Statistical Analysis of Anatomical Expression Pattern And Expression Strength

Osamu Ogasawara

September 1, 2005

Laboratory for Gene-Expression Analysis

Center for Information Biology/DNA Data Bank of Japan

Research Organization of Information and Systems

National Institute of Genetics

Table of Contents

Abstract

The advent of whole-genome sequencing and large-scale profiling of gene expression has revealed several unexpected phenomena in the genome that had never been discovered from the studies of a small number of genes. With the examination of the factors responsible for the formation of such new phenomena, it is conceivable that new relationship between underlying biological processes will be elucidated. In this thesis, I examine such phenomena in the transcriptome of a large variety of organisms, using data from public databases and those obtained in our laboratory and I report unexpected relationships in the transcriptome evolution.

Zipf's law of transcriptome is one of the phenomena that have been revealed by genome-wide studies of gene expression. This law states that there is a relationship between the transcript frequency (f) and abundance rank (r) represented as $f=k/r^b$, where k is a constant and b is a constant parameter that represents the absolute value of the slope in a log-log plot of transcriptome frequencies. I reported in my published paper that this law was applicable to all human normal tissues that I observed. Further, in muscle and liver, which are primarily composed of a homogeneous population of differentiated cells, the slope parameter b was nearly equal to 1. In cell lines, epithelial tissue and compiled transcriptome data, only high-rankers deviate from the law. In addition to my work, several other authors also reported this law in other species. It has been known that this law is applicable to a large variety of species, such as vertebrates (*Homo sapiens, Mus musculus* and *Rattus norvegicus*), invertebrates (*Drosophila melanogaster* and *Caenorhabditis elegans*), other eukaryotes (*Saccharomyces cerevisiae*

and *Arabidopsis thaliana*) and even to bacteria (*Escherichia coli*). It is remarkable that the observed slope parameter *b* is almost unique (*b≈1*) irrespective of the species investigated.

To explain the factors responsible for the formation of the law, I proposed an evolutionary model of Zipf's law of transcriptome. In this model, Zipf's law could be replicated based on three assumptions. The first assumption states that the baseline expression level of each gene in a cell is coded in the genome sequence such as in the *cis*-elements or enhancer regions; therefore, the expression levels of genes are affected by mutations, and these are inherited to the offspring. This assumption is supported by the fact that in a large variety of organisms, the expression levels of genes have abundant natural variation and familial aggregation. In addition, it is known that the location of the *cis*-element and/or the *trans*-factor of the genes can be determined with the quantitative trait locus (QTL) analysis in which the expression level of each gene is treated as a quantitative trait. The second assumption states that the expression level changes in stochastic proportion to its intensity. This assumption is supported by the observation that expression differences accumulate at a constant ratio in primates and rodents. The third assumption states that the number of expressed genes in a cell type is nearly constant throughout the evolutionary process and that any functional gene is prohibited from losing its gene expression ability. By the Monte-Carlo simulation of the model, I showed that a stable distribution of f=0.1/r was obtained from these three assumptions, regardless of the initial distribution. To demonstrate that the uniqueness of slope parameter *b* among variety of species can be replicated from the evolutionary model, I conducted a Monte-Carlo simulation to determine the condition for converging the distribution with the slope parameter *b≈1*. In my model, the slope parameter *b* depends on the number of mRNA molecules in a cell (M), the number of genes expressed in the cell (G), and the permissible lower limit of expression level (L). When the value of parameter M is fixed to 300,000, as in the case of a typical human cell, and L is set to a sufficiently small value, i.e., 1–2 copies/cell, the distribution converged to $b \approx 1$ over a wide range of values of parameter G, i.e., from 10,000 to 50,000 genes in a cell. This is the reason for the universality of *b*, which is predicted by the evolutionary model of Zipf's law.

At approximately the same time, several authors (Kuzunetosov 2003, Frusawa and Kaneko 2003, Ueda 2004) proposed other models of Zipf's law of transcriptome independently (see reference in the thesis 20, 21, 22). All of their models attributed the Zipf's law of transcriptome to the gene expression dynamics in each cell (the dynamics model, hereafter). They argued that the change in gene expression level in each cell follows the formulation of geometrical Brownian movement. However, in addition to the lack of reliable biological evidence for the dynamics model, I pointed out that the dynamics model cannot replicate the observed Zipf's law of transciptome even in mathematical sense, contrary to the authors' assertion. From the dynamics model, it follows that the rank order of expression level in each cell independently diverged at random, even if the same type of cell was considered. It is noteworthy that the observations of expression level distribution were obtained from a mixture of millions of cells. The central limit theorem of probability theory states that the distribution obtained from such a mixture of a large number of completely diverged samples should be a normal distribution. Therefore, if the dynamics model is valid, the observed distribution of the expression level of genes should follow a normal distribution, not a Zipf's law distribution. Such divergence does not occur in my evolutionary model; hence, Zipf's law is replicated.

Obviously, the determination of the correct cause of Zipf's law of transcriptome critically influences the direction of further investigation. Based on the dynamics model of Zipf's law, Ochiai et al. (2004) derived a formula that describes the elementary process of gene expression dynamics in a cell. Determining such a formula is crucial for estimating gene regulatory networks from time course data of gene expression profiles. However, if my assertion is valid, the proposed formula will lose its ground. If my evolutionary model is accepted, Zipf's law of transcriptome would be related to the neutral model of transcriptome evolution, proposed by Khaitovich et al. (2004) (53). They discovered a clocklike accumulation of gene expression divergence within primates and rodents (53). These results were in agreement with the observation of Rifkin et al. (2003), who reported that differences in gene expression were consistent with phylogenetic relationships among Drosophila species(12). Zipf's law of transcriptome can be viewed as a new support for the clocklike accumulation of expression diversity, because the assumption of my evolutionary model is nearly equivalent to the neutral model of transcriptome evolution.

In the evolutionary model of Zipf's law, I focused only on the evolutionary change in expression levels of genes in a cell. Next, I tried to extend my study to the expression patterns in various tissues (anatomical expression pattern, hereafter). To investigate the evolution of anatomical gene expression patterns, I focused on housekeeping genes as a special set of genes that were definitely expressed and function in all cell types.

Identification of housekeeping genes from large-scale expression profiles was first exemplified by Velculescu et al. (1999) who used the SAGE method(24). This was followed by Warrington (2000) and Hsiao (2001) who used oligonucleotide microarrays (25, 26). These studies opened up new opportunities to explore the relationship between expression patterns and other features of genes, such as gene length, sequence divergence, location in the chromosomes, and so on. Several sets of housekeeping genes were published along with such studies, but it has rarely been well discussed whether or not the analyzed set of genes is a non-biased representative of housekeeping genes. In fact, by comparing the two published screenings for housekeeping genes, one based on the GeneChip method and the other based on the SAGE method, I found that there was a low concordance between the results of the two screening methods. I also found that, in both processes, there was poor sensitivity in the identification of housekeeping genes. Therefore, I examined the causes of this inconsistency, and by tuning the parameters for housekeeping gene selection, I compiled a more reliable set of housekeeping genes. In this study, I found a good correlation between the observed breadth of gene expression (the number of organs in which gene expression was detected) and the expression level of genes, even in a set of known housekeeping genes. Based on this, I concluded that the expression level of a gene seriously affects the apparent breadth of its expression. This was particularly manifested in the result where I succeeded in doubling the number of housekeeping gene candidates (from 2,792 to 5,537) without losing specificity. The newly identified housekeeping genes (new HK) and the previously identified housekeeping genes (old HK) shared features in terms of constancy of expression abundance among tissues (expression evenness), cellular localization of products, and the fraction of genes that have CpG islands at their transcription start sites. Estimated contaminants, which comprise approximately 12%–20% of either new or old HK, were genes that were unique to widely distributed cells rather than those that were common to a wide variety of cells.

Main points of this thesis are summarized as follows.

Part I

1. I reported that mRNA frequencies in human normal tissues obeyed the Zipf's law. Especially, in the organs that are primarily composed of a homogeneous population of differentiated cells, the slope parameter was nearly equal to 1.
2. I proposed a new theoretical model for explanation of the factors responsible for the formation of the Zipf's law. It is the evolutionary model of the Zipf's law of transcriptome. Further, I gave several experimental supports for the each assumption of the model.
3. I concluded that the gene expression dynamics models for the Zipf's law of transcriptome are not valid because they can not replicate the Zipf's law even in mathematical sense.

Part II

In order to extend my study from the gene expression strength in a tissue type to the expression patterns in various tissues (anatomical expression pattern), I focused on housekeeping genes as a special set of genes that were definitely expressed and function in all cell types.

1. By comparing the two representative large scale screenings for housekeeping genes in the human genome, I found that there was a low concordance between the results of the two screening methods and there was poor sensitivity in the identification of housekeeping genes.
2. I demonstrated that the cause of the low concordance was that the expression level

of a gene seriously affects the apparent expression breadth (the number of tissues in which the gene was expressed), because there was a good correlation between the observed breadth of gene expression and the expression level of genes, even in a set of known housekeeping genes.

I compiled a new and more reliable set of housekeeping genes, and I succeeded in doubling the number of housekeeping gene candidates (from 2,792 to 5,537) without losing specificity.

Acknowledgments

This thesis is based on the collaborative works of many researchers in Okubo laboratory in DDBJ and Odaiba JBIRC, whereby it is a pleasant aspect that I have now the opportunity to express my gratitude for all of them.

First of all, I am deeply indebted to my direct supervisor Professor Kousaku Okubo for supporting this work. His comments, suggestions, corrections and discussions greatly improved this work. He gave me a chance to learn genome science in this institute therefore I could have learned a lot of things from him about interpretation of expression data, presentation technique and so on.

I would like to thank members of Functional Genomics Group of JBIRC (Biological Information Research Center, Japan) in Odaiba; Teruyoshi Hishiki, Koji Arikawa, Katsuji Murakawa, Masae Maeda, Mitsuhiro Miyazaki, Yuka Onuma, Fumiaki Meguro, Nasa Takahashi, Atsunobu Inoue, Sumiyo Takiguchi, Goichi Tanaka and Mitsuhiro Umeda. I also want to thank Takuro Tamura, President of a bioinfomatics company, BITS Co., Ltd. This thesis is based heavily on their effort of comprehensive and bulky expression data measurement and laborious data analysis.

My colleagues of the laboratory all gave me the feeling of being at home at work; Koichi Itoh, Hitoshi Adatchi, Makiko Otsuji, Isao Kubota, Sumiyo Sugimoto, Miya Shiojima ,Takayasu Iizuka, Koji Watanabe, Hiroaki Imamura. Many thanks for being your colleague.

General Introduction

Recent genomic sequencing efforts followed by transcriptome and comparative genomics studies have elucidated many universal phenomena in the genomes which had never been discovered from the studies of small number of genes. Some example of such phenomena are the conservation of nucleotide sequences in the housekeeping genes(1), patterns of gene expression level and expression profiles(2-4), chromosomal clustering of transcriptionally related genes (2-10), the relationship between expression patterns and nucleotide sequences (11-13) and so on. Some of these new and occasionally unexpected phenomena are likely to be a refrection of hidden relationship between some biological processes. Therefore, with the examination of the factors responsible for formation of such new phenomena, it is expected that new knowledge can be obtained about the biological process itself or new causal relationships between biological processes.

Zipf's law of transcriptome is one such statistical phenomenon in the genome that has been revealed by genome-wide studies of gene expression. When the frequency of each mRNA extracted from a tissue sample f and its abundance rank in the sample r ($r = 1$ for the gene of the most frequent mRNA, $r = 2$ for the gene of the second most frequent mRNA, and so on) follows the formula $f = k/r^b$, it is referred to as Zipf's law of transcriptome. In that formula, k is a constant coefficient and b is a constant parameter that represents the absolute value of the slope in a log-log plot of transcriptome frequencies (slope parameter, hereafter). Zipf's law (14) is also known as the power law or the Pareto distribution (15). In phenomena that obey Zipf's law, a few parts occur many times and most parts occur only a few times. Therefore, this is an instance of the so-called heavy-tailed distribution.

Several other researchers have reported Zipf's law of transcriptome in direct or indirect form. Using the serial analysis of gene expression (SAGE) method, Zhang et al. (1997) first suggested this phenomenon in an incomplete form based on measurements of expression abundance in human colorectal epithelium and pancreatic cells(16). They reported that the bulk of the mRNA mass (approximately 75%) consisted of transcripts that were on an average expressed at more than five copies per cell. However, in contrast, most transcripts (approximately 86%) were expressed at less than five copies per cell, but in aggregates, this low-abundance class represented only 25% of the mRNA mass. Such heavy-tailed behavior is consistent with the expected results of Zipf's law of transcriptome. Therefore, it can be regarded as the first report of this law. However, they did not mention the explicit form of distribution followed by transcript abundance.

Determination of the form of distribution was first reported in 2002. Luscombe (2002) reported that many types of genomic features—occurrence of the DNA word, gene families, protein folds, pseudogene families, and abundance of transcript—obeyed Zipf's law in worm and yeast genomes (17). To construct the best stochastic model describing the distribution of transcript occurrence, Kuznetsov (2002) used the SAGE method to analyze the distribution of human, mouse, and yeast transcript occurrence. He concluded that the profiles followed a Pareto-like distribution model (18). I reported that the law was applicable to all human normal tissues that I observed, and the values of parameters k and b approached 0.1 and 1, respectively, in organs that were primarily composed of a homogeneous population of differentiated cells (19).

Presently, it is well known that Zipf's law of transcriptome is applicable to a large variety of organisms such as vertebrates (humans, mice, and rats), invertebrates (*Drosophila melanogaster* and *Caenorhabditis elegans*), other eukaryotes (yeast, *Arabidopsis thaliana*), and even to prokaryotes (*Escherichia coli*) (18,20-23). It is remarkable that the slope parameter b is almost unique ($b \approx 1$) among all species investigated.

It is not realistic to attribute this phenomenon to any bias in the measuring process because in humans, mice, and yeast, the law was confirmed using all three methods—GeneChip, SAGE, and EST (17-23). These methods are based on different principles of measurement. Therefore, this phenomenon should not be an artifact but should be reflective of some type of biological process.

What biological process causes Zipf's law of transcriptome? In 2003, I published a paper describing an evolutionary model that explains the cause of this phenomenon (19). In that study, I assumed that mutations in the cis-element or other sequences on the genome that codes for the expression strength cause a stochastic change in the expression level and that the change is proportional to the original expression level. I provided evidence that Zipf's law was derived from these assumptions. The evolutionary model insists that the phenomenon was the result of an "evolutionary time scale" process, such as that which occurs over a span of several million years or more.

At approximately the same time, several other models were proposed for explaining this phenomenon. All these models attributed the causes of Zipf's law of transcriptome to the gene expression dynamics in each cell. In other words, the gene expression dynamics model insisted that the phenomenon was the result of a "cell-life time scale" process, such as that which occurred over one day or several days. Furusawa and Kaneko (2003) proposed an explanation of the phenomena on the basis of a deterministic model of reaction kinetics. In this model, the phenomenon is explained by a balance between the rate of newly synthesized mRNA, the rate of mRNA degeneration, and the rate of supply of "nutrient" from the environment (21). Kuznetsov (2003) formulated the dynamics of mRNA synthesis and degradation with a "birth-and-death" stochastic process. He then insisted that Zipf's law of transcriptome can be explained on the basis of the formulation (20). Using time course data of expression levels of genes, Ueda (2004) stated that from *E. coli* to humans, gene expression dynamics follows the same and surprisingly simple principle where gene expression changes are proportional to their expression level, and this proportional dynamics can replicate the observations of Zipf's law of transcriptome (22).

In this thesis, I compared these two models from both the experimental and theoretical points of view and determined the model that is more plausible.

In the evolutionary model of Zipf's law, I only focused on the evolutionary change in expression level or strength of genes. Next, I tried to extend the study to anatomical expression patterns or profiles. To investigate the evolution of anatomical gene expression patterns, I focused on housekeeping genes as a set of genes that definitely express and function in all cell types.

Identification of housekeeping genes from large-scale expression profiles was first exemplified by the work of Velculescu et al. (1999) who used the SAGE method, and this was followed by Warrington (2000) and Hsiao (2001) who used oligonucleotide microarrays (GeneChip) (24-26). These studies opened up new opportunities to explore the relationship between expression patterns and structural features, and based on these, new constraints on the organization of our genome have been discovered (1,3,11,27,28). Several sets of housekeeping genes were published along with such studies Several sets of housekeeping genes were published along with such studies, but it has rarely been well discussed whether the analyzed set of genes is a non-biased representative of housekeeping genes. In fact, on comparing the published screenings for housekeeping genes, one based on the GeneChip method and the other based on the SAGE method, I found low concordance between the results of the two screening methods, and in both processes, there was poor sensitivity in the identification of housekeeping genes. Therefore, I examined the cause for this inconsistency, and by tuning the parameters for housekeeping gene selection, I compiled a more reliable set of housekeeping genes in Part II.

Part I.

The Evolutionary Model of Zipf's Law of Transcriptome

Introduction

Zipf's law, which is also known as the power law or Pareto distribution, has been observed in many different population distributions. In 1896, Vilfredo Pareto demonstrated that income distribution can be described by the so-called Pareto distribution (15). In 1949, George Kingsley Zipf discovered that the occurrence of words in text documents obeyed a heavy-tailed distribution, and this is the origin of the name of Zipf's law (14). Several other examples that follow Zipf's law can be listed, for example, the relative size of cities (14), the connectivity of nodes in large networks (29) such as the World Wide Web (30), the magnitude of earthquakes (31), and so on.

With regard to genomic biology, Mantegna et al. (1994) discussed that the usage of short base sequences in DNA also follows Zipf's law (32-34). Further instances cited in genomic biology include the occurrence of protein families or protein folds (17,35-41), the occurrence of pseudogenes and pseudomotifs (17), the connectivity within metabolic pathways (42), and the number of molecular interactions made by proteins (43).

I found that the human transcriptome also follows Zipf's law (17,19). In this part, I report the distribution of the human transcriptome in detail and propose a model that explains the reason for the occurrence of these phenomena. In this model, Zipf's law of transcriptome is related to the evolutionary process of the transcriptome. It provides a novel interpretation of transcriptome data and of evolutionary constraints on gene expression (19). To date, some authors had proposed gene expression dynamics as an alternative model for explaining the phenomenon (20,22). I compared the two models based on several experimental and theoretical evidences.

Results

1. Relationship between transcript frequency and abundance

In genetics-linguistics analogy, a transcriptome is a text in which a life plan is "expressed" with a genomic vocabulary. By analyzing SAGE tag and EST data, I found that the human transcriptome follows the statistical constraints that are characteristic of natural languages. In a corpus of texts, Zipf's law dictates that the frequency of each word f and its abundance rank r ($r = 1$ for the most frequent word, $r = 2$ for the second most frequent word, and so on) are related by the formula $f = k/r^b$ for all languages (14). In a double logarithmic axis plot, such a relationship is represented by the linear dependence of f as a function of r with a slope of –1. In muscle and liver, which are organs that are primarily composed of a homogeneous population of differentiated cells, the frequency f of each transcript and its abundance rank r are very closely distributed to the line $f = 0.1/r$ (Figure 1a). In other sources, such as cell lines and epithelial tissue, only high rankers ($r < 100$), which comprise less than 1% of the transcript variety, are derived from this trend (Figure 1b). Reduction of tissue-specific transcripts in cell lines and their dilution in complex cell populations partially explains such deviations. Compiling data for different transcriptomes affected plot similarity (Figure 1c). In normalized libraries, the Zipf-like structure was completely lost (Figure 1d).

2. Evolutionary model of Zipf's law of transcriptome

In this study, based on the following three simple assumptions, I propose an evolutionary model that explains linearity in a log-log plot and the uniqueness of slope parameters ($b \approx 1$) in various tissue samples from different species.

The first assumption states that the baseline expression level of each gene in a cell is coded in the genome sequence such as the cis-element or the enhancer region; therefore, the expression levels of genes are affected by the mutation, and this effect is inherited by the offspring (Assumption 1). Assumption 1 was supported by many studies, which reported that the expression level of genes have abundant natural variation and familial aggregation in a large variety of organisms such as yeast (44,45), killifish (46), mice (47), and humans (48-50). In addition, a few authors reported that the location of the cis-element and/or the trans-factor of the genes can be determined by using the quantitative trait locus (QTL) analysis in which the expression level of each gene was treated as a quantitative trait (expression phenotypes) (51,52).

Schadt et al. (2003) performed interval mapping of expression phenotypes using 111 F_2 mice constructed from two standard inbred strains, C57BL/6J and DBA/2J(51). Of the 23,574 genes represented on the microarray, they reported that there were 3,701 genes with LOD scores greater than 4.3 and 11,021 genes with an LOD score greater than 3.0. Morley et al. (2004) measured the baseline expression levels of 3,554 genes from a sample obtained from 14 Centre d'Etude du Polymorphisme Humain (CEPH) pedigrees. Based on the QTL analysis of the sample, they concluded that for approximately 1,000 expression phenotypes, there was significant evidence of linkage to specific chromosomal regions(52).

The second assumption states that the expression level changes in stochastic proportion to its intensity (Assumption 2). In the formula, this assumption is described as follows:

$$f_i(t+1) = c_i(t) f_i(t)$$

where $f_i(t)$ is the expression level of the gene *i* (number of mRNA occurrences of gene *i*/total number of mRNA molecules) at the *t*-th generation. It is also assumed that the total number of mRNA molecules in a cell (M) is constant throughout the evolutionary process. The proportion coefficient $c_i(t)$ is a random variable sampled from a time-independent distribution at time *t*. The expectation value of the distribution possibly approaches 1.0, and the variance is small because in the natural population of organisms, it is unrealistic to expect drastic changes in the expression levels of genes during one generation. An experimental proof of the assumption was obtained by Khaitovich et al. (2004) (53). From Assumption 2, it follows that the squared difference of the logarithm of the expression level (expression divergence) is expected to linearly increase with the divergence time in the case where expression divergence was defined as follows:

$$\text{expression divergence} \equiv E[(\log(f_i) - \log(f_i'))^2]$$

where f_i and f_i' are the expression levels of gene *i* in one species and another species, respectively. Using oligonucleotide microarrays, Khaitovich et al. (2004) studied differences in the expression levels of approximately 12,000 genes (1,998 genes after spot masking for excluding the influence of DNA sequence differences on the hybridization result) in the prefrontal cortex of several primates such as humans, chimpanzees, orangutans, and rhesus macaques. They confirmed that the expression divergence represents an approximately linear function of time over at least 20 million years. Such clocklike accumulation of expression divergence was also confirmed in rodents (Appendix A).

Based on these two assumptions, it follows that the distribution of the expression levels of genes converges to a form of lognormal distribution.

If we repeatedly apply the formula of Assumption 2 and take the logarithm of the formula, we obtain

$$\log f_i(t) = \log f_i(0) + \sum_{\tau-0}^{t-1} \log c_i(\tau)$$

The central limit theorem states that $\sum_{\tau=0}^{t-1} \log c_i(\tau)$ converges to a normal distribution with infinitely large variance (σ); hence, for a sufficiently large value of t, $f_i(t)$ will approach a form of lognormal distribution. In general, a random variable X is referred to as having a lognormal distribution if the random variable Y = log X has a normal distribution. When t becomes larger, the relative effect of the initial condition $\log f_i(0)$ becomes negligible. After all the distribution of $f_i(t)$ converges to a lognormal distribution. The cumulative distribution function of the lognormal distribution is graphed on a Zipfian plot. It is well known that when σ is sufficiently large, the function will appear to be almost linear for a large range of values. However, if we considered only these two assumptions, the uniqueness of the slope parameter b in Zipf's law of transcriptome cannot be explained because according to the central limit theorem, when, $t \to \infty$, $\sigma \to \infty$ follows. This results in an infinitely large value of the slope parameter b. In other words, any stable distribution of the expression level of genes cannot be obtained from only these two assumptions, and an additional assumption is required.

In this model, when the absolute value of slope parameter b increases, the number of expressed genes in a cell type greatly decreases because the total occurrence of mRNA in a cell M is constant (Figure 2). Obviously, this is disadvantageous for maintaining the function of cells and individuals; hence, it should be prohibited on the basis of purifying selection. Due to this reason, I introduced a third assumption in which the number of expressed genes in a cell type is nearly constant throughout the evolutionary process, and any functional gene is prohibited from losing its gene expression ability. In mathematical terms, lowering the expression level below 1 copy/cell is prohibited by purifying selection.

By Monte-Carlo simulation of the model, I confirmed that a stable distribution of $f = 0.1/r$ was obtained from these three assumptions, regardless of the initial distribution (Figure 2). It was also confirmed that no stable distribution can be obtained without Assumption 3. Ueda (2004) insisted that proportional gene expression dynamics that are equivalent to Assumption 2—not in a biological sense but in a mathematical one—can generate the observed power law transcriptional organization(22). However, it became clear that this statement is completely wrong, at least in a mathematical sense.

To demonstrate that the universality of slope parameter b can be replicated from the evolutionary model, I used the Monte-Carlo simulation to determine the condition for converging the distribution with the slope parameter $b \approx 1$. In my model, the slope parameter b depends on the following three parameters: the number of mRNA molecules in a cell (M), the number of genes expressed in the cell (G), and the permissible lower limit of expression level (L). From the simulation in which the value of parameter M was fixed to 300,000, which is the case in typical human cells, when the lower limit (L) was 1 or 2 copies/cell, the distribution converged to $b \approx 1$ over a very wide range of values for parameter G, from 10,000 to 50,000 genes in a cell. This is the reason for the universality of *b*, which is predicted by the evolutionary model of Zipf's law. When the lower limit is set to a slightly larger value such as 5 to 10 copies/cell, the value of parameter G should fall in a narrow range, from 5,000 to 10,000 genes in a cell, in order to obtain the observed distribution. When the lower limit is set to 0.1 to 0.5 copies/cell, more than 50,000 genes were required to obtain the observed distribution. Such values for parameter G is unrealistic; therefore, it was concluded that the lower limit of parameter L should be in the range of 1–2 copies/cell (Figure 3).

3. Comparison between the evolutionary model and the gene expression dynamics model

Several other models of Zipf's law of transcriptome have been published, and all of these have attempted to provide an explanation for the phenomenon based on gene expression dynamics. Although the evolutionary model and gene expression dynamics model are completely different from each other in a biological sense, from the mathematical point of view, these models share one underlying idea, i.e., the multiplicative process. In the multiplicative process, the relative amount of transcripts changes in stochastic proportion to the previous "state," and the change is independent of the absolute expression level of the gene. The evolutionary model formulated the multiplicative process as a modified version of geometrical Brownian movement (GBM).

Ueda (2004) also formulated the multiplicative process as a GBM in the context of a gene expression dynamics model(22). However, contrary to his assertion, his model cannot lead to any stable distribution because he did not assume a lower limit of gene expression as described above (Figure 2). Kuznetsov (2003) formulated the multiplicative process as a birth-death process that is a discrete state approximation of GBM. His formula (forward Kolmogorov equation) was as follows(20):

$$\frac{dp_0(t)}{dt} = -\lambda_0 p_0(t) + \mu_1(t) p_1(t)$$
$$\frac{dp_m(t)}{dt} = \lambda_{m-1}(t) p_{m-1}(t) - (\lambda_m(t) + \mu_m(t)) p_m(t) + \mu_{m+1}(t) p_{m+1}(t)$$

where $p_m(t)$ is the frequency of a gene with transcript occurrence of m at time t. $\lambda_m(t)$ and $\mu_m(t)$ are the transition probabilities of changing a transcript's occurrence at time t from m to $m+1$ and m to $m-1$, respectively.

As shown in the formula, the lower limit of expression levels of genes was introduced in his forward Kolomogorv equation without any explanation of the biological concepts. Therefore, in a mathematical sense, the Kuznetsov model is a discrete state version of my model.

The model proposed by Furusawa and Kaneko(2002) is based on the balance between the effect of upregulating genes and downregulating genes(21). Hence, it can be regarded as a deterministic equation version of birth-death process formulation. The formula proposed by him is as follows (Appendix B).

$$\frac{dn_i}{dt} = \sum_{i,l} \frac{n_j n_l}{N^2} - \sum_{i,l'} \frac{n_i n_{l'}}{N^2} + D\sigma\left(\frac{\overline{n}}{V} - \frac{n_i}{N}\right)$$

In the Furusawa model, the parameter D (the rate of supply of "nutrient" from the environment) plays the role of adjustment of slope parameter b in the Zipfian plot.

However, the gene expression dynamics model has at least three fundamental problems in explaining the law. First, there is no reliable experimental support for their assumption. Second, any reason for the universality of the slope parameter $b \approx 1$ was not explained in the gene expression dynamics model. The third and most critical problem is that contrary to the authors' assertion, the gene expression dynamics model cannot replicate the observations of Zipf's law of transcriptome. The authors might overlook the issue that the expression level of each gene changes at random even after the distribution converged to the stable form, i.e., $f = 0.1/r$, in so far as it is based on the multiplicative process (Figure 4). Based on the gene expression dynamics model, it follows that the expression level of each gene changes independently among the cells, even if the same type of cell is considered. This leads the rank order of the expression level in each cell into completely different orders from each other. It is noteworthy that the observations of expression level distribution were obtained from samples that consisted of a mixture of millions of cells. The central limit theorem of probability theory states that the distribution obtained from such mixed samples should be a normal distribution (Figure 5). Therefore, the gene expression dynamics model does not explain Zipf's law of transcriptome, at least in a mathematical sense. It is unrealistic to expect that the rank order of expression level in the same type of cells randomly diverges to completely different orders. In order to avoid this unrealistic conclusion, the gene expression dynamics model should accept an additional assumption that "each gene has its own dynamic range of expression level in each cell type." In such a case, the following questions should be answered: How are the ranges of expression level determined, and how are they determined in the form of Zipf's law?

The evolutionary model answers these questions as follows: the ranges of expression levels are determined by the genome code, and under Zipf's law, they are determined by the multiplicative process in the evolutionary process, i.e., Assumptions 2 and 3.

In the evolutionary model, the profile of expression level in the same cell type will not diverge in the same species; hence, “the mixing effect” will not appear in the observations. Based on this, I concluded that the evolutionary model is more plausible than the gene expression dynamics model.

Discussion

Determination of the correct cause of Zipf's law of transcriptome critically influences the direction of further investigation.

If the gene expression dynamics model is embraced, it is almost inevitable to relate Zipf's law of transcriptome to the derivation of the exact formula of gene expression dynamics in a cell. On the basis of Ueda's model, Ochiai et al.(2004) reported a derivation of the formula that represents the elementary process of gene expression dynamics(54). Determination of such formula is crucial for the estimation of gene regulatory networks from time course data of gene expression profiles. Therefore, if the gene expression dynamics model is valid, it should be evaluated in detail as it plays a central role in the field. In this thesis, I showed that the cause should be attributed to the evolutionary process. If my assertion is valid, the proposed formula loses its ground.

If we accept the evolutionary model, Zipf's law of transcriptome is related to the evolution of gene expression. Khaitovich et al. (2004) discovered a clocklike accumulation of expression differences within primates and rodents, and they termed this phenomenon as the neutral model of transcriptome evolution(53). These results were in agreement with the observation of Rifkin et al. (2003), who reported that differences in gene expression were consistent with phylogenetic relationships among Drosophila species(12). Zipf's law of transcriptome can be viewed as evidence of the neutral theory of transcriptome evolution because the evolutionary model for Zipf's law of transcriptome is nearly equivalent to the neutral model of transcriptome evolution.

It should be noted that my model could not determine whether the stochastic change formulated in Assumption 2 was neutral with respect to organismic fitness or whether it was the result of some kind of selection. The reason for this was that as Khaitovich also pointed out, under certain selection scenarios, the selected change would also accumulate linearly with time (55).

Based on a comparison of gene expression divergence and CDS divergence in human and mouse orthologs, Jordan et al. (2005) concluded that although there may be a neutral component in the evolution of expression, much if not most of the changes in expression are subject to purifying selection(56). They proposed a model of neutral evolution with selective constraint in which they intended that the functionally important component of gene expression is held constant by purifying selection, while the functionally irrelevant component evolves neutrally. Such views of gene expression evolution will make an impact on data interpretation for comparative studies of gene expression profiles, such as in the identification of critical genes in primate brains which distinguish the ability to think in human and other primates.

Figure Legends

Figure 1. Log (frequency) log (rank) plot (Zipf's plot) of transcriptome data. The frequency of occurrence (*f*) of each transcript in 3′ EST and SAGE tag collections representing various transcriptomes was plotted against the abundance rank. The broken line represents $f = 0.1/r$. (**a**) Organs with homogeneous populations of differentiated cells. For example, the most abundant transcript ($r = 1$) in liver, i.e., albumin, occurred at a level of approximately 12% in the EST data for liver. Gene names are given for $r = 1$–6 in liver. (**b**) Cell lines and complex tissues. (**c**) Compiled data from 51 human EST sets, 31 mouse EST sets, and 64 SAGE tag sets. Gene names are given for $r = 1$–6 in the compiled human transcriptome (3′ EST). (**d**) Occurrence of 3′ EST in normalized libraries. The total tag occurrence for each data set is given in parentheses. The frequency data were obtained from http://bodymap.ims.u-tokyo.ac.jp/datasets (3′-EST) and ftp://ncbi.nlm.nih.gov/pub/sage (SAGE). The data for liver are combined with data for two human liver libraries. The frequencies of total SAGE tags are obtained from reanalysis of all available human SAGE tags. Clustering 3′ ESTs for two representative normalized libraries in dbEST, 1N1B and 2NbHM, generated the data for normalized libraries.

Figure 2. Replication of Zipf's law of transcriptome by computer simulation.

Red points represent the expression levels of each gene. The number of expressed genes in the initial population was 20,000. Black line: $f = 0.1/r$. t: the number of generations from the initial state. Regardless of the distribution of the initial population (which have a uniform distribution in this figure), the population converged according to Zipf's law (left column). In the absence of Assumption 3, the slope of the distribution in the log-log plot was not steady at approximately –1, and the number of expressed genes decreased (right column). The simulation program was developed using the Java computer language. Random numbers were generated by the Mersenne-Twister algorithm implemented in Colt library version 1.2.0 that was developed at CERN (http://dsd.lbl.gov/~hoschek/colt/).

Figure 3. Relationship among the permissible lower limit of expression level (L), the slope parameter, and the number of genes expressed in the cell.

In this simulation, the total number of mRNA molecules in the cell was fixed at 300,000. Each point represent the average magnitude of the final state slope in 50 repetitions of the simulation with various value of L (Circle: L=0.1, triangle: L= 0.5, cross-shape(+): L=1.0, X-shape: L=2.0, diamond: L=5.0, inverted triangle: L=10.0). Estimation of slope parameter b with linear regression from the results of computer simulation was performed using the R statistical package (http://www.cran.org).

Figure 4. Stochastic change (random walk) in the expression levels of genes simulated after the distribution converged to $f = 0.1/r$.

The simulation condition was the same as in the left column of Figure 2. The x-axes represents the number of generations after the convergence, i.e., $t = 20{,}000$.

Figure 5. Distribution of the expression levels of genes in the mixed cell sample.

The distribution in each cell independently diverged from each other. N: number of cells mixed. Black line: $f = 0.1/r$.

Figure 1.

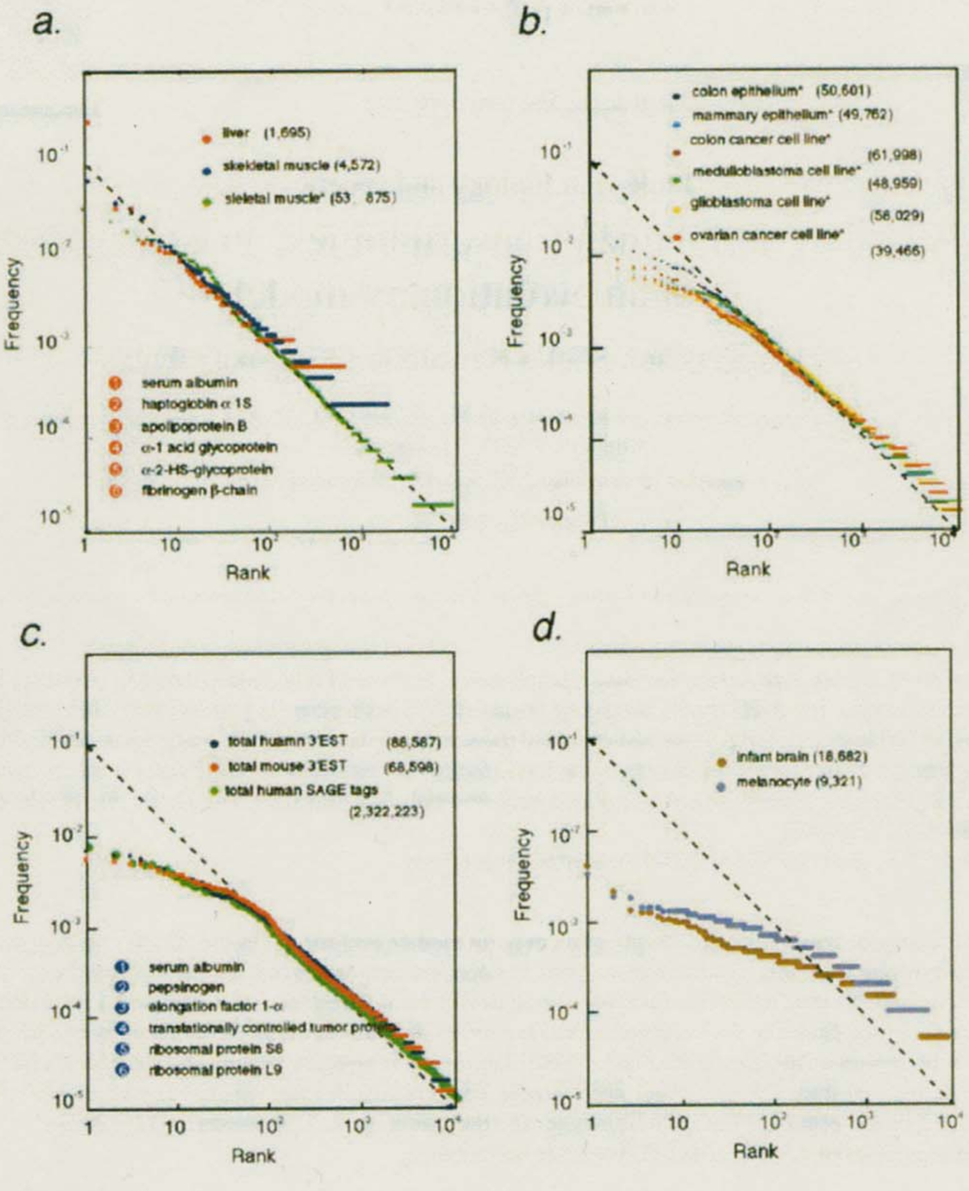

a.
b.
c.
d.
Frequency
Rank
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
1
10
10^2
10^3
10^4
liver (1,695)
skeletal muscle (4,572)
sleletal muscle* (53 875)
1 serum albumin
2 haptoglobin α 1S
3 apolipoprotein B
4 α-1 acid glycoprotein
5 α-2-HS-glycoprotein
6 fibrinogen β-chain
colon epithelium* (50,601)
mammary epithelium* (49,762)
colon cancer cell line* (61,998)
medulloblastoma cell line* (48,959)
glioblastoma cell line* (58,029)
ovarian cancer cell line* (39,466)
total huamn 3'EST (86,567)
total mouse 3'EST (68,598)
total human SAGE tags (2,322,223)
1 serum albumin
2 pepsinogen
3 elongation factor 1-α
4 translationally controlled tumor protein
5 ribosomal protein S6
6 ribosomal protein L9
infant brain (18,682)
melanocyte (9,321)

Figure 2.

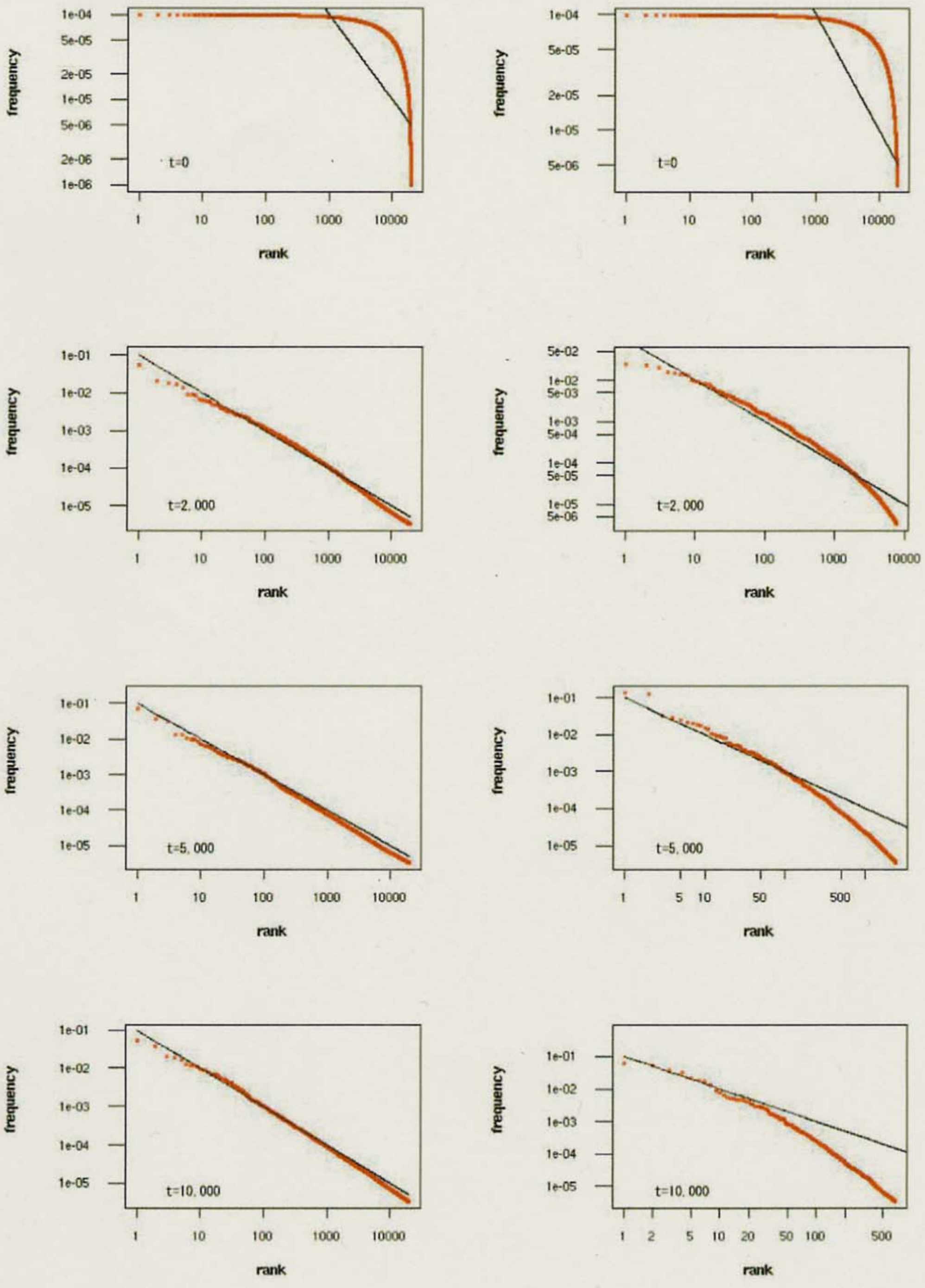
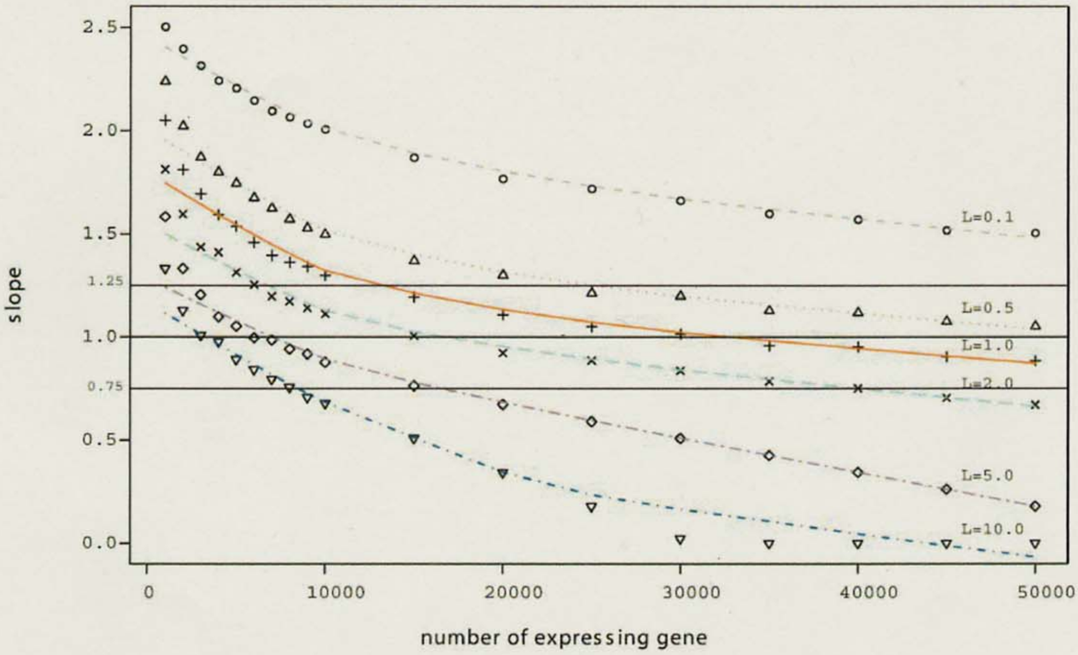

frequency
rank
t=0
t=0
t=2, 000
t=2, 000
t=5, 000
t=5, 000
t=10, 000
t=10, 000
1e-04
5e-05
2e-05
1e-05
5e-06
2e-06
1e-06
5e-02
1e-01
1e-02
5e-03
1e-03
5e-04
1e-04
5e-05
1e-05
5e-06
1
2
5
10
20
50
100
500
1000
10000

Figure 3.



slope
2.5
2.0
1.5
1.25
1.0
0.75
0.5
0.0
0
10000
20000
30000
40000
50000
number of expressing gene
L=0.1
L=0.5
L=1.0
L=2.0
L=5.0
L=10.0

Figure 4.

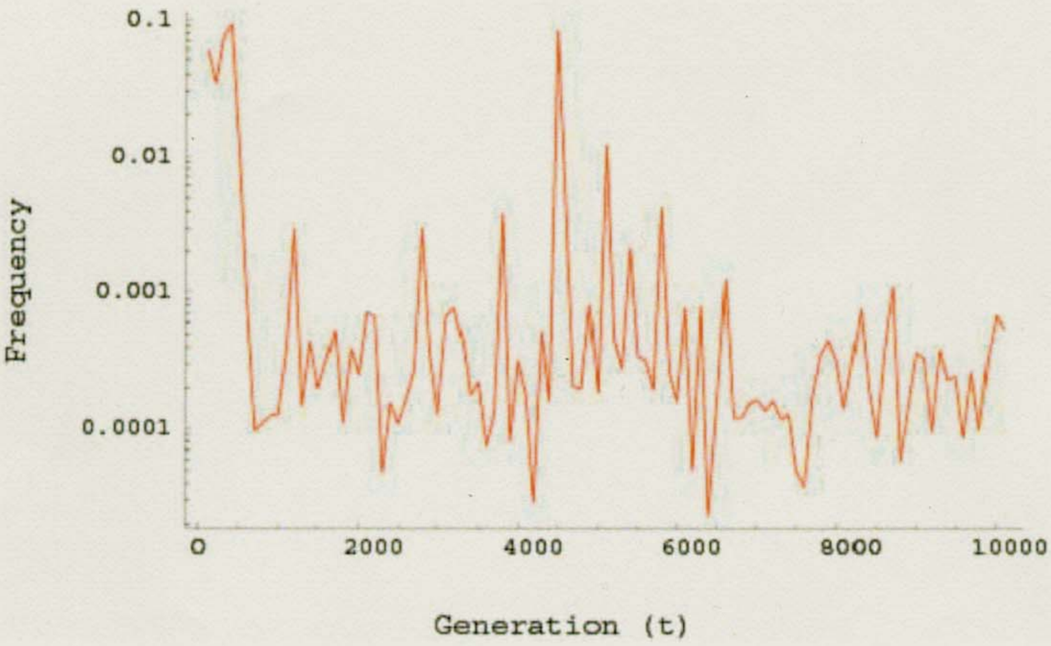
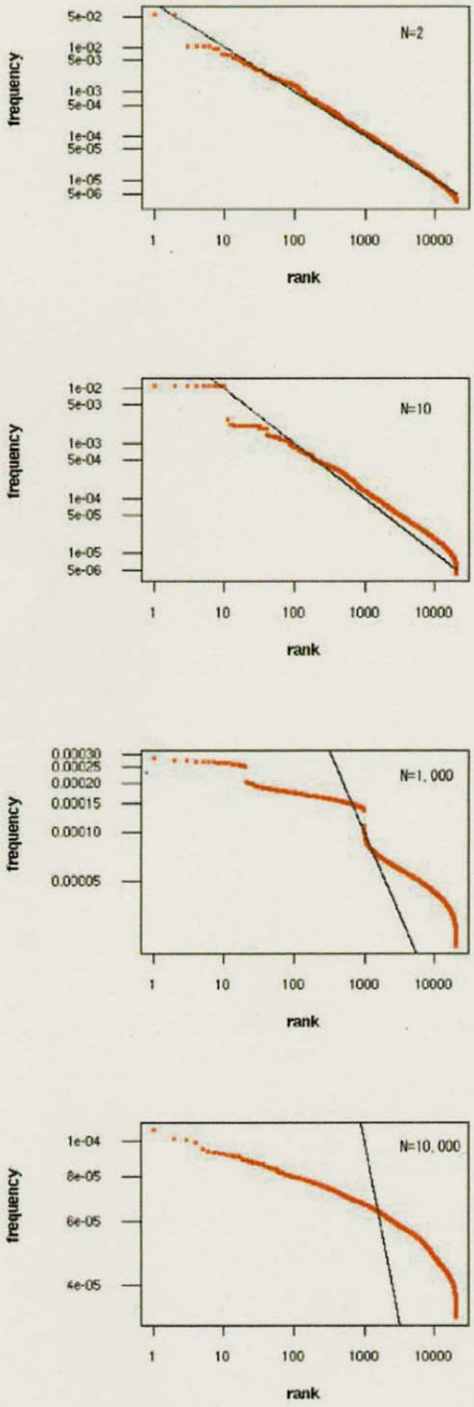

0.1
0.01
0.001
0.0001
Frequency
0
2000
4000
6000
8000
10000
Generation (t)

Figure 5.


N=2
frequency
5e-02
1e-02
5e-03
1e-03
5e-04
1e-04
5e-05
1e-05
5e-06
rank
1
10
100
1000
10000
N=10
frequency
1e-02
5e-03
1e-03
5e-04
1e-04
5e-05
1e-05
5e-06
rank
1
10
100
1000
10000
N=1,000
frequency
0.00030
0.00025
0.00020
0.00015
0.00010
0.00005
rank
1
10
100
1000
10000
N=10,000
frequency
1e-04
8e-05
6e-05
4e-05
rank
1
10
100
1000
10000

Appendix A: Clocklike accumulation of gene expression divergence

Excerpt from Khaitovich (2004) (53)

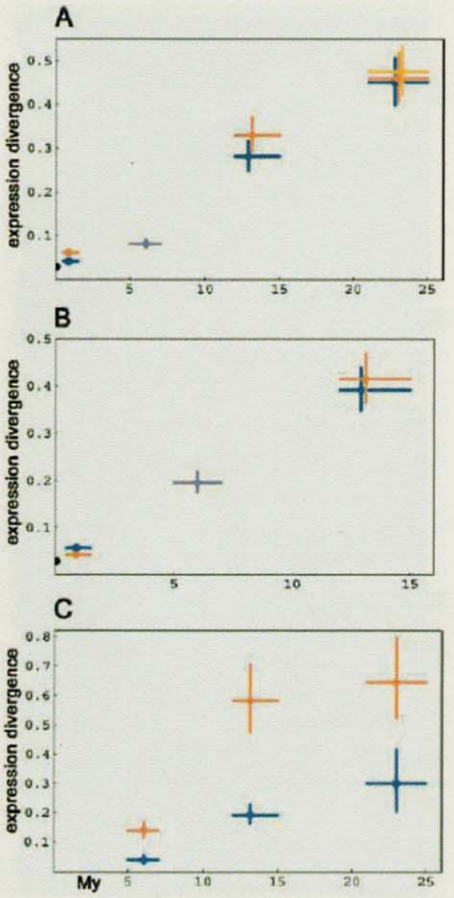


Figure 1. Brain and Liver Transcriptome Change among Primates as a Function of Time

Average expression differences within and between primates in brains (A), in liver (B), and for genes in brain for genes with high (red) and low (blue) variation among six humans (C). Colors: red, comparisons between and with humans; blue, comparisons between and with chimpanzees; purple, comparisons between humans and chimpanzees; orange, comparisons between orangutan and rhesus macaque; black, comparisons between experimental duplicates. Vertical error bars for expression indicate 95% confidence intervals calculated by 10,000 bootstraps over genes. Divergence times are according to Glazko and Nei (2003).

DOI: 10.1371/journal.pbio.0020132.g001

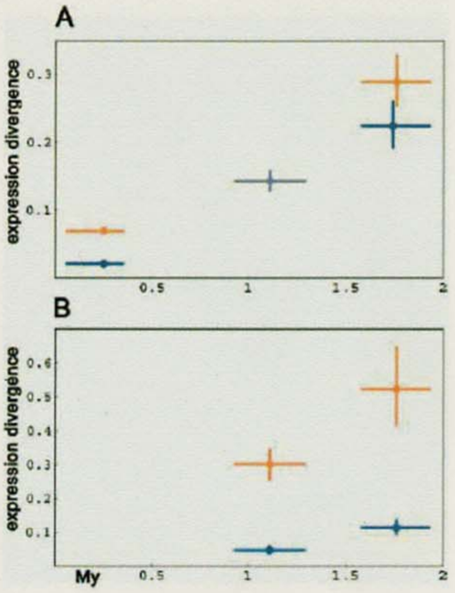


Figure 3. Brain Transcriptome Change among Mice as a Function of Time

Average expression differences within and between the mouse species (A) and for genes with high (red) and low (blue) variation among *M. musculus* individuals (B). Colors: red, comparisons between and with *M. musculus;* blue, between and with *M. spretus;* purple, between *M. musculus* and *M. spretus*. Vertical error bars for expression indicate 95% confidence intervals calculated by 10,000 bootstraps over genes. Divergence times are according to She et al. (1990).

DOI: 10.1371/journal.pbio.0020132.g003

Appendix B: Furusawa model

Excerpt from Furusawa and Kaneko (2003) (21)

The abundance of each protein is the result of a complex network of chemical reactions that is influenced by possibly a large number of factors including other proteins and genes. Then, why is Zipf's law universally observed, and what class of reaction dynamics will show the observed power-law distribution?

In order to investigate the above questions, we adopt a simple model of cellular dynamics that capture only its basic features. It consists of intracellular catalytic reaction networks that transform nutrient chemicals into proteins. By studying a class of simple models with these features, we clarify the conditions under which the reaction dynamics leads to a power law distribution of the chemical abundances.

Consider a cell consisting of a variety of chemicals. The internal state of the cell can be represented by a set of numbers, where n_i is the number of molecules of the chemical species i with i ranging from $i=1$ to k. For the internal chemical reaction dynamics, we chose a catalytic network among these k chemical species, where each reaction from some chemical i to some other chemical j is assumed to be catalyzed by a third chemical ℓ, i.e., $i + \ell \rightarrow j + \ell$ [7]. The rate of increase of n_j (and decrease of n_i) through this reaction is given by $\varepsilon n_i n_l / N^2$ where ε is the coefficient for the chemical reaction. For simplicity all the reaction coefficients were chosen to be equal, and the connection paths of this catalytic network were chosen randomly such that the probability of any two chemicals i and j to be connected is given by the connection rate ρ.

Some resources (nutrients) are supplied from the environment by diffusion through the membrane (with a diffusion coefficient D), to ensure the growth of a cell. The nutrient chemicals have no catalytic activity in order to prevent the occurrence of catalytic reactions in the environment. Through the calaytic reactions, these nutrients are transformed into other chemicals. Some of these chemicals may penetrate the membrane and diffuse out while others will not. With the synthesis of the unpenetrable chemicals that do not diffuse out, the total number of chemicals $N = \sum_i n_i$ in a cell can increase, and accordingly the cell volume will increase. We study how this cell growth is sustained by dividing a cell into two when the volume is larger than some threshold. For simplicity the division is assumed to occur when the total number of molecules $N = \sum_i n_i$ in a cell exceeds a given threshold N_{max}. Chosen randomly, the mother cell's molecules are evenly split among the two daughter cells.

In our simulations, we randomly pick up a pair of molecules in a cell and transform them according to the reaction network. In the same way, diffusion through the membrane is also computed by randomly choosing molecules inside and outside the cell. In the case with $N >> k$ (i.e., continuous limit), the reaction dynamics is represented by the following rate equation:

$$\frac{dn_i}{dt} = \sum_{j,\ell} Con(j,i,\ell) \frac{\varepsilon n_j n_\ell}{N^2} - \sum_{j,\ell} Con(i,j',\ell') \frac{\varepsilon n_i n_{\ell'}}{N^2} + D\sigma_i \left(\frac{\overline{n}_i}{V} - \frac{n_i}{N} \right)$$

where $Con(i,j,\ell)$ is 1 if there is a reaction $i + \ell \rightarrow j + \ell$ and 0 otherwise, whereas σ_i takes 1 if the chemical i is penetrable, and 0 otherwise. The third term describes the transport of chemicals through the membrane, where n_i is a constant, representing the number of the ith chemical species in the environment and V denotes the volume of the environment in units of the initial cell size. The number ni is nonzero only for the nutrient chemicals.

Part II.

Reliable Identification of Housekeeping Genes: Evaluation of Concordance and Performance of the Identification Processes

Introduction

Since the complete human genomic sequence is now known, it is essential to obtain a comprehensive set of physiological gene expression profiles from for both the functional and evolutionary analyses of the human genome. Moreover, the availability of such data opened up new opportunities to explore the relationship between expression patterns and structural features, through which we may discover new constraints on the organization of our genome (1,3,11,27,28). In fact, some genome-wide structural features of housekeeping genes have been reported over the past few years. Lercher et al. (2002) concluded that housekeeping genes tend to form clusters on chromosomes (3,28,57). Eisenberg and Levanon (2003) suggested that housekeeping genes are shorter than other genes. Zhang and Li (2004) argued that housekeeping genes evolved more slowly than tissue-specific genes. Various hypotheses have been proposed to explain these structural features of housekeeping genes (2,10,57-62,63). However, it should be acknowledged that all these arguments implicitly assumed that the sets of housekeeping genes selected for characterization were a majority or were non-biased representatives of their kind. Regardless of the importance of this aspect, no attempt has been made to evaluate the processes of selecting housekeeping genes for analysis. In this paper, I conducted an in depth evaluation of published literature on the processes for the identification of housekeeping genes. For this purpose, I integrated the GeneChip, SAGE, and EST data from public sources, and added some highly sensitive PCR-based expression measurements for confirmation. Based on the evaluation, I proposed a considerably wider selection of housekeeping genes from the same data set. The performance of these old and new selections of housekeeping genes was evaluated using control genes and various features that may reflect gene function.

Results

1. Comparison of GeneChip data and SAGE data.

Sets of housekeeping genes subjected to recent structural characterizations were identified using anatomical expression profiles, which were chiefly captured by either microarray or SAGE. I started our analysis by reproducing one representative process from microarray- and SAGE-based selections. Among microarray-based selections, I reproduced the work of Su et al. (64), in which the expression profiles of 25 independent tissues were generated by GeneChip hybridization of 85 human samples (GDS181). The array used by them (GeneChip U95A) interrogates 8,147 (36.4%) of 22,361 independent RefSeq sequences with respect to their loci (genes, hereafter). As an example of SAGE-based selections, I reproduced the work of Lercher et al. (3). They used 779,068 tags from 36 samples representing 14 normal human tissues (GDS217). According to the NCBI SAGEmap, 16,318 (73.0%) of 22,361 genes are measurable by the SAGE method.

In both studies, the expression pattern of a gene was indexed on the basis of "expression breadth," which is the number of tissue types where the transcript of a gene is called "present," according to the authors' criteria (positive tissues). Genes with breadths no less than a certain breadth threshold were then identified as housekeeping genes. In this study, for comparison, I used the ratio of positive tissues to tested tissues as the "breadth."

I first indexed the expression of genes according to the "present" calls in the original studies (Figure 1a). Unexpectedly, for the 7,592 genes measured in both data, the concordance between Chip breadth and SAGE breadth was found to be very poor.

Unless otherwise noted, I focused on these 7,592 genes in order to compare the two processes. Gene frequencies were high in the smallest and the largest Chip breadths, whereas a decrease in gene frequencies was observed with an increase in SAGE breadth. A better correlation was observed for genes with small breadths; however, even in such cases, agreement between the responsible tissues with regard to Chip and SAGE breadth was rarely observed. For example, among 116 genes with Chip breadth = 1 and SAGE breadth = 1, no agreement was observed with regard to 60% (70/116) of the positive tissues.

These two breadth indexes were then probed with a set of well-known genes that should have the largest breadth according to their reported functions (HK (house keeping genes) controls). These well-known genes included 132 genes that encode components of the three big complexes essential to cell life: ribosome, proteasome, and RNA polymerase. All these genes were included in both data sets.

In general, the breadths of these genes were underestimated to various extents (Table 1). The average Chip breadth for the three complexes was 0.97, 0.78, 0.35, respectively, and the SAGE breadth was 0.84, 0.69, and 0.36, respectively. The Chip breadth threshold of 1.0 (25/25), used in the original study (64), identified 55% of the HK controls as housekeeping genes. On the other hand the original SAGE breadth threshold of 0.64 (9/14) identified 60% of the HK controls as housekeeping genes. The ubiquitous expression of all the HK controls, as assumed from their function, was confirmed by iAFLP (65), a competitive RT-PCR method, with mRNAs from 21 major human organs and tissues. The comparison of the expression signals in iAFLP with those in the two data sets demonstrated that both GeneChip and SAGE failed to detect a substantial fraction of the HK controls. However, even with the same data sets there appeared to be some room for improvement in breadth indexing at the "present" calling step (Figure 2).

2. Signal thresholds for "present" calling.

In SAGE analysis, tags that were isolated only once (f = 1) from a large population of tags were often neglected because of the possibility of their having originated from sequencing errors of different tags. Lercher et al. (2002) also ignored the tags that occurred once, which amounted to 12% of the total tag occurrence. This is a reasonable precaution when the entire set of true tags is not available. However, in the present study, I can filter out sequencing errors by comparing tags with all predicted tags from the UniGene sequences. Having achieved this, the only reason for concern would be the conversion between true tags caused by reading errors. The expected possibility of such conversions is negligible. Based on this idea, I called genes tagged only once (f = 1) as "present" and re-calculated the SAGE breadth. Among the 7,952 genes, the genes called as "present" in each tissue increased from 878–5,484 (average 2,628) to 1,993–6,285 (average 3,779). Consequently, a substantial number of genes were indexed with larger SAGE breadths. The average SAGE breadth for the three HK control complexes increased to 0.89, 0.82 and 0.57, respectively, and agreement with Chip breadth improved for high- and middle-breadth genes (Figure. 1b).

In Gene Chip analysis, the signal threshold for "present" calling was set to AD = 200. AD stands for the average of the differences among multiple pairs of fluorescent signals from full-match and mismatch oligonucleotides on the chip. Based on this threshold, 1,976–2,646 of 7,592 genes were called as "present" in each tissue. Underestimation of the breadth of HK controls suggests that the signal threshold for "present" calling is not sufficiently sensitive. However, lowering the signal threshold for "present" calling in Chip data is not readily justified because limited information is available with regard to the signals below this threshold, especially regarding non-specific signals for non-expressing genes.

3. Breadth threshold for housekeeping genes.

Since several HK controls were not always called as "present", the breadth threshold should be lower than 1.0 in order to identify these controls as housekeeping genes. To determine the lower limit of the breadth threshold, I used 36 well-known peptide hormone genes (TS controls). The Chip breadth threshold can be lowered to 0.84 (21/25) without the inclusion of any TS control in the Chip HK, except for one TS control (human chorionic gonadotropin) with full Chip breadth as an artifact. In the SAGE data, with an enhanced "present" calling that takes f = 1 into account, the original threshold of 0.64 (9/14) is the lower limit that excludes any TS control from SAGE HK selection (Figure 1b).

By lowering the breadth threshold from 1.0 to 0.84, the number of Chip HK increased from 700 to 1,333, and 67% of the HK controls were identified. Under the same SAGE breadth threshold (0.64), enhancement in "present" calling increased the SAGE HK from 1,633 to 3,036, and 77% of the HK controls were identified. The union of Chip HK and SAGE HK increased the number of HK from 1,904 to 3,402 of 7,592 genes, and 85% of the HK controls were identified (Figure 3). Besides the 7,592 genes, the old criteria identified 888 genes, and the new criteria identified an additional 1,247 genes as housekeeping genes based on either data set.

In total, the number of housekeeping genes identified by the old criteria was 2,792 (old

HK, previously identified housekeeping genes), and this number was increased to 5,537 (expanded HK, old HK + new HK) (Table 2).

4. Sensitivity and specificity evaluation of housekeeping gene identification.

The persistent and substantial difference in the observed breadth of HK controls suggests that the strength of gene expression seriously affects the apparent breadth of its expression. In fact, I found a good correlation between the observed breadth and total cognate tag counts of HK controls in dbEST. The correlation coefficient (r) between breadth and abundance indexed by the total EST count was as follows: for SAGE breadth, r = 0.48 and P-value = 5.7×10^{-9}; for Chip breadth, r = 0.73 and P-value < 2.2×10^{-16} (Figure 4). As a consequence, the genes with weaker expression are always less likely to be identified as housekeeping genes. Regardless of the breadth of expression, the transcript abundance of a gene in a tissue of its activity can be estimated by the "peak EST rate (ρ)", which is the highest EST frequency calculated for each tissue (see Materials and Methods). There was a significant correlation between the peak EST rate and expression breadth as well; for SAGE breadth: r = 0.45 and P-value = 7.9×10^{-8}; for Chip breadth: r = 0.70 and P-value < 2.2×10^{-16} (Figure 4). Using the peak EST rate obtained by dividing the human EST data into 10 tissue categories, 16,892 genes, measurable by SAGE or GeneChip, were binned into high ($\rho \geq 10^2$; 5,334 genes), middle ($10^2 > \rho \geq 10^{1.5}$; 6,732 genes) and low ($\rho < 10^{1.5}$; 4,826 genes) abundance classes (Figure 3). In the high abundance class, the sensitivity was satisfactory in the old HK selection, and a small improvement was observed with an expansion in selection. In contrast, for the middle class genes, expansion resulted in a dramatic increase in sensitivity, from 0.13 to 0.54 (Table 3). The number of identified housekeeping genes in the three abundance classes was 2,100, 630, and 62 in the old HK and 3,304, 2,022, and 211 in the expanded HK. Based on the class specific sensitivities in the expanded HK identification, the expected number of total housekeeping genes is 3,553 and 3,744 in the high and middle abundance classes, respectively. Almost no information is available about the numbers of housekeeping genes in the low abundance class. For estimation of the specificity, with the same principle, the selection of negative controls is crucial. Since I set the breadth threshold by using TS controls, the misclassification of other tissue specific genes would not be large. However, there is another category of genes that is not meant to be included in the selection of housekeeping genes, regardless of their breadth (broad non-HK); these are genes unique to widely distributed cells, such as fibroblasts, and genes common to a class of cells, such as epithelial cells. Considering the various structures that are widely distributed in our body, and the variety of subcellular structures that are common to a class of cells, the number of broad non-HK would not be negligible. In fact, I often found genes in this category in the candidate lists of housekeeping genes, such as collagens, fibronectins, and globins. As representatives of broad non-HK, I used all the members of four gene families—collagens, laminins, cadherins, and claudins—that amounted to 118 genes altogether. The false positive rate, i.e., the fraction of broad non-HK identified as housekeeping, in old HK and expanded HK was 0.085 (10/118) and 0.15 (18/118), respectively. Assuming that the broad non-HK has the same gene population size as housekeeping genes, specificities, the fractions of true housekeeping genes in old HK and expanded HK, is 0.79, 0.84, respectively (Table 2, 3). In summary, I demonstrated that the published set of housekeeping genes represented mainly abundantly expressed genes, and I have added less abundant housekeeping genes in the expanded set. On the other hand, either selection will include as many as 20% of broad non-HK, a class of genes that have been ignored in relevant studies. The available data provides no reliable anatomical expression information about the less frequently transcribed genes that comprise one- fourth of our gene set.

5. Features of the housekeeping genes.

I have compared three groups of genes, old HK (2,792), new HK (expanded HK minus old HK; 2,745), and the remaining genes (non-HK; measurable minus expanded; 11,355), with regard to various features that may reflect their function.

One of the features was the evenness of transcript distribution. Evenness was indexed on the basis of the value of average over maximum signal across tissues, using both SAGE data (SAGE evenness) and GeneChip data (GeneChip evenness). In both data, the evenness of the HK controls showed higher values when compared with that of the TS controls. The distribution of both the new HK and old HK agreed with that of the HK controls. The evenness of non-HK showed a similar distribution to that of the TS controls (Figure 5).

The second feature was subcellular localization of their products. The subcellular localization descriptions of 6,288 genes was present in the corresponding SwissProt entries. The prominent features of old and new HK were scarcity in extracellular proteins and richness in mitochondrial proteins. The extracellular fractions were two- to three-fold smaller in housekeeping genes (non-HK/old HK = 3.30 and new HK/non-HK = 2.54), and mitochondrial fractions were two- to three- fold greater in housekeeping genes (old HK/non-HK = 3.21 and new HK/non-HK = 2.09) (Figure 6).

The third feature was the number of genes that have CpG islands over their transcription start sites (start CGI). Several studies have shown that most of the housekeeping genes have a start CGI, whereas only a small number of tissue-specific genes possesses it. Ponger et al. (2001) reported that a start CGI was detected in 90% of the housekeeping genes and in 42% of the tissue-specific genes (66). These figures were based on 51 housekeeping and 274 tissue specific genes, which they identified according to their criteria, among 864 Genbank proteins with reliably annotated transcription start sites. These 864 proteins with reliable start sites collapsed to 517 genes in the RegSeq based data, and the fraction of housekeeping genes with start CGI was revised to 85% (29/35). With respect to the 517 genes, the fraction of start CGI positives was 86% in the old HK and 80% in the new HK, whereas it was 42% in the rest (Table 3).

The average EST frequency of the new HK (279.6 tags per million) was much smaller than that of the old HK (711.4). Tight correlation (r=0.69) between breadth and strength of expression, represented by logarithm of tag frequency, is a recently proposed feature of human genes (3). This has been interpreted as housekeeping genes being more abundantly expressed in a cell than other genes (3). The similar extent of correlation was observed in our study, either with regard to Chip or SAGE breadth. However, this might simply reflects a tendency of underestimation of the breadth of weakly expressed genes that I demonstrated among HK controls (Figure 4). In fact, the correlation between the peak EST rate and expression breadth in the abundant class, where breadth of genes are correctly indexed, was significantly weaker than that in the total population. The Chip- and SAGE- based correlation coefficient was 0.246 and 0.120, respectively, in abundant class, whereas, it was 0.462 and 0.576, respectively, in total.

Discussion

The term "housekeeping gene" has three slightly different meanings. Originally, it refers to constitutively expressed genes, which are in contrast to genes with regulated expression (67). It also refers to genes expressed in any type of cells. Genes expressed in a wide variety of tissues are also practically referred to as housekeeping genes. In addition, they include genes with functions that are essential for cell life (68). Genes with one of these features are often assumed to have the additional features as well. For example, housekeeping genes have been used to standardize the amount of mRNA. Warrington et al. (2000) expressed their motivation for measuring anatomical expression profiles as "an effort to identify the subset of genes required for cell maintenance." Zhang and Li (2004) reasoned that the fast rate of evolution among widely expressed genes by stronger purifying pressure for genes essential to cells. With the advent of functional genomics, we are encountering increasing numbers of genes that do not possess all these features together. For example, the regulated expression of several genes with housekeeping functions have been reported in the search for ideal normalization probes used in microarray hybridization (69,70). The contamination of broad non-HK in widely distributed transcripts, shown in this work, encompasses a new class of such exceptions. More accurate terminology will be necessary to prevent confusion among them.

According to the tissue-categorized EST data, one-fourth of our genes are classified as rarely expressed. The peak EST rate for this class corresponds to approximately less than10 copies per cell. Most of the genes in this class were indexed by the smallest breadths, but obviously these low breadths are not reliable because even true housekeeping genes (HK control) have the smallest breadths in this class.

Nevertheless, this class seemed to be actually rich in tissue specific genes because compared with other abundance classes, genes in this class were poor in start CGI (34%; 37/109) and rich in integral membrane proteins (30%; 458/1553). This refractory fraction of breadth-abundance correlation also seemed spurious, at least in part, because genes unique to minor populations of highly complex tissue, such as the neurotransmitter receptor in the brain, were accumulated in this class, and the abundance of the transcripts for these genes was probably still underestimated by the peak EST rate calculated based on our organ level categorization of EST.

In conclusion, I would like to emphasize that expression information for about half of our genes is not yet reliable. Enhancing the sensitivity by almost two orders of magnitudes in profiling methods would be necessary for the reliable characterization of the entire population. In previous works, the characterized set of housekeeping genes ranges from 451(1,26) to 1,927 (3). According to our results, these are only fractions of genes that are biased in favor of abundantly expressed genes. Any statistical characterization of the features of housekeeping genes would be worth revisiting by selecting control populations more carefully. Finally, I would like to point out that in the analyses of genome-wide measurements, I are bound to be misguided if I treat genes as anonymous Eliminating errors and the evaluation of each step by established molecular biological knowledge in general and for individual genes appears to be the only solution to reduce such misguidance.

Materials and Methods

The human gene set and transcript abundance.

RefSeq sequences (NM and XM) mapped in the NCBI human genome map view (BUILD 34.2) were obtained from NCBI and regarded as genes after collapsing them according to their locus. RefSeq sequences annotated as pseudogenes, the T cell receptor, and immunoglobulin were excluded from the analysis to avoid confusion. EST data and RefSeq-UniGene-EST correspondence were obtained from NCBI:UniGene (Build 166). 8,209 EST libraries constructed from normal tissues were selected and manually classified into 10 tissue categories, as described(71). For each gene, the EST frequency was calculated in each tissue category, and the maximum tissue specific frequency for the gene was used as the 'peak EST rate' of expression. Subcellular localization annotation of UniProt Release 2.3 was assigned to the genes using a relation table in Ensemble Release 22.34d for mapping UniProt ID onto RefSeq ID.

The table of start CGI identification data of GenBank proteins was obtained from the Ponger's Web site http://pbil.univ-lyon1.fr/datasets/Ponger2001/data.html (66). The GenBank protein ID was converted into the RefSeq ID using a relation table provided at LocusLink.

Gene expression data

1. GeneChip and SAGE data.

The GeneChip data set representing 85 hybridization results was obtained from the GEO database (GDS181). Data representing normal tissues were selected and tissue-categorized in accordance with the methodology described by Su et al. (2002).

SAGE tag frequency data (GDS217) was obtained from the GEO database and they were clumped together based on tissue categories in accordance with the methodology described by Lercher et al. (2002). The tag frequency data for each tissue category were converted to UniGene-frequency relation using the tag-Unigene (Build166) correspondence table (reliable tag file) generated for UniGene provided on the SAGE map site.

2. iAFLP profiling of control genes.

iAFLP is a method for the relative quantitation of target sequences across multiple sources (65). In brief, the cDNAs obtained from six mRNA sources were cleaved with MboI and adapted with mutually different oligomer cassettes. These casettes have the same 20 nucleotide sequences at the free end but mutually differ in length at the ligating ends. PCR amplification of an even pool of six differentially adapted cDNAs with one gene specific primer and one dye-labeled adapter primer yields competitively amplified specific fragments having small size differences. The ratio across six fragments, analyzed on an auto sequencer, represents the concentration ratio of the target sequence among the six mRNA sources. For comparing more than six cDNA sources, multiple pools were developed with a universal reference mRNA included in each pool. The RNAs used were all poly A RNA from a pooled specimen purchased from a commercial supplier (Clontech) and universal RNA was also purchased (Clontech).

Estimation of specificity.

The specificity of housekeeping gene selection was defined as the proportion of true housekeeping genes in the selected housekeeping gene set. The specificity was estimated as a function of the ratio of true broad non-HK to true tisssue-specific genes under the assumption that the entire gene set is composed of housekeeping, broad non-HK, and tissue-specific genes. When the ratio is large, the specificity estimate approaches the maximum value and when the ratio is small, the estimate approaches the minimal value.

Table 1: Statistics of RefSeq Genes measured by both GeneChip and SAGE (7,592 genes)

	GeneChip		SAGE		Union of GeneChip HK and SAGE HK	
	Original HK	Expanded HK	Original HK	Expanded HK	Original HK	Expanded HK
Count	700	1,333	1,633	3,036	1,904	3,402
Positive Gene Count	1,976—2,646 (av. 2450.)	—	878—5,484 (av. 2628)	1,993—6,285 (av. 3779)	—	—

		GeneChip				SAGE				Union of GeneChip HK and SAGE HK	
		Original		Expanded		Original		Expanded		Original	Expanded
Positive control genes	Average EST count	Average breadth	Sensitivity	Average breadth	Sensitivity	Average breadth	Sensitivity	Average breadth	Sensitivity	Sensitivity	Sensitivity
Ribosomal Protein	2,610	24.2/25	0.92 (56/61)	24.2/25	0.95 (58/61)	11.8/14	0.84 (51/61)	12.4/14	0.90 (55/61)	0.95 (58/61)	0.97 (59/61)
Proteasome	637.3	19.6/25	0.38 (13/34)	19.6/25	0.56 (19/34)	9.7/14	0.65 (22/34)	11.5/14	0.82 (28/34)	0.70 (24/34)	0.88 (30/34)
RNA Polymerase	241.3	8.8/25	0.10 (4/37)	8.8/25	0.30 (11/37)	5.0/14	0.16 (6/37)	8.0/14	0.51 (19/37)	0.27 (10/37)	0.62 (23/37)
Negative control genes	Average EST count	Average breadth	Sensitivity	Average breadth	Sensitivity	Average breadth	Sensitivity	Average breadth	Sensitivity	Sensitivity	Sensitivity
Blood Cell and Epithelial cell specific genes	411.9	6.4/25	0.030 (2/66)	6.4/25	0.045 (3/66)	3.3/14	0.12 (8/66)	5.2/14	0.17 (11/66)	0.14 (9/66)	0.20 (13/66)

Table 2: Sensitivity and Specificity of old and improved housekeeping gene set

Abundance bins		Old HK	Expanded HK	Non-HK	Total
Abundant (on-rate ≥ 100)	Count	2,100	3,304	2,030	5,334
	Sensitivity	0.84 (89/106)	0.93 (99/106)		
	Specificity	0.87 (0.71—1.0)	0.91 (0.77—1.0)		
	Sensitivity to broad non-HK	0.17 (8/48)	0.29 (14/48)		
	Start CGI	0.87 (88/101)	0.78 (119/152)	0.35 (36/102)	
Middle (100 > on-rate ≥ $10^{1.5}$)	Count	630	2,022	4,710	6,732
	Sensitivity	0.13 (3/24)	0.54 (13/24)		
	Specificity	0.87 (0.61—1.0)	0.88 (0.75—1.0)		
	Sensitivity to broad non-HK	0.065 (2/31)	0.13 (4/31)		
	Start CGI	0.84 (11/13)	0.86 (37/43)	0.56 (62/111)	
Rare (on-rate < $10^{1.5}$)	Count	62	211	4,615	4,826
	Sensitivity	0.0 (0/2)	0.0 (0/2)		
	Specificity	— (e.r: 0.0 = 0/39)	— (e.r: 0.0 = 0/39)		
	Sensitivity to broad non-HK	0.0 (0/39)	0.0 (0/39)		
	Start CGI	0.0 (0/1)	0.75 (3/4)	0.32 (34/105)	
All bins	Count	2,792	5,537	11,355	16,892
	Sensitivity	0.70 (92/132)	0.85 (112/132)		
	Specificity	0.79 (0.55—1.0)	0.84 (0.66—1.0)		
	Sensitivity to broad non-HK	0.085 (10/118)	0.15 (18/118)		
	Start CGI	0.86 (99/115)	0.80 (159/199)	0.42 (132/318)	

Table 3. Sensitivity and Specificity of old HK and expanded HK.

Abundance class	Total genes	Old HK selection			Expanded HK selection		
		gene number	sensitivity	specificity	gene number	sensitivity	specificity
High ($\rho \geq 100$ cpm)	5,334	2,100	0.84 (89/106)	0.87 (0.71—1.0)	3,304	0.93 (99/106)	0.91 (0.77—1)
Middle ($100 > \rho \geq 10^{1.5}$)	6,732	630	0.13 (3/24)	0.87 (0.61—1.0)	2,022	0.54 (13/24)	0.89 (0.75—1.0)
Low ($\rho < 10^{1.5}$)	4,826	62	0.0 (0/2)	—	211	0.0 (0/2)	—
Total	16,892	2,792	0.70 (92/132)	0.79 (0.55—1.0)	5,537	0.85 (112/132)	0.84 (0.66—1.0)

ρ: peak EST rate.

Figure Legends

Figure 1 Comparison of GeneChip breadth and SAGE breadth.

(a) Comparison of the two expression breadths determined by previous studies.(3,64) and (b) according to the improved "present" calling. The number of genes with each combination of breadths is shown in the matrix format where a value is represented by the red intensity in each cell. Among the 7,592 genes represented in the matrix, 37 RNA polymerase complex components (yellow dots) and 33 peptide hormone genes (blue crosses) are overlaid as HK control and TS control. For example, there are 97 genes with a GeneChip breadth = 25 and SAGE breadth = 0 in the panel (a). One of them is a peptide hormone (a blue cross).

Figure 2 Three platform comparison of anatomical expression patterns for RNA polymerase components.

Anatomical expression patterns of 36 positive control genes (RNA polymerase and general transcription factors) based on iAFLP, GeneChip, and SAGE. The tissues used in the iAFLP profiles are shown in panel (a) through (e) in the order of peak positions.

(a) pituitary gland, skin, adipose tissue, retina, spleen, Human Universal Reference

(b) brain, kidney, bone marrow, testis, ovary, Human Universal Reference

(c) brain, skeletal muscle, liver, spleen, testis, Human Universal Reference

(d) corpus callosum, thymus, uterus, prostate, skeletal muscle, Human Universal Reference

(e) heart, appendix, adrenal cortex, adrenal medulla, thyroid, Human Universal Reference

(f) GDS181(GeneChip U95A) expression profile. The line on the chart indicates the presence call threshold (AD = 200). Each bar represents the AD value of DRG (dorsal root ganglion), adrenal gland, amygdala, caudate nucleus, cerebellum, corpus callosum, cortex, heart, kidney, liver, lung, ovary (pooled), pancreas, pituitary gland, placenta, prostate, salivary gland, spinal cord, spleen, testis, thalamus, thymus, thyroid, trachea, and uterus.

(g) GDS217(SAGE) expression profile. The line on the chart is the presence call threshold of Lercher et al. (2002), tag count = 2. Each bar represents tag frequencies of blood, brain, breast, colon, endothelium, heart, kidney, liver, lung, ovary, pancreas, prostate, skin, and stomach.

Figure 3 Relationships between peak EST rate and sensitivity of housekeeping gene identification.

In the histogram of gene numbers binned according to the peak EST rate, 16,892 genes profiled by SAGE or GeneChip were classified into three categories—red: 2,792 housekeeping genes (old HK) identified by Su et al. (2002) or Lercher et al. (2002), green: 2,745 housekeeping genes newly identified in this work (new HK), and blue: remaining 11,355 genes with no positive identification.

Distribution of the peak EST rate of the HK control gene set is shown in the three "rug representations" using the histogram as the log scaled on-rate axis. Orange and blue rugs indicate whether each gene is identified as a housekeeping gene in the expanded selection. (a) 61 ribosomal proteins (b) 34 proteasome components (c) 37 RNA polymerase components (GO:0003899), and general RNA polymerase II transcription factor activities (GO:0016251).

Figure 4 Correlation between expression abundance and expression breadth in 132 HK control genes.

Scatter plot consists of 132 HK control genes (61 ribosomal proteins, 34 proteasomes, and 37 general RNA polymerase II transcription factors). The transcript abundance was assessed by the total EST count and the peak EST rate (see Materials and Methods). The abundances were transformed to common logarithm.

Figure 5 Comparison of evenness among old HK, new HK, and the remaining genes.

The frequency of evenness in the old HK (red), new HK (green), non-HK (blue), HK controls (orange dashed line), and TS controls (blue dashed line) were plotted as frequency polygons.

(a) Distribution of evenness based on the SAGE profiles and (b) based on the GeneChip profiles.

Figure 6 Comparison of the subcellular localization distribution among old HK, new HK, and non-HK.

4,238 RefSeq genes that have subcellular localization information in SwissProt were subdivided into six classes based on the localization. The proportion was then compared between old HK, new HK, and non-HK.

Figure 1.

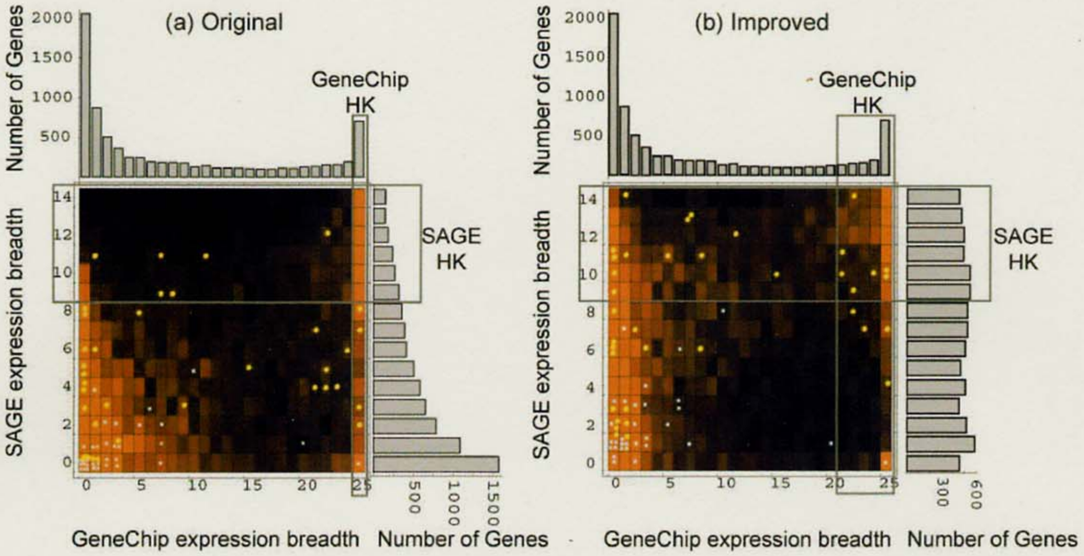

(a) Original
(b) Improved
Number of Genes
2000
1500
1000
500
GeneChip
HK
SAGE
HK
SAGE expression breadth
14
12
10
8
6
4
2
0
0
5
10
15
20
25
500
1000
1500
300
600
GeneChip expression breadth
Number of Genes

Figure 2.

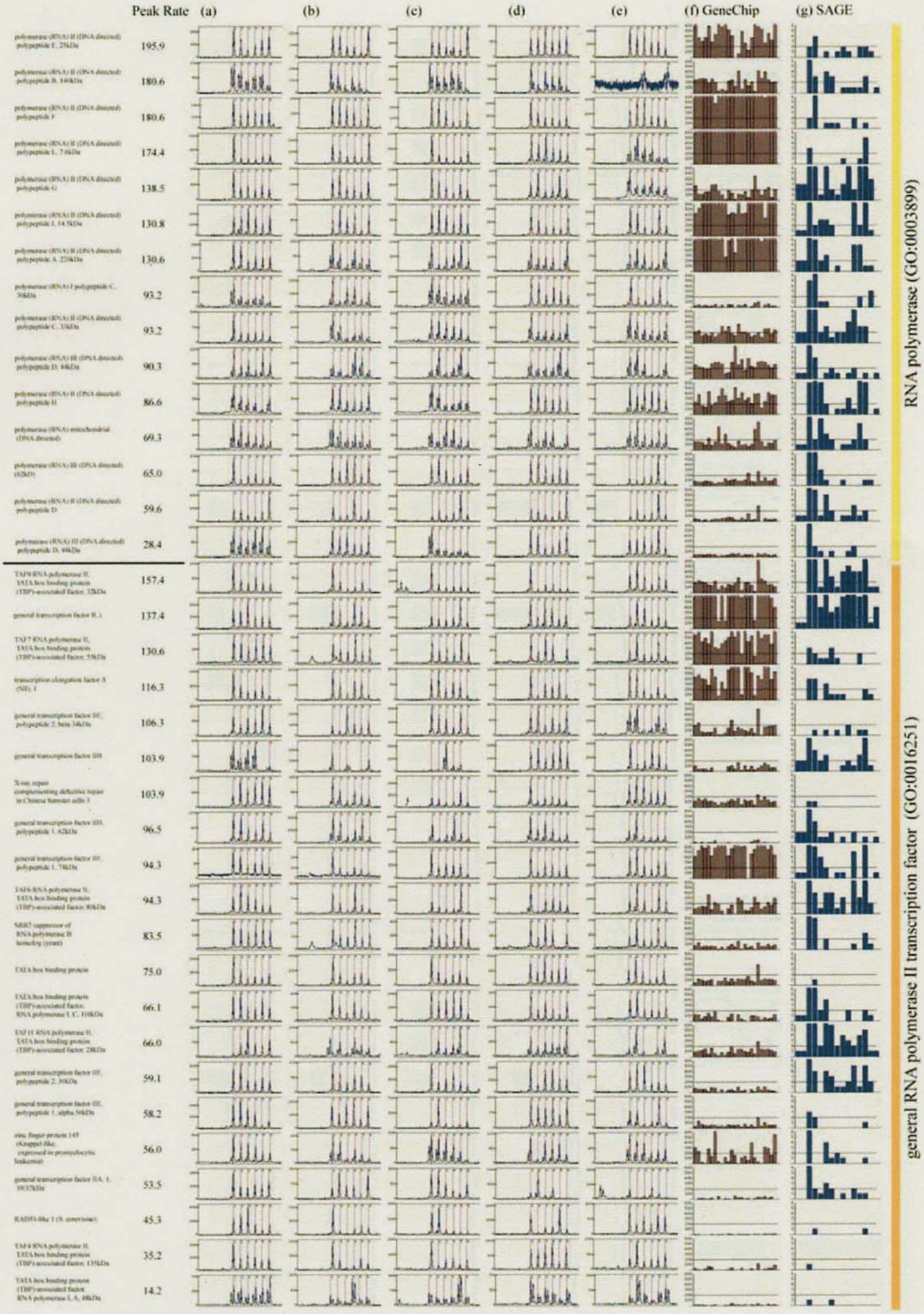

Peak Rate
(a)
(b)
(c)
(d)
(e)
(f) GeneChip
(g) SAGE
polymerase (RNA) II (DNA directed) polypeptide E, 25kDa
195.9
polymerase (RNA) II (DNA directed) polypeptide B, 140kDa
180.6
polymerase (RNA) II (DNA directed) polypeptide F
180.6
polymerase (RNA) II (DNA directed) polypeptide L, 7.6kDa
174.4
polymerase (RNA) II (DNA directed) polypeptide G
138.5
polymerase (RNA) II (DNA directed) polypeptide I, 14.5kDa
130.8
polymerase (RNA) II (DNA directed) polypeptide A, 220kDa
130.6
polymerase (RNA) I polypeptide C, 30kDa
93.2
polymerase (RNA) II (DNA directed) polypeptide C, 33kDa
93.2
polymerase (RNA) III (DNA directed) polypeptide D, 44kDa
90.3
polymerase (RNA) II (DNA directed) polypeptide H
86.6
polymerase (RNA) mitochondrial (DNA directed)
69.3
polymerase (RNA) III (DNA directed) (62kD)
65.0
polymerase (RNA) II (DNA directed) polypeptide D
59.6
polymerase (RNA) III (DNA directed) polypeptide D, 44kDa
28.4
TAF9 RNA polymerase II, TATA box binding protein (TBP)-associated factor, 32kDa
157.4
general transcription factor II, i
137.4
TAF7 RNA polymerase II, TATA box binding protein (TBP)-associated factor, 55kDa
130.6
transcription elongation factor A (SII), 1
116.3
general transcription factor IIE, polypeptide 2, beta 34kDa
106.3
general transcription factor IIB
103.9
X-ray repair complementing defective repair in Chinese hamster cells 3
103.9
general transcription factor IIH, polypeptide 1, 62kDa
96.5
general transcription factor IIF, polypeptide 1, 74kDa
94.3
TAF6 RNA polymerase II, TATA box binding protein (TBP)-associated factor, 80kDa
94.3
SRB7 suppressor of RNA polymerase B homolog (yeast)
83.5
TATA box binding protein
75.0
TATA box binding protein (TBP)-associated factor, RNA polymerase I, C, 110kDa
66.1
TAF11 RNA polymerase II, TATA box binding protein (TBP)-associated factor, 28kDa
66.0
general transcription factor IIF, polypeptide 2, 30kDa
59.1
general transcription factor IIE, polypeptide 1, alpha 56kDa
58.2
zinc finger protein 145 (Kruppel-like, expressed in promyelocytic leukemia)
56.0
general transcription factor IIA, 1, 19/37kDa
53.5
RAD51-like 1 (S. cerevisiae)
45.3
TAF4 RNA polymerase II, TATA box binding protein (TBP)-associated factor, 135kDa
35.2
TATA box binding protein (TBP)-associated factor, RNA polymerase I, A, 48kDa
14.2
RNA polymerase (GO:0003899)
general RNA polymerase II transcription factor (GO:0016251)

Figure 3.

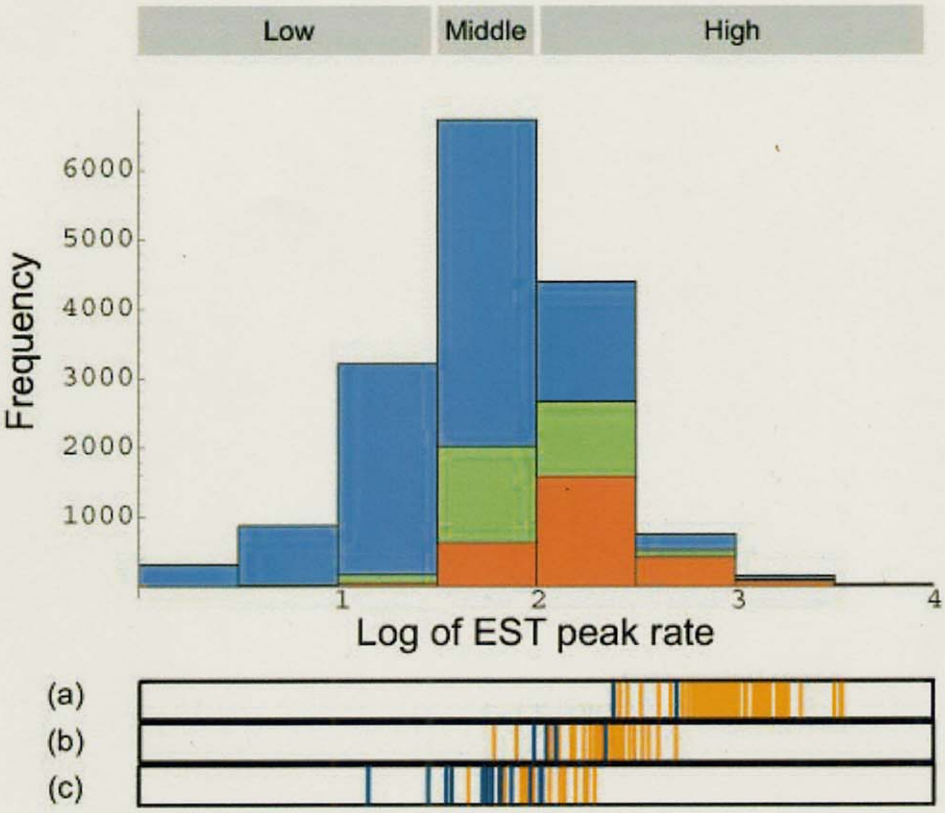

Low
Middle
High
Frequency
6000
5000
4000
3000
2000
1000
1
2
3
4
Log of EST peak rate
(a)
(b)
(c)

Figure 4.

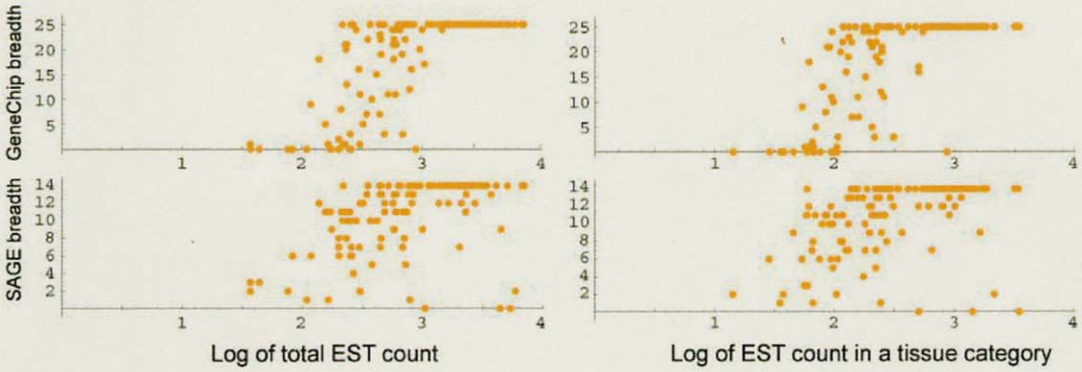

GeneChip breadth
25
20
15
10
5
1
2
3
4
SAGE breadth
14
12
10
8
6
4
2
Log of total EST count
Log of EST count in a tissue category

Figure 5.

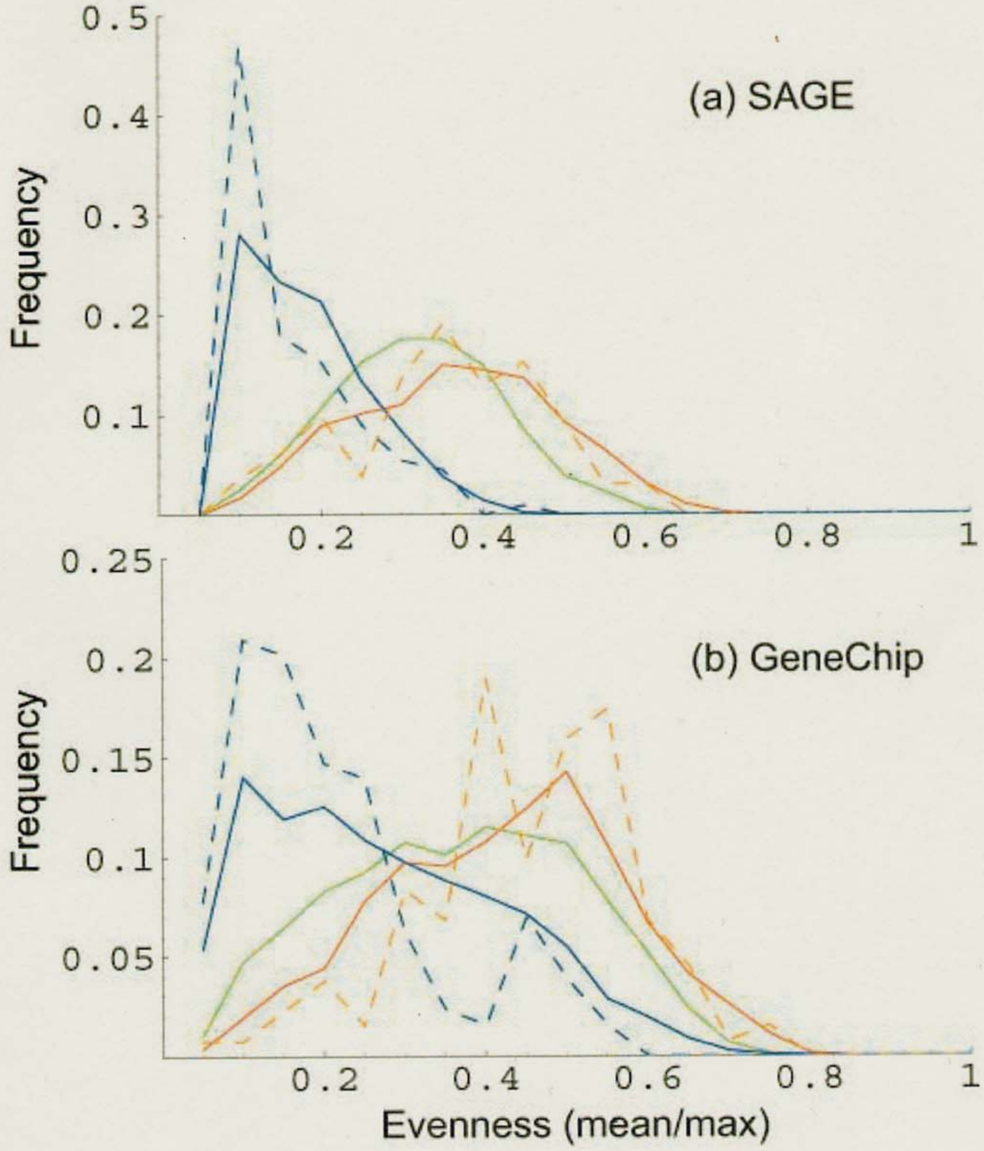
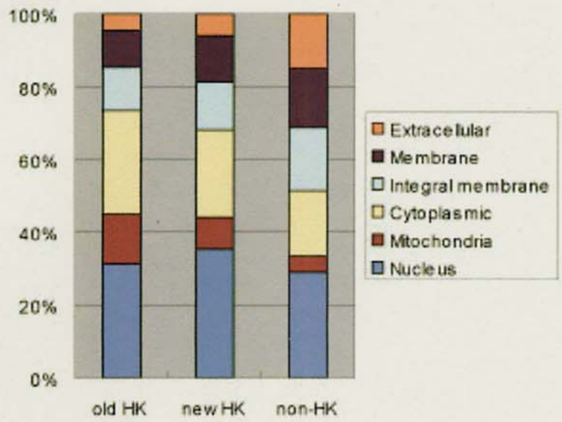

(a) SAGE
Frequency
0.5
0.4
0.3
0.2
0.1
0.2
0.4
0.6
0.8
1
(b) GeneChip
Frequency
0.25
0.2
0.15
0.1
0.05
0.2
0.4
0.6
0.8
1
Evenness (mean/max)

Figure 6.


100%
80%
60%
40%
20%
0%
old HK
new HK
non-HK
Extracelular
Membrane
Integral membrane
Cytoplasmic
Mitochondria
Nucleus

Conclusion

It is expected that hidden causal relationships between some biological processes are unraveled by a close examination of newly revealed phenomena in genomics. In this thesis, I examined Zipf's law of transcriptome, which is one of the phenomena revealed by genome-wide studies of gene expression, and I found a new relationship describing the evolutionary change of the expression levels of each gene in each cell type.

This relationship was stated as follows. (This is called the evolutionary model of Zipf's law of transcriptome)

1. The baseline expression level of each gene in a cell is coded in the genome sequence such as in the cis-element or enhancer region; therefore, the expression levels of genes are affected by the mutation, and the effects are inherited by the offspring.
2. Some of the mutations in the genome caused the stochastic change of expression level of some genes, proportional to the current expression levels of the genes.
3. The number of expressed genes in a cell type is nearly constant throughout the evolutionary process, and any functional gene is prohibited from losing its gene expression ability by purifying selection.

I showed that the Zipf's law of transcriptome could be replicate from these three statements using the Monte-Carlo simulations. In addition, I pointed out that if the three statements are valid, the expression divergence will accumulated with constant ratio. Khaitovich et.al. (2004) reported that the clock like accumulation of expression divergence. This is a strong evidence of the evolutional model of Zipf's law.

At approximately the same time when I proposed the evolutionary model, several other models of Zipf's law of transcriptome were published. All of these models (the dynamics models) attributed the cause of Zipf's law to the gene expression dynamics in each cells. All of the dynamics models states that the change of gene expression levels in each cell during each cell life could be formulated by a kind of geometrical Brownian movement. However, based on the dynamics model, it follows that the expression level of each gene will diverge independently among the cells, even if the same type of cell is considered. It is obviously unrealistic. In order to avoid this unrealistic conclusion, the dynamics model should accept an additional assumption that "each gene has its own dynamic range of expression level in each cell type." In such a case, the following questions should be answered: How are the ranges of expression level determined, and how are they determined in the form of Zipf's law? The evolutionary model answers these questions as follows: the ranges of expression levels are determined by the genome code, and they change proportional to the current expression levels of the genes in the evolutionary process.

My evolutionary model of Zipf's law of transcriptome arose following two new questions about the evolution of the gene expression.

1. What is the cause of the stochastic change of gene expression in the evolutionary process? Is it a result of some kinds of selection or of the neutral evolutionary change?
2. What is the relationship between the evolutions of expression levels in different type of cells in the same species? Do the gene expression levels evolve independently among different cell types or do they have any correlation among some cell types?

One of the promising approaches to access these questions was to investigate the evolution of anatomical gene expression patterns (expression patterns in various tissues) of each gene. Therefore I focused on housekeeping genes as a set of genes that definitely express and function in all cell types. Although several sets of housekeeping genes were published previously, but I found that they had serious problem in their identification performance of housekeeping genes. By comparing the published screening results for housekeeping genes, one based on the GeneChip method and the other based on the SAGE method, I found low concordance between the results of the two screening methods, and in both screening process, there was poor sensitivity in the identification of housekeeping genes. Therefore, in this thesis, I constructed a more reliable set of housekeeping genes from publicly available expression data, with the estimation of the performance of the identification. As a result, I succeeded in doubling the number of candidate housekeeping genes without any evidence of losing specificity. Estimated contaminants, which comprise approximately 12%–20% of either new or old HK, were genes unique to widely distributed cells rather than those common to a wide variety of cells. Finally, I pointed out that all previously characterized sets of housekeeping genes are considerably biased in favor of abundantly expressed genes. Although the questions that were arisen from my evolutionary model of the Zipf's law have been remained unsolved, the newly identified sets of housekeeping genes will be give some insight of the answer to the questions.

References

1. Zhang, L. and Li, W.H. (2004) Mammalian housekeeping genes evolve more slowly than tissue-specific genes. ***Mol Biol Evol***, 21, 236-239.
2. Caron, H., van Schaik, B., van der Mee, M., Baas, F., Riggins, G., van Sluis, P., Hermus, M.C., van Asperen, R., Boon, K., Voute, P.A. ***et al.*** (2001) The human transcriptome map: clustering of highly expressed genes in chromosomal domains. ***Science***, 291, 1289-1292.
3. Lercher, M.J., Urrutia, A.O. and Hurst, L.D. (2002) Clustering of housekeeping genes provides a unified model of gene order in the human genome. ***Nat Genet***, 31, 180-183.
4. Lercher, M.J., Blumenthal, T. and Hurst, L.D. (2003) Coexpression of neighboring genes in Caenorhabditis elegans is mostly due to operons and duplicate genes. ***Genome Res***, 13, 238-243.
5. Cohen, B.A., Mitra, R.D., Hughes, J.D. and Church, G.M. (2000) A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. ***Nat Genet***, 26, 183-186.
6. Blumenthal, T., Evans, D., Link, C.D., Guffanti, A., Lawson, D., Thierry-Mieg, J., Thierry-Mieg, D., Chiu, W.L., Duke, K., Kiraly, M. ***et al.*** (2002) A global analysis of Caenorhabditis elegans operons. ***Nature***, 417, 851-854.
7. Boutanaev, A.M., Kalmykova, A.I., Shevelyov, Y.Y. and Nurminsky, D.I. (2002) Large clusters of co-expressed genes in the Drosophila genome. ***Nature***, 420, 666-669.
8. Florens, L., Washburn, M.P., Raine, J.D., Anthony, R.M., Grainger, M., Haynes, J.D., Moch, J.K., Muster, N., Sacci, J.B., Tabb, D.L. ***et al.*** (2002) A proteomic view of the Plasmodium falciparum life cycle. ***Nature***, 419, 520-526.
9. Kamath, R.S., Fraser, A.G., Dong, Y., Poulin, G., Durbin, R., Gotta, M., Kanapin, A., Le Bot, N., Moreno, S., Sohrmann, M. ***et al.*** (2003) Systematic functional analysis of the Caenorhabditis elegans genome using RNAi. ***Nature***, 421, 231-237.
10. Pal, C. and Hurst, L.D. (2003) Evidence for co-evolution of gene order and recombination rate. ***Nat Genet***, 33, 392-395.
11. Eisenberg, E. and Levanon, E.Y. (2003) Human housekeeping genes are compact. ***Trends Genet***, 19, 362-365.
12. Rifkin, S.A., Kim, J. and White, K.P. (2003) Evolution of gene expression in the Drosophila melanogaster subgroup. ***Nat Genet***, 33, 138-144.
13. Lande, R. (1998) Risk of population extinction from fixation of deleterious and

reverse mutations. *Genetica*, 102-103, 21-27.

14. Zipf, G.K. (1949) ***Human Behavior and the Principles of Least Effort.*** Addison-Wesley, Cambridge, UK.

15. Pareto, V. (1896) Cours d'Economie Politique. Geneva, Switzlerland.

16. Zhang, L., Zhou, W., Velculescu, V.E., Kern, S.E., Hruban, R.H., Hamilton, S.R., Vogelstein, B. and Kinzler, K.W. (1997) Gene expression profiles in normal and cancer cells. *Science*, 276, 1268-1272.

17. Luscombe, N.M., Qian, J., Zhang, Z., Johnson, T. and Gerstein, M. (2002) The dominance of the population by a selected few: power-law behaviour applies to a wide variety of genomic properties. *Genome Biol*, 3, RESEARCH0040.

18. Kuznetsov, V.A., Knott, G.D. and Bonner, R.F. (2002) General statistics of stochastic process of gene expression in eukaryotic cells. *Genetics*, 161, 1321-1332.

19. Ogasawara, O., Kawamoto, S. and Okubo, K. (2003) Zipf's law and human transcriptomes: an explanation with an evolutionary model. *C R Biol*, 326, 1097-1101.

20. Kuznetsov, V.A. (2003) Family of skewed distributions associated with the gene expression and proteome evolution. *Signal Processing*, 83, 889-910.

21. Furusawa, C. and Kaneko, K. (2003) Zipf's law in gene expression. *Phys Rev Lett*, 90, 088102.

22. Ueda, H.R., Hayashi, S., Matsuyama, S., Yomo, T., Hashimoto, S., Kay, S.A., Hogenesch, J.B. and Iino, M. (2004) Universality and flexibility in gene expression from bacteria to human. *Proc Natl Acad Sci U S A*, 101, 3765-3769.

23. Konishi, T. (2004) Three-parameter lognormal distribution ubiquitously found in cDNA microarray data and its application to parametric data treatment. *BMC Bioinformatics*, 5, 5.

24. Velculescu, V.E., Madden, S.L., Zhang, L., Lash, A.E., Yu, J., Rago, C., Lal, A., Wang, C.J., Beaudry, G.A., Ciriello, K.M. *et al.* (1999) Analysis of human transcriptomes. *Nat Genet*, 23, 387-388.

25. Warrington, J.A., Nair, A., Mahadevappa, M. and Tsyganskaya, M. (2000) Comparison of human adult and fetal expression and identification of 535 housekeeping/maintenance genes. *Physiol Genomics*, 2, 143-147.

26. Hsiao, L.L., Dangond, F., Yoshida, T., Hong, R., Jensen, R.V., Misra, J., Dillon, W., Lee, K.F., Clark, K.E., Haverty, P. *et al.* (2001) A compendium of gene expression in normal human tissues. *Physiol Genomics*, 7, 97-104.

27. Bortoluzzi, S., Rampoldi, L., Simionati, B., Zimbello, R., Barbon, A., d'Alessi, F., Tiso, N., Pallavicini, A., Toppo, S., Cannata, N. *et al.* (1998) A comprehensive,

high-resolution genomic transcript map of human skeletal muscle. ***Genome Res***, 8, 817-825.

28. Lercher, M.J., Urrutia, A.O., Pavlicek, A. and Hurst, L.D. (2003) A unification of mosaic structures in the human genome. ***Hum Mol Genet***, 12, 2411-2415.

29. Barabasi, A.L. and Albert, R. (1999) Emergence of scaling in random networks. ***Science***, 286, 509-512.

30. Albert, R., Jeong, H. and Barabasi, A.L. (2000) Error and attack tolerance of complex networks. ***Nature***, 406, 378-382.

31. Gutenberg, B. and Richter, C.F. (1956) Magnitude and energy of earthquakes. ***Ann Geophys***, 9, 1.

32. Mantegna, R.N., Buldyrev, S.V., Goldberger, A.L., Havlin, S., Peng, C.K., Simons, M. and Stanley, H.E. (1994) Linguistic features of noncoding DNA sequences. ***Phys Rev Lett***, 73, 3169-3172.

33. Mantegna, R.N., Buldyrev, S.V., Goldberger, A.L., Havlin, S., Peng, C.K., Simons, M. and Stanley, H.E. (1995) Systematic analysis of coding and noncoding DNA sequences using methods of statistical linguistics. ***Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics***, 52, 2939-2950.

34. Martindale, C. and Konopka, A.K. (1996) Oligonucleotide frequencies in DNA follow a Yule distribution. ***Computer Chem***, 20, 35-38.

35. Gerstein, M. (1998) Patterns of protein-fold usage in eight microbial genomes: a comprehensive structural census. ***Proteins***, 33, 518-534.

36. Gerstein, M. (1997) A structural census of genomes: comparing bacterial, eukaryotic, and archaeal genomes in terms of protein structure. ***J Mol Biol***, 274, 562-576.

37. Gerstein, M. and Levitt, M. (1997) A structural census of the current population of protein sequences. ***Proc Natl Acad Sci U S A***, 94, 11911-11916.

38. Gerstein, M. (1998) How representative are the known structures of the proteins in a complete genome? A comprehensive structural census. ***Fold Des***, 3, 497-512.

39. Huynen, M.A. and van Nimwegen, E. (1998) The frequency distribution of gene family sizes in complete genomes. ***Mol Biol Evol***, 15, 583-589.

40. Qian, J., Luscombe, N.M. and Gerstein, M. (2001) Protein family and fold occurrence in genomes: power-law behaviour and evolutionary model. ***J Mol Biol***, 313, 673-681.

41. Koonin, E.V., Wolf, Y.I. and Karev, G.P. (2002) The structure of the protein universe and genome evolution. ***Nature***, 420, 218-223.

42. Jeong, H., Tombor, B., Albert, R., Ol

interactions: intramolecular and intermolecular protein family interaction repertoires in the PDB and yeast. ***J Mol Biol***, 307, 929-938.

44. Brem, R.B., Yvert, G., Clinton, R. and Kruglyak, L. (2002) Genetic dissection of transcriptional regulation in budding yeast. ***Science***, 296, 752-755.

45. Townsend, J.P., Cavalieri, D. and Hartl, D.L. (2003) Population genetic variation in genome-wide gene expression. ***Mol Biol Evol***, 20, 955-963.

46. Oleksiak, M.F., Churchill, G.A. and Crawford, D.L. (2002) Variation in gene expression within and among natural populations. ***Nat Genet***, 32, 261-266.

47. Sandberg, R., Yasuda, R., Pankratz, D.G., Carter, T.A., Del Rio, J.A., Wodicka, L., Mayford, M., Lockhart, D.J. and Barlow, C. (2000) Regional and strain-specific gene expression mapping in the adult mouse brain. ***Proc Natl Acad Sci U S A***, 97, 11038-11043.

48. Cheung, V.G., Conlin, L.K., Weber, T.M., Arcaro, M., Jen, K.Y., Morley, M. and Spielman, R.S. (2003) Natural variation in human gene expression assessed in lymphoblastoid cells. ***Nat Genet***, 33, 422-425.

49. Li, W. and Yang, Y. (2002) Zipf's law in importance of genes for cancer classification using microarray data. ***J Theor Biol***, 219, 539-551.

50. Yan, H. and Zhou, W. (2004) Allelic variations in gene expression. ***Curr Opin Oncol***, 16, 39-43.

51. Schadt, E.E., Monks, S.A., Drake, T.A., Lusis, A.J., Che, N., Colinayo, V., Ruff, T.G., Milligan, S.B., Lamb, J.R., Cavet, G. ***et al.*** (2003) Genetics of gene expression surveyed in maize, mouse and man. ***Nature***, 422, 297-302.

52. Morley, M., Molony, C.M., Weber, T.M., Devlin, J.L., Ewens, K.G., Spielman, R.S. and Cheung, V.G. (2004) Genetic analysis of genome-wide variation in human gene expression. ***Nature***, 430, 743-747.

53. Khaitovich, P., Weiss, G., Lachmann, M., Hellmann, I., Enard, W., Muetzel, B., Wirkner, U., Ansorge, W. and Paabo, S. (2004) A neutral model of transcriptome evolution. ***PLoS Biol***, 2, E132.

54. Ochiai, T., Nacher, J.C. and Akutsu, T. (2004) A constructive approach to gene expression dynamics. ***Physics Letter A***, 330, 313-321.

55. Felsenstein, J. (2004) ***Inferring Phylogenies***. Sinauer Associates., Sunderland(Massachusetts).

56. Jordan, I.K., Marino-Ramirez, L. and Koonin, E.V. (2005) Evolutionary significance of gene expression divergence. ***Gene***, 345, 119-126.

57. Hurst, L.D., Pal, C. and Lercher, M.J. (2004) The evolutionary dynamics of eukaryotic gene order. ***Nat Rev Genet***, 5, 299-310.

58. Williams, E.J. and Hurst, L.D. (2002) Clustering of tissue-specific genes underlies much of the similarity in rates of protein evolution of linked genes. ***J Mol Evol***, 54, 511-518.

59. Castillo-Davis, C.I. and Hartl, D.L. (2003) Conservation, relocation and duplication in genome evolution. ***Trends Genet***, 19, 593-597.

60. Megy, K., Audic, S. and Claverie, J.M. (2003) Positional clustering of differentially expressed genes on human chromosomes 20, 21 and 22. ***Genome Biol***, 4, P1.

61. Versteeg, R., van Schaik, B.D., van Batenburg, M.F., Roos, M., Monajemi, R., Caron, H., Bussemaker, H.J. and van Kampen, A.H. (2003) The human transcriptome map reveals extremes in gene density, intron length, GC content, and repeat pattern for domains of highly and weakly expressed genes. ***Genome Res***, 13, 1998-2004.

62. Lercher, M.J., Chamary, J.V. and Hurst, L.D. (2004) Genomic regionality in rates of evolution is not explained by clustering of genes of comparable expression profile. ***Genome Res***, 14, 1002-1013.

63. Vinogradov, A.E. (2004) Genome size and extinction risk in vertebrates. ***Proc R Soc Lond B Biol Sci***, 271, 1701-1705.

64. Su, A.I., Cooke, M.P., Ching, K.A., Hakak, Y., Walker, J.R., Wiltshire, T., Orth, A.P., Vega, R.G., Sapinoso, L.M., Moqrich, A. ***et al.*** (2002) Large-scale analysis of the human and mouse transcriptomes. ***Proc Natl Acad Sci U S A***, 99, 4465-4470.

65. Kawamoto, S., Ohnishi, T., Kita, H., Chisaka, O. and Okubo, K. (1999) Expression profiling by iAFLP: A PCR-based method for genome-wide gene expression profiling. ***Genome Res***, 9, 1305-1312.

66. Ponger, L., Duret, L. and Mouchiroud, D. (2001) Determinants of CpG islands: expression in early embryo and isochore structure. ***Genome Res***, 11, 1854-1860.

67. Smith, A.D., Datta, S. P., Howard Smith, G. Campbell, P. N. Bentley, R, McKenzie, H. A. (1997) ***Oxford Dictionary of Biochemistry and Molecular Biology***. Oxford University Press, New York.

68. Lewin, B. (1974) ***Gene Expression***. John Wiley and Sons Ltd, New York.

69. Thellin, O., Zorzi, W., Lakaye, B., De Borman, B., Coumans, B., Hennen, G., Grisar, T., Igout, A. and Heinen, E. (1999) Housekeeping genes as internal standards: use and limits. ***J Biotechnol***, 75, 291-295.

70. Lee, P.D., Sladek, R., Greenwood, C.M. and Hudson, T.J. (2002) Control genes and variability: absence of ubiquitous reference transcripts in diverse mammalian expression studies. ***Genome Res***, 12, 292-297.

71. Tanino, M., Debily, M.A., Tamura, T., Hishiki, T., Ogasawara, O., Murakawa, K., Kawamoto, S., Itoh, K., Watanabe, S., de Souza, S.J. ***et al.*** (2005) The Human

Anatomic Gene Expression Library (H-ANGEL), the H-Inv integrative display of human gene expression across disparate technologies and platforms. *Nucleic Acids Res*, 33 Database Issue, D567-572.