.383	2.423	−5.052	2.09	0	5	NP_055598
Serine/threonine-protein kinase RIP	0.317	0.347	2.446	−4.845	1.98	4		Q13546
Neuron-specific X11L protein (APBA2)	0.046	0.904	2.475	−22.959	9.28	2	3	Q99767
Peroxisomal 3,2-trans-enoyl-CoA isomerase	0.235	0.485	2.654	−6.060	2.28	2	6	O75521
GTP cyclohydrolase I	0.188	0.586	2.681	−7.162	2.67	4	1	P30793
Toll-like receptor 5	0.281	0.368	2.746	−5.292	1.93	2		O60602
Cathepsin B	0.180	0.593	2.767	−7.307	2.64	1	0	P07858
WD repeat and HMG-box DNA-binding protein 1	0.210	0.520	2.800	−6.519	2.33	1		O75717
Pin2-interacting protein X1	0.176	0.598	2.804	−7.394	2.64	5	0	Q96BK5
Cat eye syndrome critical region protein 5	0.224	0.477	2.879	−6.204	2.15	1	0	Q9BXW7
F-box only protein 34	0.274	0.355	2.904	−5.386	1.85	1	2	Q9NWN3
Cell death regulator Aven	0.282	0.318	2.988	−5.282	1.77	1	0	Q9NQS1

BP, biological processes and EB, expression broadness.

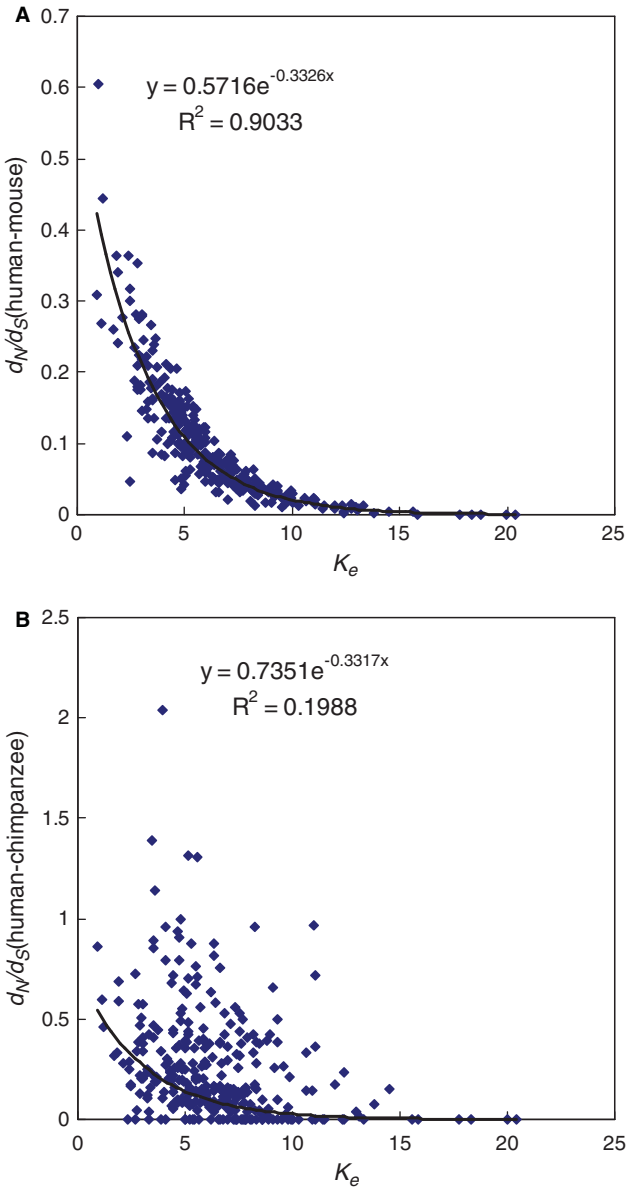

Figure 2. The negative exponential relationship between the sequence conservation (d_N/d_S) and gene pleitropy (K_e). Best-fit exponential curves and associated *R*-square values are shown. (**A**) Decline in the d_N/d_S ratio between human and mouse as a function of increasing K_e. (**B**) Decline in the d_N/d_S ratio between human and chimpanzee as a function of increasing K_e.

As the d_N/d_S (between the human and mouse in our study), together with the *H*-measure, has been involved in the estimation of gene pleiotropy K_e, one may question whether the inverse relationship between d_N/d_S-K_e is an artifact caused by the estimation procedure. To evaluate this problem, we estimated the d_N/d_S ratio between the human and chimpanzee and then to examine whether the d_N/d_S–K_e correlation still holds. As shown in Fig. 2, though the correlation of coefficient is relatively low, $r^2 = 0.20$, the d_N/d_S–K_e correlation holds high significance ($P < 10^{-8}$). In other words, a high pleiotropy estimated by the protein sequences from the human and other nonprimate mammal/vertebrates predicts a slow evolutionary rate between the human and chimpanzee.

As a slowly evolved gene may be because of either high degree of pleiotropy (K) or strong baseline selection intensity (B_0), the prediction of d_N/d_S-K correlation would not be tautological as it can be nullified by its alternative d_N/d_S-B_0. A statistically highly significant, negative d_N/d_S-K correlation among genes may suggest that the degree of pleiotropy (K) varies considerably among genes, whereas the baseline selection intensity (B_0) remains a rough constant (Table 3). The rate–pleiotropy conjecture has been shown critical in the evolutionary study of complexity and adaptation (Wagner, '89), which was also predicted at the molecular evolution and polymorphisms (Waxman and Peck, '98). Meanwhile, the baseline selection intensity varies among lowly pleiotropic genes that generally evolve fast (Table 4). Five genes (MBIP, NOP2, ADIPOR2, Reticulon-1, and APBA2) have low pleiotropic effects but remain extraordinary high baseline selection intensities. In particular, the reason why gene APBA2 evolves slowly ($d_N/d_S = 0.046$) is apparently because of a high degree of baseline selection intensity rather than gene pleiotropy.

Our preliminary analysis is consistent with previous genomic data analyses. For example, previous works showed that more highly connected proteins were under stronger selection (Hahn and Kern, 2005; He and Zhang, 2006). As Promislow (2004) showed that more highly connected proteins had more GO categories associated with them, one would expect that genes with more GO categories (BPs) should be under stronger selection (d_N/d_S). The same argument can be used to explain the correlation between the tissue broadness and sequence conservation (Yang et al., 2005; Gu and Su, 2007).

In spite of the conceptual compatibility, we have to point out that the meaning of pleiotropy differs among these studies. For instance, gene pleiotropy was measured by the connectivity (Hahn and Kern 2005), GO categories (Promislow, 2004; Salathe et al., 2006), regulatory effect of gene deletion (Dudley et al., 2005), as well as affected QTLs (Wagner et al., 2008). In our study, gene pleiotropy is defined by means of affected number of fitness components, a standard definition in the theoretical study. The gap between these updated empirical analyses and the rich body of theoretical studies can be appropriately filled by Gu (2007a) and current work to show that the theoretical definition of pleiotropy can be estimated from the protein sequence analysis.

ACKNOWLEDGMENT

The authors are grateful to Yong Huang for his assistance in data analysis and valuable discussions. They thank Mohamed Noor and Günter Wagner for helpful and constructive comments.

LITERATURE CITED

Albert AYK, Sawaya S, Vines TH, Knecht AK, Miller CT, Summers BR, Balabhadra S, Kingsley DM, Schluter D. 2008. The genetics of adaptive shape shift in stickleback: pleiotropy and effect size. Evolution 62:76–85.

Barton NH. 1990. Pleiotropic models of quantitative variation. Genetics 124:773–782.

Dudley AM, Janse DM, Tanay A, Shamir R, Church GM. 2005. A global view of pleiotropy and phenotypically derived gene function in yeast. Mol Syst Biol 1:2005;0001.

Fisher RA. 1930. The genetical theory of natural selection. Oxford: The Clarendon Press.

Griswold CK, Whitlock MC. 2003. The genetics of adaptation: the roles of pleiotropy, stabilizing selection and drift in shaping the distribution of bidirectional fixed mutational effects. Genetics 165:2181–2192.

Gu X. 2007a. Evolutionary framework for protein sequence evolution and gene pleiotropy. Genetics 175:1813–1822.

Gu X. 2007b. Stabilizing selection of protein function and distribution of selection coefficient among sites. Genetica 130:93–97.

Gu X, Su Z. 2007. Tissue-driven hypothesis of genomic evolution and sequence-expression correlations. Proc Natl Acad Sci USA 104:2779–2784.

Gu X, Zhang J. 1997. A simple method for estimating the parameter of substitution rate variation among sites. Mol Biol Evol 14:1106–1113.

Gu X, Fu YX, Li WH. 1995. Maximum likelihood estimation of the heterogeneity of substitution rate among nucleotide sites. Mol Biol Evol 12:546–557.

Hahn MW, Kern AD. 2005. Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. Mol Biol Evol 22:803–806.

He X, Zhang J. 2006. Toward a molecular understanding of pleiotropy. Genetics 173:1885–1891.

Hubbell E, Liu WM, Mei R. 2002. Robust estimators for expression analysis. Bioinformatics 18:1585–1592.

Kasprzyk A, Keefe D, Smedley D, London D, Spooner W, Melsopp C, Hammond M, Rocca-Serra P, Cox T, Birney E. 2004. EnsMart: a generic system for fast and flexible access to biological data. Genome Res 14:160–169.

Orr HA. 2000. Adaptation and the cost of complexity. Evolution 54:13–20.

Otto SP. 2004. Two steps forward, one step back: the pleiotropic effects of favoured alleles. Proc Biol Sci 271:705–714.

Promislow DEL. 2004. Protein networks, pleiotropy and the evolution of senescence. Proc R Soc (London) 271:1225–1234.

Salathe M, Ackermann M, Bonhoeffer S. 2006. The effect of multifunctionality on the rate of evolution in yeast. Mol Biol Evol 23:721–722.

Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, Cooke MP, Walker JR, Hogenesch JB. 2004. A gene atlas of the mouse and human protein-encoding transcriptomes. Proc Natl Acad Sci USA 101:6062–6067.

The Gene Ontology Consortium. 2002. Gene Ontology: tool for the unification of biology. Nat Genet 25:25–29.

Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673–4680.

Wagner GP. 1989. Multivariate mutation-selection balance with constrained pleiotropic effects. Genetics 122:223–234.

Wagner GP, Kenney-Hunt JP, Pavlicev M, Peck JR, Waxman D, Cheverud JM. 2008. Pleiotropic scaling of gene effects and the ''cost of complexity''. Nature 452:470–472.

Waxman D, Peck JR. 1998. Pleiotropy and the preservation of perfection. Science 279:1210–1213.

Williams GC. 1957. Pleiotropy, natural selection, and the evolution of senescence. Evolution 11:398–411.

Wright S. 1968. Evolution and the genetics of populations; a treatise. Chicago: University of Chicago Press.

Yang J, Su AI, Li WH. 2005. Gene expression evolves faster in narrowly than in broadly expressed mammalian genes. Mol Biol Evol 22:2113–2118.

Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 13:555–556.