RESEARCH

The Significance of Digital Gene Expression Profiles

Stéphane Audic and Jean-Michel Claverie[1]

Laboratory of Structural and Genetic Information, Centre National de la Recherche Scientifique–E.P.91,
Marseille 13402, France

Genes differentially expressed in different tissues, during development, or during specific pathologies are of foremost interest to both basic and pharmaceutical research. ''Transcript profiles'' or ''digital Northerns'' are generated routinely by partially sequencing thousands of randomly selected clones from relevant cDNA libraries. Differentially expressed genes can then be detected from variations in the counts of their cognate sequence tags. Here we present the first systematic study on the influence of random fluctuations and sampling size on the reliability of this kind of data. We establish a rigorous significance test and demonstrate its use on publicly available transcript profiles. The theory links the threshold of selection of putatively regulated genes (e.g., the number of pharmaceutical leads) to the fraction of false positive clones one is willing to risk. Our results delineate more precisely and extend the limits within which digital Northern data can be used.

Very large-scale, single-pass partial sequencing of cDNA clones from a large number of libraries has led to the identification of ~50,000 human genes (Adams et al. 1995; Aaronson et al. 1996; Hillier et al. 1996). However, a precise function or a complete transcript sequence are known for <5000 of these (Adams et al. 1995; Boguski and Schuler 1995). In the absence of functional clues for most of the newly identified genes, evidence of differential expression is the most important criteria to prioritize the exploitation of anonymous sequence data in both basic and pharmaceutical (Nowak 1995; Adams 1996; Bains 1996; Editorial 1996) research. For example, the study of expression profiles in various tumors is central to the new Cancer Genome Anatomy project (Kuska 1996; O'Brien 1997). In contrast to functional assays, the quantitative analysis of gene expression level lends itself to large-scale implementation. Two main approaches have been proposed (1) ''analog'' methods based on hybridization to arrayed cDNA libraries (Lennon and Lehrach 1991; Gress et al. 1992; Nguyen et al. 1995; Schena et al. 1995; Zhao et al. 1995) or oligonucleotide ''chips'' (Fodor et al. 1991; Southern et al. 1992; Guo et al. 1994; Matson et al. 1995); and (2) ''digital'' methods, based on the generation of sequence tags. This paper focuses on the latter. The sequence tag-based method (Okubo et al. 1992; Matsubara and Okubo 1994) consists of generating a large number (thousands) of expressed sequence tags (ESTs) (Adams et al. 1991; Wilcox et al. 1991; Adams et al. 1992; Khan et al. 1992) from 3′-directed regional non-normalized cDNA libraries. Recently, Velculescu et al. (1995) have introduced the serial analysis of gene expression (SAGE). Although tags are 100–300 nucleotides in length in the original EST approach, the SAGE method only requires nine nucleotides, therefore allowing a larger throughput. In both protocols, the number of tags is reported to be proportional to the abundance of cognate transcripts in the tissue or cell type used to make the cDNA library. The variation in the relative frequency of those tags, stored in computer databases, is then used to point out the differential expression of the corresponding genes: This is the concept of a ''digital Northern'' comparison. In the absence of a sound theoretical framework, the validity of the method has only been verified for a handful of genes in the context of two cellular differentiation systems (Lee et al. 1995; Okubo et al. 1995) inducible in vitro. Yet, with a total number of human genes of ~80,000 or more, it is not intuitive that sequencing a mere few thousand tags (a typical experiment) from highly redundant non-normalized cDNA libraries can produce a useful picture, or realistic ''transcript profile,'' of a given tissue, development stage, or cell type. What variations in tag numbers allow for a reliable inference about differential expression? How many tags should be generated? Here we present the statistical framework required to answer those questions and analyze transcript profiles in a quantitative manner.

[1] Corresponding author.
E-MAIL jmc@igs.cnrs-mrs.fr; FAX 334 91 16 45 49.

7:986–995 ©1997 by Cold Spring Harbor Laboratory Press ISSN 1054-9803/97 $5.00

RESULTS

In Methods we establish the probability distribution governing the occurrence of the same rare event in duplicate experiments. This probability distribution is a general result applicable to a wide variety of experimental situations, although this paper focuses on its use to analyze digital gene expression patterns. The main and only mathematical assumption behind the derivation is that the observed events are rare and part of a large population of possible outcomes (the distribution of which is not specified). In the context of a digital Northern, one event is the observation of a given cDNA sequence tag, and the experiment consists of the random picking and partial sequencing of a number N of cDNA clones. Given the usual complexity (i.e., the number of different genes expressed) of cDNA libraries, observing a given cDNA qualifies as a rare event, as the abundance of most individual messages is of the order of a few percents or less.

Random Fluctuation vs. Significant Change in Tag Number: When to Infer Differential Expression

Let us randomly pick $N = 1000$ clones from a cDNA library and generate the corresponding sequence tags; a given message (e.g., interleukin-2) will be picked x (e.g., two) times, with x in a typical (0–10) range. If we now redo this experiment, that is, again pick 1000 clones and generate the tags, the same message will now be picked y (e.g., 3) times. If the experiments have been duplicated correctly and the clones selected at random, we expect x and y to be close, albeit often different because of random fluctuations. In the Methods section, we show that the expected probability of observing y occurrences of a clone already observed x times is given by the simple formula:

$$p(y|x) = \frac{(x+y)!}{x!y!2^{(x+y+1)}} \qquad (1)$$

Equation 1 can be used to compute a confidence interval $[y_{min}, y_{max}]_\epsilon$ within which we expect to find y with a given probability, noted $1–2\epsilon$, where 2ϵ is the significance level. For ϵ small (e.g., 2.5% or less), y values falling outside the $[y_{min}, y_{max}]_\epsilon$ interval correspond to $p(y \mid x) << 1$, therefore pointing out very unlikely random fluctuations between the two experiments. The confidence intervals for the usual 1% and 5% significance levels are given in Table 1.

The same confidence intervals listed in Table 1 can in fact be used to analyze the results of sampling N clones from two different libraries. Provided all experimental factors are well replicated, significant discrepancies between x (from one library) and y (from the other) will now characterize differentially expressed genes, for example, the relative abundance of which is unlikely to be the same in the two libraries. Simply reading Table 1, we see that variations in counts such as 7 → 0, or 2 → 12 are significant ($P < 0.01$) evidence of regulated gene expression, whereas variations such as 3 → 0 or 8 → 16 are not ($P > 0.05$). However, we do not advocate the use of rigid significance thresholds to analyze digital transcript profiles, as discussed below.

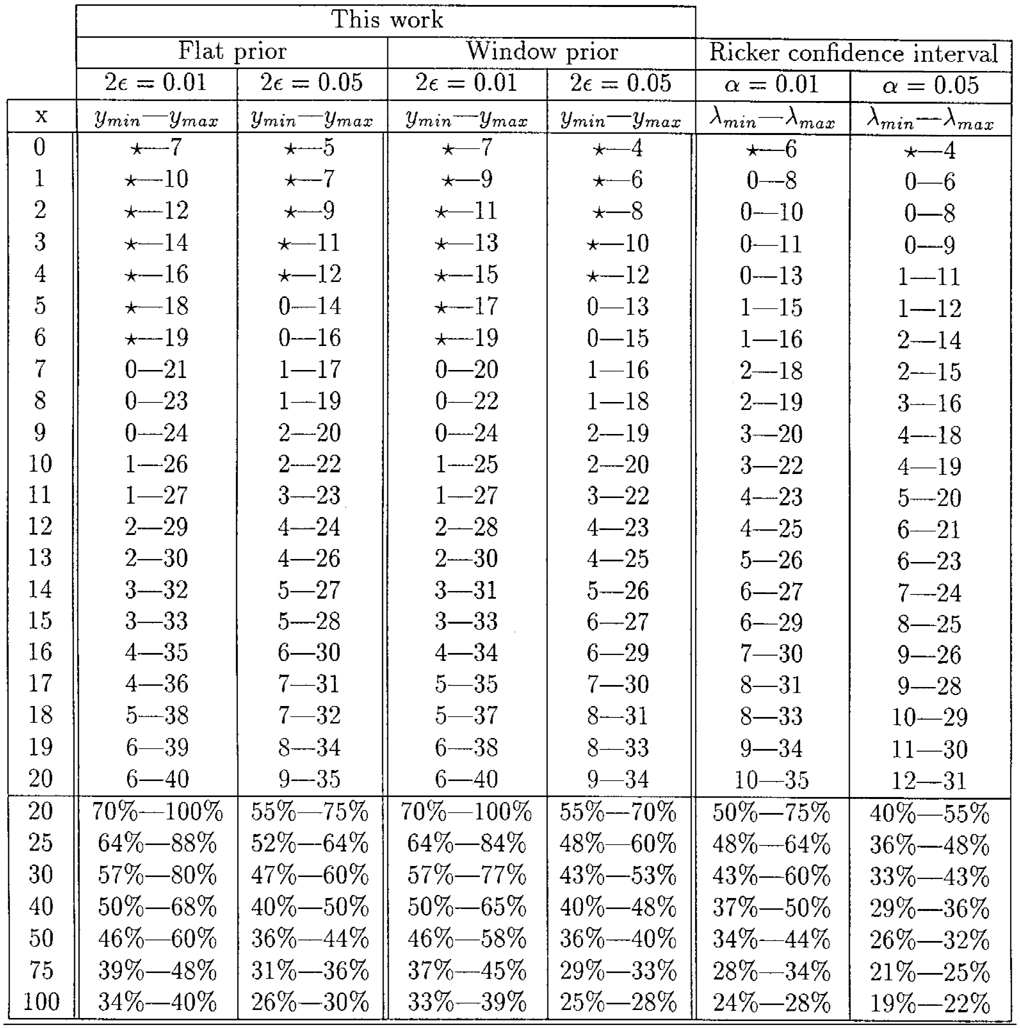
Influence of the Sampling Size

Surprisingly at first, $p(y \mid x)$ in Equation 1 does not involve the sampling size N, that is, the total number of picked clones. The fluctuation probabilities, and confidence intervals, depend only on the values of the observed counts. To understand why, we must remember that Equation 1 governs the results of strictly duplicated experiments. Given N clones are sampled, the most likely tags to be picked up are, intuitively, those corresponding to cDNA, the abundance of which is of the order of $1/N$, or larger (according to Equation 3, the probability of finding a given cDNA with $1/N$ abundance while picking up N clones is 0.63, see also Equation 13). Choosing a sampling size therefore corresponds to targeting a given subset of genes, the level of expression of which allows their tags to occur at reasonable frequencies.

As expected, more reliable inferences can be made on clones corresponding to larger absolute frequencies (i.e., the ones more often picked up). For example (see Table 1), a variation in counts from 1–3 (threefold increase) is not indicative of a significant ($P < 0.05$) increase, whereas a variation from 4–12 is significant at $P < 0.05$, and a variation from 7–21 is significant at $P < 0.01$. For a gene expressed at a given rate, increasing the sampling size N leads to higher tag counts, and allows more stringent statistical inference to be made, for the same proportional variation.

Most often in practice one wishes to compare digital Northerns or gene profiles that have been computed from the random picking of different numbers of clones, N_1 and N_2. The mathematical problem is now to establish the probability for a given cDNA (e.g., interleukin-2) to be picked up x times when the sampling size was N_1 and y times when the sampling size was N_2. Equation 1 then becomes (see Methods):

Table 1. Confidence Intervals in Function of the Value of ***x***



	This work					
	Flat prior		Window prior		Ricker confidence interval	
	$2\epsilon = 0.01$	$2\epsilon = 0.05$	$2\epsilon = 0.01$	$2\epsilon = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$
x	y_{min}—y_{max}	y_{min}—y_{max}	y_{min}—y_{max}	y_{min}—y_{max}	λ_{min}—λ_{max}	λ_{min}—λ_{max}
0	⋆—7	⋆—5	⋆—7	⋆—4	⋆—6	⋆—4
1	⋆—10	⋆—7	⋆—9	⋆—6	0—8	0—6
2	⋆—12	⋆—9	⋆—11	⋆—8	0—10	0—8
3	⋆—14	⋆—11	⋆—13	⋆—10	0—11	0—9
4	⋆—16	⋆—12	⋆—15	⋆—12	0—13	1—11
5	⋆—18	0—14	⋆—17	0—13	1—15	1—12
6	⋆—19	0—16	⋆—19	0—15	1—16	2—14
7	0—21	1—17	0—20	1—16	2—18	2—15
8	0—23	1—19	0—22	1—18	2—19	3—16
9	0—24	2—20	0—24	2—19	3—20	4—18
10	1—26	2—22	1—25	2—20	3—22	4—19
11	1—27	3—23	1—27	3—22	4—23	5—20
12	2—29	4—24	2—28	4—23	4—25	6—21
13	2—30	4—26	2—30	4—25	5—26	6—23
14	3—32	5—27	3—31	5—26	6—27	7—24
15	3—33	5—28	3—33	6—27	6—29	8—25
16	4—35	6—30	4—34	6—29	7—30	9—26
17	4—36	7—31	5—35	7—30	8—31	9—28
18	5—38	7—32	5—37	8—31	8—33	10—29
19	6—39	8—34	6—38	8—33	9—34	11—30
20	6—40	9—35	6—40	9—34	10—35	12—31
20	70%—100%	55%—75%	70%—100%	55%—70%	50%—75%	40%—55%
25	64%—88%	52%—64%	64%—84%	48%—60%	48%—64%	36%—48%
30	57%—80%	47%—60%	57%—77%	43%—53%	43%—60%	33%—43%
40	50%—68%	40%—50%	50%—65%	40%—48%	37%—50%	29%—36%
50	46%—60%	36%—44%	46%—58%	36%—40%	34%—44%	26%—32%
75	39%—48%	31%—36%	37%—45%	29%—33%	28%—34%	21%—25%
100	34%—40%	26%—30%	33%—39%	25%—28%	24%—28%	19%—22%

The value of *x* (first column), one of the occurrence numbers. The intervals are given for the 95% (2 ε = 0.05) and 99% (2ε = 0.01) confidence levels. Up to *x* = 20, the exact boundaries, immediately outside the confidence interval (first significantly different values) are indicated. A star is used when none are possible. For larger values, the boundaries are given as percentages to be subtracted or added to *x*. Ricker's confidence interval characterizes the value of λ, not *y* (see Methods). The use of a flat *p* (λ) prior distribution results in the most stringent test, as expected. Although the number (*N*) of clones sampled does not appear in the expression of *p*(*y*|*x*) (Equation 1), its influence shows in the fact that the confidence interval becomes proportionally smaller as *x* (and *y*) increases (e.g., 1 → 7 has the same statistical significance as 40 → 60). For the same expression level, larger *N* will result in larger absolute values for *x* and *y*, making the detection of significant differential expression more sensitive.

$$p(y|x) = \left(\frac{N_2}{N_1}\right)^y \frac{(x+y)!}{x!y!\left(1+\frac{N_2}{N_1}\right)^{(x+y+1)}} \qquad (2)$$

Whereas Equation 1 applied to the analysis of fluctuation in counts in strictly identical experiments, Equation 2 now applies to the analysis of counts in experiments only differing by the total number of clones randomly picked up. In practice, Equation 2 will be used to analyze experiments performed on two different libraries, using different sampling sizes. As for Equation 1, small *p*(*y* | *x*) are expected to characterize the genes exhibiting regulated expression, the relative abundance of which is unlikely to be the same in the two libraries.

Comparison with Fisher's (2 × 2) Exact Test

The (2 × 2) contingency tables arising from treatment versus control experiments are traditionally analyzed with Fisher's exact test (Siegel 1956; Agresti 1996). Differential EST count data can be presented in a tabulated form so as to suggest the use of this test, as follows:

	Brain cDNA library	Liver cDNA library
Number of actin ESTs	2	11
Number of other ESTs	998	1189
Total clones sampled	1000	1200

The statistical significance according to Fisher's exact test for such a result is 4.6% (two-tail *P*-value, i.e., the probability for such a table to occur in the hypothesis that actin EST frequencies are independent of the cDNA libraries). In comparison, the *P*-value computed from the cumulative form (Equation 9, see Methods) of Equation 2 (i.e., for the relative frequency of actin ESTs to be the same in both libraries, given that at least 11 cognate ESTs are observed in the liver library after two were observed in the brain library) is 1.6%. Fisher's (2 × 2) exact test is always more conservative than our test (e.g., Fisher's *P*-value of 1.6% requires a 2 → 13 EST count transition in the above setting). Besides being too conservative, there is a more fundamental difficulty in using this test to analyze EST count data. The sampling scheme assumed by Fisher's exact test in principle requires the total number of data values in the contingency table to be fixed, as well as both the row marginal total and the column marginal totals. In our prospective experimental situation, only the column marginals (i.e., the numbers of clones sampled from each library) are fixed. The extension of Fisher's exact test to cases where only one set of marginal totals is fixed (Tocher 1950) is still controversial. In the context of the above EST counting results, there is an additional problem with the lack of homogeneity in the definition of the ''other EST'' category. This category represents different subsets of transcripts for different libraries.

The use of Fisher's (2 × 2) exact test is more natural for a different type of EST data analysis: the study of library-dependent alternative transcripts of the same gene (i.e., splice or polyadenylation variants) (D. Gautheret, O. Poirot, F. Lopez, S. Audic, and J.-M. Claverie, in prep.). Here, the results for an hypothetical gene G_1 may look as follows:

	G_1-related transcripts in brain library	G_1-related transcripts in liver library
Long-form mRNA	2	10
Short-form mRNA	8	3
Total G_1-related clones	10	13

where the alternative categories are unambiguously defined and refer to the same objects. For example, the above results constitute good evidence that G_1 is expressed in different forms in those tissues (Fisher's exact test two-tail *P*-value = 1.2%).

False Leads in the Selection of Candidate Genes

A crucial measure of the power of statistical significance tests is their rate of false alarm, that is, how often random fluctuations are expected to be mistaken for significant differences in the results. When analyzing the transcript profiles from two different libraries, a false alarm would cause a gene to be deemed differentially transcribed, whereas in fact it is not. The rate of false alarm is therefore a direct estimate of the fraction of false leads, when searching for differentially expressed genes on the basis of differences in tag counts. The rates of false alarm associated with the $P < 0.01$ and $P < 0.05$ confidence intervals listed in Table 1 have been computed by Monte-Carlo simulation on the basis of two experimental sequence tag distributions (Table 2; Fig. 1). The rate of false alarms associated with the use of Equation 1 (in fact, its cumulative form Equation 9, see Methods) is very small for genes represented by small tag counts and slowly increases for higher tag counts, without ever exceeding the selected significance level. Such good behavior validates the use of the confidence intervals (Table 1) computed from Equation 1 and Equation 9 to assess the statistical significance of variations in digital Northern data. The curves labeled ''window'' characterize the very similar behavior of a slightly less conservative derivation of the same test (see Methods, Equation 15). For comparison, Figure 1 also presents the behavior of another test, based on an inappropriate application of Ricker's confidence intervals (Ricker 1937) (see Methods).

DISCUSSION

An appropriate statistical test is now at our disposal to begin analyzing digital gene expression profiles

Table 2. Publicly Available Distributions of Sequence Tags

Tags per class	Percentage %	Number of classes
64	7.6	1
46	5.5	1
37	4.4	1
31	3.7	1
20	2.4	1
16	1.9	2
14	1.7	2
12	1.4	1
11	1.3	1
9	1.1	1
8	1.0	2
6	0.7	1
5	0.6	8
4	0.5	7
3	0.3	15
2	0.2	32
1	0.1	351

ESTs per class	Percentage %	Number of classes
22	2.2	1
21	2.1	1
12	1.2	1
9	0.9	1
8	0.8	1
7	0.7	5
6	0.6	6
5	0.5	4
4	0.4	7
3	0.3	31
2	0.2	115
1	0.1	468

(*Left*) Data from Velculescu et al. (1995): Frequency of occurrence of each of the 428 transcript species represented in 840 SAGE tags randomly generated from a 3′-directed cDNA library from human pancreas. (*Right*) Data from Okubo et al. (1992): Frequency of occurrence of each of 641 transcript species represented in 982 randomly sequenced clones from a 3′-directed cDNA library from human liver cell line HepG2.

in a more quantitative way. For example, the test can be used to determine how many genes appear regulated at various confidence levels using the data from a typical experiment (e.g., sampling a thousand clones). We analyzed the data gathered by Okubo et al. (1995) on the human promyelocytic leukemia cell line HL60 induced by dimethylsulfoxide (DMSO) or tetradecanoylphorbolacetate (TPA). Table 3 shows the 21 EST classes the occurrences of which exhibit significant variations at the 1% level. Most of the corresponding genes make biological sense in term of differentiation along the granulocyte or monocyte pathways.

This example serves to discuss a subtle point in the interpretation of the *P* values computed from Equation 1, 2, and 9. Rigorously, these equations apply to the case where a given gene (e.g., lipocortin) would have been selected for scrutiny before looking at the differences in cognate tag counts between libraries. When comparing two libraries without specifying in advance the transcripts we want to follow, and then focusing a posteriori on any of those exhibiting significant variations, the average number of expected false positive N_{false} is $N_{\mathrm{false}} = PN_{\mathrm{species}}$, where N_{species} is the number of different transcript species encountered and *p* is a given significance level. For instance, in the experiment analyzed in Table 3, N_{species} is of the order of 600 (Okubo et al. 1995). It is therefore possible that up to four ($600 \times 7 \times 10^{-3}$) out of the 21 transcript species listed in Table 3 are not truly differentially expressed.

Therefore, when two libraries are compared without prior gene selection, the use of a predetermined significance threshold is not advisable. The *P* values computed from Equation 1, 2, and 9 should simply be used to rank all observed variations by order of decreasing statistical significance (analogous to how "similarity hits" are listed after database searches). The end-users can then make their own choice about the number of candidate target genes to be retained from the top of the list, bearing in mind the corresponding number of expected false positives.

Although the present interpretation of a digital Northern focuses on the genes exhibiting the most spectacular differential expressions, there is already ample evidence that small changes can cause drastic

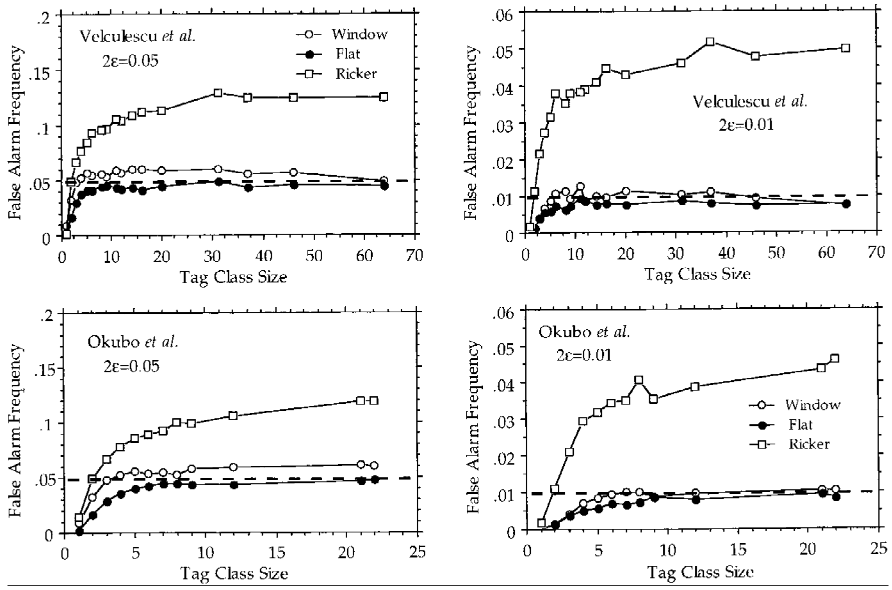


Figure 1 Rate of false alarm computed according to the confidence intervals listed in Table 1. (*Top*) Monte-Carlo simulation of the random sampling of 840 tags distributed according to the data from Velculescu et al. (1995; see Table 2). (*Bottom*) Monte-Carlo simulation of the random sampling of 982 ESTs distributed according to the data from Okubo et al. (1992; see Table 2). The frequency of false alarm was computed for two significance levels (2ϵ = 5%, *left* and 2ϵ = 1%, *right*) and plotted in function of the tag class size (from 1–64 for Velculescu et al., from 1–22 for Okubo et al.). In all cases, the rate of false alarm increases up to a plateau for larger class sizes. The test (cumulative form of Equation 1) derived from the flat $p(\lambda)$ prior shows perfect behavior with a maximal rate of false alarm always less than the significance levels (broken lines). The test (cumulative form of Equation 15) derived from the window $p(\lambda)$ prior exhibits a slightly higher rate of false alarms. Both versions of the test exhibit conservative behaviors for class size <5, with a false alarm rate even less than expected. In contrast, Ricker's confidence intervals (Equation 12) are grossly inadequate and lead to false alarm rates up to four times the significance level. Graphs are computed from the analysis of 1000 repetitions of each experiment.

effects. Disease states caused by haploinsufficiency and trisomy suggest that $2 \rightarrow 1$ or $2 \rightarrow 3$ proportional changes in expression level may be of biological significance. Table 1 shows that there is no theoretical limit to the detection of such small variations from the comparison of digital expression patterns. Simply, the sampling size has to be increased enough for the required numbers of cDNA tags to reach a significance threshold (for instance $40 \rightarrow 60$, for a confidence level of 95%).

Analog hybridization-based methods (Fodor et al. 1991; Lennon and Lehrach 1991; Gress et al. 1992; Southern et al. 1992; Guo et al. 1994; Matson et al. 1995; Nguyen et al. 1995; Schena et al. 1995; Zhao et al. 1995) are traditionally opposed to digital tag-counting methods (Okubo et al. 1992; Matsubara and Okubo 1994; Lee et al. 1995; Okubo et al. 1995; Velculescu et al. 1995) for the analysis of differential gene expression. Both types of methods are sensitive to the quality of the original messenger RNA preparation and/or cDNA libraries. Analog methods promise higher throughput, lower cost, and have the capacity of studying transcripts on a much wider scale of abundance. They are therefore expected to supersede digital methods. On the down side, however, hybridization signals are not easily reproducible, and can be affected by many unknown properties such as the cDNA library complexity, as well as clone and sequence specific features (e.g., insert size, nucleotide composition, presence of repeats, secondary structure, triple helix interaction, etc.). Therefore, the hybridization-based methods require an estimation of the dispersion of the signal associated with each clone (i.e., enough repetitions of each experiment), and multiple standardization and calibration procedures to allow the meaningful comparison of hybridization patterns obtained from various sources (tissues, cell types, etc.) or from different membranes or chips. This is far from routine and has yet to be worked out. In contrast, and thanks to the unique properties of the Poisson distribution, digital methods have the capacity of providing a quantitative assessment of differential expression without the repetition or the standardization of individual tag-counting experiments. The statistical analysis presented here provides an objective method to analyze digital transcript profile data, and adapts it to fit (1) the number of leads one wants to be followed; (2) the fraction of false clues to be tolerated; and (3) the level of modulation in gene expression considered of biological interest.

A program is available on our web site (http://igs-server.cnrs-mrs.fr) to compute the confidence

Table 3. List of ESTs Exhibiting Significant (***P*** < 0.01) Differences in Abundance in the HL60 Cell Line Induced by DMSO or TPA

EST ID	HL60	HL60 + TPA	HL60 + DMSO	Significance
418	22	10	1	3×10^{-7}
211	24	10	2	4×10^{-7}
19	8	23	2	8×10^{-7}
356	16	2	0	3×10^{-6}
380	12	1	0	6×10^{-5}
135	4	12	0	6×10^{-5}
285	14	8	1	1×10^{-4}
2015	0	11	0	2×10^{-4}
244	0	1	14	3×10^{-4}
293	13	6	1	3×10^{-4}
292	11	0	1	5×10^{-4}
650	14	5	2	5×10^{-4}
335	15	3	3	9×10^{-4}
444	10	4	1	2×10^{-3}
1674	0	8	1	4×10^{-3}
155	0	8	3	4×10^{-3}
861	6	1	0	7×10^{-3}
305	6	2	0	7×10^{-3}
1806	0	6	0	7×10^{-3}
1808	0	6	0	7×10^{-3}
1766	0	6	0	7×10^{-3}

Only the probability (computed according to Equations 7 and 8) corresponding to the most significant transition (numbers in bold) is listed (Okubo et al. 1995). The total EST numbers sampled from the HL60, HL60 + TPA and HL60 + DMSO cDNA libraries are 845, 845, and 1058, respectively. ESTs 418, 211, 356, 285, 293, 292, 650, 335, 444, 861, 305 corresponding to ribosomal proteins, and EST 380, a tag to an unkown gene, exhibit a marked reduction of expression level in the DMSO- and/or TPA-induced differentiated states. In constrast, ESTs 135 (ferritin), 2015 (LD78/macrophage inflammatory protein), 1674 (methionine adenosyl-transferase), 155 (thymosin β-4), 1806 (lipocortin), 1808 (thymosin β-10), and 1766 (a metallothionein) appear more abundant in the TPA-induced state, also highly enriched in EST 19 (the ubiquitous elongation factor 1-α). β-Actin (EST 244), is the only markedly increased tag in the DMSO-induced state. EST numbers, abundance data, and protein assignments are from the ''body map'' public expression data repository at http://www.imcb.osaka-u.ac.jp (K. Okubo and K. Matsubara).

intervals corresponding to arbitrary significance levels and sampling size N_1 and N_2.

METHODS

Let us denote $p(x)$ the probability to observe x sequence tags of the same gene (i.e., from the 3′ end of the same transcript) when N cDNA clones are picked randomly. For each transcript representing a small (i.e., less than 5%) fraction of the library and $N \geqslant 1000$, $p(x)$ will closely follow the Poisson distribution:

$$p(x) = \frac{e^{-\lambda}\lambda^{x}}{x!} \tag{3}$$

where λ is the actual (albeit unknown) number of transcript of this type per N clones in the library. If we duplicate this experiment (i.e., once again randomly pick N clones of the same library and generate sequence tags), we will now observe y occurrences of the same transcript. What is the probability of the various y values? An approximate solution consists in using x as the maximum likelihood estimate for λ and compute the probability for y occurrences given a Poisson distribution of mean $\lambda = x$:

$$p(y|x) = \frac{e^{-x}x^{y}}{y!} \tag{4}$$

Equation 4 is not symmetrical in x and y. This is an obvious flaw as the probability should not depend on which of the x or y values were observed first. $p(y \mid x) = p(x \mid y)$ should hold provided that an equal number N of clones is sampled in both experiments. Equation 4 is not the correct formula, because we have not yet taken into account the fluctuation of x around the unknown mean λ. To account for the fact that the actual value of λ is unknown, we have to integrate Equation 4 over all possible λ values:

$$p(y|x) = \int_{0}^{\infty} d\lambda p(d = \lambda|x)p(y|d = \lambda) \tag{5}$$

$p(d = \lambda \mid x)$ in Equation 5 is the probability that the actual abundance of a given transcript is λ given that *x* occurrences of a cognate tag have been observed in one experiment. The second term in the integral is the probability of drawing *y* occurrences given a Poisson distribution of mean λ:

$$p(y|d=\lambda) = \frac{e^{-\lambda}\lambda^{y}}{y!} \qquad (6)$$

Using Bayes' theorem $p(d = \lambda \mid x)$ can be written as

$$p(d=\lambda|x) = \frac{p(x|d=\lambda)p(d=\lambda)}{\int_0^\infty d\lambda'\ p(x|d=\lambda')\ p(d=\lambda')} \qquad (7)$$

To evaluate Equation 7, we need to define the prior distribution $p(d = \lambda)$. The least constrained hypothesis (i.e., with the least information content), is to attribute an equal a priori probability to all λ values in the [0, ∞] range. Incorporating such a flat prior in Equation 5 leads to

$$p(y|x) = \frac{1}{x!y!}\int_0^\infty d\lambda e^{-2\lambda}\lambda^{(x+y)} \qquad (8)$$

From the definition of the Γ function for integer arguments we observe that

$$\int_0^\infty d\lambda e^{-2\lambda}\lambda^{(x+y)} = \frac{(x+y)!}{2^{(x+y+1)}}$$

and finally obtain the expression given in Results:

$$p(y|x) = \frac{(x+y)!}{x!y!2^{(x+y+1)}} \qquad (1)$$

This equation can be used in a wide variety of experimental situations. Equation 1 defines the probability of observing *x* and *y* occurrences of the same rare event in duplicated experiments, regardless of the detailed probability distribution of those events among the set of possible outcomes. In particular, in the context of transcription profiles, $p(y \mid x)$ can be evaluated regardless of the distribution of each transcript (provided it is rare) within a cDNA library.

To compute the confidence intervals listed in Table 1, we made use of the cumulative distributions:

$$C(y \leqslant y_{\min}|x) = \sum_{y=0}^{y \leqslant y_{\min}} p(y|x) \qquad (9a)$$

$$D(y \geqslant y_{\max}|x) = \sum_{y=y_{\max}}^{\infty} p(y|x) \qquad (9b)$$

These equations allow the computation of an interval $[y_{\min}, y_{\max}]_\epsilon$ such as $C(y \leqslant y_{\min} \mid x) \leqslant \epsilon$ and $D(y \geqslant y_{\max} \mid x) \leqslant \epsilon$. Given that an event is observed *x* times in one experiment, the number *y* of occurrences of this event in a duplicate experiment is expected to fall within the interval $[y_{\min}, y_{\max}]_\epsilon$ with a probability of 1–2ϵ. Equation 9, a and b, can therefore serve as a significance test when comparing, for instance, the results of sampling *N* clones from two different libraries. For 2ϵ small (e.g., 5% or less), *y* values falling outside the $[y_{\min}, y_{\max}]_\epsilon$ interval correspond to $p(y \mid x) << 1$, and point out significant differences between the two experiments. They should include differentially expressed genes, for example, for which λ is different in the two libraries.

Generalization to Different Sampling Sizes

When different numbers of clones N_1 and N_2 are sequenced from the same library, Equation 5 becomes

$$p(y|x) = \int_0^\infty d\lambda_2 \int_0^\infty d\lambda_1 p(d_1=\lambda_1|x)p(y|d_2=\lambda_2)\delta\left(\lambda_2 - \frac{N_2}{N_1}\lambda_1\right) \qquad (10)$$

where the two abundance values λ_1 and λ_2 are forced in the same ratio as N_1 and N_2. Using the same bayesian argument as before (Equation 7) leads to

$$p(y|x) = \frac{1}{x!y!}\left(\frac{N_2}{N_1}\right)^{y}\int_0^\infty d\lambda_1 e^{-\lambda_1\left(1+\frac{N_2}{N_1}\right)}\lambda_1^{(x+y)} \qquad (11)$$

the last integral is simply

$$\frac{(x+y)!}{\left(1+\frac{N_2}{N_1}\right)^{(x+y+1)}}$$

leading to the formula presented in the Results section:

$$p(y|x) = \left(\frac{N_2}{N_1}\right)^{y}\frac{(x+y)!}{x!y!\left(1+\frac{N_2}{N_1}\right)^{(x+y+1)}} \qquad (2)$$

Ricker's Confidence Interval

The confidence interval computed from Equation 1 (and its cumulative form, Equation 9, a and b) is different from one introduced previously by Ricker (1937) although, at first, the two may appear to be related.

Given *x* occurrences of a sequence tag, Ricker's formula defines a confidence interval $[\lambda_{\min}, \lambda_{\max}]_x$ for λ (again the actual number of transcripts of this type per *N* clones in the library) such as

$$p(k \leqslant x) = \sum_{k=0}^{x}\frac{e^{-\lambda_{\max}}\lambda_{\max}{}^{k}}{k!} \leqslant \frac{\alpha}{2} \qquad (12a)$$

and

$$p(k \geqslant x) = \sum_{k=x}^{\infty}\frac{e^{-\lambda_{\min}}\lambda_{\min}{}^{k}}{k!} \leqslant \frac{\alpha}{2} \qquad (12b)$$

where α is typically 5% or 1%. Ricker's confidence intervals for various values of *x* are given in Table 1. Those intervals are close to those computed from Equation 1, but delineate the range of likely λ values, not *y* (the number of occurrences of the same event in a duplicated experiment). It is possible for *x* and *y* to fall outside each other's Ricker's confidence interval $[\lambda_{\min}, \lambda_{\max}]$, while still being nonsignificant fluctuations around the same λ value. The confidence intervals computed from Equation 12, a and b, are therefore too narrow to properly define significant discrepancies between *x* and *y*. The false alarm rate associated with the use of Ricker's confidence intervals is too high (Fig. 1).

However, an interesting use of Equation 12, a and b, is the estimation of the range of possible frequencies $[\lambda_{\min}, \lambda_{\max}]_x = 0$ for cDNAs not yet encountered after picking *N* clones. For example, the 95% confidence interval is given by:

$$0 < N\lambda < 3.7 \qquad (13)$$

That is, the abundance of a cDNA not picked up among one thousand clones is unlikely (5% chance) to be larger than 3.7/1000.

Influence of the Prior Distribution

In the bayesian context, it is prudent to assess the influence of the prior hypothesis used to derive Equation 1 and Equation 2. The flat $p(\lambda)$ prior allowing equiprobable λ values in the $[0, \infty]$ range might appear too broad and unrealistic. Nevertheless, it is the most intuitively neutral distribution one can use. The quick convergence of the Poisson distribution rends the contribution of extreme λ values negligible as soon as $|\lambda - x|$ or $|\lambda - y|$ increase. To verify this point, more reasonable distributions for $p(\lambda)$ can be constructed by confining the accessible λ values within a window $[\lambda_{\mathrm{min}}, \lambda_{\mathrm{max}}]_x$ centered around the already observed value x. Such a window can for instance be Ricker's confidence intervals as defined in the previous section (Equation 12, a and b). We then confine the only permitted values of λ to be in this interval, with an equal probability; therefore,

$$p(d = \lambda) = \frac{H(\lambda - \lambda_{\mathrm{min}}) \times [1 - H(\lambda - \lambda_{\mathrm{max}})]}{\lambda_{\mathrm{max}} - \lambda_{\mathrm{min}}} \qquad (14)$$

where H denotes the Heaviside function, the value of which is 1 for positive argument and 0 otherwise. Equation 1 then becomes

$$p(y|x) = \frac{F_{2\lambda_{\mathrm{min}},2\lambda_{\mathrm{max}}}(x+y)}{F_{\lambda_{\mathrm{min}},\, \lambda_{\mathrm{max}}(x)y!2^{(x+y+1)}}} \qquad (15)$$

with

$$F_{\lambda_{\mathrm{min}},\, \lambda_{\mathrm{max}}}(x) = \int_{\lambda_{min}}^{\lambda_{max}} d\lambda e^{-\lambda}\lambda^x \qquad (16)$$

When $\lambda_{\mathrm{min}} \rightarrow 0$ and $\lambda_{\mathrm{max}} \rightarrow \infty$, we recover the initial Equation 1. We note in passing that the other limit, $\lambda_{\mathrm{min}} = \lambda_{\mathrm{max}} = x$, corresponds to the most stringent ''Dirac prior'':

$$p(d = \lambda|x) = \delta(x - \lambda) \qquad (17)$$

forcing $\lambda = x$ and turning $p(y \mid x)$ into the simple Poisson distribution (Equation 4).

The confidence intervals for the usual 1% and 5% significance levels are given in Table 1 for both the flat and the window prior $p(d = \lambda)$. There is little difference, with the test derived from using a flat prior being a bit more conservative, as expected. On the down side, Figure 1 shows that the test derived from the window $p(\lambda)$ prior gives rise to a higher rate of false alarm.

ACKNOWLEDGMENTS

We thank Drs. J. Weissenbach, D. Gautheret, R. Ewing, and C. Abergel for critically reading the manuscript. This work was sponsored by a collaborative research grant from Incyte Pharmaceuticals, Inc.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked ''advertisement'' in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Aaronson, J.S., B. Eckman, R.A. Blevins, J.A. Borkowski, J. Myerson, S. Imran, and K.O. Elliston. 1996. Toward the development of a gene index to the human genome: An assessment of the nature of high-throughput EST sequence data. *Genome Res.* 6: 829–845.

Adams, M.D. 1996. Progress towards a complete set of human genes. In *Genomes, molecular biology and drug discovery* (ed. M.J. Browne and P.L. Thurby). Academic Press, London, UK.

Adams, M.D., J.M. Kelley, J.D. Gocayne, M. Dubnick, M.H. Polymeropoulos, H. Xiao, C.R. Merril, A. Wu, B. Olde, R.F. Moreno et al. 1991. Complementary DNA sequencing: Expressed sequence tags and human genome project. *Science* 252: 1651–1656.

Adams, M.D., M. Dubnick, A.R. Kerlavage, R. Moreno, J.M. Kelley, T.R. Utterback, J.W. Nagle, C. Fields, and J.C. Venter. 1992. Sequence identification of 2,375 human brain genes. *Nature* 355: 632–634.

Adams, M.D., A.R. Kerlavage, R.D. Fleischmann, R.A. Fuldner, C.J. Bult, N.H. Lee, E.F. Kirkness, K.G. Weinstock, J.D. Gocayne, O. White et al. 1995. Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. (The Genome Directory, Suppl.) *Nature* 377: 3–174.

Agresti, A. 1996. *An introduction to categorical data analysis.* John Wiley, New York, NY.

Bains, W. 1996. Virtually sequenced: The next genomic generation. *Nature Biotechnol.* 14: 711–713.

Boguski, M.S. and G.D. Schuler. 1995. ESTablishing a human transcript map [News]. *Nature Genet.* 10: 369–371.

Editorial. 1996. Capitalizing on the genome. *Nature Genet.* 13: 1–5.

Fodor, S.P.A., J.L. Read, M.C. Pirrung, L. Stryer, A.T. Lu, and D. Solas. 1991. Light-directed, spatially addressable parallel chemical synthesis. *Science* 251: 467–470.

Gress, T.M., J.D. Hoheisel, G.G. Lennon, G. Zehetner, and H. Lehrach. 1992. Hybridization fingerprinting of high-density cDNA-library arrays with cDNA pools derived from whole tissues. *Mamm. Genome* 3: 609–619.

Guo, Z., R.A. Guilfoyle, A.J. Thiel, R. Wang, and L.M. Smith. 1994. Direct fluorescence analysis of genetic polymorphisms by hybridization with oligonucleotides arrays on glass supports. *Nucleic Acids. Res.* 22: 5456–5465.

Khan, A.S., A.S. Wilcox, M.H. Polymeropoulos, J.A. Hopkins, T.J. Stevens, M. Robinson, A.K. Orpana, and J.M. Sikela. 1992. Single pass sequencing and physical and genetic mapping of human brain cDNAs [see Comments]. *Nature Genet.* 2: 180–185.

Kuska, B. 1996. Cancer genome anatomy project set for take-off. *J. Natl. Cancer Inst.* 88: 1801–1803.

Hillier, L., G. Lennon, M. Becker, M.F. Bonaldo, B. Chiapelli, S. Chissoe, N. Dietrich, T. DuBuque, A. Favello, W. Gish et al. 1996. Generation and analysis of 280,000 human expressed sequence tags. *Genome Res.* 6: 807–828.

Lee, N.H., K.G. Weinstock, E.F. Kirkness, J.A. Earle-Hugues, R.A. Fuldner, S. Marmaros, A. Glodek, J.D. Gocayne, M.D. Adams, A.R. Kerlavage et al. 1995. Comparative expressed-tag analysis of differential gene expression profiles in PC-12 cells before and after nerve growth factor treatment. *Proc. Natl. Acad. Sci.* 92: 8303–8307.

Lennon, G.G. and H. Lehrach. 1991. Hybridization analyses of arrayed cDNA libraries. *Trends Genet.* 7: 314–317.

Matson, R.S., J. Rampal, S.L. Pentoney Jr., P.D. Anderson, and P. Coassin. 1995. Biopolymer synthesis on polypropylene supports: Oligonucleotide arrays. *Anal. Biochem.* 224: 110–116.

Matsubara, K. and K. Okubo. 1994. Identification of new genes by systematic analysis of cDNAs and database construction. *Curr. Opin. Biotechnol.* 4: 672–677.

Nguyen, C., D. Rocha, S. Granjeaud, M. Baldit, K. Bernard, P. Naquet, and B.R. Jordan. 1995. Differential gene expression in the murine thymus assayed by quantitative hybridization of arrayed cDNA clones. *Genomics* 29: 207–216.

Nowak, R. 1995. Entering the postgenome era. *Science* 270: 368–369.

O'Brien, C. 1997. Cancer genome anatomy project launched. *Mol. Med. Today* 3: 94.

Okubo, K., N. Hori, R. Matoba, T. Niiyama, A. Fukushima, Y. Kojima, and K. Matsubara. 1992. Large scale cDNA sequencing for analysis of quantitative and qualitative aspects of gene expression. *Nature Genet.* 2: 173–179.

Okubo, K., K. Itoh, A. Fukushima, J. Yoshii, and K. Matsubara. 1995. Monitoring cell physiology by expression profiles and discovering cell type-specific genes by compiled expression profiles. *Genomics* 30: 178–186.

Ricker, W.E. 1937. The concept of confidence or fiducial limits applied to the Poisson frequency distribution. *J. Am. Statist. Assoc.* 32: 349–357.

Siegel, S. 1956. *Nonparametric methods for the behavioral sciences.* McGraw-Hill, New York, NY.

Schena, M., D. Shalon, R.W. Davis, and P.O. Brown. 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 270: 467–470.

Southern, E.M., U. Maskos, and J.K. Elder. 1992. Analyzing and comparing nucleic acid sequences by hybridization to arrays of oligonucleotides: Evaluation using experimental models. *Genomics* 13: 1008–1017.

Tocher, K.D. 1950. Extension of the Neyman-Pearson theory of tests to discontinuous variates. *Biometrika* 37: 130–144.

Velculescu, V.E., L. Zhang, B. Vogelstein, and K.W. Kinzler. 1995. Serial analysis of gene expression. *Science* 270: 484–487.

Wilcox, A.S, A.S. Khan, J.A. Hopkins, and J.M. Sikela. 1991. Use of 3′ untranslated sequences of human cDNAs for rapid chromosome assignment and conversion to STSs: Implications for an expression map of the genome. *Nucleic Acids Res.* 19: 1837–1843.

Zhao, N., H. Hashida, N. Takahashi, Y. Misumi, and Y. Sakaki. 1995. High-density cDNA filter analysis: A novel approach for large-scale, quantitative analysis of gene expression. *Gene* 156: 207–213.

Received March 12, 1997; accepted in revised form August 22, 1997.