BIOINFORMATICS

Vol. 18 no. 4 2002
Pages 576–584

Making sense of microarray data distributions

David C. Hoyle[1,*], Magnus Rattray[2], Ray Jupp[3] and Andrew Brass[1,2]

[1]School of Biological Sciences, University of Manchester, Stopford Building, Oxford Rd, Manchester M13 9PT, UK, [2]Department of Computer Science, University of Manchester, Kilburn Building, Oxford Rd, Manchester M13 9PL, UK and [3]Aventis Pharmaceuticals, 4041 Route 202-206, PO Box 6800, Bridgewater, NJ 08807, USA

Received on September 5, 2001; revised on October 31, 2001; accepted on November 16, 2001

ABSTRACT

Motivation: Typical analysis of microarray data has focused on spot by spot comparisons within a single organism. Less analysis has been done on the comparison of the entire distribution of spot intensities between experiments and between organisms.

Results: Here we show that mRNA transcription data from a wide range of organisms and measured with a range of experimental platforms show close agreement with Benford's law (Benford, *Proc. Am. Phil. Soc.*, **78**, 551–572, 1938) and Zipf's law (Zipf, *The Psycho-biology of Language: an Introduction to Dynamic Philology*, 1936 and *Human Behaviour and the Principle of Least Effort*, 1949). The distribution of the bulk of microarray spot intensities is well approximated by a log-normal with the tail of the distribution being closer to power law. The variance, σ^2, of log spot intensity shows a positive correlation with genome size (in terms of number of genes) and is therefore relatively fixed within some range for a given organism. The measured value of σ^2 can be significantly smaller than the expected value if the mRNA is extracted from a sample of mixed cell types. Our research demonstrates that useful biological findings may result from analyzing microarray data at the level of entire intensity distributions.

Contact: david.c.hoyle@man.ac.uk

INTRODUCTION

Microarray experiments provide a way of studying the RNA expression levels of tens of thousands of genes simultaneously. Typically these experiments compare different cell types, for example normal versus diseased cells, to identify genes which are differentially expressed. Robust statistical methodologies are required to determine which genes are differentially expressed, and which sets of genes behave in similar ways—for example for use in guilt-by-association studies—and this has been the focus of extensive research (see, for example, Quackenbush (2001) for a recent review of the issues surrounding computational analysis of microarray data). The information that can be obtained from examining the distribution of spot intensities itself is a much less studied area. Figure 1 shows a typical microarray spot intensity distribution. The data was supplied to us by Aventis and obtained using Affymetrix oligonucleotide chips and mRNA extracted from human tissue. Several features are immediately obvious, for example the distribution is heavily skewed with most spots having a low intensity, whilst a few have very high intensities.

A number of questions arise naturally:

- Is there a generic form for the distribution, independent of chip technology or the species being studied?
- If there is a generic form, what are the appropriate statistics to describe it?
- Can we use a knowledge of the spot intensity distribution to assist in tasks such as quality control?
- By quantifying the generic features of the spot intensity distributions can we uncover biological behaviour that may not be apparent to more conventional analysis tools?

We will focus primarily upon the first two questions. In this paper we have examined spot intensity distributions obtained from microarray analyses of a wide range of species and tissues (listed in Table 1). Importantly, the data analysed has been generated using a wide range of different microarray technologies. Addressing the first question we demonstrate that microarray data belongs to the large class of systems showing good agreement with Benford's law (Benford, 1938), and that the bulk of the data from a microarray experiment generally has a log-normal distribution. The tail of the distribution of microarray data shows good agreement with Zipf's law (Zipf, 1936, 1949), suggesting a power law tail. The width of the distribution is positively correlated with the number of genes in the genome of the organism being studied.

*To whom correspondence should be addressed.

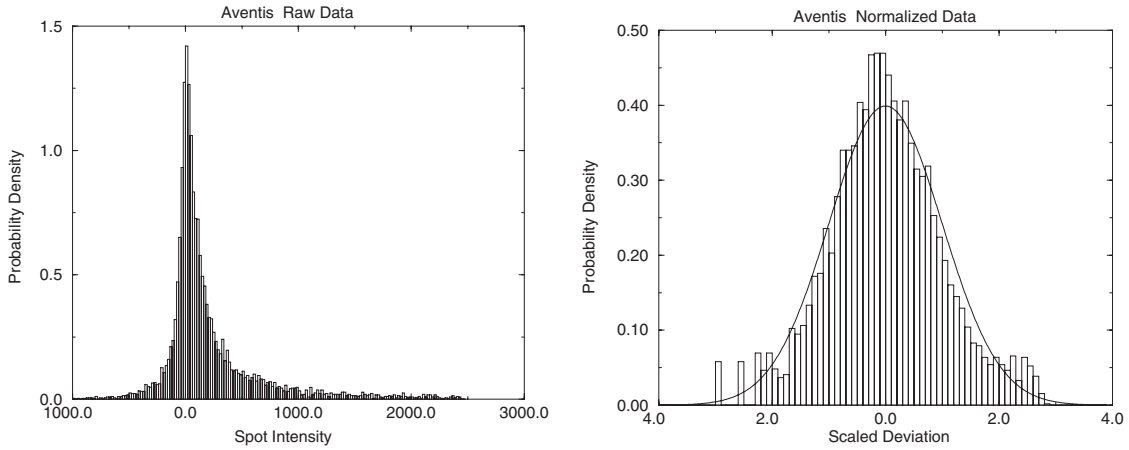

Fig. 1. Distribution of corrected spot intensities for one of the human data sets supplied by Aventis. The left hand plot shows the distribution of raw values $\hat{s}$ and the right hand plot shows the distribution of $(\log \hat{s} - \mu)/\sigma$. $\hat{s}$ is the average difference between positive matches and mis-matches in the Affymetrix system. μ and σ^2 are mean and variance respectively of $\log(\hat{s})$ evaluated only over positive values of $\hat{s}$. The solid line in the right hand plot is the standard normal $\mathcal{N}(0, 1)$.

Table 1. Table showing microarray data sets analysed and their agreement with Benford's law

Data set reference	Organism	Type of array	Typical no. of spots on array	No. of samples	χ^2 1st digit	Average variance logged data ± 1SD	Average inter-quartile range logged data ± 1SD
Aventis[a]	Human	Oligonucleotide	7 129	36	1.677×10^{-3}	2.82 ± 0.24	1.99 ± 0.18
Alon *et al.* (1999)	Human	Oligonucleotide	7 464	62	5.120×10^{-3}	2.29 ± 0.12	2.06 ± 0.10
Aventis[a]	Rat	Oligonucleotide	8 822	570	1.222×10^{-3}	2.72 ± 0.37	2.18 ± 0.20
Renovo[b]	Rat	Oligonucleotide	8 806	3	9.158×10^{-4}	2.94 ± 0.15	2.13 ± 0.04
Brutsche *et al.* (2001)	Human	Membrane	588	49	2.611×10^{-2}	1.66 ± 0.44	1.46 ± 0.22
Diehn *et al.* (2000)	Human	Glass slide	6 720	1	1.302×10^{-2}	1.38 ± 0.01	1.51 ± 0.06
Perou *et al.* (1999)	Human	Glass slide	5 777	26	2.026×10^{-1}	0.88 ± 0.35	1.35 ± 0.42
Ross *et al.* (2000)	Human	Glass slide	10 000	66	4.318×10^{-3}	2.26 ± 0.32	2.14 ± 0.17
Gracey *et al.* (2001)	Fish	Glass slide	5 472	26	2.269×10^{-2}	2.73 ± 0.75	2.49 ± 0.46
Schaffer *et al.* (2001)	*Arabidopsis*	Glass slide	12 619	17	4.015×10^{-3}	1.85 ± 0.33	1.73 ± 0.18
Reinke *et al.* (2000)	*C. elegans*	Glass slide	13 323	29	2.882×10^{-3}	2.26 ± 0.31	1.67 ± 0.17
White *et al.* (1999)	*Drosophila*	Glass slide	6 240	19	3.676×10^{-3}	1.47 ± 0.31	1.54 ± 0.16
Hayes[c]	Yeast	Glass slide	6 272	2	1.526×10^{-2}	1.43 ± 0.14	1.41 ± 0.08
DeRisi *et al.* (1997)	Yeast	Glass slide	6 153	7	7.165×10^{-2}	0.64 ± 0.05	1.03 ± 0.06
Diehn *et al.* (2000)	Yeast	Glass slide	8 448	1	5.032×10^{-2}	1.28 ± 0.35	1.08 ± 0.09
Gasch *et al.* (2000)	Yeast	Glass slide	8 990	159	9.115×10^{-3}	1.39 ± 0.40	1.46 ± 0.24
SMD	*E. coli*	Glass slide	4 807	64	3.921×10^{-2}	1.09 ± 0.37	1.27 ± 0.33

[a]Data from Affymetrix oligonucleotide chips supplied by Aventis.
[b]Data from Affymetrix oligonucleotide chips supplied by Renovo Ltd.
[c]Yeast data from PCR-generated oligonucleotides spotted onto glass slides, supplied by Professor Stephen Oliver and Dr Andrew Hayes.

THEORY AND RESULTS

Characterization of microarray spot intensity distributions

Benford's law. Ever since Newcomb (1881) noted that books of log tables were always much grimier at the start than the end it has been known that the distribution of the first significant digit of many data sets does not follow a uniform distribution. Later Benford (1938) also conjectured that the occurrence of the first significant digit follows a particular probability distribution such that the number 1 comes up about 30% of the time, whereas 9 only occurs 5% of the time. It is now known that this distribution—Benford's law—is found in many data sets from American league baseball statistics (Benford, 1938), to areas of rivers (Benford, 1938) and financial accounts

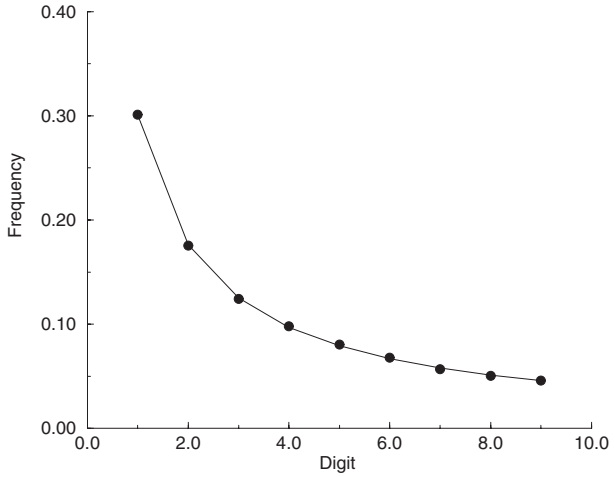

Fig. 2. Plot of 1st significant digit frequencies. The data consists of 20 samples of 7129 spot intensities from Affymetrix oligonucleotide chips. The mRNA was extracted from Human cells. The solid line is Benford's law. The circles are the experimental result.

(Nigrini, 1996). The idea that data from complex processes naturally satisfy Benford's law has more recently been put on a firm theoretical basis by Hill (1995).

The measurement of microarray spot intensities is the end product of a set of complex biological and experimental processes. It is therefore reasonable to ask whether there is any correspondence with Benford's law in raw intensity values from microarray experiments. The distribution of 1st significant digits is given by Benford's law (in base 10) as,

$$P(D) = \log_{10}(1 + D^{-1}). \tag{1}$$

Figure 2 shows 1st significant digit frequencies f_D averaged over 20 samples from Affymetrix oligonucleotide chip experiments (supplied by Aventis). Agreement with Benford's law invariably improves when significant digit frequencies are averaged over several samples (Raimi, 1976). However we have found almost the same degree of correspondence with Benford's law for each of the individual samples studied in this data set, as for the total.

We have also examined several other data sets of microarray spot intensities for correspondence with Benford's law. The data sets are a mixture of those supplied to us and publicly available data sets mainly obtained from Stanford University's MicroArray Database (SMD; http://genome-www4.stanford.edu/MicroArray/SMD). The results of our analysis are summarized in Table 1 by quoting the average (over samples) of the $\chi^2_{\text{1st digit}}$ statistic. For a single sample $\chi^2_{\text{1st digit}}$ is given as,

$$\chi^2_{\text{1st digit}} = \sum_{D=1}^{D=9} \frac{(\log_{10}(1 + D^{-1}) - f_D)^2}{\log_{10}(1 + D^{-1})}. \tag{2}$$

For the two-label data we have treated the intensity values (corrected for background) from each channel as separate samples. Where it is possible to identify them, we have eliminated from our analysis those spots which are used for control purposes, empty, or flagged as being suspect. For data from Affymetrix oligonucleotide chips we have used the average difference values and ignored the Present/Absent call. This is an attempt to make the treatment of Affymetrix data comparable to that from two-label experiments.

A natural question to ask is—what distribution gives rise to Benford's law? Since Benford's law (1) is scale free any underlying distribution must also be scale free, e.g. power law. Pietronero *et al.* (2001) demonstrate that for power law distributions $P(x) \sim x^{-\alpha}$ then Benford's law (1) results automatically for $\alpha = 1$, and a generalized Benford's law for $\alpha \neq 1$. However as the example distribution in Figure 1 clearly shows it cannot be power law throughout its entire range. Indeed computer simulation reveals that the $\chi^2_{\text{1st digit}}$ values calculated here, although small, are highly statistically significant. Sampling $N = 6000$ digits from the Benford distribution (1) gives the probability of observing $\chi^2_{\text{1st digit}} > 9.1 \times 10^{-3}$ (the average value obtained from the data sets of Gasch *et al.* (2000)) as $p < 0.0002$. The observed $\chi^2_{\text{1st digit}}$ values indicate that typically the observed distributions of 1st significant digits are genuinely distinguishable from the Benford distribution (1), but very close to it. We are then led to ask what scale dependent distributions show approximate but close agreement with Benford's law. Leemis *et al.* (2000) show that for any random variable W whose fractional part is uniformly distributed, $U(0, 1)$, then 10^W will satisfy Benford's law; for example if W is distributed symmetrically about an integer and has a probability density function that is piecewise linear between successive integers. If one requires a smooth distribution then if W has a symmetric distribution of large variance σ^2, one can see that 10^W will approximate Benford's law well, with exact agreement in the limit $\sigma^2 \to \infty$. If one constrains the distribution of W no further then the obvious (maximum entropy) choice of smooth distribution is a Gaussian, and thus $Y = 10^W$ is distributed log-normally. In Figure 1 we have also plotted the observed distribution of scaled logged spot intensities. The comparison to the standard normal $\mathcal{N}(0, 1)$ is good. We conclude that microarray data is at least consistent with having come from a log-normal distribution. In the absence of a generative model one cannot categorically

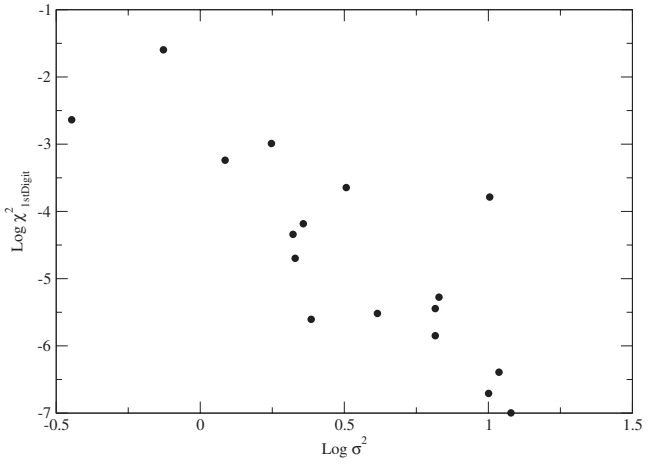

Fig. 3. Plot of $\log \chi^2_{\text{1st digit}}$ against $\log \sigma^2$ for the data sets listed in Table 1 (average values of $\chi^2_{\text{1st digit}}$ and σ^2 have been used).

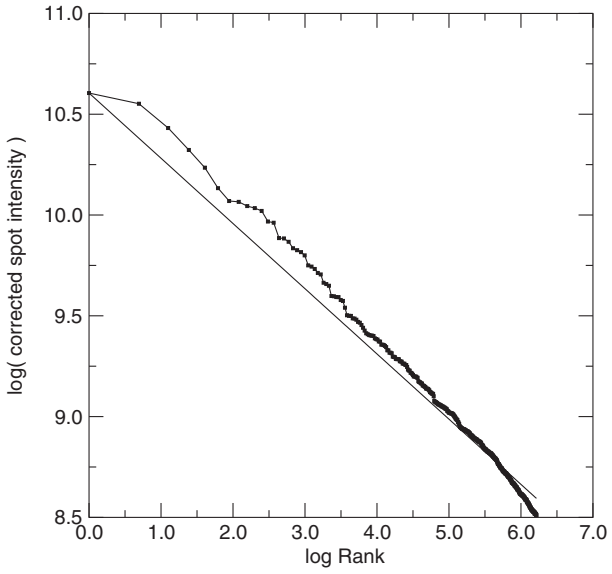

Fig. 4. Plot of log(corrected spot intensity) against log(rank) for the top 500 spots in one channel in one of the data sets of Ross *et al.* (2000). The solid line is the fit to Zipf's law.

say that the distribution of spot intensities is definitely log-normal. However we stress that a log-normal is an *extremely good* approximation to the bulk of the data, particularly for the higher eukaryotes. In general the agreement with a log-normal distribution improves in going from lower eukaryotes to higher eukaryotes.

Increasing agreement with Benford's law is expected with increasing variance, σ^2, of the logged data. Plotted in Figure 3 are the average $\chi^2_{\text{1st digit}}$ values against average σ^2 values for each data set. A clear negative correlation is present. It should be noted that the variances of logged data are only calculated using spots with strictly positive intensities.

As a final cautionary note it is worth stating that the generic statistical features we have focused on will only be present in microarray data sets if spot intensities are taken from an unbiased sample of genes. Biased samples of genes can occur by focusing only on genes of particular interest, e.g. genes known to be associated with a particular clinical condition, or if genes only from a limited number of functional groups or biochemical pathways are expressed by the organism.

Zipf's law. The agreement of microarray data with Benford's law suggests the log-normal distribution as a potential distribution for normalization of the bulk of the corrected spot intensities. Certainly microarray data should be analysed on a log scale. However we have already noted that power law distributions are also capable of reproducing Benford's law and such distributions may be a more appropriate description of the tails of spot intensity distributions. Many examples exist where real data sets show a log-normal like distribution of values for the bulk of the data but also show power law tails (Montroll and Shlesinger, 1982; Stanley *et al.*, 1999). When spot intensities I_r are ordered by rank r, from the highest I_1 to the lowest, we observe approximate Zipf's law behaviour for the highest intensities in microarray data sets from various human tumour samples. Zipf's law is a linear relation,

$$\log I_r = \log I_1 + \nu \log r, \nu < 0 \qquad (3)$$

noted by Zipf (1936, 1949) to apply to various data sets including word frequencies in passages of text and to sizes of cities. It is a trivial matter to show that a Zipf's law tail and a power law tail, $P(x) \sim x^{-\alpha}$, are inter-changeable with $\nu = 1/(1-\alpha)$. Plotted in Figure 4 is log spot intensity against log rank for the 500 largest spot intensities from one of the data sets of Ross *et al.* (2000). Approximate Zipf's law behaviour is clearly seen for the largest spot intensities and a value of $\nu = -0.32$ is extracted by fitting (3). Similar plots can be made for all the data sets we have analysed.

Some curvature in the plot shown in Figure 4 is apparent to the eye. For all the data sets of Ross *et al.* (2000) we have found the magnitude of the curvature, determined by fitting $\log I_r = \log I_1 + \nu \log r + \zeta (\log r)^2$ to the data, to be statistically significant ($p < 0.001$) but always much smaller than the term linear in $\log r$. Thus the data displays only an approximate, although very good, agreement with Zipf's law.

One must be careful when looking for power law behaviour in the tail of a distribution close to log-normal.

As has often been noted (Montroll and Shlesinger, 1982; Sornette, 2000) the tail of a log-normal distribution can do a very good impression of a power law. The density function of a log-normal can be written (Sornette, 2000),

$$P(x)\,\mathrm{d}x = \frac{1}{\sqrt{2\pi\sigma^2}}x^{-1}\exp\left(-\frac{1}{2\sigma^2}(\log x-\mu)^2\right)\mathrm{d}x$$
$$= \frac{\mathrm{e}^{-\mu}}{\sqrt{2\pi\sigma^2}}(x\mathrm{e}^{-\mu})^{-1-\eta(x)}\,\mathrm{d}x \qquad (4)$$

where $\eta(x) = \frac{1}{2\sigma^2}(\log x - \mu)$. With $\eta(x)$ being a slowly varying function of x due to the log, the log-normal can approximate a power law for $\log x > \mu$. Typically the full log-normal structure of the tail would only be apparent when examining the probability density over a range of $\log x$ extending more than 2σ standard deviations beyond μ (i.e. $\eta(x)$ changing by $\mathcal{O}(1)$ over this range). With the average value of $\sigma \simeq \sqrt{2.26} \simeq 1.5$ for the data sets of Ross *et al.* (2000), 2σ takes us well into the tail of any log-normal and would require sampling of many more points than there are on a typical microarray to accurately estimate the probability density in this region. The data shown in Figure 4 has $\sigma \simeq 1.49$ and extends from about 1.47–2.88 standard deviations of log spot intensity above the mean value. Thus distinguishing by eye between the tail of a log-normal distribution and a power law is difficult for the typical number of data points available to us from a microarray experiment. However the variation in slope observed in Figure 4 is actually considerably less than would be expected from a log-normal with the same mean and variance of log spot intensity. Secondly the local effective Zipf's law exponent from a log-normal tail would be $\nu \sim -1/\eta(x)$. Thus to observe an exponent $\nu \simeq -0.32$ would require $\eta(x) \gg 1$ i.e. $\sigma^{-1}(\log x - \mu) \gg 2\sigma$. With $\sigma \simeq 1.49$ for the data in Figure 4, then to observe such a small Zipf's law exponent ν from a log-normal tail would require values of log spot intensity approximately 3 or more standard deviations above the mean value. This is certainly not the case. Therefore we conclude that the tail of the spot intensity distribution is something close to a genuine power law and is less likely to be the tail of a log-normal. The detected curvature may be due to: (i) Mandelbrot's modified form of Zipf's law (Mandelbrot, 1966) may be a more appropriate description of the data; or (ii) a genuine power law tail may be the asymptotic form of some limiting process that has not been reached due to the finite genome sizes and finite number of microarray spots.

Extraction of biological information from characteristics of spot intensity distributions

So far we have been discussing the characteristics of the distribution of spot intensities from microarray experiments. How do these characteristics relate to the characteristics of the underlying mRNA abundance distribution? Individual spot intensities cannot be taken as a precise measure of mRNA abundance, although work by Ishii *et al.* (2000) reveals that corrected spot intensities from oligonucleotide chips can be taken as reasonable estimates of mRNA abundance. In a two fluor microarray experiment we can take the spot intensity I_i, for gene i, as being a mixture of biological signal B_i and systematic but gene specific effects E_i. Thus we can model, $I_{i,G} = B_{i,G}E_i$, $I_{i,R} = B_{i,R}E_i$, where the subscripts G, R refer to the two fluorescent labels. A similar (single label) model can be used for oligonucleotide spot intensities. The systematic gene effect E_i is typically assumed to be independent of, or only weakly dependent on, the fluorescent label so that typically the factor E_i can be eliminated by considering the ratio $I_{i,G}/I_{i,R}$. We see that, $\mathrm{Var}(\log I_G) = \mathrm{Var}(\log B_G) + \mathrm{Var}(\log E) + 2\mathrm{Cov}(\log B_G, \log E)$ and likewise for the other label. E is considered to be systematic, determined largely by the specific gene sequence. However, Wagner (2000) has found little correlation between differential gene expression and sequence similarity. We may therefore reasonably take $\mathrm{Cov}(\log B_G, \log E)$ to be small in comparison to the other two contributions to $\mathrm{Var}(\log I_G)$. We assume that $\mathrm{Var}(\log B_G)$ is the dominant contribution to $\mathrm{Var}(\log I_G)$ (and similarly for $\mathrm{Var}(\log I_R)$). Ishii *et al.* (2000) have compared data from Affymetrix oligonucleotide chips and SAGE (serial analysis of gene expression), using identical mRNA samples obtained from human blood monocytes and granulocyte-macrophage colony-stimulating factor induced macrophages. A correlation of $r = 0.817$ was found between log(corrected spot intensity) of the oligonucleotide chips and log(tag frequency) of the SAGE analysis. If we consider SAGE analysis to provide a more quantitatively accurate estimate of mRNA abundance then the analysis of Ishii *et al.* (2000) suggests an upper estimate of around 33% for the contribution of $\mathrm{Var}(\log E)$ to $\mathrm{Var}(\log I)$ in the case of oligonucleotide chips. The consistency between the values of $\mathrm{Var}(\log I)$ obtained from Affymetrix chips and those obtained from arrays of spotted clones for Human data suggests a similar sized contribution of $\mathrm{Var}(\log E)$ to $\mathrm{Var}(\log I)$ may be valid for two fluor experiments.

The major contributions to the systematic effect E are often attributed to: (a) the specific secondary structure of the mRNA interfering with its ability to hybridize to the probe on the array. However it has been suggested (Southern *et al.*, 1999) that this is not an important effect for arrays of spotted clones or PCR products where the hybridization conditions are stringent enough to melt most secondary structure. For hybridization to oligonucleotide arrays this may not be true. Most of our analysis has concentrated on arrays of spotted clones, although for the Human data, as we have already noted, there is reasonable

consistency between the values of $\mathrm{Var}(\log I)$ obtained from Affymetrix chips and those obtained from arrays of spotted clones. (b) The number of label molecules attached to the mRNA being proportional to length of the reverse transcription products since label molecules are attached internally and not just at the sequence ends. We might therefore expect a possible correlation between spot intensity and length of spliced open reading frame. Reverse transcription of the mRNA is not always fully complete and so any correlation would be expected to be strongest for the shortest genes. For the data sets of Gasch *et al.* (2000) the correlation between spot intensity and gene length is negligible. For spliced open reading frames of *Saccharomyces cerevisiae* with a length less than 250 bp the average (over both channels and all the samples) correlation coefficient for this data set is $\overline{|r|} = 0.087$.

Given the above considerations we continue our analysis of the data sets viewing $\sigma^2 = \mathrm{Var}(\log I)$ as a noisy estimate of Var(log mRNA abundance) for the sample from which the mRNA was extracted. Any large scale changes in σ^2 we consider to be due to large scale changes in $\mathrm{Var}(\log B)$. Certainly we have no *a priori* reason to believe that when using similar experimental protocols $\mathrm{Var}(\log E)$ would significantly change between organisms. Note that we are not assuming a direct correspondence between spot intensity and mRNA abundance. The contribution of $\mathrm{Var}(\log E)$ to $\mathrm{Var}(\log I)$ is non-negligible. Therefore for any specific gene the size of the systematic effect E will most likely be too large to consider spot intensity as a precise (up to a global scale factor) estimate of mRNA abundance for that gene. However a well defined and significant statistical correlation can still exist between spot intensity and mRNA abundance. Therefore we consider the Zipf's law tail of the spot intensity distribution to infer a Zipf's law tail for the underlying mRNA abundance distribution. The approximate log-normal shape of the spot intensity distribution raises the possibility of the underlying mRNA abundance distribution being approximately log-normal, although we acknowledge that at this stage we cannot discount the possibility that the left hand tail of the spot intensity distribution is a measurement artefact. Where an explicit form for the underlying mRNA abundance distribution is required we shall assume a log-normal form. We have avoided using $\mathrm{Var}(\log(I_G/I_R))$ to estimate $\mathrm{Var}(\log B)$ since with few genes being highly up or down regulated between different samples $\mathrm{Cov}(\log B_G, \log B_R)$ is typically on the same scale as $\mathrm{Var}(\log B_G)$ and $\mathrm{Var}(\log B_R)$, with the consequence that $\mathrm{Var}(\log(I_G/I_R)$ will be small and not well correlated with $\mathrm{Var}(\log B_G) + \mathrm{Var}(\log B_R)$.

Having obtained values of σ^2 for several data sets and several organisms we wish to ascertain if there is any biological information in these values. Is the value of σ^2 unique to a given organism and is there any general trend in σ^2 as one moves from lower to higher eukaryotes? There is some support for these ideas if we look at individual data sets. For example, for the data set of Ross *et al.* (2000) the average variance of logged data across the 66 different chips is approximately 2.26 with a standard deviation of only 0.32, despite the 66 chips representing cell lines derived from tumours and normal tissues of widely different origin (breast, colon, etc.). This average value of 2.26 is clearly distinct from the average values obtained for lower eukaryotes such as *S. cerevisiae* or *Drosophila melanogaster*.

Variance of the log-normal and the effect of mixed cell types. At first sight the differing variances observed in the data sets of Perou *et al.* (1999) and Ross *et al.* (2000) is perplexing given that both studies included mRNA derived from human tumour specimens. However the work of Ross *et al.* (2000) used cell lines derived from tumours, whilst that of Perou *et al.* (1999) used primary tumour tissue directly. As noted by Perou *et al.* (1999) their samples could potentially contain not just carcinoma cells but also epithelial cells, stromal cells, adipose cells, endothelial cells and infiltrating lymphocytes. If the experimental sample consists of mixed cell-types then the observed spot intensities will consist of sums of several approximately log-normally distributed variables, one for each cell-type in the sample, which may have widely differing means and variances. We can easily simulate this situation by writing $R = \sum_{i=1}^{M} r_i$ with $\log r_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$. We take $M = 10$ and $\mu_i \sim U[3, 8]$, $\sigma_i^2 \sim U[0.6, 2.6]$ to reflect the range of mean and variances observed in the data sets in Table 1. Sampling 1000 points for R and repeating this process 1000 times gives $\overline{\mathrm{Var}(\log R)} = 0.464$. Thus in this simple simulation the observed variance of the logged data from a mixed cell-type sample is considerably less than the average variance (over the individual cell types) of 1.6, and is even less than the lower bound of 0.6 on the variance from a single cell-type. A similar reduction in variance may also occur during the process of actually scanning the hybridized array since, when reading the intensity value from a given spot, neighbouring spots can also contribute to some degree.

Can we quantify this effect of having mRNA extracted from mixed cell types? Consider $R = \sum_{i=1}^{M} r_i$ with $\log r_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$. We take $\mu_i, \sigma_i, i = 1, \ldots, M$ to be *i.i.d* random variables. In general M will not be sufficiently large for the Central Limit Theorem (CLT) to apply. Sums such as these, of log-normal distributed random variables, frequently occur in the field of mobile communications (Fenton, 1960; Schwartz and Yeh, 1982). Typically R is considered to be well approximated by another log-normal (Fenton, 1960; Schwartz and Yeh, 1982). Thus we take $\log R$ to have mean μ_{mc} and variance

$\sigma^2_{\rm mc}$ (the subscript mc denoting mixed cell type values) and approximate $\log R \sim \mathcal{N}(\mu_{\rm mc}, \sigma^2_{\rm mc})$. One can match first and second moments of R and $\sum_{i=1}^{M} r_i$ to give (Fenton, 1960),

$$\sigma^2_{\rm mc} = \log\left[1 + \frac{\sum_i \mathrm{e}^{2\mu_i+\sigma_i^2}(\mathrm{e}^{\sigma_i^2}-1)}{\left(\sum_i \mathrm{e}^{\mu_i+\frac{1}{2}\sigma_i^2}\right)^2}\right]. \qquad (5)$$

If the $r_i, i = 1, \ldots, M$ contributing to R are in fact *i.i.d*, i.e. $\mu_1 = \mu_2 = \cdots = \mu_M = \mu_{\rm sc}$ and $\sigma_1 = \sigma_2 = \cdots = \sigma_M = \sigma_{\rm sc}$ (the subscript sc denoting single cell type), then one has,

$$\sigma^2_{\rm mc} = \log[1 + M^{-1}(\mathrm{e}^{\sigma^2_{\rm sc}} - 1)] \qquad (6)$$

$$\simeq M^{-1}(\mathrm{e}^{\sigma^2_{\rm sc}} - 1) + \mathcal{O}(M^{-2}) \; M \to \infty. \qquad (7)$$

The accuracy of determining $\mu_{\rm mc}$ and $\sigma^2_{\rm mc}$ by matching first and second moments decreases with increasing M. More accurate recursive numerical procedures exist (Schwartz and Yeh, 1982) however the above formula is useful from an illustrative point of view. From (7) we can see that there is an increasing reduction in the variance of the logged data ($\log R$) as the number of contributing log-normal distributions, M, increases. Strictly speaking the justifications for (7) will not be valid as $M \to \infty$, and at any rate if the $r_i, i = 1, \ldots, M$ are *i.i.d* then the approximation of R by a log-normal distribution must be replaced by a Gaussian as $M \to \infty$. One still obtains $\mathrm{Var}(\log R) \simeq M^{-1}(\mathrm{e}^{\sigma^2_{\rm sc}} - 1)$ if the CLT is used for $M \to \infty$, primarily because in applying the CLT we are still focusing on the first and second moments of R. The pre-factor of M^{-1} can be derived through a simple perturbative argument, irrespective of the form assumed for the distributions of the components $r_i, i = 1, \ldots, M$ that comprise R (other than having finite first and second moments). We can take the value of $\sigma^2_{\rm sc} \simeq 2.26$ obtained from the data of Ross *et al.* (2000) as being the appropriate underlying value for the samples of Perou *et al.* (1999) for which we set $\sigma^2_{\rm mc} \simeq 0.88$. Applying (6) we obtain $M \simeq 6$, i.e. 6 different cell types present in the samples of Perou *et al.* (1999). Whilst we consider this estimate of M to be very approximate it is certainly not an unrealistic figure, and intriguingly is the number of cell types listed by Perou *et al.* (1999) as possibly contributing to their extracted mRNA. The fact that we using σ^2, the variance of log(spot intensity) as a crude measure of log(mRNA abundance) will obviously also limit the accuracy of our estimate of number of cell-types present. Although the derivation may appear complex there is a clear implication of the above result. The ratio,

$$\frac{\exp(\sigma^2_{\rm mc}) - 1}{\exp(\sigma^2_{\rm sc}) - 1} \qquad (8)$$

provides an estimate of the number of cell types contributing to the extracted mRNA. Obviously in using (8) we are assuming that each cell type present contributes an equal amount of mRNA. This is unlikely to be the case and so (8) should be viewed as the effective number of cell types present. Many more small scale contaminants may actually be present than is estimated by (8) but if their contribution to the total mRNA is negligible we can effectively consider them to be absent.

Variance of the log-normal and genome size. If we concentrate on data sets where mRNA has not been extracted from mixed cell types then is any trend in the values of σ^2 discernible? Plotted in Figure 5 is σ^2 against approximate genome size (in terms of number of genes). Where data sets are available for the same organism from different laboratories we have taken a simple average, weighted only by the number of samples in each data set. If the source of the mRNA in a given data set is known to be from a mixed cell type population then we have omitted that data set from the average. Thus in calculating a value of σ^2 for Human we have used only the data sets supplied to us by Aventis and those of Alon *et al.* (1999) and Ross *et al.* (2000). We have used approximate genome size since for most of the organisms studied here a precise value is not known. However this does not affect the clear underlying trend present in Figure 5 which shows increasing σ^2 with increasing genome size. With σ^2 generally increasing with genome size we would expect better agreement with Benford's law for organisms with larger genomes. A large amount of the scatter in Figure 5 is due to the fact that σ^2 has been estimated from the corresponding sample variance and that estimation of the underlying signal for low expressed gene is significantly affected by the noise in the spot and background intensity values. More accurate estimation of σ^2 can be done by fitting a presumed parametric form for the underlying distribution of spot intensities to the data from just the highest expressed genes. However to avoid using a presumed distribution we have kept the naive estimates of σ^2, which, as Figure 5 shows, are still capable of revealing the underlying biological trend. A similar trend is obtained if one plots IQD^2 against genome size, where IQD is the inter-quartile distance of the sample distribution of log spot intensities, again calculated using only positive values. The inter-quartile distance is often considered to be a more robust estimator of the scale of a distribution (Huber, 1981).

CONCLUSIONS

Our research in this paper has focused on the analysis of microarray experiments, not at the level of multiple spot-by-spot comparisons, but at the level of entire spot intensity distributions. We have started the process of

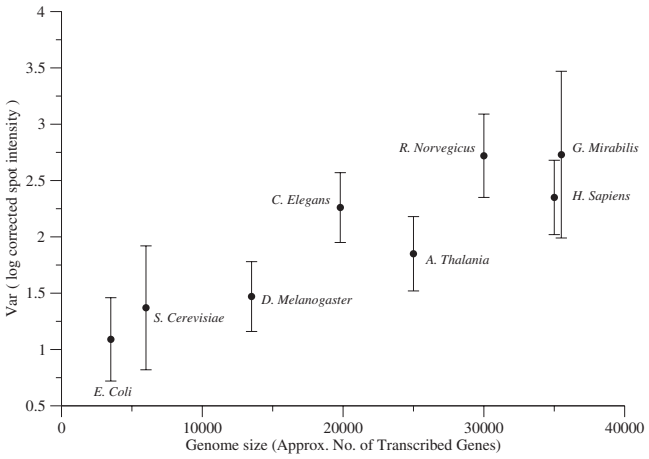

Fig. 5. Plot of σ^2 against N (approximate genome size).

comparing microarray data and spot intensity distributions between organisms. Analysis of microarray data sets covering many different organisms and different chip technologies has shown that microarray data generically shows good agreement with the laws of Benford and Zipf. Do such discoveries help us to answer the questions posed at the beginning of this paper?

- Analysis of several microarray data sets shows that a log-normal distribution is a good approximation for the distribution of the large majority of the spot intensity values. In general the tails of spot intensity distributions show good agreement with Zipf's law, suggesting a power law may be a more appropriate description for the tail. From this we infer an approximate Zipf's law tail and possible log-normal shape for the underlying mRNA abundance distribution.

- From microarray data one should calculate $\chi^2_{\text{1st digit}}$, quantifying the agreement with Benford's law. From positive spot intensities one should calculate the variance σ^2 and inter-quartile distance IQD, of log spot intensity. From positive values the Zipf's law exponent ν can also be calculated. The central region of the spot intensity distribution can be characterized by σ^2 or IQD. The information in the tails of the distribution can be characterized by the Zipf's law exponent ν.

- σ^2 appears to be a roughly fixed characteristic for a given organism and given experimental protocol. This raises the possibility using σ^2 as a measure for quality control. The true value of σ^2 can be affected by a variety of different factors such as whether the mRNA has been extracted from mixed cell types, but in a manner that is well understood. Data sets with a weak underlying signal compared to the background intensity can lead to a higher than expected proportion of intensities close to zero after correcting for the background. On logging the data this leads to a heavy left hand tail and consequently a value of σ^2 significantly greater than that expected. Therefore differing chip technologies, with differing noise profiles, can produce differing values of σ^2 even for the same biological sample.

- Our analysis of σ^2 across several different organisms reveals a general trend of increasing width, to the log(spot intensity) distribution, with increasing genome size. Such a general trend is unlikely to be have been uncovered using more conventional analysis of microarray data. From this we infer a general trend of increasing variance of log(mRNA abundance) with genome size.

ACKNOWLEDGEMENTS

We would like to thank Professor Stephen Oliver and Dr Andrew Hayes for supplying data on *S. cerevisiae* and Professor Mark Ferguson of Renovo Ltd for supplying data on *R. norvegicus*. D.C.H. would like to acknowledge the receipt of an MRC (UK) Special Training Fellowship in Bioinformatics.

REFERENCES

Alon,U., Barkai,N., Notterman,D.A., Gish,K., Ybarra,S., Mack,D. and Levine,A.J. (1999) Broad patterns of gene expression revealed by clustering analysis of tumour and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl Acad. Sci. USA*, **96**, 6745–6750.

Benford,F. (1938) The law of anomalous numbers. *Proc. Am. Phil. Soc.*, **78**, 551–572.

Brutsche,M.H., Brutsche,I.C., Wood,P., Brass,A., Morrison,N., Rattray,M., Mogulkoc,N., Simler,N., Craven,M., Custovic,A., Egan,J.J. and Woodcock,A. (2001) Apoptosis signals in atopy and asthma measured with cDNA arrays. *Clin. Exp. Immunol.*, **123**, 181–187.

DeRisi,J.L., Iyer,V.R. and Brown,P.O. (1997) Exploring the metabolic and genetic control of expression on a genomic scale. *Science*, **278**, 680–686.

Diehn,M., Eisen,M.B., Botstein,D. and Brown,P.O. (2000) Large scale identification of secreted and membrane-associated gene products using DNA microarrays. *Nature Genet.*, **25**, 58–62.

Fenton,L.F. (1960) The sum of log-normal probability distributions in scatter transmission systems. *IRE Trans. Commun. Syst.*, **8**, 57–67.

Gasch,A.P., Spellman,P.T., Kao,C.M., Carmel-Harel,O., Eisen,M.B., Storz,G., Botstein,D. and Brown,P.O. (2000) Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell.*, **11**, 4241–4257.

Gracey,A.Y., Troll,J.V. and Somero,G.N. (2001) Hypoxia-induced gene expression profiling in the euryoxic fish *Gillichthys mirabilis*. *Proc. Natl Acad. Sci. USA*, **98**, 1993–1998.

Hill,T.P. (1995) A statistical derivation of the significant-digit law. *Stat. Sci.*, **10**, 354–363.

Huber,P.J. (1981) *Robust Statistics*. Wiley, New York.

Ishii,M., Hashimoto,S., Tsutsumi,S., Wada,Y., Matsushima,K., Kodama,T. and Aburatani,H. (2000) Direct comparison of Genechip and SAGE on the quantitative accuracy in transcript profiling analysis. *Genomics*, **68**, 136–143.

Leemis,L.M., Schmeiser,B.W. and Evans,D.L. (2000) Survival distributions satisfying Benford's law. *The American Statistician*, **54**, 1–6.

Mandelbrot,B. (1966) Information theory and psycholinguistics: a theory of word frequencies. In Lazarsfeld,P.F. and Henry,N.W. (eds), *Readings in Mathematical Social Science*. Science Research Associates Inc., Chicago, pp. 350–368.

Montroll,E.W. and Shlesinger,M.F. (1982) On 1/f noise and other distributions with long tails. *Proc. Natl Acad. Sci. USA*, **79**, 3380–3383.

Newcomb,S. (1881) Note on the frequency of use of the different digits in natural numbers. *Am. J. Math.*, **4**, 39–40.

Nigrini,M. (1996) A taxpayer compliance application of Benford's law. *J. Am. Taxation Assoc.*, **18**, 72–91.

Perou,C.M., Jeffrey,S.S., Van De Rijn,M., Rees,C.A., Eisen,M.B., Ross,D.T., Pergamenschikov,A., Williams,C.F., Zhu,S.X., Lee,J.C.F., Lashkari,D., Shalon,D., Brown,P.O. and Botstein,D. (1999) Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proc. Natl Acad. Sci. USA*, **96**, 9212–9217.

Pietronero,L., Tossati,E., Tossati,V. and Vespignani,A. (2001) Explaining the uneven distribution of numbers in nature: the laws of Benford and Zipf. *Physica* A, **293**, 297–304.

Quackenbush,J. (2001) Computational analysis of microarray data. *Nature Rev. Genet.*, **2**, 418–427.

Raimi,R. (1976) The first digit problem. *Am. Math. Monthly*, **83**, 521–538.

Reinke,V., Smith,H.E., Nancd,J., Wang,J., Van Doren,C., Begley,R., Jones,S.J.M., Davis,E.B., Scherer,S., Ward,S. and Kim,S.K. (2000) A global profile of germline gene expression in *C. elegans*. *Mol. Cell*, **6**, 605–616.

Ross,D.T., Scherf,U., Eisen,M.B., Perou,C.M., Rees,C., Spellman,P., Iyer,V., Jeffrey,S.S., De Rijn,M.V., Waltham,M., Pergamenschikov,A., Lee,J.C.F., Lashkari,D., Shalon,D., Myers,T.G., Weinstein,J.N., Botstein,D. and Brown,P.O. (2000) Systematic variation in gene expression patterns in human cancer cell lines. *Nature Genet.*, **24**, 227–235.

Schaffer,R., Landgraf,J., Accerbi,M., Simon,V., Larson,M. and Wisman,E. (2001) Microarray analysis of diurnal and circadian-regulated genes in *Arabidopsis*. *Plant Cell*, **13**, 113–123.

Schwartz,S.C. and Yeh,Y.S. (1982) On the distribution function and moments of power sums with log-normal components. *Bell Syst. Tech. J.*, **61**, 1441–1462.

Sornette,D. (2000) *Critical Phenomena in Natural Sciences*. Springer, Heidelberg.

Southern,E., Mir,K. and Shchepinov,M. (1999) Molecular interactions on microarrays. *Nature Genet.*, **21**, (supplement: *The Chipping Forecast*, 5–9).

Stanley,H.E., Amaral,L.A.N., Canning,D., Gopikrishnan,P., Lee,Y. and Liu,Y. (1999) Econophysics: can physicists contribute to the science of economics? *Physica* A, **269**, 156–169.

Wagner,A. (2000) Decoupled evolution of coding region and mRNA expression patterns after gene duplication: implications for the neutral-selectionist debate. *Proc. Natl Acad. Sci. USA*, **97**, 6579–6584.

White,K.P., Rifkin,S.A., Hurban,P. and Hogness,D.S. (1999) MicroArray analysis of *Drosophila* development during metamorphosis. *Science*, **286**, 2179–2184.

Zipf,G.K. (1936) *The Psycho-biology of Language: an Introduction to Dynamic Philology*. Routledge, London.

Zipf,G.K. (1949) *Human Behaviour and the Principle of Least Effort*. Addison-Wesley, Cambridge, MA.