© 1994 Oxford University Press Nucleic Acids Research, 1994, Vol. 22, No. 5 799–806

Evaluation of quantitative variation in gene expression

Emmanuel Spanakis* and Danièle Brouty-Boyé
Institut d'Oncologie Cellulaire et Moléculaire Humaine, 129 rue de Stalingrad, F-93000 Bobigny, France

Received November 18, 1993; Revised and Accepted January 28, 1994

ABSTRACT

We investigate the behaviour of the gene-expression rate as a statistical variable using autoradiographic data for 39 transcripts from a heterogeneous set of 80 breast-tissue cultures. Despite standardization, the data distributions of all transcripts showed intervals of normality and intervals of systematic departure from normality which most frequently resulted in a significant skewness and/or kurtosis. Non-normal shapes are attributed to modulation of gene expression. This statistical particularity creates difficulties in the evaluation of differences among specimens. Using classical parametric and non-parametric procedures for normal and non-normal variation, respectively, we demonstrate that large differences in optical density are neither necessary nor sufficient for associating expression rates with biological factors. The transcripts coding for the metalloprotease stromelysin-3 (ST3) and for the receptor to insulin-like growth factors (IGFR) are used as examples and their variation is presented in detail. *ST3* expression appeared to be specifically associated with mammary stroma fibroblasts derived from post-radiation fibrosis lesions. *IGFR* was expressed at higher rates in mammary gland and skin fibroblasts than in mammary epithelial cells and was subject to frequent and strong modulation.

INTRODUCTION

The identification of causes and effects of modulation of gene expression is a common task in diverse fields of biological research. In any field of research, progress depends not only on the quantity and quality of available data and but also on adequate evaluation. By 'evaluation' we mean an estimation of the probability that an observation is due to chance. The sequence-specificity of transcript detection techniques (1) has been a major concern of molecular biologists. The most primitive technique, a 'dot' or 'slot' blot analysis, is to load a certain amount of an RNA extract on a filter and allow a probe to hybridize with, or attach to, anything it can, i.e. exactly homologous transcripts, other partially homologous transcripts, contaminant DNA and other impurities. In northern blot analysis, the sequence-specificity problem is solved by an electrophoresis prior to blotting. The transcript of interest migrates away from any related transcripts and any impurities and only signals that correspond to molecules of a recognized size are considered. But partially homologous transcripts may also have very similar sizes. So, technical development towards increased specificity has always been welcomed. However, in transcript detection technology specificity increases at the expense of sensitivity. The sensitivity of a northern blot is practically lower than that of a slot blot because some molecules are not intact or are differently spliced and are not taken into account. Also, because some molecules are not successfully charged on the filter but remain in the gel. For this reason, a northern analysis requires larger specimen volumes than a dot analysis. Every step introduced between RNA extraction and autoradiography (mRNA purification, RNase treatment, and so on) yields less than 100% of the 'true' transcript and may so reduce the sensitivity of detection as well as add artificial variation between specimens. Losses may add up to a false negative result. The probability that a negative score is artificial increases with the complexity of the protocol. Of course, techniques that employ sequence-specific amplification by polymerase chain reaction are very specific and extremely sensitive, but are of little use for quantification. Firstly, because amplification introduces enormous artificial variation between specimens and, secondly, because the minute concentrations of transcripts that can be so detected (few molecules in a whole population of cells) probably have analogous, minor phenotypic importance. Complex methods are also less efficient with respect to the number of specimens one can process at a time. As for the quality of the data, it would seem that we have to deal with false positive or negative scores anyway; or have we?

The idea behind a recently proposed method (2) is to keep the power and the simplicity of a slot blot and to solve the specificity problem mathematically. A multi-gene slot blot analysis combined with a principal components analysis is not merely a very productive method but is, also, a very sensitive and an extremely specific one. The total signal (of optical nature, or of any nature) is considered as the sum of a gene-specific part and a non-specific part. The non-specific signal is, by definition, independent of the sequence of the probe. Whether this signal is due to non-specific attachment of the probes to the loaded material or to variation in the overall transcription rate, it may be accurately calculated and removed from the data using simple arithmetics. The returned values, referred to as specific signals, are standardized measures of gene-, and cell-type-specific modulation.

Like any measure of gene expression, and any measurable variable, specific-modulation estimates are subject to experimental

*To whom correspondence should be addressed error. There are, now, statistical procedures for evaluating practically any type of data but statistics are desperately missing from gene expression literature. Quantitative differences in the intensity of autoradiographic signals between specimens are almost always evaluated by magnitude and intuition, and explanatory models are too often based on single observations. References 3 through 6 provide some examples of this serious methodological negligence; by glancing through any journal reporting quantitative autoradiographs the reader will find many more examples from his/her own field. Estimation, comparison and explanation of gene-expression rates are particularly liable to experimental error not only because they rely upon complex measuring techniques but also because gene expression is by nature a complex phenomenon. There are several possible hypotheses about gene expression rates. (i) A gene is expressed at a constant rate; any variation of the experimental estimates of this rate is artificial and such estimates should be randomly distributed around one theoretical mean. (ii) A gene is essential, or non-essential, but is modulated and can have two possible states (expression / stimulation, or expression / inhibition), (iii) three states (inhibition, normal expression, stimulation) or (iv) a continuously variable rate of expression, between zero and a maximum. According to each hypothesis there should be one, two, three or an infinite number of theoretical mean expression-rates. Although the last hypothesis would seem, today, to be generally valid for most genes in nature, for a particular gene in a particular biological sample all the four hypotheses have equal chances to be rejected and what is true for one gene may not be true for another. Testing these hypotheses is, in principle, simple. One may count the number of significantly different group-means in the sample. However, before computing means and standard errors on everything in sight one must examine how the data are distributed. The validity of any statistical comparison depends primarily on the validity of the assumptions about the distributions being compared. The aim of this article is to demonstrate potential artifacts associated with the method of evaluation of quantitative autoradiographic data and to point out the necessity for a statistical theory of gene expression. The examples are drawn from our own field of research, i.e. breast cancer.

Breast cancer is a complex disease, one of the best studied cancers and, still, a very poorly understood one. Abnormal growth of tumoral epithelial cells is apparently supported by an also abnormal micro-environment (stroma) consisting of fibroblasts, vascular elements and extracellular matrix. There is substantial evidence (7–11), and good theoretical reason to believe, that epithelial and fibroblastic cells communicate by means of differential gene expression in order to establish normal function of the mammary gland or to grow into a tumour. We are, therefore, studying the expression of genes that are generally related to cell growth, differentiation and cancer in epithelial cells and stromal fibroblasts with various pathological backgrounds, under various physiological conditions, in order to select those genes that would show significant cell-, and gene-specific modulation. Out of 39 sequences, probed in 80 cell cultures, the transcript coding for the metalloprotease stromelysin-3 (ST3; 12) presented the smallest specific variation, and the transcript coding for the receptor to insulin-like growth factors (IGFR) varied very widely (2). We demonstrate that detection of a 'strong signal', even if this is proved to be gene-specific, is neither sufficient nor necessary for associating a gene with a phenotype. It all depends on data distributions.

MATERIALS AND METHODS

Cells and transcript detection

The cells, the culture treatment with cholera toxin (CT) and/or 12-O-tetradecanoylphorbol-13-acetate (TPA) as well as the RNA quantification method, from extraction to densitometry, have been described (2). The specific signal of transcript i in specimen j (SS_{ij}) was calculated as

$$SS_{ij} = [(d_{ij} - m_i) / s_i] - PC_j \text{ (equation 1)}$$

where d is the optical *d*ensity of the transcript in the specimen, m and s are, respectively, the *m*ean and the *s*tandard deviation of the optical densities of the transcript in 80 specimens, and PC is the first-principal-component score of the specimen, calculated by principal components analysis of 39 transcripts (2). The term in brackets is the z-value of the observed optical density of probe i in specimen j whereas the PC_j is the z-value of the expected non-specific signal. The latter represents the cumulative effects of variation in the amount and purity of the loaded material, in the transcriptional activity of the cells and in any conceivable factor that may influence densities non-specifically. The unit of the specific signal scale is the standard deviation of raw optical densities. ST3 was probed with the *e* *Eco*RI 1.7kb fragment of the pBSIISK*ZIV* plasmid (13) kindly provided by Dr P Basset. The *Eco*RI 0.7kb fragment of the p*IGF-I-R*.8 plasmid was used for IGFR detection; this, and all other probes used in this study were obtained from the American Tissue Culture Collection (Rockville, Maryland, USA).

Statistics

The Kolmogorov–Smirnov one-sample, two-tail test was employed for comparing the specific-signal distributions to a standard normal distribution; the reported Lilliefors probabilities refer to the shape of a distribution and are independent of its location or scale. The difference of a mean from zero was evaluated by the one-sample, two-tail t-test. Stepwise analysis of variance, without covariates, was performed as described (2). Differences between group means were evaluated by the Bonferroni pairwise comparison procedure (particularly recommended for comparisons among few groups) when the analysis of variance resulted in a significant ($P < 0.05$) F statistic. Commercial statistical software (SYSTAT version 5.2; SYSTAT Inc, Evanston IL, USA) was used for all computations.

RESULTS

Normal and non-normal distributions of specific signals

The statistical procedures for evaluating differences between means, and the mean itself as a descriptive statistic, are appropriate only to normal distributions. The assumption of normality can be tested by fitting a normal-distribution function to the data and examining the linearity of the observed values with the expected values in a so-called normal probability plot. If the data are normally distributed, then the plotted values should fall on an approximately straight line. The skewness and kurtosis values provide additional, quantitative information about the departure of the distribution from normality. Figure 1 shows normal probability plots for the observed ST3 and IGFR specific signals. The distribution of ST3-specific signals in 74 out of 80 specimens was almost perfectly normal. The three highest values were abnormally high and the three lowest values were abnormally low (Figure 1a). It should be pointed out that the

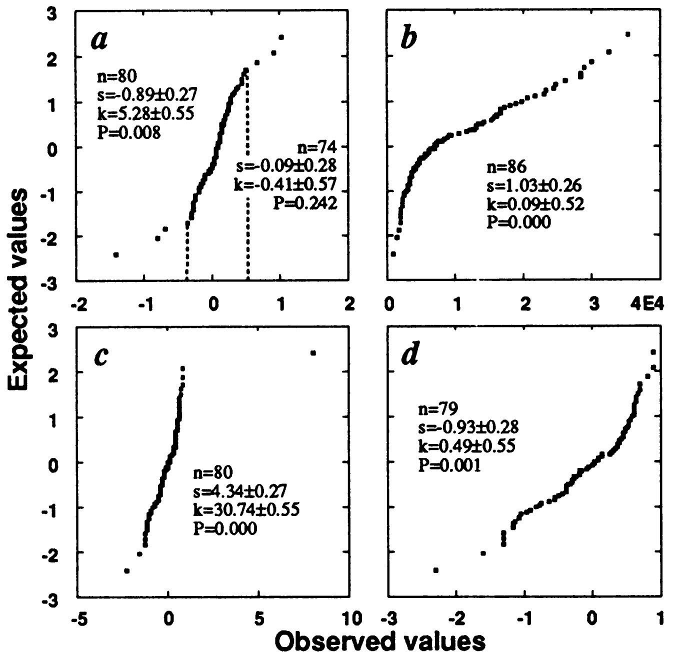

Figure 1. Normal probability plots and descriptive statistics of (**a**) ST3-specific signals, (**b**) ST3 optical densities, (**c**) IGFR-specific signals including or (**d**) excluding an extreme positive value. n = number of observations; s = skewness with √(6/n) standard error; k = kurtosis with √(24/n) standard error; P is the two-tail probability that the observed distribution, after Lilliefors standardization, is a normal one calculated by the Kolmogorov–Smirnov one-sample procedure using a standard normal distribution. The statistics on the left side of curve *a* refer to the entire sample, and those on the right, to the values in the interval of ±0.5.

specific-signal transformation always improved normality; optical densities were a lot less normally distributed (Figure 1b). This is to say that departure from normality was not at all an artefact of equation 1 but seems to be a real problem associated with any autoradiographic data and, as we argue, with the very nature of gene expression. The distribution of IGFR presented severe departure from normality. This was mainly due to an extremely abnormal positive value (Figure 1c) and also to a considerable irregularity in the distribution's main body (Figure 1d). Typically, the distributions of all the 39 examined sequences presented intervals of normality and intervals of departure from normality which most frequently resulted in a significant value for total skewness and/or kurtosis and in a Lilliefors probability below 0.05. Examples of spectacular systematic departures from normality are shown in Figure 2.

Evaluation of normal variation

The variance may be partitioned and group means may safely be compared if the distributions remain normal within groups. Concerning the ST3 distribution in mammary tissues, these conditions were met when the 6 extreme values were excluded and the data were split by tissue and pathology. Analysis of variance showed no significant differences between skin and mammary gland, nor between mammary epithelial and stromal cells. The only significant difference was between the non-pathological stroma and the post-radiation fibrosis groups (P=0.002). This difference corresponded to some 20% of the total variance (R^2=0.196). As shown in Figure 3, only the

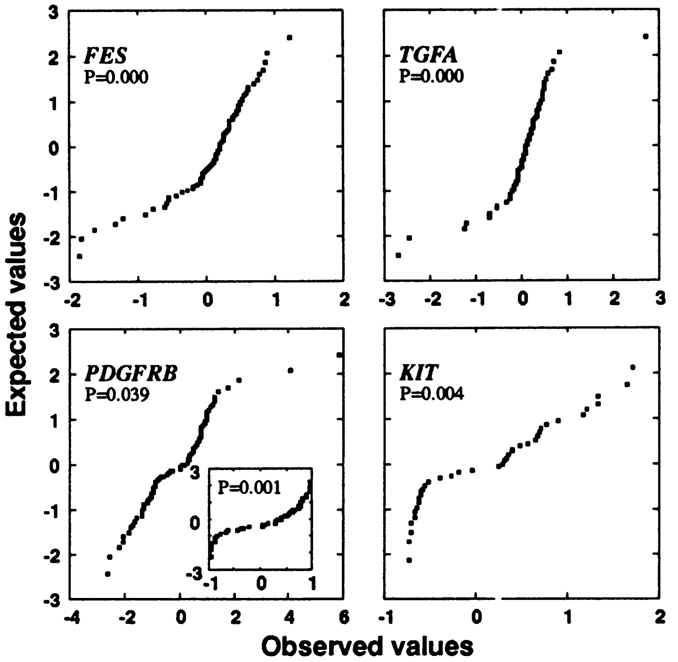

Figure 2. Examples of transcript distributions presenting systematic departure from normality in 80 specimens (FES, TGFA, PDGFRB) or in 39 specimens (KIT). P is the Lilliefors probability that the observed distribution is normal (as in Figure 1). The distribution of PDGFRB values around zero (magnified in the window) shows significant negative kurtosis.

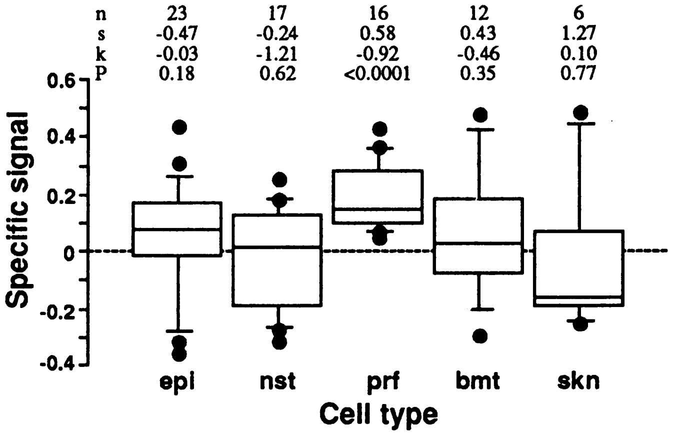

Figure 3. Box-plots describing the distributions of ST3-specific signals between 0.5 and −0.5 in epithelial cells (epi), non-pathological stromal fibroblasts including fibroblasts from tissues adjacent to tumours (nst), in fibroblasts from post-radiation fibrosis lesions (prf), in fibroblasts from benign or malignant tumours (bmt) and in skin fibroblasts (skn). Sub-groups were pooled only if their distributions were normal and their means not significantly different. The plots show the median (central horizontal line in the box), the upper and lower quartiles (the box), the 10th and 90th percentiles (the bars) and extreme values. The number of observations (n) and the skewness (s) and kurtosis (k) of the distributions within groups are also shown. P is the two-tail probability that the group's mean is zero, calculated by the one-sample t-test. A specimen of doubtful histopathology was excluded from this analysis.

fibrosis group mean departed significantly from 0. In skin fibroblasts, however, the ST3 data presented a significant departure from normality even when an extreme negative value

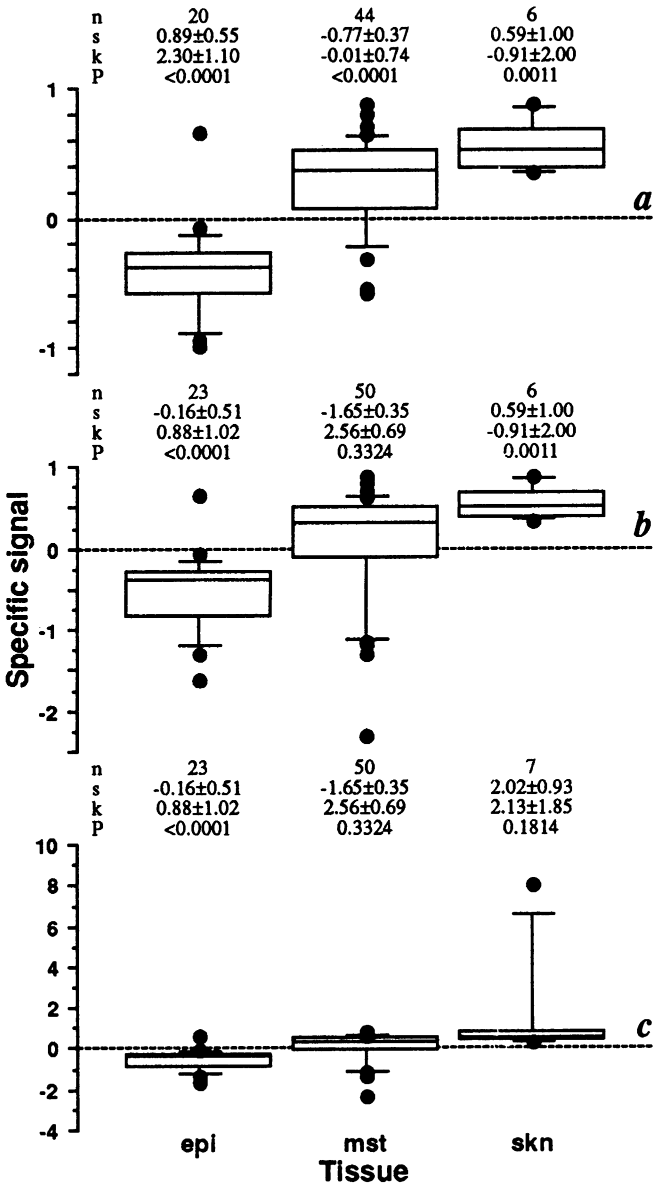

Figure 4. Box plots describing the distributions of IGFR-specific signals (**a**) between 1 and −1, (**b**) excluding only one extreme positive observation or (**c**) without range restriction, in mammary epithelial (epi) and stromal cells (mst; all pathology groups pooled) and in skin fibroblasts (skn). The box-plot presentation, the descriptive statistics and the probability that the group-mean is zero are as in the legend of Figure 3; standard errors for skewness and kurtosis are calculated as √(6/n) and √(24/n), respectively.

was removed. For this reason, the above comparisons of the skin-group mean to other group means, or to zero, are suspect.

The distribution of IGFR is split by tissue in Figure 4. In the interval of ± 1 (Figure 4a) the data remained normal within each tissue-group and all the three means were significantly different from zero. Therefore, the usual densities of IGFR were higher in the fibroblasts, and lower in epithelial cells, than 38 control genes, together, would predict. The difference btween fibroblasts and epithelial cells was significant ($P<0.001$, $R^2=0.502$) but that between skin and stromal fibroblasts was not significant.

Table 1. Specifications of cultures presenting extreme modulation of ST3

Value	Group	Treatment
1.0	n	none
0.9	b	none
0.7	n	TPA
−0.7	c	CT+TPA
−0.8	c	TPA
−1.4	s	CT+TPA

The groups (with the number of cultures in the group) are n=non-pathological stroma (13), b=benign tumour (10), c=carcinoma (4) and s=skin (7).

Statistical effects of non-normal values

When 8 extreme negative IGFR values were considered, the skewness of the distribution in the fibrosis group increased in absolute value and the mean was no longer significantly greater than zero (Figure 4b). Expression still appeared to be specifically stimulated in the skin-cell cultures and inhibited in the epithelial lines. The difference between epithelial cells and fibroblasts was still significant ($P<0.001$), and that between skin and gland fibroblasts, non-significant. The reduction of a positive mean to zero when some additional negative values are considered, as in the case of mammary fibroblasts, might seem fair. When the extreme positive value belonging to CT and TPA treated skin fibroblasts was added in the analysis (Figure 4c), the skin-data distribution departed from normality and the mean was no longer significantly higher than zero. This reduction of a positive mean to zero by an additional value which is actually higher than the mean itself is obviously due to a well known statistical artefact: means of non-normal distributions are nonsense.

Evaluation of non-normally distributed data

Non-normally distributed data cannot be used for computing and comparing means, but splitting into groups may result in normal distributions within groups. Extreme non-normal values can be examined individually. Table 1 presents the specifications of the cultures that gave extreme ST3 signals. The probability of drawing by chance one of n identical items when 3 items are drawn from a pool of 80 is $3n/80$; the probability of drawing a second one is $[3n/80\times2(n-1)/79]$. We may, thus, compute the probability that one of the three abnormal values on either side of zero belonged to a benign-tumor extract ($P=0.375$) or a skin-cell extract ($P=0.263$), and that the other two extreme values belonged to non-pathological stroma extracts ($P=0.148$) or to carcinoma stroma extracts ($P=0.011$), simply by chance. These probabilities, and the fact that the extreme values were not really very far from the expected 'normal' ones, would suggest that the discrepancies from normality were rather erroneous (i.e. not explained by tissue and pathology contrasts alone), with the exception of the two ductal carcinoma cultures in which TPA seems to have had a real inhibitory effect on ST3. The probability of the extreme positive IGFR value being observed by chance in a skin fibroblast culture was 0.086. This low probability, the very large absolute value of this observation and the significantly positive mean of IGFR in other cultures of the skin strain, considered together, would indeed suggest that IGFR was up-regulated in these cells. Had we not examined the distribution of the data, however, we would have no statistical support to suggest this, because the mean specific signal in this strain, from 7 cultures, would not be significantly different from

Table 2. Common errors of evaluation due to inappropriate analysis or experimental design

Analysis	Defects	Suggestions
Parametric comparisons of non-normal distributions	miss real differences; may create artificial differences	use a normalizing transformation; omit non-normal data; use non-parametric statistics
Non-parametric comparisons of normal distributions	miss fine quantitative differences	try a parametric analysis first
binary transformation of continuous data	creates artificial stability or artificially abrupt differences	useful when detection is difficult; disastrous in between-gene comparisons; assign non-parametric statistics
comparisons to a single specimen	have limited explanatory power	use one-sample t-test
comparisons between single specimens	have no explanatory power	evaluation not possible

zero and because the probability that the one 'really' positive value belonged to a skin specimen by chance was higher than the conventional level of significance. The absolute value itself provides information about the strength of the presumed modulation, compared to usual variation, but this information is insufficient for attributing the modulation to skin-specific factors. In the example of ST3, we have no statistical evidence for the biological importance of the 3 highest specific signals whereas we do have evidence for the importance of signals with lower positive values in the post-radiation fibrosis group. Large differences between specimens are, therefore, not necessarily more significant than smaller ones. Information about gene expression is to be found in the distribution of the data and not in individual specimens.

DISCUSSION

Specificity and significance of differences in gene expression

The issues of specificity and significance are too frequently mixed up in the literature of gene expression. As mentioned in the introduction, innumerable are the papers reporting differences in the expression of experimental genes between single specimens of different types and claiming biological importance only because these differences are gene-specific, i.e. obtained with a specific detection protocol and/or not observed in a control gene. Authors employing highly specific detection techniques, such as the RNase protection assay, rarely use control genes (e.g. 4, 5). A difference in a gene's expression may be thought to have a biological importance only if it is both, gene-specific and statistically significant. These conditions are totally independent of each other. They must, therefore, be tested separately using appropriate experimental designs. The problems arising from an inappropriate specificity test—using a single, constitutively expressed control gene—have been discussed (2). Here, we appear to use no control genes but the data we present have already been controlled for specificity, and corrected, using a large number of randomly chosen control genes. So we may concentrate this discussion on the question of significance. The arguments that follow are summarized in Table 2.

Gene expression data are frequently distributed in a non-normal manner

Non-normal distributions may arise from pooling normal distributions with different means. Consider i cell populations in which a gene is expressed at rates μ_i. The experimental estimates of each μ_i can be expected to be normally distributed around that theoretical rate. Though, for the whole set of μ_i-estimates to be normally distributed, the theoretical μ_i rates should, themselves, be normally distributed around their own mean. To give an example from our experimental material, we could expect that our ST3 data be normally distributed in the stromal fibroblasts if the rate of expression of ST3 in each such culture was a normal variant of the mean rate in mammary stroma. Similarly, our whole set of data for a gene in the entire experimental material could be expected to present a normal distribution only if the mean expression rates in breast epithelial, stromal and skin cells were normal variants of *one* real expression rate in all breast tissues. When measuring a biological constant with a simple instrument (a morphometric ratio with a ruler or the optical density of a solution with a photometer), the hypothesis that sample means are normally distributed around a single theoretical mean is reasonable, but regarding a potentially variable gene-expression rate measured with a sophisticated multi-step protocol the assumption of normality must be routinely tested.

The first cause of trouble that comes to mind when analysing autoradiographic data is film saturation. Saturation may, indeed, severely distort the distribution of optical densities causing *negative* skewness. Recent technologies for quantifying radiation directly (e.g. PhosphorImager analysis; 13) have advantages over autoradiography, in this respect. However, saturation is not the only possible cause of negative skewness, it cannot cause positive skewness (as was the case in our complete data sets) and it can easily be avoided by differential exposure. When autoradiographic densities *are* linear functions of expression rates, then the observation of a non-normal distribution is, in itself, a reason for rejecting the hypothesis that the specimens derive from a pool with a single mean; or, in other words, that the gene is expressed at a constant rate in the sample. The observation of a normal distribution is not proof of constancy, because the sum of distinct distributions may not necessarily depart significantly from normality. The method of evaluation of gene expression depends, therefore, on whether one is dealing with an apparently constant or an apparently variable expression rate.

Skewness and kurtosis

The skewness index is a measure of the relative frequency of observations on each side of the mean and, as such, it indicates tendencies of modulation with reference to the observed mean rate. A significant positive skewness (e.g. IGFR in skin cells) indicates that the gene is usually expressed at rates below the group-mean and that, occasionally, it is up-regulated (long tail on the positive side). A negative skewness (IGFR in stromal fibroblasts) indicates that the gene is generally over-expressed and frequently down-regulated. Kurtosis indicates the relative frequency of values near, and away from, the mean. A positive kurtosis (IGFR in pooled fibroblasts; 27.3 ± 0.6) indicates frequent or strong modulation, on either side. A negative kurtosis does not indicate absence of modulation—this would result in a normal distribution—but, rather, a stabilizing mechanism which restricts extreme fluctuation of the expression rate. The negative kurtosis of all IGFR data in the interval of ± 1 pooled (-0.8 ± 0.6) was rather artificial, since we intentionally cut off many extreme values. PDGFRB values presented a genuinely negative kurtosis around zero (see detail in Figure 2). We do not analyze this case here but we may speculate that a short-tailed distribution may arise as a result of various experimental artefacts (load calibration or a narrow range of linear autoradiography) and, perhaps, also as a result of natural regulatory mechanisms stabilizing gene expression. Data presenting significant skewness or kurtosis may mean a lot, or nothing, and should be evaluated with great caution.

Explained variation

Statistical methods can be very powerful. If our statistics are correct, we may claim to have traced a significant difference (between post-radiation-fibrosis and non-pathological fibroblasts) that corresponds to 20% of the ST3-specific variation, i.e. only 2.4% of the total variation of the optical signals obtained with this probe; and yet ST3 was the transcript that presented the narrowest specific variation of all the 39 transcripts we studied. We may also claim to have traced a significant effect of TPA on ST3 in a ductal-carcinoma fibroblast isolate, and a combined effect of CT and TPA on IGFR in a skin fibroblast culture, without repeating the treatment; in some research areas, like in biopsy identification, repetition may be practically or theoretically impossible. But do such 'minor' or 'single' observations carry any weight? The percentage of variance that a model explains (the power of the model) should not be confused with the level of significance of the results. In the context of a distribution extreme values are not single observations. Together with all the other observations such values determine the shape of the distribution. It is the shape of data distributions that we examine, evaluate and try to interpret and not individual signals, as is usually the case in gene expression literature.

Nevertheless, we do worry about statistical artefacts; statistics do not prove anything, they only suggest and test hypotheses. The beauty of the specific-signal transformation is that it enables direct comparisons of data obtained in different detection and quantification sessions, on different scales, for different genes. If the difference in *ST3* expression between non-pathological and post-radiation-fibrosis stroma is real then its statistical significance should persist as more such specimens become available. If CT and TPA really affected *IGFR* in the skin fibroblasts then one should expect to observe changes in the expression of other genes, given that the effect of these drugs on transcription is indirect and pleiotropic (14, 15). This prediction has been confirmed in this study: a transcription-factor gene related to differentiation (*p53*), an interleukin gene (*IL4*) and several growth-factor (*TGFA*, *TGFB*, *IGF2*) and other growth-factor-receptor genes (*EGFR*, *PDGFRB* and *FES*) were also strongly stimulated while a transcription/replication-factor gene related to growth (*MYC*), a ribosomal gene (*28S rRNA*) and growth itself were inhibited. In contrast, the specific signals of a metabolic transcript (GAPD), of a cytoskeleton transcript (ACTB) and of 19 other transcripts, homologous or unrelated to the above, remained very close to zero. Consequently, the distributions of the modulated transcripts presented significant skewness whereas the great majority of the examined genes presented normal distributions in the skin fibroblasts (data not shown).

Random variation

When statistics are used incorrectly, or not at all, autoradiographic data can be very misleading. The example of ST3, seen from the pessimist's point of view, shows that as much as 97.6% of the total autoradiographic variation, including extreme observations, can have no recognizable biological importance. This means that unless significance is appropriately documented, the odds are that observed differences in optical density represent random error. A 100-fold, or an even greater difference in a gene product between tissues (e.g. as claimed by Glatt and Snyder; 3, 4) may be non-significant while a 5-fold, or an even smaller difference (as reported by Pieroni *et al.* for the same gene in the same tissues; 4) may be significant, depending on the within-tissue variation (not reported in articles 3 and 4). Here, a difference corresponding to nearly 8 standard deviations of IGFR optical densities hardly approached the conventional 95% confidence level whereas a difference corresponding to about 0.2 standard deviations of ST3 optical densities was highly significant. The magnitude of a difference, by itself, tells nothing about statistical significance and biological importance. We cannot assign any importance to observations that we have not been able to explain statistically; like, or instance, to ST3 signals in epithelial specimens.

Statistics are equally necessary whether gene expression is measured on a continuous scale (raw or transformed optical densities) or on a nominal scale (presence − absence). The example of IGFR in skin cells demonstrates how an unwise use of the mean for describing non-normal data may lead to an erroneous rejection of the modulation hypothesis. Non-normal data should rather be described by distribution-free statistics (skewness, kurtosis, percentiles, frequencies etc.) and evaluated by non-parametric procedures. On the other hand, unthoughtful transformation of continuous normal data to a nominal scale may result in a massive waste of information leading to all types of errors. We have no examples from our own work since we have used a continuous scale precisely for this reason. We selected for this purpose an article by Cullen *et al.* (11) of which the subject and the biological material are very similar to ours and in which, exceptionally, some statistics are reported. We have no intention to criticize these particular authors, because the procedures they followed are the same classical procedures of evaluation of gene expression data as in hundreds of other articles.

Errors associated with current methodologies

The binary transformation. Typically, when a transcript is traced in all examined specimens it is considered as constant and all the between-specimens variation is ignored. This constancy is merely the result of a transformation of continuous densities to a binary scale. Cullen *et al.* claimed that expression of *TGFB*, *PDGFA*, *FGF2* and *FGF5* did not differ between their benign and malignant breast-tumor fibroblasts simply because they were able to trace these transcripts in both cell types. Why should easily detectable gene products be unrelated to phenotypic characters? Analysis of variance of randomly chosen abundant transcripts has demonstrated the opposite possibility (2). Easy detection does not wave aside the need of evaluation.

When the sensitivity threshold of the detection technique approaches the median intracellular abundance of a gene product, some of the specimens may present no detectable signals. In this case, a binary scale and non-parametric procedures are unavoidable but, even then, the distribution of the data is important for selecting the right test. Cullen *et al.* detected IGF2 signals in 1 out of 9 benign-tumour fibroblasts and in 5 out of 9 malignant-tumour fibroblasts. They assumed that their negative scores were real. They erroneously applied the χ^2 approximation, which results in a probability of 0.045 for that table, and concluded that the difference was significant. Fisher's exact probability test, which is the appropriate procedure for their table, would result in a probability of 0.131 and would lead to the opposite conclusion even if the possibility of false negatives were ruled out.

The binary transformation becomes particularly troublesome when applied to an abundant control gene-product, so as to give all specimens a positive score, and to a scarcer experimental gene-product that is detectable only in some specimens. This creates the impression that the binary variation of the experimental gene is 'gene-specific' and that statistics are no longer needed. For evaluating the specificity of the difference in TGFA between the control epithelial line and the experimental fibroblasts, in Cullen's *et al.* experiment, one should ask whether there was an analogous difference in other transcripts, rather than whether TGFB was detectable in the fibroblasts. But, was the difference in TGFA significant in the first place?

The use of a single, non-random control specimen. This is probably the most common error concerning the evaluation of significance. In electrophoretic analysis, the use of a sure positive marker is indeed recommended. This specimen is to indicate, roughly, the power of the probe and the position at which the wanted transcript must have migrated. The 'control' cells are purposely selected for their *known* ability to express the gene in question at a detectable rate; perhaps, at an unusually, or pathologically, high rate. Non-random specimens, however, do not represent any group and have no statistical meaning. The experiments of Cullen *et al.* may suggest that diploid fibroblasts express less *TGFA* and *PDGFB* than what an immortal, pathological, hypertriploid epithelial cell line is known to be able to express; or simply confirm, in a tautological manner, that the control specimen expressed more *TGFA etc* than random specimens do. By no means can such results be interpreted as meaning that fibroblasts express these genes less than random epithelial isolates usually do, or not at all. A sound evaluation of any difference between two types of cells would require more than one randomly chosen specimen of each type. When the frequency of negatives, or positives, within a group of specimens is close to 100%, there may be too little left for non-parametric statistics to explain; one should consider using a more sensitive detection method, or, for positive scores, a finer scale.

CONCLUSIONS

We investigated the variation of expression of several genes in breast tissues. The solution to the problems of sensitivity and specificity of transcript detection adopted here was to examine total RNA extracts deposited on filters without further manipulation and finely calculate the overall transcriptional activity of each cell population (the mRNA content of the specimens) from a large sample of mRNA species. Without adding any artificial variation, the 'specific-signal' transformation accounts for non-specific attachment of probes to loaded material and standardizes the transcriptional activities of cell populations and those of genes. Genes may, thus, be compared as if they were all expressed at the same average rate in nature, and cell populations, as if they all produced the same total amount of mRNA. This transformation also improves the normality of optical density distributions of and facilitates, thus, a parametric statistical analysis.

Our results suggest that *ST3* and *IGFR* may be involved in mammary gland biology and/or pathology and deserve further study. ST3 was confirmed to be a TPA-modulated transcript specific to pathological fibroblasts, as Basset *et al.* originally proposed (12). Although the effect of 100nM TPA on our particular ductal carcinoma strain (inhibition) was the opposite to what Basset *et al.* had observed with 10nM TPA on other strains (stimulation), we know that different concentrations of TPA can have quite different—and even opposite—effects on a gene's expression (14). Moreover, we do not expect all ductal carcinomas to be of the same type nor to respond to treatment in the same manner. It is also important to emphasize that a negative 'specific signal' means a *relative* inhibition with reference to other genes (38 in this case) and not necessarily an absolute reduction of the gene's expression rate. Basset *et al.* (12) did not compare *ST3* to other TPA-sensitive genes. Expression of *ST3* seems, nevertheless, to be also associated with another pathological state of mammary stroma, the post-radiation fibrosis. IGFR appeared to be a fibroblast-specific transcript subject to frequent and strong modulation. This result suggests that stromal fibroblasts are able to modulate their sensitivity to IGF-mediated messages. Cullen's *et al.* results concerning *IGF1* (11), as well as our own observations on *IGF1* and *IGF2* expression in mammary tissues (E. Spanakis and D. Brouty-Boyé, in preparation), indicate that the *IGF-IGFR* system is involved in the communication between mammary epithelium and stroma and, perhaps, also in breast pathogenesis.

The most important point this article intends to make, however, is that explicit assumptions and rigorous—if not advanced—statistical procedures are imperative in gene expression research because gene expression rates as well as optical densities of gene products are complex variables and their evaluation presents multiple technical and theoretical difficulties. This point merits the attention of molecular biologists since practically all we know about gene expression is, thus far, based on autoradiographic data of which the probabilities have not been estimated. The multi-gene slot blot analysis can provide massive amounts of high quality data. Distribution descriptive statistics assist in the selection of the right evaluation procedure.

ACKNOWLEDGEMENTS

We thank V. Magnien for technical assistance, L Israël for comments, A. Kahn for providing image analysis facilities at INSERM Unit 129 (Paris), B. Azzarone, B. Coulomb for providing cells and P. Basset for providing the *ST3* probe. This work was supported by the 'Groupement d'Etudes et de Recherches Cancerologiques', France and the 'Association pour la Recherche sur le Cancer', France.

REFERENCES

1. Sambrook,J., Fritch,E.F. and Maniatis,T. (1989) *Molecular Cloning, A Laboratory Manual.* 2nd edition. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
2. Spanakis,E. (1993) *Nucleic Acids Res.*, **21**, 3809–3819.
3. Glatt,C.E. and Snyder,S.H. (1993) *Nature*, **361**, 536–538.
4. Pieroni,J.P., Miller,D., Premont,R.T. and Iyengar,R. (1993) *Nature*, **363**, 679–680, with a reply by Glatt, C.E. & Snyder, S.H.
5. Dupriez,V.J., Darville,M.I., Antoine,I.V., Gegonne,A., Ghysdael,J., and Rousseau,G.G. (1993) *Proc Natl. Acad. Sci. USA* , **90**, 8224–8228.
6. Luo,Y., Raible,D. and Raper,J.A. (1993) *Cell*, **75**, 217–227.
7. Van den Hooff,A. (1988) *Adv. Cancer Res.*, **50**, 159–196.
8. Camps,J.L., Chang,S.M., Hsu,T.C., Freeman,M.L., Hong,S.J., Zhau,H.E., Eschenbach,A.C. and Chung,L.W.K. (1990) *Proc. Nat. Acad. Sci. USA*, **87**, 75–79.
9. Chomette,G., Auriol,M., Tranbaloc,P. and Blondon,J. (1990) *Pathol. Res. Pract.*, **186**, 70–79.
10. Ferguson,J.E., Schor,A.M., Howel,A. and Ferguson,M.W. (1990) *Differentiation*, **42**, 199–207.
11. Cullen,K.J., Smith,H.S., Hill,S., Rosen,N. and Lippman,M.E. (1991) *Cancer Res.*, **51**, 4978–4985.
12. Basset,P., Bellocq,J.P., Wolf,C., Stoll,I., Hutin,P., Limacher,J.M., Podhajcer,O.L., Chenard,M.P., Rio,M.C. and Chambon,P. (1990) *Nature*, **348**, 699–704.
13. Arkins,S., Rebeiz,N., Biragyn,A., Reese,D.L. and Kelley,K.W. (1993) *Endocrinology*, **133**, 2334–2343.
14. Nishizuka,Y. (1988) *Nature*, **334**, 661–665.
15. Chang,F.H. and Bourne,H.R. (1989) *J. Biol. Chem.*, **264**, 5352–5357.