General Statistics of Stochastic Process of Gene Expression in Eukaryotic Cells

V. A. Kuznetsov*,[1] G. D. Knott† and R. F. Bonner*

**Laboratory of Integrative and Medical Biophysics, National Institute of Child Health and Human Development, National Institutes of Health, Bethesda, Maryland 20892-5772 and †Civilized Software, Silver Spring, Maryland 20906*

Manuscript received January 5, 2002
Accepted for publication March 31, 2002

ABSTRACT

Thousands of genes are expressed at such very low levels (≤1 copy per cell) that global gene expression analysis of rarer transcripts remains problematic. Ambiguity in identification of rarer transcripts creates considerable uncertainty in fundamental questions such as the total number of genes expressed in an organism and the biological significance of rarer transcripts. Knowing the distribution of the true number of genes expressed at each level and the corresponding gene expression level probability function (GELPF) could help resolve these uncertainties. We found that all observed large-scale gene expression data sets in yeast, mouse, and human cells follow a Pareto-like distribution model skewed by many low-abundance transcripts. A novel stochastic model of the gene expression process predicts the universality of the GELPF both across different cell types within a multicellular organism and across different organisms. This model allows us to predict the frequency distribution of all gene expression levels within a single cell and to estimate the number of expressed genes in a single cell and in a population of cells. A random "basal" transcription mechanism for protein-coding genes in all or almost all eukaryotic cell types is predicted. This fundamental mechanism might enhance the expression of rarely expressed genes and, thus, provide a basic level of phenotypic diversity, adaptability, and random monoallelic expression in cell populations.

GENE expression within a cell is a complex process involving chromatin remodeling, transcription, and export of RNA from the nucleus to the cytoplasm where mRNA molecules are translated into proteins. The physiological activity and cell differentiation of a mammalian cell is controlled by 10,000 or more protein-coding genes associated with ~300,000–500,000 mRNA transcripts (Bishop *et al.* 1974). The *complete* gene expression profile for a given set of cells is the list of all expressed genes, together with each gene's expression level defined as the average number of cytoplasmic mRNA transcripts per cell. Currently, gene expression profiling methods (*e.g.*, serial analysis of gene expression (SAGE; Velculescu *et al.* 1995, 1999), cDNA, or oligonucleotide microarrays (Holstege *et al.* 1998; Jelinsky and Samson 1999) measure gene transcripts from large numbers of cells (*i.e.*, not a single cell) and cannot reliably detect the thousands of genes that are expressed at very low copy numbers (less than two per cell). Many of these lower-level transcripts may be essential for determining normal and pathological cell phenotypes (Chen *et al.* 2000; Ohlsson *et al.* 2001). However, a rationale for an extreme number of rare transcripts has remained unresolved.

Determination of biologically significant expressed genes in eukaryotic cells is a challenging biological problem (Bishop *et al.* 1974; Velculescu *et al.* 1999). An important current issue for gene identification is determining the true statistical distributions of the number of genes expressed at *all* possible expression levels, both in *a single cell* and in *a population of cells.* Identification of such distributions can provide a theoretical basis for accurately counting the number of expressed genes and the total number of genes in a given cell type and for better understanding the mechanism(s) governing the expression of thousands of genes at very low levels. The similar problem of estimating the distribution of species in a population or different alleles in a population has been intensively discussed (see for references Huang and Weir 2001).

The statistics of expressed genes can be partially specified by the proportions of expressed genes that have one, two, etc. transcripts present in an associated mRNA sample (*i.e.*, a normalized histogram of gene expression levels). Analysis of such empirical histograms using large-scale gene expression databases leads to models of the underlying gene expression level probability functions (GELPF) in a cell and in a population of cells. Interestingly, similar gene expression "patterns" in different cells were observed >25 years ago by RNA-DNA hybridization (Bishop *et al.* 1974). These and more recent gene expression data sets have demonstrated very broad ranges of gene transcript levels (*i.e.*, from 0.1 to 20,000 transcripts per human cell; Bishop *et al.* 1974; Velculescu *et al.* 1999). However, a suitable theoretical model of the distribution of gene expression levels in a cell has not been previously identified due to under-

[1] *Corresponding author:* Laboratory of Integrative and Medical Biophysics, National Institute of Child Health and Human Development, NIH, Bldg. 13, Rm. 3W16, Bethesda, MD 20892-5772.
E-mail: vk28u@nih.gov

sampling, unreliable detection of many low abundance transcripts, experimental errors, and ambiguities in the identification of many transcripts.

A large body of experimental and theoretical literature on molecular mechanisms of gene expression control makes it increasingly evident that stochastic processes in transcription and translation machinery (as well as within signaling pathways and cross talk between different pathways) need to be considered to fully understand basic processes of gene expression. In particular, several experimental systems indicate that initiation of gene transcription is a discrete process in which many individual protein-coding genes existing in an off state can be stochastically switched to an on state resulting in the production of mRNAs in sporadic pulses (Ko 1992; Ross *et al.* 1994; Newlands *et al.* 1998; McAdams and Arkin 1999; Hume 2000; Sutherland *et al.* 2000; Ohlsson *et al.* 2001; Sano *et al.* 2001).

In this study we present evidence that the functional form of the GELPF is invariant among eukaryotic cell types. Stochastic and probabilistic mechanisms of the initiation of the gene expression process can help explain the observed universality of the GELPF across different cell types in a multicellular organism and across different organisms. We describe a new distribution function and derive from it a probabilistic model of the growth of a population (*e.g.*, the number of all transcripts) with many distinct classes (*e.g.*, distinct expressed genes) in a complex system (*e.g.*, observed SAGE transcriptome for a population of homogeneous cells) as sampling increases. This model allows us to predict the frequency distribution of gene expression levels for all genes and the total number of genes expressed in a representative cell averaged over time. The model exhibits predictive power even when the sequencing database is incomplete and contains ambiguity in sequence to gene assignments.

RESULTS

Distribution of the gene expression levels: We have analyzed diverse large-scale gene expression databases for different human tissues and cell lines (http://www.ncbi.nlm.nih.gov/UniLib; http://www.ncbi.nlm.nih.gov/UNIGENE; http://www.ncbi.nlm.nih.gov/CGAP/ncicgap; http://www.ncbi.nlm.nih.gov/SAGE), mouse tissues (http://www.ncbi.nlm.nih.gov/UniLib), and yeast cells (ftp://genome-ftp.stanford.edu.pub/yeast/tables/SAGE_Data; http://www.sagenet.org; http://www.hsph.harvard/geneexpression) to identify the GELPF for eukaryotic cells. These data sets have been created by three different technologies: sequencing of clones in complementary DNA (cDNA) and SAGE libraries and oligonucleotide microarray hybridization methods. These techniques involve making cDNA sequences of the less stable mRNA molecules and then using specific short-nucleotide sequence tags that match different mRNAs to quantify their relative abundance in the cell sample. For each data set, we define its *library* as the list of sequenced tags that match mRNAs associated with genes, together with the number of occurrences of each specific tag. Let M denote the size of the library, *i.e.*, the total number of tags in it, and let $n(m, M)$ denote the number of *distinct* tags that have the expression level m (occurring m times) in the given library of size M. The observed value $\tilde{n}(m, M)$ only approximates the number of expressed genes with expression level m in the cell sample due to experimental errors, nonunique tag-gene matching, and incorrect annotation of genes (see below). Let J denote the observed expression level for the most abundant tag in the library; J increases with the library size M. Then $\sum_{m=1}^{J}\tilde{n}(m, M) = N$ is the number of distinct tags in the library. The points $(m, g(m))$ for $m = 1, \ldots J$, where $g(m) = \tilde{n}(m, M)/N$, form the histogram corresponding to the empirical relative frequency distribution of expressed genes. This is a size-frequency form of the empirical GELPF and it represents an estimate of the GELPF in the cell sample.

We found that the empirical GELPF histograms, constructed for analyzed yeast SAGE libraries, mouse and human SAGE or cDNA libraries (Velculescu *et al.* 1995, 1999; Lal *et al.* 1999; Strausberg *et al.* 2000), as well as Affymetrix microarray samples for yeast cells (Jelinsky and Samson 1999; Jelinsky *et al.* 2000), exhibited similar monotonically skewed shapes with a greater abundance of rarer transcripts and more gaps among the higher-occurrence expression levels (Figures 1 and 2).

Several classes of skewed probability functions [Poisson, exponential, logarithmic series, simple power law, Pareto-like, and mixture of log-series and exponential (Johnson *et al.* 1993)] were fit (see methods) to empirical gene expression level histograms for >50 human, mouse, and yeast SAGE libraries; 30 human cDNA libraries in Cancer Genome Anatomy Project (CGAP) databases (http://www.ncbinlm.nih.gov/CGAP; http://www.ncbinlm.nih.gov/SAGE); and 30 microarrays of normal and treated yeast cells (http://www.hsph.harvard/gene expression; Holstege *et al.* 1998).

The best fit by our criteria was obtained using the discrete Pareto-like probability function,

$$f(m) = z^{-1}/(m + b)^{k+1}, \quad (1)$$

where the $f(m)$ is the probability that a randomly chosen distinct tag (representing a gene) occurs m times in the library. The function f involves two unknown parameters, k and b, where $k > 0$ and $b > -1$; the normalization factor z is the generalized Riemann zeta-function value, $z = \sum_{j=1}^{J} 1/(j + b)^{k+1}$. We call Equation 1 the generalized discrete Pareto (GDP) model. Note that J, the maximum observed expression level, is a sample-size-dependent quantity $J = J(M)$. The parameter k reflects the skewness of the probability function; the parameter b characterizes the deviation of the GDP distribution from a simple power law (with $b = 0$; see, for example, dotted

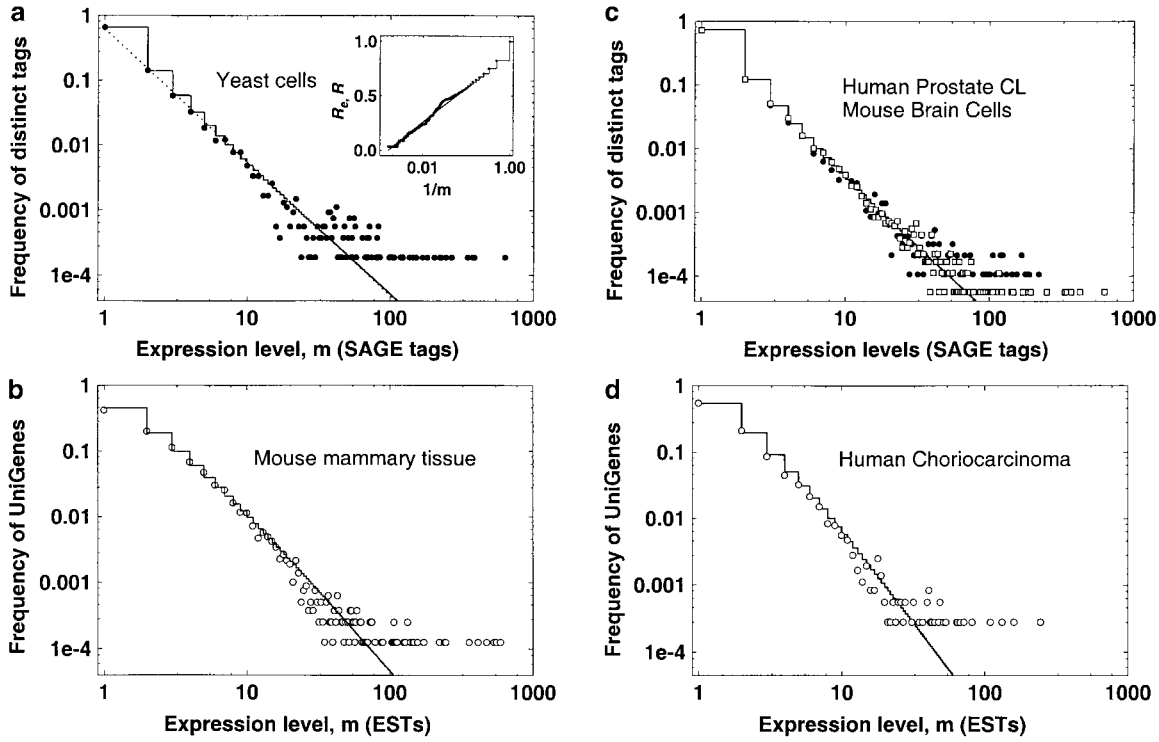

FIGURE 1.—Empirical relative frequency distributions of the gene expression levels. Log-log plots are shown. (a) ●, log-phase yeast cell growth library with 20,096 SAGE tags; solid step-function line, GDP model with $k = 0.974 \pm 0.004$, $b = -0.173 \pm 0.004$; dotted line, simple power law with $k = 1.03 \pm 0.005$. Inset plot: ●, empirical cumulative fraction function R_e values $[R_e = (\sum_{j=1}^{m} j \cdot \bar{n}(j, M))/(N \cdot M)]$for the ● histogram (main plot); the solid line is the corresponding theoretical model R $[R(m) = (\sum_{j=1}^{m} j \cdot f(j))/\sum_{j=1}^{J} j \cdot f(j)]$ computed by the GDP model (main plot). (Cumulative data reduce the apparent "noise" in the histogram data.) (b) ○, mouse mammary cells cDNA library 341 of size 36,675 ESTs; solid step-function line, GDP model with $k = 1.44 \pm 0.006$, $b = 1.34 \pm 0.002$. (c) ●, human colon cancer cells SAGE library 2892.2 of size 22,637 tags; solid step-function line, GDP model for ● data at $k = 1.08 \pm 0.030$, $b = -0.28 \pm 0.010$; □, mouse brain (primary meduloblastoma) cells SAGE library 3871 of size 43,274 tags. (d) ○, human choriocarcinoma cells cDNA library 2427 of size 10,087 ESTs; solid step-function line, GDP model with $k = 1.88 \pm 0.044$, $b = 1.34 \pm 0.005$.

line in Figure 1a). The inset plot in Figure 1a demonstrates that the fitted GDP model predicts the empirical cumulative fraction function $R_e(m)$ (for the definition of R_e see the Figure 1 legend); this demonstrates that our model fits well over the entire range of experimental values.

Figure 2 shows the frequency of the numbers of distinct open reading frames (ORFs)/genes *vs.* hybridization signal intensity values for Affymetrix microarray hybridization data obtained for normal yeast cell transcriptome (http://www.hsph.harvard/geneexpression; JELINSKY and SAMSON 1999). After subtraction of background noise, the total (digital) hybridization signal intensity, *s*, was normalized to the typical number of mRNA molecules per yeast cell (JELINSKY and SAMSON 1999). The window plot in Figure 2 is the empirical frequency distribution of the number of genes expressed at *s* copies per cell. It has a skewed form with a long right-side tail and with a left-side tail down to threshold levels of reliable detection of at least ~0.5 copies per cell. The cooccurrence of the low-expressed genes/ORFs in three or more of the six analyzed microarrays (http://www. hsph.harvard/geneexpression) allowed us to suggest that at least 45% of the known 6200 genes/ORFs are present but at <1 copy per cell. Table 1 clearly shows that the skewed GELPFs from such filtered Affymetrix data for yeast cells at different phases of cell life are all very similar and are all fitted by the GDP model down to 0.5 transcripts per cell.

Regardless of either the method used to generate the gene expression profile (SAGE, cDNA library sequencing, or oligonucleotide microarray hybridization) or the species studied, we have observed that the GDP functional form fits the observed GELPFs. The parameters of the fitted GDP, however, show a significant dependence on sample size, specific eukaryote species, and methods used to generate the library (Table 1).

Effect of library size on gene expression level distribution: Similarly sized libraries made using the same method from many different human tissues and cell lines have similar numbers of distinct gene tags and are characterized by empirical GELPFs with nearly equivalent parameters in their best-fit GDP models (see Table 1). As the size of a library increases, the shape of the empirical GELPF changes systematically: (1) p_1, the fraction of distinct tags represented by only one copy, be-

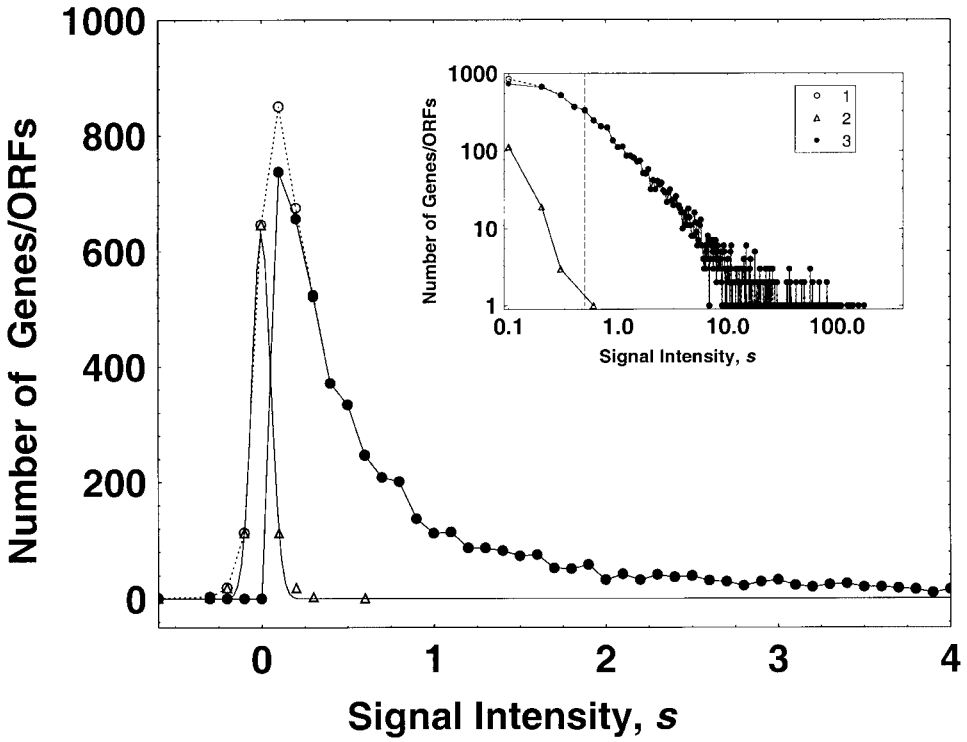

FIGURE 2.—The empirical frequency distribution of the hybridization signal intensity values for Affymetrix microarray hybridization data for normal yeast cell genes/ORFs (JELINSKY and SAMSON 1999). We checked that the background noise intensity signals (△) were distributed with an approximately normal distribution function with mean zero and standard deviation 0.054. Intensity signal distribution before (○) and after (●) subtraction of the background noise from the signal values is shown. The total hybridization signal intensity has been normalized (by JELINSKY and SAMSON 1999) with respect to the mean number of mRNA molecules per yeast cell. A point $(s, f(s))$ on the major plot shows the number of genes/ORFs, f, which have the signal intensity value, s. The vertical discontinuous line indicates the lower bound of the "reliably" detectable intensity signal values (the lower bound corresponds to 0.5 transcripts per cell). The window plot (log-log plot) shows the signal intensity distributions before and after subtraction of noise, shown over the full range of the positive signal intensity values.

comes smaller; (2) J increases in proportion to M; (3) the parameter b becomes larger; and (4) the parameter k increases and then slowly decreases (Figure 3; Table 1). Despite significant variation in human tissue types studied, the number of distinct tags, N, appears to be essentially *invariant* for the similar-size SAGE libraries (Figure 3b). Although the yeast genome is less complex, yeast SAGE libraries behave similarly (Table 1; Figure 4a). We also found that for yeast, mouse, and human SAGE and cDNA libraries, all values of the scaling parameter a ($a = J/M$, which represents the frequency of occurrence of the most common transcript within the library or cell population) fall within narrow ranges (Table 1). These observations suggest that all studied cell types have a common skewed underlying probability function form.

Importantly, in so-called "scale-free" (or self-similar, *i.e.*, any part of the system is statistically similar to the whole) biological and physical systems, described by a simple power law ($b = 0$; $j = \infty$; k, z are the positive constants), the parameter $b = 0$ and the parameter k is assumed to be independent of the size of the system (STANLEY *et al.* 1999; JEONG *et al.* 2000; GOMEZ *et al.* 2001). We did not observe such properties in the GELPFs; they display a nonlinear, rather than linear trend in log-log coordinates and a sample-size dependence. For example, Table 1 shows that the parameter b in the GDP model is significantly different from 0 for most data sets, and b becomes larger as the library size increases (see, for example, libraries 2892.1 and 2892.2 or libraries 154 and 154.1 in Table 1 and Figure 3a). In the case of SAGE libraries for different human cell types, the Pearson correlation coefficient between the library sizes and the values of parameter b equals 0.9. We observed that the values of parameter b approach 0 as M increases at relatively small sample sizes (for example, ~10,000 SAGE tags, Table 1).

Although the Pareto-like models appear to fit empirical GELPFs down to the least transcript abundance observed (~0.2–0.5 copies per cell in yeast microarray experiments, Figure 2), theoretically these models demonstrate an unlimited increase in the number of species (*i.e.*, different expressed genes) as the sample size approaches infinity. This contradicts the fact that there is a finite number of different mRNAs (different expressed gene products). Thus these models must be considered at best empirical approximations of an underlying probability. We have developed a construction model (see METHODS and APPENDIX) for the underlying probability distribution. When this distribution is finitely sampled, the results fit by Pareto-like GELPFs. The model explicitly exhibits the observed sample size dependence but retains a finite limit to the number of different classes as the sample size increases. Importantly, this model assumes that each *expressed* gene has a positive probability of being observed in any given sample and also that the expression level for this gene is statistically independent of the expression levels for other genes. The expression of those small groups of genes that are regulated by common sets of transcription factors would be

TABLE 1

Fitting of the empirical frequency distributions of the gene expression levels

Methods and libraries	Library size, M	No. of distinct tags, N	M/N	p_1	J	J/M	k	±SE	b	±SE	Ψ
SAGE (yeast)											
Log-phase	20,096	5,324	3.78	0.66	636	0.032	0.97	0.004	−0.173	0.004	9.3
S-phase	19,871	5,785	3.44	0.67	561	0.028	0.98	0.004	−0.197	0.004	9.8
G2/M-phase	19,527	5,303	3.68	0.67	519	0.027	0.96	0.006	−0.195	0.006	8.8
Pooled library	59,494	11,329	5.25	0.62	1,716	0.029	0.94	0.008	−0.108	0.008	7.7
Total true tags	47,393	5,819	8.14	0.46	1,716	0.036	0.97	0.001	0.494	0.001	7.7
SAGE (mouse)											
19018	43,274	17,754	2.43	0.72	630	0.015	1.19	0.001	−0.165	0.001	8.6
20427	61,240	24,796	2.46	0.73	425	0.007	1.14	0.001	−0.195	0.001	8.0
Human											
154	81,516	19,137	4.26	0.53	1,598	0.02	1.25	0.012	0.57	0.016	7.1
144	61,245	17,323	3.53	0.56	521	0.009	1.39	0.005	0.62	0.006	9.8
143	51,949	13,589	3.82	0.56	370	0.007	1.27	0.006	0.49	0.007	9.5
153	51,906	16,257	3.19	0.59	659	0.013	1.39	0.008	0.48	0.009	8.9
161	49,334	15,182	3.25	0.59	832	0.017	1.44	0.010	0.57	0.007	8.3
122	45,911	15,243	3.01	0.61	450	0.010	1.37	0.023	0.40	0.024	7.1
160	42,978	13,394	3.21	0.59	338	0.008	1.36	0.014	0.44	0.015	8.2
146	37,512	13,033	2.88	0.64	370	0.010	1.38	0.028	0.30	0.027	9.3
145	27,229	9,452	2.88	0.64	326	0.012	1.30	0.009	0.30	0.009	9.3
123	26,669	10,182	2.62	0.65	323	0.012	1.42	0.010	0.24	0.009	9.3
2892.2	22,637	9,348	2.42	0.74	221	0.010	1.08	0.030	−0.28	0.010	11
171	20,050	7,702	2.6	0.72	440	0.022	1.40	0.034	−0.027	0.025	8.3
167	16,361	5,900	2.77	0.69	561	0.034	1.27	0.063	0	0	9.1
166	14,616	5,383	2.72	0.7	462	0.032	1.28	0.010	0.015	0.015	10
172	8,936	4,507	1.98	0.76	210	0.024	1.36	0.026	−0.144	0.017	8.5
154.1	8,936	4,590	1.95	0.76	181	0.030	1.36	0.023	−0.123	0.013	9.2
2892.1	6,313	3,531	1.79	0.81	78	0.012	1.05	0.015	−0.480	0.008	10
1698	2,861	1,961	1.46	0.83	19	0.007	1.09	0.110	−0.500	0.058	8.3
cDNA (LifeTech)											
Mouse, Lib. 341	36,675	8,019	4.57	0.42	1,641	0.045	1.44	0.010	0.90	0.06	5
Mouse, Lib. 946	12,309	4,023	3.06	0.56	427	0.035	1.49	0.001	0.75	0.003	7.8
Human, Lib. 2427	10,087	3,586	2.81	0.54	246	0.029	1.88	0.040	1.34	0.05	7.1
Affymetrix arrays (yeast)											
Log-phase	16,762	3,000	5.59	0.46	144	0.009	0.86	0.001	0.37	0.003	7.4
G1-phase	17,408	2,862	6.08	0.45	178	0.010	0.85	0.001	0.36	0.004	6.7
S-phase	16,440	2,903	5.66	0.47	151	0.009	0.85	0.001	0.32	0.004	7.0
G2/M-phase	17,036	2,900	5.87	0.45	156	0.009	0.84	0.001	0.36	0.004	6.9

Characterization of the empirical frequency distributions of the gene expression levels for yeast, mouse, and human cell-type libraries and goodness-of-fit analysis using the generalized discrete Pareto (GDP) model. M is the number of tags (a size of the library); N is the number of distinct tags. p_1 is the fraction of distinct tags represented by one copy in the library. J is the maximum observed gene expression level in the library. k and b are the parameters of the GDP model. Ψ is the goodness-of-fit criterion (see METHODS). Ψ ranges between excellent (11–8), very good (8–6), and satisfactory (6–4). Yeast SAGE libraries: cells on G2/M-, S-, and log-phase stages of cell life; a pool of these three libraries; and a true tags library. Mouse SAGE libraries (Unilib IDs): 19018 (brain, meduloblastoma), 20427 (brain, normal, purified granular cell precursors). Human SAGE libraries (Unilib IDs): 154 (normal brain cells, >95% white matter), 144 [H1110, glioblastoma (GBM)], 143 [H392, GBM cell line (CL)], 153 (pooled GBMs), 161 (pooled normal brain), 122 (HCTT116, colon cancer CL), 160 (NHA, normal astrocyte CL), 146 (RKO, colon cancer CL), 145 (SW837, colon cancer CL), 123 (Caco2, colon cancer CL), 2892.2 (LNCaP, the prostate cancer CL library 2892 after 1 year), 171 (primary colon cancer), 167 (normal colon), 166 (normal colon), 172 (primary colon cancer), 154.1 (normal brain tissue, sublibrary taken from library 154), 2892.1 (LNCaP, initial library 2892), and 1698 (ovary carcinoma). cDNA libraries (Life Technologies method): mouse mammary cell library 341, mouse normal kidney library 496, and human choriocarcinoma cell library 2427. Affymetrix microarrays: normal yeast cells on log-, G1-, S-, and G2/M-phases of cell life (data from database http://www.hsph.harvard/geneexpression).

expected to show correlations within any given cell. However, the average correlation between expression events for a given gene and all other (thousands of) expressed genes would likely be statistically insignificant. Furthermore, since expression profiles are obtained from cells at one instant, specific transcription

events driven by the same transcription factors would have weaker correlations due to temporal and spatial fluctuations (*e.g.*, chromatin dynamics) within a given cell and certainly when averaged over a large population of cells. This assumption of essentially statistical independence among all transcription events in a population of cells is consistent with experimental observations (Chelly *et al.* 1989; Ko 1992; Ross *et al.* 1994; Newlands *et al.* 1998; Fiering *et al.* 2000; Sano *et al.* 2001). Even for synchronized yeast cells arrested in G1-, S-, and G2/M-phases of cell life, the empirical GELPFs in microarray experiments were very similar to each other and to the GELPF observed for yeast cells in log-phase of cell growth (see Table 1). These results suggest that the shape of the empirical GELPFs is relatively robust to different correlations between expressed genes at least in normal yeast cells at different phases of cell life. In addition, our analyses of the seven microarray (Jelinsky and Samson 1999; Jelinsky *et al.* 2000) data sets for normal yeast cell samples and pooled three SAGE libraries (Velculescu *et al.* 1997) of normal yeast cell samples show that at least 2000 yeast genes/ORFs are expressed, but at <0.5 copies per cell on average (see, for example, Figures 2 and 4b). At the level of the individual cell, these rare transcription events can be treated as stochastic events. A mathematical description of our model of the GELPF is presented in methods and the appendix.

Analysis of the empirical GELPF using SAGE databases: Using the LG model (see Equations 3 and 4 and methods for a definition of the LG model) to fit empirical population growth curves like those presented in Figure 3b, we can predict the frequencies of the gene expression levels p_1, p_2, . . . , for a given cDNA or SAGE library size (Figure 3c); p_i is the probability that a random gene has i transcripts. Application of the LG model to human SAGE databases results in extremely large estimates (138,000 distinct tags expressed in brain and 127,000 expressed in a "typical" human tissue) compared to the total number of genes in the genome (30,000–40,000 genes; International Human Genome Sequencing Consortium 2001; Venter *et al.* 2001; Kuznetsov 2002). This demonstrates the well-known discrepancy between the numbers of different expressed sequences in SAGE or Unigene libraries and the number of human genes. This large discrepancy can be attributed to a variety of sources including sequencing errors, multiple restriction sites on the same transcripts leading to multiple tags per gene, and alternative splicing (Lal *et al.* 1999; Velculescu *et al.* 1999;

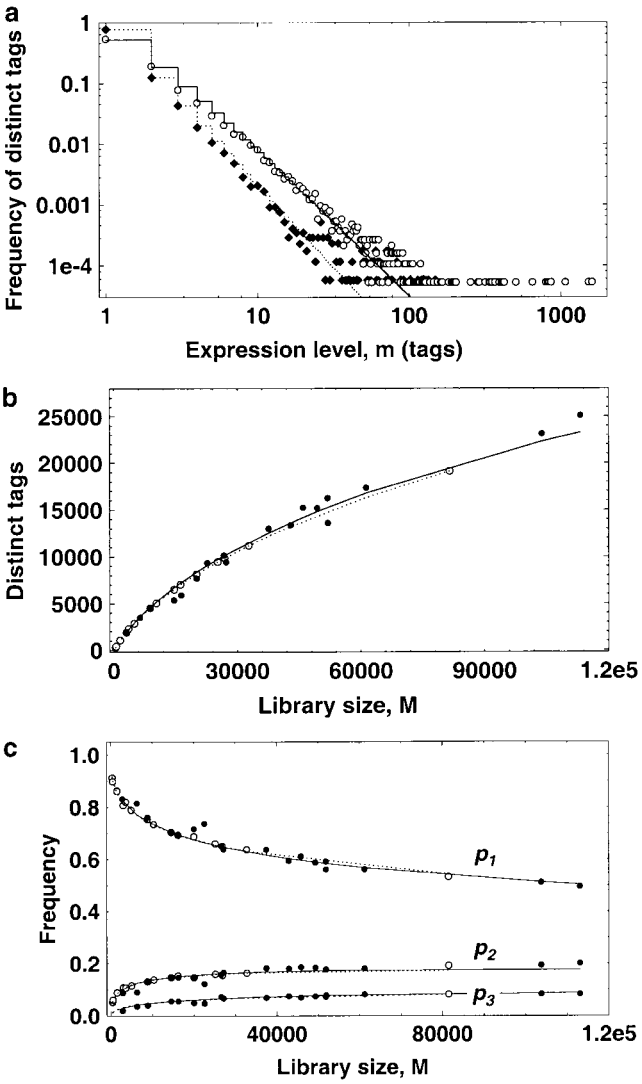

Figure 3.—Effects of library size on the empirical GELPF, on the number of distinct tags, and on the frequencies of low-abundance transcripts. (a) Log-log plot. ○, the empirical GELPF for human white matter brain tissue library 154 of size 81,516 tags; solid step-function line, the GDP model for ○ data; ◆, average frequency of expression levels for 10 sublibraries of size 6313 tags taken at random without replacement from library 154 and represented in an average by 3497 distinct tags; dotted step-function line, GDP model (at $k = 1.62 \pm 0.07$, $b = 0.01 \pm 0.004$) for ◆ data. (b) The number of distinct tags in SAGE libraries. ●, a library presented in Table 1; solid line, logarithmic growth (LG) model (Equation 3 in methods) with $d = 112{,}786 \pm 4343$, $c = 0.41 \pm 0.007$ for the ● data set. The two highest ● points present a pool of libraries 143 and 153 and a pool of libraries 144 and 153, respectively. ○, the number of distinct tags in library 154 and in sublibraries sampled without replacement from library 154; dotted line, LG model with $d = 119{,}627 \pm 1072$, $c = 0.39 \pm 0.001$) for the ○ data set. (c) Prediction of frequencies of distinct tags occurred one, two, and three times in human tissue libraries. ○, a sublibrary of library 154 formed by choosing tags randomly without replacement from brain tissue SAGE library 154; dotted lines link values for functions $p_1(M)$, $p_2(M)$, and $p_3(M)$ predicted by the BD model for corresponding ○ data sets. ●, human cell library; solid lines link values for functions $p_1(M)$, $p_2(M)$, and $p_3(M)$ predicted by the BD model for corresponding ● data sets.

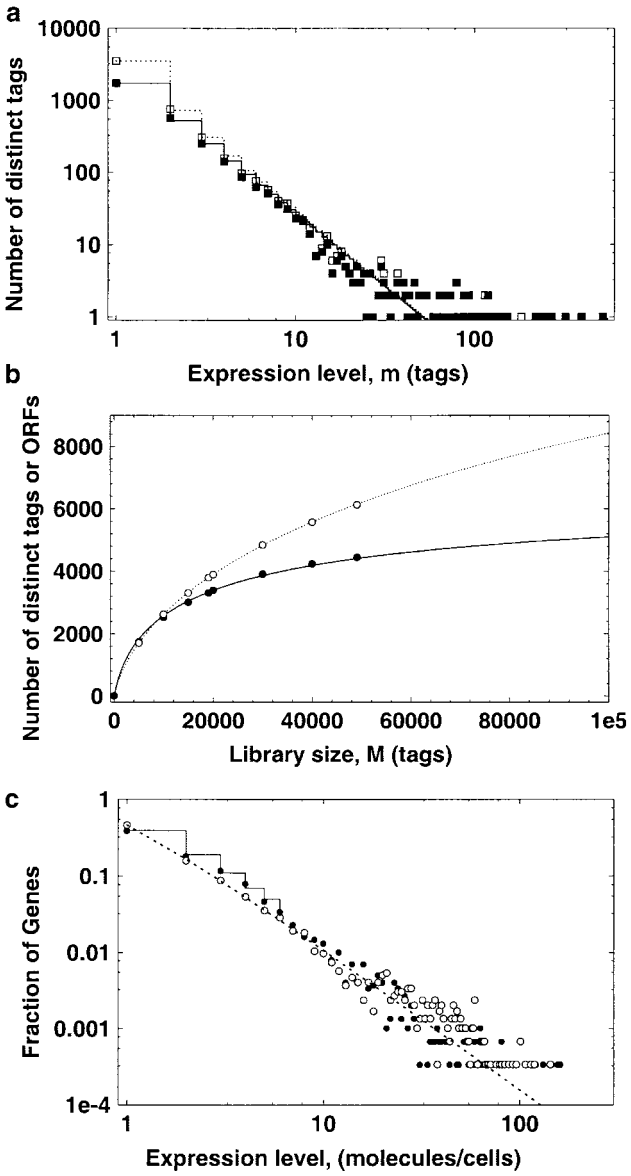

FIGURE 4.—Correction of the empirical histogram for yeast cell SAGE library, the population growth curves, and the GELPF for a single yeast cell. (a) Log-log plot. □, number of distinct tags of 5303 distinct tags represented by 19,527 tags in a G2/M phase-arrested cell library; dashed step-function line, the GDP model with $b = -0.195 \pm 0.005$, $k = 0.96 \pm 0.006$ for □ data; ■, number of true tags of the same library after removing erroneous tags; solid step-function line, GDP model with $b = 0.207 \pm 0.013$, $k = 0.991 \pm 0.011$ for ■ data. (b) Population growth curves. ○, number of true distinct tags of sublibraries from pooled yeast library of 47,393 true tags; dashed line, LG model with $d = 20{,}000 \pm 1946$, $c = 0.356 \pm 0.02$ for ○ data; ●, number of genes/ORFs observed in these sublibraries; LG model with $d = 6575 \pm 185$, $c = 0.579 \pm 0.01$ for ● data. (c) Log-log plot. Solid step-function line, the fraction of genes/ORFs estimated by the BD model for a single yeast cell *vs.* expression level; ●, relative frequency *vs.* expression level generated from the fitted GDP model (with $k = 1.56$, $b = 2.17$) after 3000 random generations of occurrence number value m (see *Simulation of theoretical histograms* in METHODS); ○, relative frequency of 3000 genes/ORFs *vs.* expression level in a single log-phase yeast cell, estimated from Affymetrix microarray hybridization data (http://www.hsph.harvard/geneexpression). This histogram was constructed as follows. For each ORF/gene, the scaled hybridization intensity signal value, I, in the yeast GeneChip database (http://www.hsph.harvard/geneexpression, JELINSKY *et al.* 2000), was converted to a transcript count per cell using the empirical formula $m = (I - 20)/165$, rounded to the nearest integer ($[0.5, 1.5), [1.5, 2.5), \ldots$). The correction factor 20 is the average background intensity signal; the scaling factor 165 was estimated by comparing the hybridization signal intensity values with expression levels of 18 genes, whose absolute mRNA levels were reliably determined by quantitative hybridization experiments (IYER and STUHL 1996) and observed in several normalized microarray hybridization data sets (HOLSTEGE *et al.* 1998; JELINSKY and SAMSON 1999). These genes are RPS4BA, GCN4, SPT15, RAS2, RPO21, FAR1, DED1, HIS3, CDC28, TRP3, GAL11, HAT1, POL12, NUP157, PRP4, PRP3, PEX14, and RAD52. Dashed line, the fitted GDP model graph with $k = 0.86 \pm 0.01$, $b = 0.37 \pm 0.003$ for ○ data. In this analysis, we excluded the genes/ORFs whose expression levels (after subtraction of the noise and normalization of the signal intensity to transcript count per cell) were <0.5 copy per cell (see Figure 2).

CHEN *et al.* 2000; STOLLBERG *et al.* 2000; CARON *et al.* 2001).

Analysis of errors in SAGE cell libraries and prediction of the number of expressed genes in a population of cells: Without removing experimental errors in SAGE libraries one cannot obtain an accurate estimate of the number of expressed genes, N_t, and the GELPF. Using our probabilistic model (Equations 2–5, METHODS), we developed a computational methodology to estimate the true GELPF for a SAGE library, even when the SAGE library is incomplete (contains only a fraction of all expressed genes for the sample cell type). Our methodology is as follows.

1. We selected only tags whose location on the yeast chromosome map coincided with protein-coding gene or ORF regions (called here "true tags").
2. We constructed population growth curves for the numbers of different genes/ORFs found in the tag location database.
3. We fitted the growth curve for the numbers of distinct genes/ORFs by Equations 3 and 4.
4. We calculated N_t, using Equation 5, and, finally, calculated the true underlining GELPF, using Equation 2.

To validate this approach, we analyzed 11,329 yeast cell distinct SAGE tags representing 59,494 SAGE tags of the three yeast cell SAGE libraries (VELCULESCU *et al.* 1997, Table 1). Since almost all yeast protein-coding genes/ORFs and their location on chromosomes are known, we can obtain the true distinct tags and their expression levels in a yeast SAGE library by eliminating erroneous tags that fail to match known 3′ *NLa*III genes/ORFs regions and adjacent 3′ end regions presented in the chromosome tag location database (http: genome-http://www.stanford.edu/Saccharomyces). This database was generated by VELCULESCU *et al.* (1997) and currently

contains 8480 distinct tags, matching 4735 of ~6200 known yeast genes/ORFs. We found that 25% (2849 distinct tags) of the 11,329 analyzed distinct tags failed to match the yeast genome and these tags were associated with sequencing errors, 23.5% (2661 distinct tags) failed to match ORFs and adjacent 3′ end regions, and 51.4% (5819 distinct tags) were classified as the true distinct tags. Also, we found that 1689 distinct tag sequences of the 2661 distinct failed tags matched antisense sequences within 1504 genes/ORFs.

Figure 4a shows empirical GELPF for distinct SAGE tags in the G2/M phase-arrested yeast cell library contains 19,527 tags with 5303 distinct tags. By filtering these tags by matching in the tag location database, we discovered 3239 erroneous tags (16.6% of the 19,527 tags in the library) corresponded to 2103 distinct tags. Most of the 2103 distinct erroneous sequences found in the set of 3239 erroneous tags occur only one or two times. The remaining 16,288 tags corresponded to 3200 distinct tags with matched 2936 genes/ORFs in the tag location database.

By sampling randomly from a pooled library containing all the observed true tags from the three SAGE libraries, we constructed population growth curves for both the number of distinct tags chosen and the corresponding number of different genes/ORFs found in the tag location database (Figure 4b). Sample size-dependent LG model (Equations 3 and 4) fits detected numbers of both distinct true tags and different genes/ORFs. In the case of distinct true tags (○, Figure 4b), our estimator (Equation 5) once again predicts a very large value of 25,103 ± 2000 distinct true tags compared to the total number of known yeast genes. For genes/ORFs (●, Figure 4b), a more reasonable estimate of the total number of expressed genes, $N_t = 7025 \pm 200$, was obtained. After minor corrections (see *Accuracy of an estimate of the number of expressed genes for yeast cells* in METHODS), this estimate is consistent with CANTOR and SMITH (1999) and JOHNSON (2000) estimates. Thus, we can suggest that all or almost all yeast genes are expressed in a growing normal yeast cell population, *i.e.*, $N_t \approx G$, where G is the total number of genes in the entire yeast genome.

Using the estimated parameters $c = 0.579$ and $d = 6580$ in the LG function (Equations 3 and 4, METHODS) and an estimate $M_{cell} = 15{,}000$ of the number of mRNAs per yeast cell (VELCULESCU *et al.* 1997), Equation 5 (METHODS) predicts 3009 genes/ORFs/cell. This estimate is consistent with our estimate for a single yeast cell in the G2/M phase-arrested state (2936 genes/ORFs by SAGE data) and with our estimates for yeast cells by Affymetrix microarray data (Table 1).

The GELPF in a single yeast cell: The GELPF for a single yeast cell was estimated for corrected data (Figure 4a), using both the BD and GDP models (see METHODS); the results are presented in Figure 4b. To validate our mathematical models used to analyze SAGE data, we also determined the GELPF on the basis of Affymetrix microarray data sets (JELINSKY and SAMSON 1999; JELINSKY *et al.* 2000). Figure 4c shows a histogram constructed for the microarray hybridization experiment (JELINSKY *et al.* 2000) for normal yeast cells. We converted the hybridization intensity signal values to gene expression values with 0.5 transcripts per cell chosen as a reasonable low-limit cutoff point (see also Figure 2). In this case, 3000 more highly expressed genes/ORFs representing ~16,000 transcripts per cell were found. Figure 4c shows that the GELPF for the Affymetrix microarray data follows the GDP model ($k = 0.86 \pm 0.001$, $b = 0.37 \pm 0.003$) and is consistent with the GELPF for corrected SAGE data. Similar skewed frequency distributions were also observed (see examples in Table 1) in 30 other microarray experiments using normal and stressed yeast cells (HOLSTEGE *et al.* 1998; JELINSKY and SAMSON 1999; JELINSKY *et al.* 2000).

METHODS

Binomial differential distribution and an estimator of the total number of expressed genes: Let M denote the total number of transcripts in a given "error-free" library and let N denote the number of distinct gene tags (or tag/signals converting to genes) for that library. Let p_m denote the probability that a randomly chosen gene is represented by m associated transcripts in the library for $m = 1, 2, \ldots$. On the basis of a multinomial distribution model for sampled transcripts, when M is large enough, we obtain the discrete probability function p_m, in terms of M and N, as

$$p_m = (-1)^{m+1} \frac{1}{N} \frac{M!}{m!(M-m)!} \frac{d^m N}{dM^m} \tag{2}$$

(see the APPENDIX), where $m = 1, 2, \ldots$. Note that N is treated as a function of M, so p_m is a function of M. We call this function the binomial differential (BD) function. Taking $m = 1$ in Equation 2, we obtain the differential equation

$$dN/dM = p_1 N/M, \tag{3}$$

with $N(1) = 1$. Equation 3 defines the "logarithmic growth" (LG) function $N(M)$. p_1 is a decreasing function of M (see Figure 3c). We use the empirical approximation (KUZNETSOV 2001)

$$p_1 = \frac{1 + 1/d^c}{1 + (M/d)^c}, \tag{4}$$

where c and d are positive constants. Using an explicit specification of p_1 allows us to fit the BD and LG models to empirical histograms. With p_1 defined by Equation 4, Equation 3 has an exact solution for $N(M)$ in the limit as $M \to \infty$:

$$N(\infty) = N_t = (1 + d^c)^{(1+1/d^c)/c}. \tag{5}$$

N_t is an estimator of the number of expressed genes in a large population of cells. Using Equation 2 with fitted values of the parameters d and c provides a mean of computing p_1, p_2, . . . at a given library size M.

Estimation of the GELPF for a single cell: First, we use the BD model (Equation 3) with fitted parameters c and d in $p_1(M)$ to compute the probability values p_1, p_2, . . . , p_6 for 3009 yeast genes/ORFs corresponding to the library size 15,000 transcripts. Because the GDP model is a good approximation of BD distribution at fixed M (see Figure 3, Figure 4c, and METHODS), it is acceptable to use the GDP model to estimate p_m for larger m. (This use of the GDP model was necessary because there are no readily available numerical algorithms that do not accurately compute values of high-order derivatives.) We fit the GDP model (Equation 1) to the six points predicted by the BD probability distribution at constraints $M \approx 15{,}000$, $J \approx 0.028*M$ and extrapolate the fitted GDP model to estimate values of p_m for $m > 6$ (solid step line in Figure 4c). To check the self-consistency of our predictions, we estimated the total number of transcripts, M, from the fitted GDP model and noted that the result was 15,000.

Accuracy of the estimated number of expressed genes for yeast cells: Our estimate, $N_t = 7024$ genes/ORFs, is ~4–10% higher than current estimates of the total number of distinct ORFs in the yeast genome (6200–6760 genes/ORFs; CANTOR and SMITH 1999; JOHNSON 2000). This relatively small difference could be due to the existence of erroneous and redundant tags that nevertheless match genes/ORFs and their adjacent genomic regions. Our analysis does not take into account nonannotated ORFs and overlapping ORFs that match the same tag. Additionally, ~1–3% of transcripts would be expected to lack an *Nla*III site and would therefore be missing in the database.

In the case of yeast, ~5% of the genes show alternative splicing. Furthermore, splice variants might have the same primary tag, alternative tags, or become SAGE silent, depending on the restriction sites remaining. We summed all SAGE tags that matched ORFs to obtain the GELPF, so the only effect would be to miss the small number of splice variants lacking a *NLa*III restriction site. We do not know the frequency distribution of alternative splicing transcripts in yeast and human cells. With respect to our model of GELPF, we might assume that splice variants for yeast cells will have a skewed form of the probability distribution and not have a significant effect on our estimate of the true GELPF.

Simulation of theoretical histograms: Given the number of expressed genes, N, and the best-fit parameters k and b of the GDP distribution, we sample the values of m at random on the basis of the function $f(m)$ (Equation 1) N times (once for each gene). Then we count the occurrence numbers of generated values m in the intervals (0–1], (1, 2], . . . and construct the simulated gene expression level frequency histogram for a given value N. Note the corresponding value M is randomly determined by our sampling (see Figure 4c).

Goodness-of-fit analysis methods, numerical calculations, and software: Parameters in models were estimated MLAB mathematical modeling software (Civilized Software, Silver Spring, MD, www.civilized.com). For goodness-of-fit analysis, we used the modified Akaike information criterion [or model selection criterion (MSC)],

$$\Psi = \log\left(\sum_{m=1}^{J} (g(m) - E(g))^2 / \sum_{m=1}^{J} (g(m) - f(m))^2\right) - 2v/J,$$

where $m = 1, 2, \ldots, J$. In our case m is the expression level value and J is the maximum observed gene expression level in the library; g is the empirical relative frequency distribution, f is the theoretical probability distribution function with v unknown parameters, and $E(g)$ is the mean value of observed data. Note, the Ψ is independent of the scaling of data points. Ψ ranges between excellent (11–8), very good (8–6), satisfactory (6–4), and poor (4–1).

We also used the cumulative fraction function R (see Figure 1), as well as several regular goodness-of-fit criteria (sum of squares for deviations, the Wilcoxon two-sample rank-order test).

By our goodness-of-fit criteria, the GDP model is superior to simple power law, as well as many other skew probability functions (Poisson, log-series, and exponential) and mixed logarithmic series + exponential distribution. For example, for library sizes >40,000 SAGE tags, the values of the Ψ ranged between 3 and 6 (satisfactory or poor); however, Ψ values ranged in (11–7) for the GDP model. Similar superiority of the GDP model (measured by the R and Ψ criteria) was observed after goodness-of-fit analysis of the distribution models to microarray data.

Symbolic differentiation and subsampling were performed using MLAB. Monte Carlo experiments were performed using MLAB and programs written in Fortran-90. Data-mining tools of the Cancer Research Anatomy Project including X profiling and SAGE/map (LAL *et al.* 1999; http://www.ncbi.nlm.nih.gov/SAGE) were also used.

DISCUSSION

Even with their large differences in genome organization yeast, mouse, and human cells all demonstrate similar skewed long-tail Pareto-like gene-expression level distributions. The observed distributions have the following characteristics in common: *There are few redundant and many rare transcripts.* The universality of the empirical GELPF form for different eukaryotic cells suggests a common underlying *probabilistic* mechanism associated with the gene expression process conserved in eukaryote evolution. Similar distributions have been observed for the connectivity numbers of metabolic net-

works (Jeong *et al.* 2000), for the rates of protein synthesis of prokaryotic organisms (Ramsden and Vohradsky 1998), in different DNA-related phenomena (see Li 1999; Stanley *et al.* 1999; Gomez *et al.* 2001 for references), and in many models of the self-organized systems (http://linkage.rockfeller.edu/wli/zipf). All such systems exhibit a strong stochastic component.

Both oligonucleotide microarray hybridization and construction of SAGE libraries allow large-scale characterization of gene expression profiles including low-abundance transcripts. However, in both technologies, the determination of expression levels at one or fewer transcripts per cell is limited by issues of limited sensitivity and erroneous measurements. These limitations become more severe with increasing size of the transcriptome. We developed a comprehensive statistical approach to analyzing the empirical distribution of expressed genes for large transcriptomes obtained by the SAGE method by first removing sequence errors using chromosome location maps for SAGE tags and then applying statistical modeling and the BD model (Equations 2–5) to filtered data (Figure 4). Our resulting SAGE data were similar to the GELPF data that we obtained for normalized oligonucletide microarray hybridization data sets for yeast cells. Our methodology of construction of the correct underlying probability distribution could be used to analyze large SAGE transcriptomes for different mammalian cells and cell types, including human transcriptomes, and for evaluation of the new modifications of the SAGE method using, for example, 21-mer tags.

Our statistical modeling approach provides a justifiable way to compare the GELPFs using samples (cDNA or SAGE libraries) with different sizes. This approach also can be used to permit the use of exact statistical tests for different transcriptomes (*i.e.*, obtained for normal and cancerous cell tissues). Our novel numerical estimator of the number of species (*i.e.*, expressed genes; Equation 5) can be used to estimate the number of expressed genes in a single cell and in a population of the cells by SAGE or cDNA data sets, even if data are incomplete and exhibit severe experimental errors.

On the basis of our analysis of the GELPF in the normalized yeast microarray databases (Jelinsky and Samson 1999; Jelinsky *et al.* 2000) 1330 ± 45 genes/ORFs are represented on average by a single mRNA molecule per cell, at least 2000 genes/ORFs are expressed at 0.1–0.5 molecules per cell on average, and $\sim$47% (2917) of all yeast genes are expressed at less than one transcript per cell. Our estimate for yeast cell SAGE data is consistent with these estimates. Approximately 1200 genes/ORFs are represented on average by a single mRNA molecule per yeast cell (Figure 4b). The population growth curve for genes/ORFs (Figure 4b) predicts that 3800 additional genes/ORFs ($\sim$55% of all yeast genes) are expressed at less than one transcript per cell. Even in proliferating mammalian cells, the majority of genes are thought to be transcribed over a short period of time (4–7 min per transcript) and infrequently, less than once per hour (Jackson *et al.* 2000). On the basis of a similar methodological approach for estimating the GELPFs for a SAGE transcriptome, which we used in this article for yeast SAGE data sets, the analysis of a large ($\sim$600,000 SAGE tags) human transcriptome data set (Velculescu *et al.* 1999) indicated that $\sim$70% of all protein-coding human genes are expressed with less than one transcript per cell on average (Kuznetsov 2002). Such low numbers of transcripts in a cell population may be due to the action of a random transcription process in individual cells (McAdams and Arkin 1999; Fiering *et al.* 2000; Hume 2000; Sutherland *et al.* 2000).

Initiation of transcription has been observed to occur sporadically and randomly both in time and location on chromosomes in a variety of cell systems (Ko 1992; Ross *et al.* 1994; Newlands *et al.* 1998; Sano *et al.* 2001). In this study, we present additional data and arguments supporting our hypothesis that at the level of the individual cell the transcription events for a given gene at an instant appear to be statistically independent of expression levels for thousands of other genes. The existence of such a random transcription process would imply that all or almost all protein-coding genes in a genome should have a small but positive probability to be transcribed in any given cell during any fixed time interval. This suggestion is consistent with the observation that small transcript copy numbers occur even for various tissue-specific genes in human cells of different type, such as fibroblasts, lymphocytes, etc. (Chelly *et al.* 1989). Although not all cells of a population would have a copy of a specific transcript at a given moment, we would expect to see all these genes expressed, at least at a low level, in a sufficiently large cell population at any point in time. That is, *ergodicity* holds. This point is supported by the yeast expression data in the microarray database (Jelinsky *et al.* 2000): We observed that only 250 ORFs ($\sim$150 of them are "questionable" or "hypothetical" ORFs) of $\sim$6200 genes/ORFs were not expressed in any of six presented microarray samples from normal growing yeast cells.

In mammalian cells, a significant fraction of genes are silenced (transcripts are not observed); the silent state of a gene can be inherited, but later reactivated involving the stochastic, all-or-none mechanism at the level of a single cell (Sutherland *et al.* 2000). Low-probability transcription events for many genes in a cell could be regulated by its own specific transcripts. Sporadic initiation of the transcription process for rarely expressed genes could also be under dynamical control of some non-protein-coding genes associated with stress-response control (Eddy 2001). Many RNAs transcribed from these genes represent anti-sense RNA transcripts that overlap protein-coding genes on the other genomic strand. We might suspect that in response to environ-

mental changes, various stress conditions, and local fluctuation of molecular composition, the initiation of transcription events for many rarely expressed genes in each cell could be under dynamical control of these noncoding genes. Such autoregulation might tend to keep the low-expressed gene "one-half on," thus sporadically providing the mechanism of low expression for many genes in a cell population.

Physically, random "basal" transcription of genes might reflect nonlinear responses of the independent "gene transcription complexes" to internal or external fluctuations including thermal molecular motion. Noise in nonlinear dynamical systems can play a constructive role: It can, for example, improve a system's sensitivity to weak signals (Wiesenfeld and Jaramillo 1998). If it is so, noise in the gene-expression machinery could enhance weak transcription signals. This amplification mechanism for gene expression in eukaryotic cell types might provide a basic level of phenotypic diversity within a cell population and thus could facilitate adaptation of a population of cells. The stochastic variability in the case of rarely transcribed genes could also lead to changes in the genotype in which lineage commitment results from a selective rather than an instructive mechanism in cells (Ko 1992; Hume 2000; Sutherland *et al.* 2000; Ohlsson *et al.* 2001; Sano *et al.* 2001). Random initiation of gene transcription in a given cell would allow the "essential" rarely transcribed gene to provide the random switch between active and inactive states during the formation of daughter cells and thus to provide genotype diversity in the cell population. Because the distribution of the lifetimes of "switch-on" and "switch-off" states for genes in a single cell has a long right tail (McAdams and Arkin 1999), one of the two alleles of the same locus for a given low-expressed gene might be present in the same state for a long period of time. In this case, a natural clonal selection process in a population of the cells could select the clone(s) with a monoallelic gene expression, *i.e.*, a phenomenon in which only a single copy, or allele, of a given gene of diploid organisms appears (Ohlsson *et al.* 2001). Further statistical analysis and modeling of the large-scale gene expression data could help us to understand how the stochastic variability of gene expression in a single cell might lead to changes of the genotype repertoire in developing tissues and in an entire organism.

In this study, we have analyzed only "snapshot" gene expression profiles, which are based on an "averaging" of the many thousands of cells in various states and which by themselves do not provide us direct information about regulatory relationships between expressed genes. Taking multiple times the samples of transcripts from homogenous cell populations or from single cells of this population, we could study the correlation between expression levels of genes. In this case, it would be interesting to construct a distribution model that takes into account the correlations between the levels of different transcripts. For example, the multivariate multinomial distribution model can be constructed (see Johnson *et al.* 1997). Such a distribution model would be suitable, in particular, for analysis of gene-regulating network data representing gene expression profiles for individual cells sampled with replicates from an "isogenic" cell population at several periods of observation. Identification of such a multivariate multinomial distribution model may help to estimate the biologically significant correlations between genes and evaluate the stochastic component in dynamics of gene expression clusters at low expression levels.

The authors thank V. Velculescu for providing the supplementary information on the yeast cell SAGE database. We thank J. Berzofsky, I. Belyakov, K. Chumakov, R. Nossal, R. Strausberg, A. Strunnikov, T. Tatusov, and two reviewers for critical comments on the manuscript.

LITERATURE CITED

Bishop, J. O., J. G. Morton, M. Rosbash and M. Richardson, 1974 Three classes in Hela cell messenger RNA. Nature **250:** 199–204.

Cantor, C. R., and C. L. Smith, 1999 *Genomics*. John Wiley & Sons, New York.

Caron, H., B. van Schaik, M. van der Mee, F. Baas, G. Riggins *et al.*, 2001 The human transcriptome map: clustering of highly expressed genes in chromosomal domains. Science **291:** 1289–1292.

Chelly, J., J.-P. Concordet, J.-C. Kaplan and A. Kahn, 1989 Illegitimate transcription: transcription of any gene in cell type. Proc. Natl. Acad. Sci. USA **86:** 2617–2621.

Chen, J.-J., J. D. Rowley and S. M. Wang, 2000 Generation of longer cDNA fragments from serial analysis of gene expression tags for gene identification. Proc. Natl. Acad. Sci. USA **97:** 349–353.

Eddy, S. R., 2001 Non-coding RNA genes and the modern RNA world. Nat. Rev. Genet. **2:** 919–928.

Fiering, S., E. Whitelaw, D. I. K. Martin, 2000 To be or not to be active: the stochastic nature of enhancer action. BioEssays **22:** 381–387.

Gomez, S. M., S.-H. Lo and A. Rzhetsky, 2001 Probabilistic prediction of unknown metabolic and signal-transduction networks. Genetics **159:** 1291–1298.

Holstege, F. C. P., E. G. Jennings, J. J. Wyrick, T. I. Lee, C. J. Hengartner *et al.*, 1998 Dissecting the regulatory circuitry of a eukaryotic genome. Cell **95:** 717–728.

Huang, S.-P., and B. S. Weir, 2001 Estimating the total number of alleles using a sample coverage method. Genetics **159:** 1365–1373.

Hume, D. A., 2000 Probability in transcriptional regulation and implications for leukocyte differentiation and inducible gene expression. Blood **96:** 2323–2328.

International Human Genome Sequencing Consortium, 2001 Initial sequencing and analysis of the human genome. Nature **409:** 860–921.

Iyer, V., and K. Struhl, 1996 Absolute mRNA levels and transcriptional initiation rates in *Saccharomyces cerevisiae*. Proc. Natl. Acad. Sci. USA **93:** 5208–5212.

Jackson, D. A., A. Pombo and F. Iborra, 2000 The balance sheet for transcription: an analysis of nuclear RNA metabolism in mammalian cells. FASEB J. **14:** 242–254.

Jelinsky, S. A., and L. D. Samson, 1999 Global response of *Saccharomyces cerevisiae* to alkylating agent. Proc. Natl. Acad. Sci. USA **96:** 1486–1491.

Jelinsky, S. A., P. Estep, G. M. Church and L. D. Samson, 2000 Regulatory networks revealed by transcriptional profiling of damaged *Saccharomyces cerevisiae* cells: Rpn4 links base excision repair with proteasomes. Mol. Cell. Biol. **20:** 8157–8167.

Jeong, H., B. Tombor, R. Albert, Z. N. Ottvai and A.-L. Barabasi, 2000 The large-scale organization of metabolic networks. Nature **407:** 651–654.

Johnson, M., 2000 The yeast genome: on the road to the gold age. Curr. Opin. Genet. Dev. **10:** 617–623.
Johnson, N. L., S. Kotz and A. W. Kemp, 1993 *Univariate Discrete Distributions.* John Wiley & Sons, New York.
Johnson, N. L., S. Kotz and N. Balakrishnan, 1997 *Discrete Multivariate Distributions.* John Wiley & Sons, New York.
Ko, M. S. H., 1992 Induction mechanism of a single gene molecule: stochastic or deterministic. BioEssays **14:** 341–346.
Kuznetsov, V. A., 2001 Distribution associated with stochastic processes of gene expression in a single eukaryotic cell. EURASIP J. Appl. Signal Proc. **4:** 285–296.
Kuznetsov, V. A., 2002 Statistics of the numbers of transcripts and protein sequences encoded in the genome, pp. 125–171 in *Computational and Statistical Approaches to Genomics*, edited by W. Zhang and I. Shmulevich. Kluwer, Boston.
Lal, A., A. E. Lash, S. F. Altschul, V. Velculescu, L. Zhang *et al.*, 1999 A public database for gene expression in human cancers. Cancer Res. **59:** 5403–5407.
Li, W., 1999 Statistical properties of open reading frames in complete genome sequences. Comput. Chem. **23:** 283–301.
McAdams, H. H., and A. Arkin, 1999 It's a noisy business! Genetic regulation at the nanomolar scale. Trends Genet. **15:** 65–69.
Newlands, S., L. K. Levitt, C. S. Robinson, A. B. Karpf, V. R. Hodgson *et al.*, 1998 Transcription occurs in pulses in muscle fibers. Genes Dev. **12:** 2748–2758.
Ohlsson, R., A. Paldi and J. A. Marshall Graves, 2001 Did genomic imprinting and X chromosome inactivation arise from stochastic expression? Trends Genet. **17:** 136–141.
Ramsden, J. J., and J. Vohradsky, 1998 Zipf-like behavior in prokaryotic protein expression. Phys. Rev. E **58:** 7777–7780.
Ross, I. L., C. M. Browne and D. A. Hume, 1994 Transcription of individual genes in eukaryotic cells occurs randomly and infrequently. Immunol. Cell. Biol. **72:** 177–185.
Sano, Y., T. Shimada, H. Nakashima, R. H. Nicholson, J. F. Eliason *et al.*, 2001 Random monoallelic expression of three genes clustered within 60 kb of mouse t complex genomic DNA. Genome Res. **11:** 1833–1841.
Stanley, H. E., S. V. Buldyrev, A. L. Goldberger, S. Havlin, C. K. Peng *et al.*, 1999 Scaling features of noncoding DNA. Physica A **273:** 1–18.
Stollberg, J., J. Urschitz, Z. Urban and C. D. Boyd, 2000 A quantitative evaluation of SAGE. Genome Res. **10:** 1241–1248.
Strausberg, R., K. H. Buetow, M. R. Emmert-Buck and R. D. Klausner, 2000 The Cancer Genome Anatomy Project: building an annotated gene index. Trends Genet. **16:** 103–106.
Sutherland, H. G., M. Kearns, H. D. Morgan, A. P. Headley, C. Morris *et al.*, 2000 Reactivation of heritably silenced gene expression in mice. Mamm. Genome **11:** 347–355.
Velculescu, V. E., L. Zhang, B. Vogelstein and K. W. Kinzler, 1995 Serial analysis of gene expression. Science **270:** 484–487.
Velculescu, V. E., L. Zhang, W. Zhou, J. Vogelstein, M. A. Basrai *et al.*, 1997 Characterization of the yeast transcriptome. Cell **88:** 243–251.
Velculescu, V. E., S. L. Madden, L. Zhang, A. E. Lash, J. Yu *et al.*, 1999 Analysis of human transcriptomes. Nat. Genet. **23:** 387–388.
Venter, J. C., M. D. Adams, E. W. Myers, P. W. Li, R. Mural *et al.*, 2001 The sequence of the human genome. Science **291:** 1304–1351.
Wiesenfeld, K., and F. Jaramillo, 1998 Mini-review of stochastic resonance. Chaos **3:** 539–548.

Communicating editor: G. A. Churchill

APPENDIX: A STOCHASTIC MODEL OF GENE EXPRESSION

Let us assume that N_t genes 1, 2, . . . , N_t are expressed independently with M_t associated transcripts in total in the cells of a large cell population with respective probabilities $q_1, q_2, \ldots q_{N_t}$ where **Pr**(a random transcript corresponds to gene i) $= q_i$. Let the random variable s_i denote the number of transcripts for gene i in a random library of size M. Note $\Sigma_{i=1}^{N_t} s_i = M$. When $M \ll M_t$, sampling with replacement is an acceptable model of library construction; this is described by a multinomial distribution. The joint probability of observing $s_1 = y_1$ mRNA transcripts of gene 1, . . . , $s_{N_t} = y_{N_t}$ mRNA transcripts of gene N_t in a given library with size M is defined by the probability function, $f(y_1, \ldots, y_{N_t}; M) := \Pr[s_1 = y_1, \ldots, s_{N_t} = y_{N_t}]$, where

$$f(y_1, \ldots, y_{N_t}; M) = \frac{M!}{\prod_{j=1}^{N_t} y_j!} \prod_{j=1}^{N_t} q_j^{y_j}. \tag{A1}$$

The function f has the unknown parameters $q_1, q_2, \ldots q_{N_t}$ and N_t, together with the constraints $\Sigma_{j=1}^{N_t} q_j = 1$ and $\Sigma_{j=1}^{N_t} y_j = M$. The marginal probability function $f_i(m, M) := \Pr(s_i = m)$ is the probability that the unique tag for gene i occurs exactly m times in a library of size M:

$$f_i(m, M) = \frac{M!}{m!(M - m)!} q_i^m (1 - q_i)^{M-m}.$$

We can estimate the expected number of distinct genes, $n(m, M)$, which have m transcripts in our library of size M. Let $\delta_{ij} = 1$ when $i = j$ and 0 otherwise. Now,

$$n(m, M) := \sum_{i=1}^{N_t} E(\delta_{s,m}) = \sum_{i=1}^{N_t} f_i(m, M).$$

Let the random variable $G_M = \#\{i | s_i > 0\}$; G_M is the number of distinct genes represented in the library of size M. We can estimate the expected number of genes $N(M)$ in a given library as $E[G] := N(M)$, where

$$N(M) := \sum_{m=1}^{M} n(m, M) = \sum_{j=1}^{N_t} (1 - (1 - q_j)^M) = N_t - n(0, M), \tag{A2}$$

where $n(0, M)$ denotes the expected number of distinct genes that escaped detection in the given library; $n(0, M)$ is given as $n(0, M) = \Sigma_{j=1}^{N_t} (1 - q_j)^M$.

Using the formulas for $f_i(m, M)$, $n(m, M)$, and $N(M)$ we can derive the recursion formula (Kuznetsov 2001)

$$n(m, M) = (-1)^{m+1} \frac{M!}{m!(M - m)!} (\nabla^m N(M)), \tag{A3}$$

where $m \in \{1, \ldots, M\}$. Also, $n(m, M) = 0$, if $m > M$. ∇ is the backward difference operator $\nabla N(M) = N(M) - N(M-1), \ldots, \nabla^m N(M) := \nabla^{m-1} N(M) - \nabla^{m-1} N(M - 1)$. These results allow us to compute $n(m, M)$ for any given values of m and M. Recall we let $p_m = n(m, M)/N$. Then, for large M we may approximate p_m with its continuous analog and obtain the probability function called the binomial differential model (Equation 2, methods). This function has a skewed form and is approximated for large enough M and m by the power function $p_m \sim 1/m^2$ (Lotka-Zipf law).