A Correlated Noise Model for the Significance Analysis of Gene Expression Data

Joachim Theilhaber[1],[*] Sridaran Natesan[1], Karen Chandross[2],
Tim Connolly[1], Steven Goldman[3], Jean Merrill[2],
Steven Bushnell[1]and Christoph Brockel[1]

September 24th, 2002

Aventis Bioinformatics Report
Released for open publication September 2002

1. Aventis Pharmaceuticals, Cambridge Genomics Center, 26 Landsdowne Street, Cambridge, MA 02139, USA.

2. Aventis Pharmaceuticals, Cell Biology and Neuropathology, Route 202-206, Bridgewater, NJ 08807, USA.

3. Cornell University Medical Center, Neurology Dept., 1300 York Avenue, New York, NY 10021, USA.

Corresponding author:
Joachim Theilhaber
tel. (617) 768-4016
fax. (617) 374-8808
e-mail: joachim.theilhaber@aventis.com

[*] To whom correspondence should be addressed

Abstract

Motivation: The Student t-test, applied to two-tissue comparisons based on Affymetrix chip data, often results in the non-operational conundrum of “all genes have significantly different regulation”, whenever the number of samples is large. The intraclass correlation (icc) model presented here is a new statistical model for the analysis of gene expression data that addresses this problematic. The icc model includes correlated, tissue-dependent noise terms, which are essential in avoiding the global overestimation of statistical significance that otherwise compromises the process of gene selection.
Results: The icc model can be used with a standard t-test by a simple, sample-size-dependent rescaling of the t statistic, or, for applications requiring more sensitivity, with a more elaborate, biased-variance statistic t^* which we define, in conjunction with a semi-parametric resampling scheme to establish a reference distribution. Application of the icc model to the problem of selecting genes involved in specification and differentiation of neuronal and glial cells, on the basis of expression profiling of fetal brain tissues, indicates that at false-discovery rate $F_d = 25\%$, detection sensitivity for biological marker genes induced in the germinative regions examined is of order $S \sim 40\%$.
Contact: joachim.theilhaber@aventis.com

1 Introduction

In recent years much effort has been applied to analyzing gene expression data generated by DNA chips and microarrays, typically to select a group of genes involved in a biological process of special interest(Eisen et al., 1998; Alon et al., 1999; Alizadeh et al., 2000; Ross et al., 2000; Spellman et al., 1998). In this process of gene selection, the initial emphasis was not on rigorous statistical methods, but rather on using semi-quantitative criteria for establishing significance of differential gene regulation (such as a requiring a minimum fold change in gene expression over several measurements). More recently however, rigorous statistical methods have been applied to the problem of gene selection(Tusher et al., 2001; Jin et al., 2001). In this spirit, in the present paper, we present a statistical model for selecting genes on the basis of two-class comparisons, such as might occur when comparing gene expression between two different tissue panels. We have named the model the intraclass correlation model ("icc model"), because its central feature is a tissue-dependent, correlated noise term, which is added to more standard, statistically independent noise terms. The icc model can be applied in two ways: it can be incorporated into the t-test by a simple, sample-size-dependent rescaling of the t statistic; for situations requiring more sensitivity however, on can use a more elaborate, biased-variance statistic t^* which we define, in conjunction with a semi-parametric resampling scheme for establishing a reference distribution.

As shown here, when intraclass correlation is *not* taken into account, statistical tests tend to overestimate significance, and as sample sizes grow, eventually *all* genes are assigned a significant change. Because the icc model crucially avoids such a divergence in assignment, we refer to its effect as a "renormalization", in analogy to a similar process in physics.

In this paper, the icc model is applied to two problems, both based on Affymetrix chip expression data. First, we compare two large panels of lung and liver tissue expression data. This problem is not analyzed because of its intrinsic biological interest, but to systematically develop the icc model. Second, we proceed to a problem of explicit biological interest, the selection of genes involved in the specification and differentiation of neuronal and glial cells in the developing central nervous system, using expression profiles of fetal brain tissues. A preliminary validation of the results, based on a test set of biological markers, indicates that for detection of genes induced in germinative regions, at a false-discovery rate $F_d = 25\%$ we are achieving a sensitivity $S \sim 40\%$.

2 t-tests for lung-liver tissue comparisons

In order to systematically explore performance of the t-test on gene expression data, we performed comparisons of large panels of rat lung versus rat liver tissue samples. Expression data for a total of 30 rat lung and 30 rat liver tissue samples, all from different donors, was obtained by hybridization of the processed mRNA to 60 Affymetrix chips of the Rg_u34a chip design[1]. The resulting "lung-liver" data set consisted of expression profiles for the 8758 qualifiers[2] represented on the Rg_u34a chip design (excepting controls), where each profile contains 60 intensity values (Affymetrix average differences(Lockhart et al., 1996)) quantifying the relative mRNA abundances in the different samples, and denoted by

$$\text{lung samples:} \qquad x_i, \qquad i = 1, 2, \ldots, n_1 \ , \tag{1}$$

$$\text{liver samples:} \qquad y_j, \qquad j = 1, 2, \ldots, n_2 \ , \tag{2}$$

where $n_1 = 30$ is the number of lung samples and $n_2 = 30$ the number of liver samples.

To compare, on a qualifier-by-qualifier basis, the two data series in Eqs.(1,2), we used the t statistic (equal variances model, (Keeping, 1995, p.184)), given by

$$t = \frac{\bar{y} - \bar{x}}{\left((\frac{1}{n_1} + \frac{1}{n_2}) s^2 \right)^{1/2}} \ , \tag{3}$$

where $\bar{x}$ and $\bar{y}$ are the sample means of $\{x_i\}$ and $\{y_j\}$, respectively, and where s^2 is the sample variance

[1] The 30 lung samples were generated from 30 individual Brown Norway rats, used as controls in the course of a respiratory study unrelated to the present work. The 30 liver samples were derived from 30 individual Sprague-Dawley rats, similarly used in the course of toxicology studies.

[2] For Affymetrix chips, a "qualifier" refers to the set of features, also known as a probe set, which together measure the abundance of transcripts containing a given RNA sequence. The mapping of qualifiers into genes is many-to-one.

$$s^2 = \frac{1}{n_1 + n_2 - 2} \left(\sum_{i=1}^{n_1} (x_i - \bar{x})^2 + \sum_{j=1}^{n_2} (y_j - \bar{y})^2 \right) . \tag{4}$$

In the underlying model assumed here each intensity value is equal to a population mean, plus a noise term. If we furthermore assume, for each qualifier, the null hypothesis

$$H_0 : \left\{ \begin{array}{l} \text{1. equal population means for the panel of lung and} \\ \quad \text{and the panel of liver samples,} \\ \\ \text{2. normally distributed noise amplitudes,} \\ \\ \text{3. and independent noise terms in each measurement,} \end{array} \right. \tag{5}$$

then the P-value P corresponding to t under the two-tailed test is given by

$$P = 1 - A(t|\nu) , \tag{6}$$

where $A(t|\nu)$ is the integral of the Student-t distribution on the interval $(-t, t)$(Abramowitz and Stegun, 1972, p.948) and $\nu = n_2 + n_1 - 2$ are the degrees of freedom of the test.

The lung-liver data set was analyzed by performing a separate t-test on each of its 8758 expression profiles, each time comparing the lung to the liver samples. It is convenient to define the two cumulative distribution functions

$$N_f(P_0) = N(P \leq P_0 | \text{observed distribution}) , \tag{7}$$

$$N_r(P_0) = N(P \leq P_0 | \text{reference distribution}) . \tag{8}$$

In Eq.(7), $N_f(P_0)$ denotes the number of qualifiers found with P-value less than or equal to P_0, in the observed distribution of P-values (i.e. in the actual test), while in Eq.(8), $N_r(P_0)$ denotes the number of qualifiers that would be found in the *reference* distribution, the distribution that obtains if every profile is generated under some realization of the null hypothesis H_0. Under H_0, we have to a good approximation

$$N_r(P_0) \; \approx \; P_0 \; N_0 \; , \tag{9}$$

where $N_0 = 8758$ is the total number of qualifiers in the data set ($N_r(P_0)$ also embodies a sampling variance that is relatively small provided $P_0 N_0 \gg 1$, and that we choose to ignore here).

Distributions $N_f(P_0)$ were computed for a number of random subsamplings of the tissue panels, in each case choosing without replacement $m_1 = m_2 \equiv m$ samples from the lung and liver panels, respectively, with $m = 3, 4, 6, 10, 20, 30$. The distributions are displayed in Fig. 1, where they are compared to the unique reference distribution $N_r(P_0)$. Note that the figure is a log-log plot, with $\log(N(P_0))$ plotted against $-\log(P_0)$, so that the most significant data, for which $P_0 << 1$, lies on the right-hand side of the graph. Taking logarithms of both sides of Eq.(9), we obtain *Fig. 1 after here.*

$$\log(N_r(P_0)) \; = \; \log N_0 \; - \; (-\log P_0) \; , \tag{10}$$

so that as plotted in the figure, the reference distribution follows a straight line with slope of -1.

The striking feature of Fig. 1 is that all observed distributions are very different from the reference distribution, and that the difference systematically grows with sample size, with the observed distributions extending more and more to the right, into regions of very high significance. Observed and reference distributions are not even comparable for the noisiest and least significant data ($P_0 \to 1$, left-hand side of the graph), where the distributions might have been expected to merge. Examination of the distribution of the corresponding t statistic, shown in Fig. 2 for $m = 30$, shows that it is much broader than the reference Student-t distribution ($\nu = 58$), even in the central region: it is this "swelling" of the t distribution, which grows with sample size, that explains the drift of the observed distributions to ever smaller P-values. *Fig. 2 after here.*

If taken at face value, the behavior observed in Fig. 1 is surprising because it implies that eventually, nearly all of the genes represented on the Rg_u34a chip will be found to have significantly different levels of expression in the two tissues. For instance, for the $m = 30$ comparison, setting the threshold P_0 so that the false discovery rate $F_d = 0.25$ (Appendix A) selects for $N_f = 7356$ qualifiers. Of this selection, about three-quarters or 5500 should be true positives. As the Rg_u34a chip carries only 8758 qualifiers in all, this result implies that at least 63% of the genes represented on the chip have significantly different expression in the two tissues, and the trend observed in Fig. 1 indicates that with enough samples, nearly all genes will be selected.

The results presented above suggest a sample-size dependent artifact in the assignment of significance, and led us to reexamine the assumptions of the null hypothesis H_0 which underlies all of the t tests. In revisiting the null hypothesis, we focused on explaining the overall swelling of the t distribution (Fig. 2), which is at the root of the divergent distributions observed in Fig. 1.

We first explored the possibility that non-normality in the distribution of noise was the cause of the swelling of the t distribution. As a test of this hypothesis, we randomly permuted columns in each row of the lung-liver data matrix, thereby breaking correlations between expression levels and tissue type, but otherwise leaving the distribution of noise terms intact on a qualifier-by-qualifier basis. Fig. 3 shows that the distribution $N_f(P_0)$ for this randomized data set is nearly identical to the null distribution $N_r(P_0)$ expected under H_0. This result indicates that non-normality of the noise distribution is not the cause of the swelling of the t distribution.

Fig. 3 after here.

We also explored whether the variance-stabilizing transformation of (Durbin et al., 2002), designed to remove the dependence of the variance on the mean, and to symmetrize the noise distribution, would suppress the swelling of the t distributions. We found that the transformation had no such effect, with the t distributions having very similar widths before and after the transformation.

An alternative explanation for the swelling of the t distribution, which emphasizes the role of correlations, is that the noise terms intervening in the measurements are not strictly independent. In particular, if the noise terms contain a tissue-specific component, which is separately consistent across lung samples and across liver samples, but uncorrelated between lung and liver, the result will be an "intraclass" correlation of the noise terms(Keeping, 1995, p.226). This is the basis for the intraclass correlation (icc) model, which we believe correctly accounts for the swelling of the t distribution. The intraclass correlation tends to increase the difference in the sample means while maintaining the sample variance; the result is to artificially increase the observed values of t over what would be expected in the absence of correlations. The icc model is described in some detail in Appendix B. In the next section we focus on its application to the analysis of the lung-liver data set.

2.1 The intraclass correlation model: results

The icc model is completely determined by a single parameter, the intraclass correlation coefficient ρ. In Appendix B, we derive the expression

$$\rho \;=\; \frac{\sigma_\eta^2}{\sigma_\eta^2 + \sigma^2}\;, \tag{11}$$

where σ^2 is the variance of the uncorrelated noise terms, and σ_η^2 the variance of the correlated, tissue-specific noise terms.

The icc model predicts that instead of the raw t statistic, a "renormalized" statistic t' should be used,

$$t' \;=\; \beta\, t \tag{12}$$

where t is given as before (Eq.(3)), and where the renormalization constant β is given by

$$\beta \;=\; \left(1 + \frac{2\rho}{(\frac{1}{n_1}+\frac{1}{n_2})(1-\rho)}\right)^{-1/2}\;. \tag{13}$$

Under the icc model null hypothesis H_0' (Eq.(35), Appendix B), it is now t', and not t, that is distributed according to the Student-t distribution with $\nu = n_1 + n_2 - 2$ degrees of freedom. Note that for $\rho \neq 0$, we necessarily have $\beta < 1$, and this explains the swelling of the distribution of the "raw", unrenormalized statistic t. Furthermore, for a given ρ, β decreases with increasing sample size, so that the corresponding raw t distributions broadens with increasing sample size relative to a Student-t distribution, leading to the trend observed in Fig.1.

To apply Eqs.(12,13) to the lung-liver data set we actually proceed backwards, starting from the data itself rather than explicitly using Eq.(11). First, β is estimated from the data according to the estimator

$$\hat{\beta} \;=\; \frac{t_{\nu,0.75} - t_{\nu,0.25}}{t_{0.75} - t_{0.25}}\;, \tag{14}$$

where $t_{\nu,0.25}$ and $t_{\nu,0.75}$ are the 25th and 75th percentiles of the Student-t distribution with $\nu = n_1 + n_2 - 2$ degrees of freedom, respectively, and $t_{0.25}$

and $t_{0.75}$ the corresponding percentiles for the observed distribution of t, with t evaluated from Eq.(3) as before. An assumption implicit in Eq.(14) is that the distribution in the interquartile range reasonably approximates H_0', with significant departures occurring only outside that range, in the tails of the observed distribution. An estimator for the correlation coefficient is then found by solving Eq.(13) for ρ,

$$\hat{\rho} = \left(1 + \frac{2\hat{\beta}^2}{(\frac{1}{n_1} + \frac{1}{n_2})(1 - \hat{\beta}^2)}\right)^{-1} . \tag{15}$$

For instance, for the lung-liver data set with $n_1 = n_2 = 30$ samples, we find $\hat{\beta} = 0.135$, for which $\hat{\rho} = 0.641$. The cumulative distributions $N_f(P_0)$ obtained from the renormalized test statistic t', and for the same sample sizes as in Fig. 1, are shown in Fig. 4, where it can be seen that the strongly divergent behavior of the distributions has been suppressed (compare to Fig. 1). In particular, at low significance levels ($P_0 \to 1$, left-hand side of graph) the observed distributions now all smoothly merge into reference distribution $N_r(P_0)$ *Fig. 4 after here.*

2.2 Consistency of the icc model

The icc model applied to the lung-liver data set makes sense only if ultimately, a single value of ρ, specific to the lung-liver tissue pair, can be used to renormalize all distributions, irrespective of sample sizes m_1 and m_2. To verify that the estimate of $\rho \approx 0.641$ obtained above for $n_1 = n_2 = 30$ is not an artefact of fitting a single case by Eq.(14), we thus systematically investigated the dependency of $\hat{\rho}$ on sample size and sample composition. Values of $\hat{\rho}$ and $\hat{\beta}$ were generated from t-tests performed between random, equal sized subsamplings of the 30 lung and 30 liver samples. Subsamples without replacement, of size $m_1 = m_2 \equiv m$, $1 \leq m \leq 20$, were used, with 5 independent subsamplings generated for each value of m. In Fig. 5, we show the dependence of $\hat{\rho}$ and $\hat{\beta}$ on subsample size m. It can be seen that a convergent value $\hat{\rho} \approx 0.65$ is obtained provided $m \geq 4$ samples in each tissue panel. It is important to note that while $\hat{\rho} \approx$ constant for $m \geq 4$, the renormalization factor $\hat{\beta}$ is markedly decreasing with increasing m, so that an essentially constant ρ captures rescaling of the t distribution over a wide range of conditions. *Fig. 5 after here.*

Fig.(6) shows the concommittant dependency on subsample size m of the number N_F of qualifiers found after renormalization, when the selection threshold P_0 is continually adjusted to maintain a false-discovery rate $F_d = 0.25$. The figure indicates that despite a relatively large sampling variance, due to the large heterogeneity of the samples within each tissue panel, the mean value of N_F converges to the value $\bar{N}_F \sim 500$, provided $m \geq 10$. Thus ultimately, about 6% of the 8758 qualifiers on the Rg_u34a chips are found to have significantly different expression in the lung-liver panels.

Fig. 6 after here.

3 Application: detection of genes involved in neuron and glial cell development

The icc model was applied to expression data pertaining to the genesis of neuronal and glial cells in the human embryo and fetus(Keyoung et al., 2001). In this study, expression profiles were obtained, for a series of different developmental stages, for tissue samples from two specific regions of the developing fetal brain, the cerebral cortex and the ventricular zone. The ventricular zone is a germinative region, in which initially proliferating stem cells undergo successive differentiations into more specialized progenitors, which in turn undergo terminal differentiation into neurons or glial cells(Chenn et al., 1997; Goldman and Luskin, 1998). Following terminal differentiation, the neurons migrate outward into the cerebral cortex where they assume a final, mature phenotype. The aim of the analysis was to find genes with significantly different expression in the ventricular zone compared to the cerebral cortex, and in particular, to identify genes more highly expressed in the ventricular zone, as these genes are potentially inducing the pro-neuronal or pro-glial cell differentiation processes which are the focus of the investigation.

22 matched tissue samples from the ventricular zone and from the cerebral cortex were obtained from 11 donors at various stages of development, spanning a period from the 15th (E15) to the 23rd week (E23) after conception (Table 1). The processed mRNA from each of these samples was hybridized to the 5 commercial Affymetrix Hg_u95 chips (A through E chips) and to an additional, custom-made Affymetrix chip (AVTF2). In total, 73373 expression profiles were generated, each profile consisting of 22 paired intensity measurements (Affymetrix average differences) quantifying relative

Table. 1 after here.

transcript abundance as measured by each Affymetrix qualifier, and denoted by

$$\text{cortex samples:} \qquad x_i, \qquad i = 1, 2, \ldots, n_1 \ , \tag{16}$$

$$\text{ventricular zone samples:} \qquad y_j, \qquad j = 1, 2, \ldots, n_2 \ . \tag{17}$$

In Eqs.(16) and (17), $n_1 = n_2 \equiv n = 11$, the number of matched sample pairs. In what follows, the data set will be referred to as the "cortex-ventricular zone" (cortex-vz) data set.

3.1 Tests of tissue specificity: paired t-tests and biased-variance statistic for increased sensitivity

Because of the presence of matched samples, for each profile in the cortex-vz data set we applied a paired t-test(Keeping, 1995, p.185)(rather than an unpaired test as in Eq.(3)), computing the statistic

$$t \ = \ \frac{\bar{y} - \bar{x}}{\left(\frac{1}{n}s^2\right)^{1/2}} \ , \tag{18}$$

where $n = 11$ is the number of sample pairs, $\bar{x}$ is the sample mean for the cortex samples, $\bar{y}$ is the sample mean for the ventricular zone samples, and where s^2 is the sample variance of the differences between matched samples,

$$s^2 \ = \ \frac{1}{n-1} \sum_{i=1}^{n} (z_i - \bar{z})^2 \ , \tag{19}$$

where $z_i = y_i - x_i$.

For a given value of t, P-values P were computed from Eq.(6) as before, but now with $\nu = n - 1$ degrees of freedom. The resulting, unrenormalized cumulative distribution $N_f(P_0)$ is shown in Fig. 7a. Applying the methods of the previous sections through Eqs.(14) and (15), we find that the tissue intraclass correlation is characterized by coefficient $\rho = 0.083$, and that a rescaling coefficient $\beta = 0.7$ is required to renormalize t in accordance with Eq.(12). The renormalized distribution function that results is shown in Fig. 7b: unfortunately, as can be seen, the renormalized distribution nearly merges everywhere into the reference distribution, which implies that very

Fig. 7 after here.

few significant profiles can be found. For instance, for a false-discovery rate $F_d = 0.25$, only 4 significant profiles are found out of 73373. This result is in contrast to those of the lung-liver comparison, where even after renormalization, hundred of significant profiles were found for a comparable number of degrees of freedom (Fig. 6), and furthermore for a much smaller data set (8758 profiles instead of 73373); this difference in detection rate can be ascribed to the more subtle differences between the ventricular zone and cortex, which are two related neural tissues, compared to the more drastic differences between the lung and liver tissues examined above.

Because the simple renormalization scheme of Eq.(12) was not sufficient to guarantee detection of genes in the more difficult cortex-vz comparisons, to boost the sensitivity we adopted a "biased-variance" t statistic t^*, written

$$t^* = \frac{\bar{y} - \bar{x}}{\left(\frac{1}{n}(s^2 + \sigma_0^2)\right)^{1/2}} , \tag{20}$$

where the variance bias term σ_0^2 is an adjustable parameter, chosen to maximize sensitivity. It should be noted that t^* is similar to, but not identical to a statistic used by Tusher et al.(Tusher et al., 2001). The bias term σ_0^2, which reduces the value of t^* relative to that of the usual t statistic, has the greatest effect on data with overall low intensities, for which sample means and sample variances are small. The variance bias thus acts as a filter that suppresses the contributions of the low-level, noisy expression profiles. These profiles are very numerous in the data set, and act as a "background noise" that otherwise tends to mask the significant data.

For $\sigma_0 > 0$, the sampling distribution of t^* under the null hypothesis H_0' of the icc model is not a Student-t distribution, and to our knowledge cannot be obtained analytically. To gauge significance on the basis of their test statistic, Tusher et al. used a randomization procedure in which members of the two tissue panels were randomly permuted to establish an empirical, reference distribution. However, this approach is not applicable under the assumptions of the icc model, because it is crucial that the reference distribution maintain the intraclass correlations, and these correlations are destroyed if one simply mixes assignments of the different tissue samples (as was shown in Fig.3).

We resorted instead to a semi-parametric resampling scheme to generate a reference data set that conserve the intraclass correlation of the cortex-vz comparison, while approximating overall the conditions of the null hypothesis. A data set consisting only of the intensities for the cortex samples (i.e. $n = 11$ samples in all) was first assembled. For each qualifier in this data set, resampling then proceeded as follows (see Appendix C for details): for the k-th time point in the expression profile (as defined in Table 1) the sample mean μ_k and sample variance σ_k^2 were estimated from the intensities of the r_k replicates assigned to that time point. Using the estimates μ_k and σ_k^2, new intensities were then generated for all r_k replicates, by a Monte Carlo simulation which re-generates both uncorrelated and correlated noise terms, and then adds them to the values of μ_k. In this procedure, the variance of the correlated noise relative to that of the uncorrelated noise is specified by the intraclass correlation coefficient ρ, which is a fixed parameter of the simulation, and which is directly estimated from the cortex-vz data set, via Eq.(15) ($\rho = 0.083$).

The Monte Carlo resampling procedure was applied twice, with different random number seeds, to generate two independent synthetic data sets. Because these data sets incorporate different realizations of the icc noise model, but are both based on the same cortex expression data, comparison provides an instantiation of the null hypothesis. The reference distribution of t^* was thus generated by comparing on a qualifier-by-qualifier basis the two data sets, using the same biased variance statistic as used in the actual cortex-vz comparisons, Eq.(20).

3.2 Sensitivity optimization of the t^*-test

For the cortex-vz comparisons that follow, it was found practical to continue using Eq.(6) to transform the test statistic t^* into a "P-value" P. Although, for $\sigma_0 > 0$, P does not measure an absolute level of significance, it remains a convenient variable for graphical display. False-discovery rates were directly determined from the relation $F_d(P_0) = N_r(P_0)/N_f(P_0)$ (Eq.(21), Appendix A), where now both $N_f(P_0)$ and $N_r(P_0)$ are determined from the data.

Figs. 8a-h explore the effect of increasing the variance bias term σ_0 on the distribution functions $N_f(P_0)$ and $N_r(P_0)$, and the concommittant effects on the false-discovery rate as a function of P_0. Thus, the top panels (a,c,e,g) compare the observed and reference distributions (heavy dots, $N_f(P_0)$; thin lines, $N_r(P_0)$), while the bottom panels (b,d,f,h) display the corresponding false discovery rates as a function of $-\log(P_0)$, for $\sigma_0 = 0, 100, 250, 2500$, respectively. In all figures, the vertical lines signal the decision thresholds where a false-discovery rate $F_d = 0.25$ is obtained, and N_F denotes the total number of qualifiers found at that point.

Note that the first figure, Fig. 8a, for which $\sigma_0 = 0$, is actually equivalent to Fig. 7b, because $t^* = t$ in this particular case. The renormalization procedures underlying Figs. 7b and 8a are quite different however, as in Fig. 7b it is the observed distribution $N_f(P_0)$ that was renormalized by rescaling through the transformation $t' = \beta t$, while in Fig. 8a, the observed distribution in untouched, and it is the reference distribution $N_r(P_0)$ that is "adjusted" by using the semi-parametric resampling method described above to generate it. The results are nonetheless the same, with only $N_F = 4$ qualifiers found at the nominal false-discovery rate $F_d = 0.25$.

Fig. 8 after here.

Figs. 8c-h show the effects of systematically increasing the bias term σ_0 above 0. For $\sigma_0 = 100$ (Fig. 8c), an open region appears between the curves of the distribution functions $N_f(P_0)$ and $N_r(P_0)$, and as a consequence, the false discovery rate function $F_d(P_0)$ (Fig. 8d) sharply decreases for increasing $-\log(P_0)$. $N_F = 1899$ qualifiers are found at the false-discovery rate $F_d = 0.25$. Further increasing σ_0 to 250 (Fig. 8e), further increases the separation between the distribution functions, leading to an even greater drop in the false-discovery rate (Fig. 8f), with now $N_F = 3029$ qualifiers found at $F_d = 0.25$. Much larger values of the bias, however, such as $\sigma_0 = 2500$ (Fig. 8g), lead to a reversing trend ($N_F = 1652$), with now decreasing sensitivity for increasing σ_0.

The dependence of N_F, the total number of qualifiers found, on σ_0 at fixed false-dicovery rate $F_d = 0.25$ is plotted in Fig. 9. For $\sigma_0 < 80$, N_F is trivially small. For $\sigma_0 = 80$ exactly, the distributions $N_f(P_0)$ and $N_r(P_0)$ achieve a critical separation, leading to the sudden jump in the value of N_F. Maximum sensitivity (maximum value of N_F) is then obtained for $\sigma_0 = 250$, for which $N_F = 3029$. Note that although $\sigma_0 = 250$ maximizes overall sensitivity, it does not insure a strictly monotonic behavior of the false discovery rate as a function of $-\log(P_0)$. Thus, in Fig. 8f, after bottoming-out for $-\log(P_0) \approx 4$, the false discovery rate increases again as a function of $-\log(P_0)$ (right-hand side of the figure). This "contamination" of the most significant data is tied to the persistence of a long tail in the reference distribution function; it can be suppressed by choosing a larger bias, albeit at the cost of reducing the overall sensitivity somewhat. In the present case, suppression of the tail occurs for $\sigma_0 = 2500$ (Fig. 8g), for which $N_F = 1652$.

Fig. 9 after here.

3.3 Neuronal and glial marker genes

To quantify the biological relevance of gene selections obtained by the t^*-test, we examined the distribution in the overall population of qualifiers, of the P-values of a restricted set of biological markers, many of which are expected to show strong differential expression in the cortex and ventricular zone comparisons. Thus, a test set of 24 genes (Table 2), known to be markers for cells belonging to the neuronal, oligodendrocytic and astrocytic lineages was assembled on the basis of expert biological knowledge. The list included genes expressed in progenitor cells, as well as genes associated with the more mature, differentiated cell phenotypes. For instance, the oligodendrocytic markers include genes expressed in mature, myelinating cells (MAG, MOG, MBP), as well as a transcription factor (SOX10)(Wegner, 2000) and other markers (CNP, PLP) also expressed in oligodendrocyte progenitors. Similarly, neuronal markers include genes expressed in mature cells (MAP1B, MAP2, βIII tubulin, NF-L,NF-H), as well as a transcription factor (SOX2) expressed in progenitor cells. Because of redundancies in gene representation on the Affymetrix chips, altogether the 24 genes map into 36 distinct qualifiers.

Table. 2 after here.

It should be emphasized that the list of marker genes defined by Table 2 is only an an approximate test set, because it is not based on independent expression profiling of the relevant tissues, but rather, on a somewhat weaker expectation based on biological expert knowledge. Indeed, some of the genes in Table 2 may in actuality undergo no differential regulation at all. Note however that this state of affairs simply makes the task of detection artificially harder, and must necessarily result in a conservative (rather than inflated) estimate of detection sensitivity.

It should also be noted that for most of the genes in Table 2, the expected dominance of expression in one tissue over the other (vz $>$ cx or vz $<$ cx) is not known a priori; this lack of knowledge is irrelevant for the present test, as we are only concerned with detecting significant change, and not in confirming a given tissue specificity.

3.4 Marker genes distributions

Figs. 10 display the cumulative distributions of the biological markers, ranked in the global population according to P-value, with separate distributions shown for qualifiers with average expression higher in the ventricular zone (vz > cx, Fig. 10a) and for qualifiers with average expression higher in cortex (vz < cx, Fig. 10b). In each figure, a rank of 1 indicates the smallest P-value (the most significant expression profile), and the straight line denotes the average, reference cumulative distribution expected under completely random sampling of the parent population. *Fig. 10 after here.*

In Fig. 10a, displaying data for vz > cx, 24 out of the 36 biological markers are represented, out of a total of 40575 qualifiers exhibiting vz > cx expression (for most of which, it should be emphasized, the difference is not significant). The very fast rise of the cumulative distribution on the left indicates strong overrepresentation of the biological markers among significantly regulated profiles. The one-sided t^*-test, with $\sigma_0 = 2500$, selects for 2017 qualifiers with significant differential regulation out of 40575 at false-discovery rate $F_d = 0.25$ (5% of the total, decision threshold shown in the figure). 9 biological markers out of 24 (37%) are in this selection, corresponding to a 37%/5% $\approx$ 7-fold “enrichment” of markers relative to the global population. Eliminating redundancies in the qualifier to gene mapping, we find that 8 marker genes (ASH1, βIII-tubulin, GFAP, HES-1, BLBP, tubulin-α1, PLP, FAT) out of 18 are selected at the given decision threshold, indicating a detection sensitivity $S = 8/18 \approx 40\%$.

In Fig. 10b, displaying data for vz < cx, 13 markers out of the 36 are represented, for a total of 32798 qualifiers. The same t^*-test as above selects 1671 qualifiers out of the 32798 at false-discovery rate $F_d = 0.25$ (5% of the total). 7 biological markers out of 13 (46%) are in the selection, corresponding to a 46%/5% $\approx$ 9-fold enrichment of biological markers relative to the global population. 3 marker genes (MAP1B, NF-L, ELAV-3) out of 7 are selected at the given decision threshold, again very roughly indicating a detection sensitivity $S \approx 40\%$.

The overall statistical significance of the ranked distributions of biological markers can be further quantified by P-values P_{ks} obtained from the Kolmogorov-Smirnov test(Theilhaber et al., 2002)(Keeping, 1995, p.259) performed against a uniform distribution. We thus find $P_{ks} = 6.4 \times 10^{-7}$ and $P_{ks} = 8.6 \times 10^{-5}$, for Figs. 10a and b, respectively: these small P-values simply confirm the visually obvious nonuniformity of the marker genes distributions.

The strong association between the group of marker genes, selected on the basis of biological relevance, and their ranking based on the test statistic

t^*, supports the overall validity of the statistical approach in selecting genes. The estimated detection sensitivity of relevant genes at false-discovery rate $F_d = 25\%$ is $S \sim 40\%$. The ultimate validation of the methodology must come however, from an in-depth, a posteriori investigation of the biological relevance of any other genes selected by the method: such an investigation is in progress and will be reported elsewhere.

4 Conclusions

Through systematic comparisons of large panels of gene expression data, we have found that to perform meaningful significance tests it is essential that the underlying statistical model includes a class-specific, correlated noise term, in addition to the usual statistically independent term, typically assumed in simpler models. The resulting "intraclass correlation model" (icc model) is characterized by an additional parameter, the noise correlation coefficient ρ, which is furthermore directly determined from the data. The role of noise correlation is crucial, in that it corrects for potentially gross overestimation of statistical significance, and avoids the conundrum of "all genes are significantly regulated".

We have found that the icc model can be incorporated into the conventional t-test by a very straightforward, sample-size dependent rescaling of the t statistic. For comparisons requiring more sensitivity however, we used a modified, biased-variance statistic t^*, and computed significance under the icc model by a semi-parametric resampling scheme using Monte Carlo simulation. This methodology was applied to the biological problem of finding genes differentially regulated in the central nervous system during fetal development, specifically those involved in the specification and differentiation of neuronal and glial cells. A preliminary validation of the approach was obtained by using a test set of marker genes, determined on the basis of biological expert knowledge. From this test set, we estimated that the sensitivity S of detection of biologically relevant genes, at false-discovery rate $F_d = 25\%$, was in the range $S \sim 40\%$. It should be noted however that the ultimate validation of the methodology, and of the genes selected by it, will come from a detailed, biologically focused investigation, now in progress, and to be presented in another context.

Finally, it should be emphasized that while the analyses presented here focused on using t-tests or modified t-tests for basic two-tissue comparisons, the underlying icc model itself is very general. In particular, the Monte Carlo resampling scheme we have presented can always be used to generate a reference data set, which can then used in conjunction with any other type of test statistic.

Appendix A: false discovery rates

In selecting for the most significant data we retain only qualifiers with $P \leq P_0$. Rather than choose the selection threshold P_0 a priori, it is more meaningful to determine it on the basis of the false discovery rate $F_d = F_d(P_0)$, defined as the estimated average fraction of false positives in the selection. F_d is approximately given by

$$F_d \;=\; \frac{N_r(P_0)}{N_f(P_0)} \;, \tag{21}$$

where $N_f(P_0)$ is the total number of qualifiers found with $P \leq P_0$, and where $N_r(P_0)$ is the corresponding number of qualifiers found in the reference distribution. In the text, $N_r(P_0)$ is either determined from an analytical expression (Eq.(9)), or empirically, from the distribution that results from data resampling.

Appendix B: the intraclass correlation (icc) model

Consider the data for two tissue panels, as described by Eqs.(1) and (2). For a given qualifier, we write the intensities $\{x_i\}$ and $\{y_j\}$ according to

$$\begin{aligned} x_i &= \mu + \alpha_x + \epsilon^T_{xi} \;, \qquad i = 1, \ldots, n_1 \;, & (22)\\ y_j &= \mu + \alpha_y + \epsilon^T_{yj} \;, \qquad j = 1, \ldots, n_2 \;. & (23) \end{aligned}$$

In Eqs.(22,23), μ is the mean expression level, common to both tissues, α_x and α_y are non-random, tissue-specific effects (with convention $\alpha_x + \alpha_y = 0$), and ϵ^T_{xi} and ϵ^T_{yj} are the “total” noise terms, which account for all random effects in the measurements. We assume that ϵ^T_{xi} and ϵ^T_{yj} can be written as the sums

$$\begin{aligned} \epsilon^T_{xi} &= \epsilon_{xi} + \eta_x \ , \qquad i = 1, \ldots, n_1 \qquad (24) \\ \epsilon^T_{yj} &= \epsilon_{yj} + \eta_y \qquad j = 1, \ldots, n_2, \qquad (25) \end{aligned}$$

where ϵ_{xi} and ϵ_{yj} are the uncorrelated components of the noise, and where η_x and η_y are tissue-specific components, common to all samples within a given tissue, but uncorrelated between tissues.

The noise terms ϵ_{xi} and ϵ_{yj} are assumed to have zero mean[3],

$$< \epsilon_{xi} > = < \epsilon_{yj} > = 0 \ , \qquad \text{for all } i, j \ , \qquad (26)$$

and identical variance σ^2,

$$\mathrm{var}(\epsilon_{xi}) = \mathrm{var}(\epsilon_{yj}) = \sigma^2 \ , \qquad \text{for all } i, j \ , \qquad (27)$$

and to be all mutually uncorrelated,

$$< \epsilon_{xi}\epsilon_{xi'} > = < \epsilon_{yj}\epsilon_{yj'} > = < \epsilon_{xi}\epsilon_{yj} > = 0 \ , \qquad \text{for all } i \neq i', \ j \neq j' \ . \ (28)$$

Similarly, the tissue-specific noise terms η_x and η_y are assumed to satisfy

$$\begin{aligned} < \eta_x > &= < \eta_y > = 0 \ , \qquad (29) \\ \mathrm{var}(\eta_x) &= \mathrm{var}(\eta_y) = \sigma^2_\eta \ , \qquad (30) \\ < \eta_x \ \eta_y > &= 0 \ , \qquad (31) \end{aligned}$$

and are assumed uncorrelated with all terms $\{\epsilon_{xi}\}$ and $\{\epsilon_{yj}\}$. Finally, all noise terms are assumed to be sampled from normal distributions.

Because of the additive structure of Eqs.(24,25), where a single tissue-specific term intervenes in each separate data series, the complete noise terms $\{\epsilon^T_{xi}\}$ and $\{\epsilon^T_{yj}\}$ are correlated within their respective tissue panels. Specifically, let us define $\rho^{(x)}_{ii'}$ to be the correlation coefficient for two noise terms ϵ_{xi} and $\epsilon_{xi'}$ in the $\{x_i\}$ data series, with $i' \neq i$. $\rho^{(x)}_{ii'}$ is given by

[3] In what follows, $< x >$ denotes the mean of random variable x.

$$\rho^{(x)}_{ii'} \;=\; \frac{\mathrm{Cov}(\epsilon^T_{xi}, \epsilon^T_{xi'})}{[\mathrm{var}(\epsilon^T_{xi}) \cdot \mathrm{var}(\epsilon^T_{xi'})]^{1/2}} \;, \tag{32}$$

where $\mathrm{Cov}(\epsilon^T_{xi}, \epsilon^T_{xj}) \equiv < (\epsilon^T_{xi} - < \epsilon^T_{xi} >)(\epsilon^T_{xi'} - < \epsilon^T_{xi'} >) >$ is the covariance of ϵ_{xi} and $\epsilon_{xi'}$, and with a similar equation holding for $\rho^{(y)}_{jj'}$. Based on Eqs.(24-31), we find that $\rho^{(x)}_{ii'}$ and $\rho^{(y)}_{jj'}$ are constant for all pairs of indices, and are given by

$$\rho^{(x)}_{ii'} \;=\; \rho^{(y)}_{jj''} \;=\; \rho \;, \qquad \text{for all } i \neq i',\; j \neq j' \;, \tag{33}$$

where ρ, the intraclass correlation coefficient for either tissue(Keeping, 1995, p.226), is given by

$$\rho \;=\; \frac{\sigma^2_\eta}{\sigma^2_\eta + \sigma^2} \;. \tag{34}$$

We shall assume that in a given data set, a single value of ρ applies to all qualifiers. This value of ρ thus characterizes the icc model, and the null hypothesis H'_0 under the icc model can be stated as

H'_0 : (35)

1. equal population means for the panel of lung and and the panel of liver samples (i.e. α_x, $\alpha_y = 0$ in Eqs.(22, 23)).
2. normally distributed noise amplitudes,
3. and noise terms are correlated within each tissue, with qualifier-independent correlation coefficient ρ.

Distribution of the t-statistic under the icc model

Consider the t statistic as defined by Eq.(3). Using Eqs.(22) and (23), the numerator in Eq.(3) can be written

$$u \;\equiv\; \bar{y} - \bar{x} \;=\; \bar{\epsilon_x} \;-\; \bar{\epsilon_y} \;+\; \eta_y \;-\; \eta_x \;. \tag{36}$$

where $\bar{\epsilon_x}$ and $\bar{\epsilon_y}$ are the sample means of $\{\epsilon_{xi}\}$ and $\{\epsilon_{yj}\}$, respectively. Using Eqs.(26-30) and Eqs.(34,36), we find that u is a normally distributed random variable, with mean and variance given by

$$\begin{aligned} < u > &= 0 \ , && (37) \\ \mathrm{var}(u) &= \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\sigma^2 \ + \ 2\sigma_\eta^2 \ . && (38) \end{aligned}$$

Note that the variance σ_η^2 of the tissue-specific noise terms *increases* the variance of $\bar{y} - \bar{x}$ relative to the what it would be in the absence of intraclass correlations. On the other hand, the sample variance s^2 is unaffected by the tissue-specific terms η_x and η_y, because they are subtracted out in each of the sums making up Eq.(4). s^2 can be written

$$s^2 \ = \ \sigma^2 \ \frac{S^2}{\nu} \ , \tag{39}$$

where $\nu = n_2 + n_1 - 2$ and where S^2, the sum of squares in Eq.(4) divided by σ^2, is distributed as a χ^2 variable with ν degrees of freedom. Using Eqs.(37-39), we can then write

$$t \ = \ \frac{\left(\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\sigma^2 \ + \ 2\sigma_\eta^2\right)^{1/2}}{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{1/2}\sigma} \ t' \ . \tag{40}$$

In Eq.(40), t' is given by

$$t' \ = \ \frac{z}{S^2/\nu} \ , \tag{41}$$

where z is a normally distributed random variable with mean 0 and variance 1, and where S^2 is distributed as χ^2 with ν degrees of freedom: it immediately follows that t' is distributed as Student-t with ν degrees of freedom(Spiegel, 1975, p.117). Simplifying and modifying the prefactor in Eq.(40), by using Eq.(34) to eliminate explicit reference of σ^2 and σ_η^2 in favor of ρ, leads to Eqs.(12) and (13). Finally, it should be noted that the entire derivation outlined above also applies to paired t statistics (Eq.(18)), by simply setting $n_1 = n_2 = n$ where n is the number of matched sample pairs.

Physical mechanisms for intraclass correlation

We have not pursued a systematic investigation of the possible physical sources of the intraclass correlation, which in Eq.(34) is only captured phenomenologically. We shall simply note that events occurring during the multiple steps which lead to the final hybridization of processed cRNA (harvesting of samples, storage, RNA extraction and mRNA amplification, etc) may affect the final representation of cRNAs in a tissue-dependent manner. In particular, the presence of different mixes of ribonucleases(D'Allesio and Riordan, 1997) in the tissues of origin may systematically change the abundances and compositions of different cRNA fragments downstream, once, as is inevitable, partial digestion of the original mRNAs has occurred after harvesting.

A more extreme view of intraclass correlation is that it represents an actual (in vivo) regulation of gene expression. In this picture, it is then true that nearly all genes significantly change in expression from one tissue to the next: however, it is also understood that through the icc model, we are folding into the null hypothesis "uninteresting" variation, that is, that part of biological variation with variance given by Eq.(34).

To conclude, we strongly suspect that both technological and biological effects are at play in intraclass correlation; they are indistinguishable in the phenomenological model presented here.

Appendix C: Monte Carlo resampling method

In order to build a reference data set approximating the null hypothesis H_0' of the icc model (Eqs.(35)) where it is crucial to maintain the observed intraclass correlation, a semi-parametric resampling method based on Monte Carlo simulation was used. The method proceeds in several steps as described below.

For each qualifier, independently:

1. For the kth time point in the expression profile (e.g. Table 1), evaluate the sample mean μ_k and sample variance σ_k^2, using the r_k replicates belonging to that time point.

2. Compute a global estimate σ^2 of the variance of uncorrelated noise across all time points, using the formula

$$\sigma^2 \;=\; \frac{1}{n-K}\,\sum_{k=1}^{K}\sum_{j=1}^{r_k}(x_{kj}-\mu_k)^2 \;, \tag{42}$$

   where K is the total number of time points, n the total number of samples, and where x_{kj} is the jth replicate at the kth time point.

3. Compute the variance σ_η^2 of the correlated noise, by solving Eq.(34) for σ_η^2,

$$\sigma_\eta^2 \;=\; \sigma^2\frac{\rho}{1-\rho} \;. \tag{43}$$

   where ρ is estimated from the original data, Eq.(15).

4. Generate new intensities x^*_{kj} according to a Monte Carlo scheme, with

$$x^*_{kj} \;=\; \mu_k \;+\; \sigma_k\,\xi_{kj} \;+\; \sigma_\eta\,\zeta \;, \tag{44}$$

   where ξ_{kj} , $k=1,2,\ldots,K$ and $j=1,2,\ldots,r_k$ are independent, Gaussian random variables with zero mean and unit variance, and where ζ is also a Gaussian random variable with zero mean and unit variance. All random variables $\{\xi_{kj}\}$ and ζ are mutually uncorrelated.

ACKNOWLEDGMENTS

We wish to thank Paul Giresi, Fraser Sim, Min Wang, Sean Ge and Brian Moldover for comments and contributions pertaining to this work, and Neeta Roy and Han Soo who have been centrally involved in the experimental generation of the neural developmental data (used here primarily as a test bed of the methodology). The expression data sets for the lung-liver comparisons were generated by the Aventis Respiratory and Toxicology groups, respectively (in contexts unrelated to the present analysis).

References

Abramowitz, M., Stegun, I. A. (1972) *Handbook of Mathematical Functions.* Dover, New York.

Alizadeh, A. et al. (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403**, 503–512.

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

Chenn, A., Braisted, J. E., McConnell, S., O'Leary, D. M. (1997) Development of the cerebral cortex: mechanisms controlling cell fate, laminar and areal patterning, and axonal connectivity. In Cowan, W. M., Jessell, T. M., Zipursky, S. L. (eds), *Molecular and Cellular Approaches to Neuronal Development.* Oxford University Press, Oxford, pp.440–473.

D'Allessio, G., Riordan, J. F. (1997) *Ribonucleases: Structures and Functions.* Academic Press, New York.

Durbin, B. P., Hardin, J. S., Hawkins, D. M., Rocke, D. M. (2002) A variance-stabilizing transformation for gene-expression microarray data. Proc. 10th Int. Conf. on Intelligent Systems for Molecular Biology (ISMB 2002). *Bioinformatics* **18** supp. 1, S105-S110.

Eisen, M. B., Spellman, P. T., Brown, P. O., Botstein, D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* **95**, 14863–14868.

Goldman, S. A., Luskin, M. B. 1998. Strategies utilized by migrating neurons of the postnatal vertebrate forebrain. *Trends in Neuroscience* **21**, 107–114.

Jin, W., Riley, R. M., Wolfinger, R. D., White, K. P., Passador-Gurgel, G., Gibson, G. (2001) The contributions of sex, genotype and age to transcriptional variance in *Drosophila melanogaster. Nature genetics* **29**, 389–395.

Keeping, E. S. (1995) *Intoduction to Statistical Inference.* Dover, New York.

Keyoung, H. M., Roy, N. S., Benraiss, A., Louissaint, Jr., B., Suzuki, A., Hashimoto, M., Rashbaum, W. K., Okano, H., Goldman, S. (2001) High-yield selection and extraction of two promoter-defined phenotypes of neural stem cells from the fetal human brain. *Nature biotechnology* **19**, 843–850.

Lockhart, D. J. et al.(1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. Nat. Biotechnol. **14**, 1675-1680.

Ross, D. T. et al. (2000) Systematic variation in gene expression patterns in human cancer cell lines. *Nature genetics* **24**, 227–244.

Spiegel, M., R. (1975) *Theory and Problems of Probability and Statistics* McGraw-Hill, New York.

Spellman P. T., Sherlock, G., Zhang, M. Q., Iyer V. R., Anders, K., Eisen M. B., Brown, P. O., Botstein, D., Futcher, B. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol Biol Cell.* **9**, 3273–97.

Theilhaber, J., Bushnell, S., Jackson, A., Fuchs, R. (2001) Bayesian Estimation of Fold-Changes in Gene Expression: the PFOLD Algorithm. *Journal of Computational Biology* **8**:585-614.

Theilhaber, J., Connolly, T., Roman-Roman, S., Bushnell, S., Jackson, A., Call, K., Garcia, T., Baron, R. (2002) Finding Genes in the C2C12 Osteogenic Pathway by k-Nearest-Neighbor Classification of Expression Data. *Genome Research* **12**:165–176.

Tusher, V., G., Tibshirani, R., Chu, G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. USA* **98**, 5116–5121.

Wegner, M. (2000) Transcriptional control in myelinating glia: flavors and spices. *GLIA* **31**, 1–14.

TABLE CAPTIONS

Table 1: List of ventricular zone (vz) and cortex (cx) tissue samples profiled on Affymetrix chips for the neuronal development study. Each row corresponds to a matched (vz, cx) pair of tissues; the grouped time points are defined for the Monte Carlo resampling described in Section 3.1.

Table 2: The 24 biological marker genes used to represent astrocytic, neuronal, oligodendrocytic and radial glial cell lineages (mapping into 36 distinct Affymetrix qualifiers). CELL: primary cell type; QUALIFIERS: number of qualifiers for each gene.

FIGURES CAPTIONS

Fig. 1: Cumulative distributions $N_f(P_0)$ (Eq.(7)) of P-values obtained for t-tests performed on the lung-liver data set, for random subsamplings of the lung and liver tissue panels, with subsample sizes $m_1 = m_2 \equiv m$ indicated in the plot. The straight line with slope -1 (leftmost), denotes the reference distribution $N_r(P_0)$ (Eq.(8)) that would obtain under the null hypothesis. Note that the figure is a log-log plot, with the most significant data on the right-hand side.

Fig. 2: Histogram of the t statistic for the lung-liver data set (8758 qualifiers), for $m = 30$ samples in each tissue (with 50 bins equally spaced over $-50 \leq t \leq 40$). The Student-t distribution with $\nu = 58$ degrees of freedom, the expected distribution under null hypothesis H_0, is indicated in the center (arbitrary units).

Fig. 3: Cumulative distribution $N_f(P_0)$ of P-values obtained for t-tests performed on the lung-liver data set after complete randomization of column assignments on a qualifier-by-qualifier basis, for $m = 30$ samples in each tissue panel. As in Fig. 1, the straight line denotes the reference distribution expected under the null hypothesis H_0.

Fig. 4: Cumulative distributions $N_f(P_0)$ of P-values obtained for t-tests performed on the lung-liver data set, *after* renormalization of the t statistic by Eq.(12), for all the comparisons already considered in Fig. 1. Subsample sizes m for each comparison are indicated in the plot. The straight line with slope -1 (leftmost), denotes the reference distribution that obtain under the new null hypothesis H_0'.

Fig. 5: Dependence of the intraclass correlation coefficient ρ and the scaling factor β on the number of samples m used in the lung-liver comparisons. Values of the estimators $\hat{\rho}$ (diamonds) and $\hat{\beta}$ (triangles) were generated from t-tests performed between random, equal sized subsamplings of the 30 lung and 30 liver samples. Subsamples without replacement, of size $m_1 = m_2 \equiv m$, $1 \leq m \leq 20$, were used, with 5 independent subsamplings generated for each value of m.

Fig. 6: Dependence of the number N_F of qualifiers found in the lung-liver comparisons, as a function of sample size m in both tissue panels, after renormalization of the t statistic (see Fig. 4); a constant false-discovery rate $F_d = 0.25$ is imposed. 5 independent subsamplings were generated for each value of m.

Fig. 7: Distributions $N_f(P_0)$ for the cortex-vz comparisons (paired t-tests, 11 samples in each tissue panel). a) unrenormalized distribution; b) distribution after renormalization ($\hat{\rho} = 0.083$, $\hat{\beta} = 0.700$). The reference distributions $N_r(P_0)$ are indicated by the straight lines with slope -1.

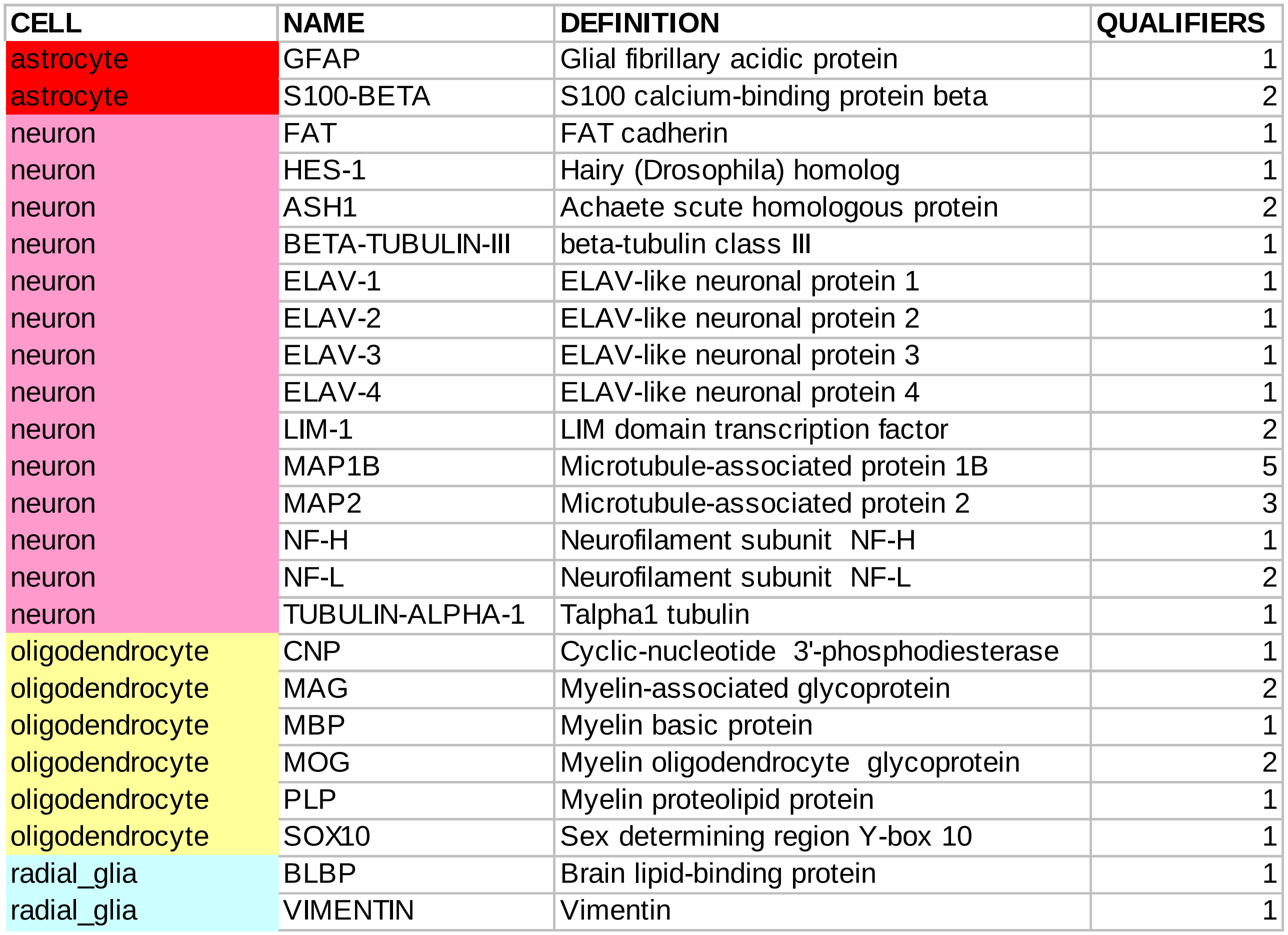
Fig. 8: Cortex-vz comparisons (11 samples in each tissue panel), using the biased-variance statistic, Eq.(20), and Monte Carlo resampling to establish the reference distributions. The plots show dependence on variance bias for $\sigma_0 = 0, 100, 250, 2500$, as indicated at the top of the figures. The top panels (a,c,e,g) compare the observed and reference distributions (heavy dots, $N_f(P_0)$; thin lines, $N_r(P_0)$); the bottom panels (b,d,f,h) display the corresponding false discovery rates as a function of P_0. The vertical lines signal the decision thresholds where a false-discovery rate $F_d = 0.25$ obtains.

Fig. 9: Dependence of the number of qualifiers N_F found in the cortex-vz comparisons, as a function of variance bias term σ_0 (Eq.(20)), for a constant false-discovery rate $F_d = 0.25$.

Fig. 10: Cumulative distributions of the 36 biological lineage markers (24 distinct genes, Table 2) in the global population of expression profiles ranked according to the biased variance statistic t^*, with $\sigma_0 = 2500$; rank 1 always denotes the most significant profile. The straight diagonal lines (red) indicate the average distribution under a random sampling of the population, and the vertical lines (blue) the decision threshold for false-discovery rate $F_d = 25\%$. a) vz > cx: 24 markers (18 genes) occur in a population of 40575 qualifiers, $P_{ks} = 6.4 \times 10^{-7}$; b) vz < cx: 12 markers (7 genes) occur in a population of 32,798 qualifiers, $P_{ks} = 8.6 \times 10^{-5}$.

Donor #	**Age**	**Grouped time point** k	**Replicates** r_k
1	E15	1	3
2	E16		
3	E16		
4	E18	2	3
5	E18		
6	E18		
7	E21	3	3
8	E22		
9	E23	4	3
10	E23		
11	E23		

Table 1.



CELL	NAME	DEFINITION	QUALIFIERS
astrocyte	GFAP	Glial fibrillary acidic protein	1
astrocyte	S100-BETA	S100 calcium-binding protein beta	2
neuron	FAT	FAT cadherin	1
neuron	HES-1	Hairy (Drosophila) homolog	1
neuron	ASH1	Achaete scute homologous protein	2
neuron	BETA-TUBULIN-III	beta-tubulin class III	1
neuron	ELAV-1	ELAV-like neuronal protein 1	1
neuron	ELAV-2	ELAV-like neuronal protein 2	1
neuron	ELAV-3	ELAV-like neuronal protein 3	1
neuron	ELAV-4	ELAV-like neuronal protein 4	1
neuron	LIM-1	LIM domain transcription factor	2
neuron	MAP1B	Microtubule-associated protein 1B	5
neuron	MAP2	Microtubule-associated protein 2	3
neuron	NF-H	Neurofilament subunit NF-H	1
neuron	NF-L	Neurofilament subunit NF-L	2
neuron	TUBULIN-ALPHA-1	Talpha1 tubulin	1
oligodendrocyte	CNP	Cyclic-nucleotide 3'-phosphodiesterase	1
oligodendrocyte	MAG	Myelin-associated glycoprotein	2
oligodendrocyte	MBP	Myelin basic protein	1
oligodendrocyte	MOG	Myelin oligodendrocyte glycoprotein	2
oligodendrocyte	PLP	Myelin proteolipid protein	1
oligodendrocyte	SOX10	Sex determining region Y-box 10	1
radial_glia	BLBP	Brain lipid-binding protein	1
radial_glia	VIMENTIN	Vimentin	1

Table 2.

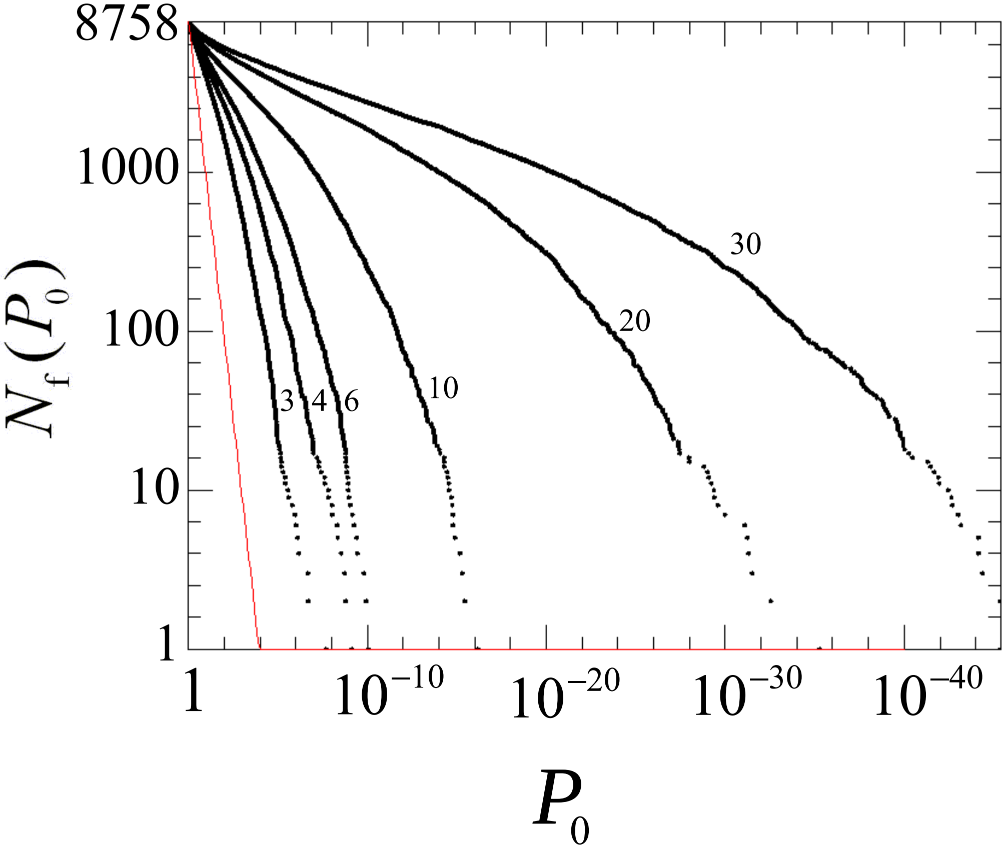

8758
1000
100
10
1
$N_\mathrm{f}(P_0)$
1
10^{-10}
10^{-20}
10^{-30}
10^{-40}
P_0
3
4
6
10
20
30

Fig. 1

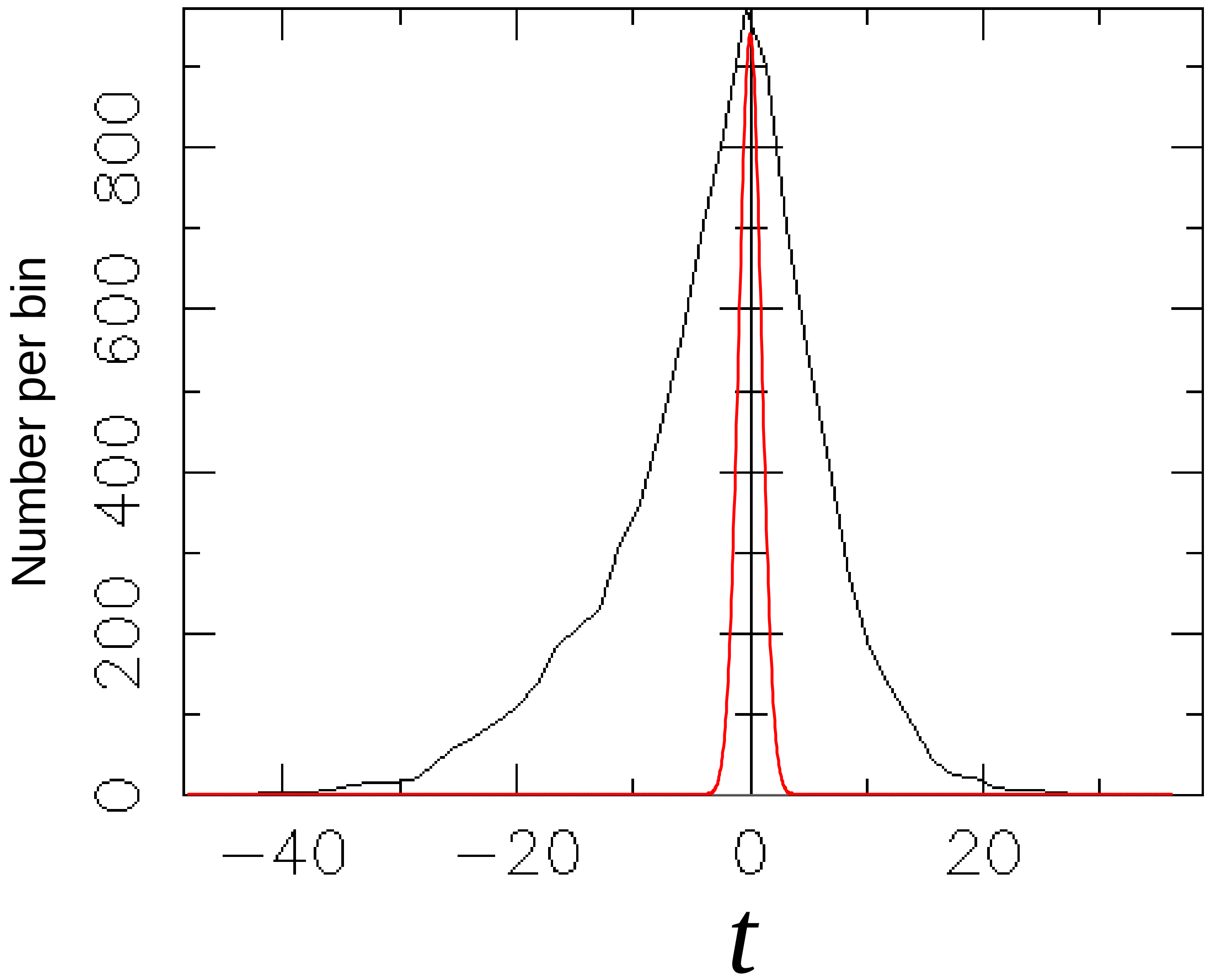


Fig. 2

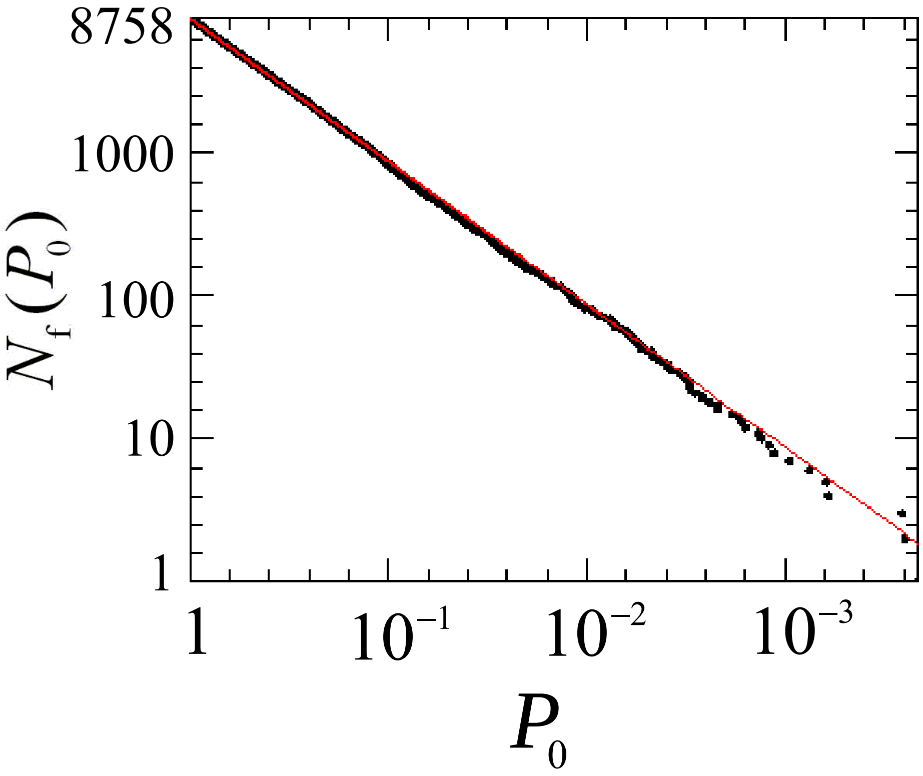


Fig. 3

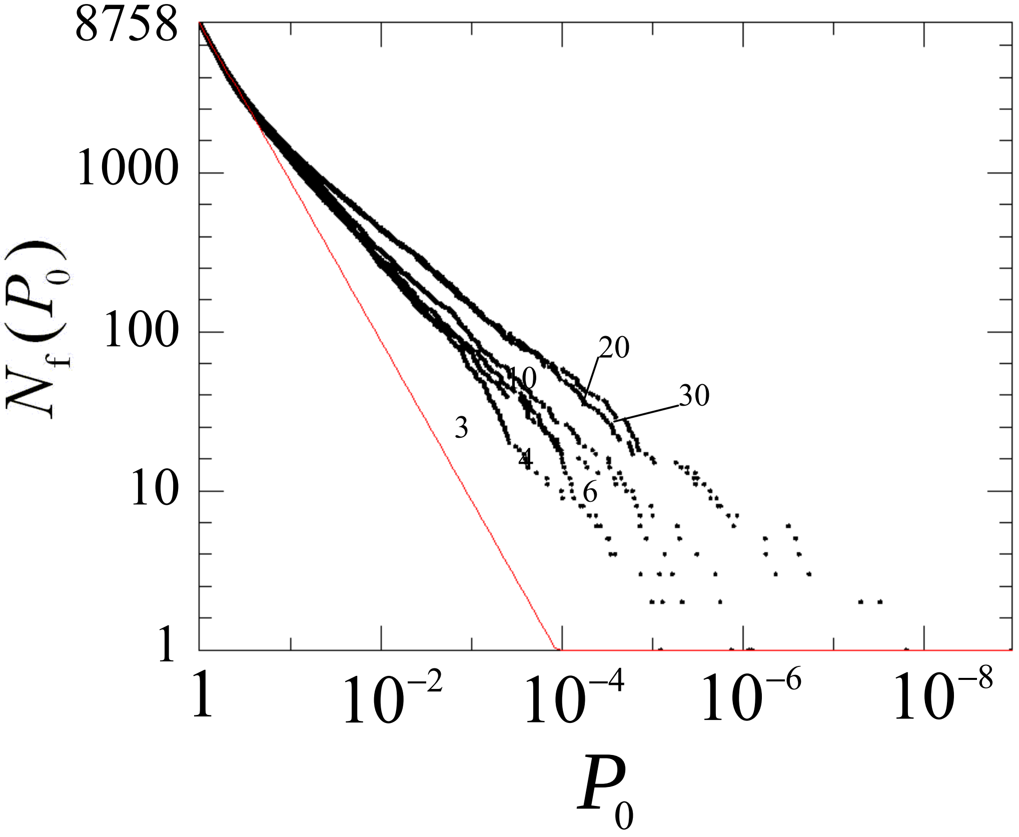


Fig. 4

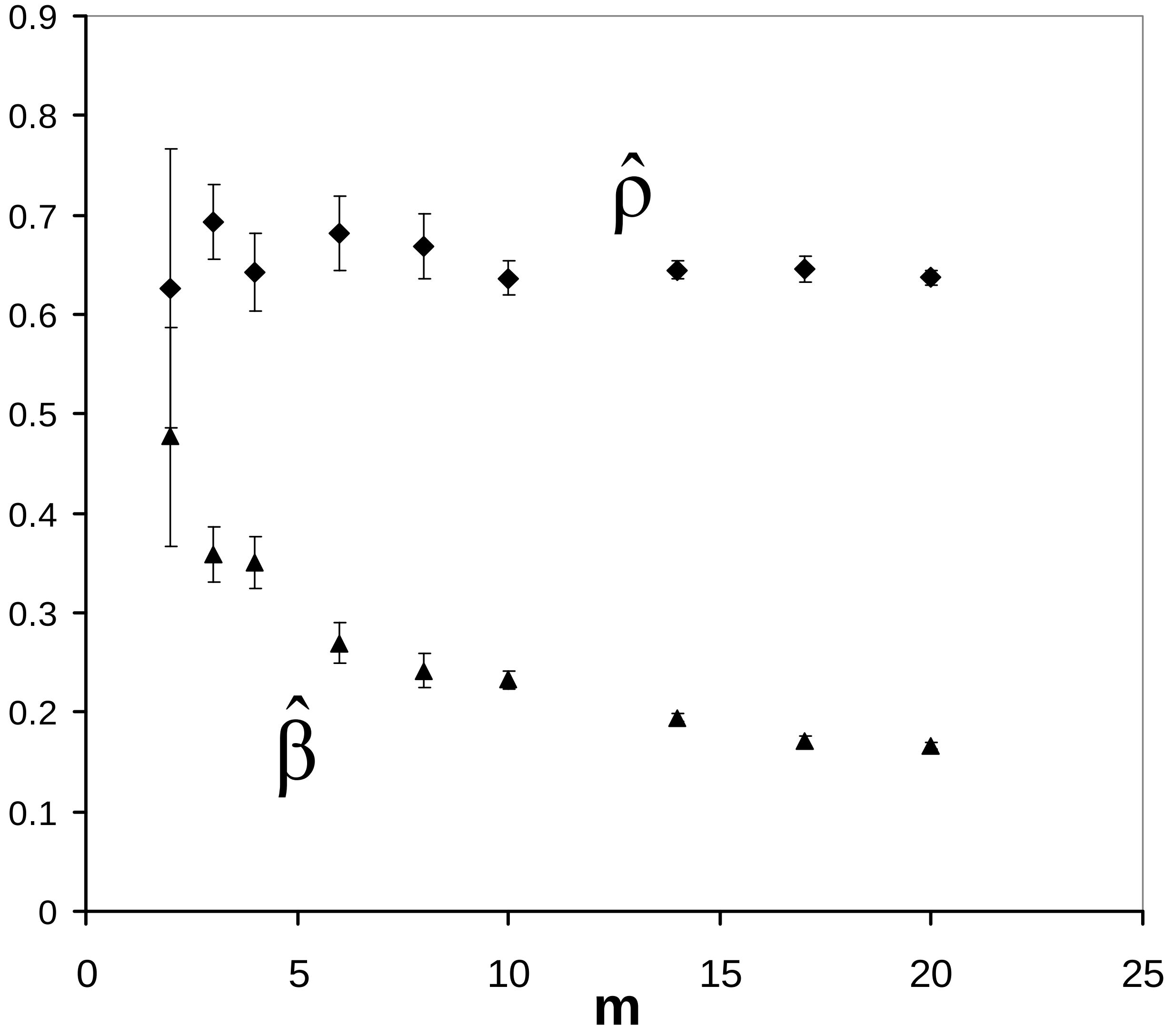


Fig. 5

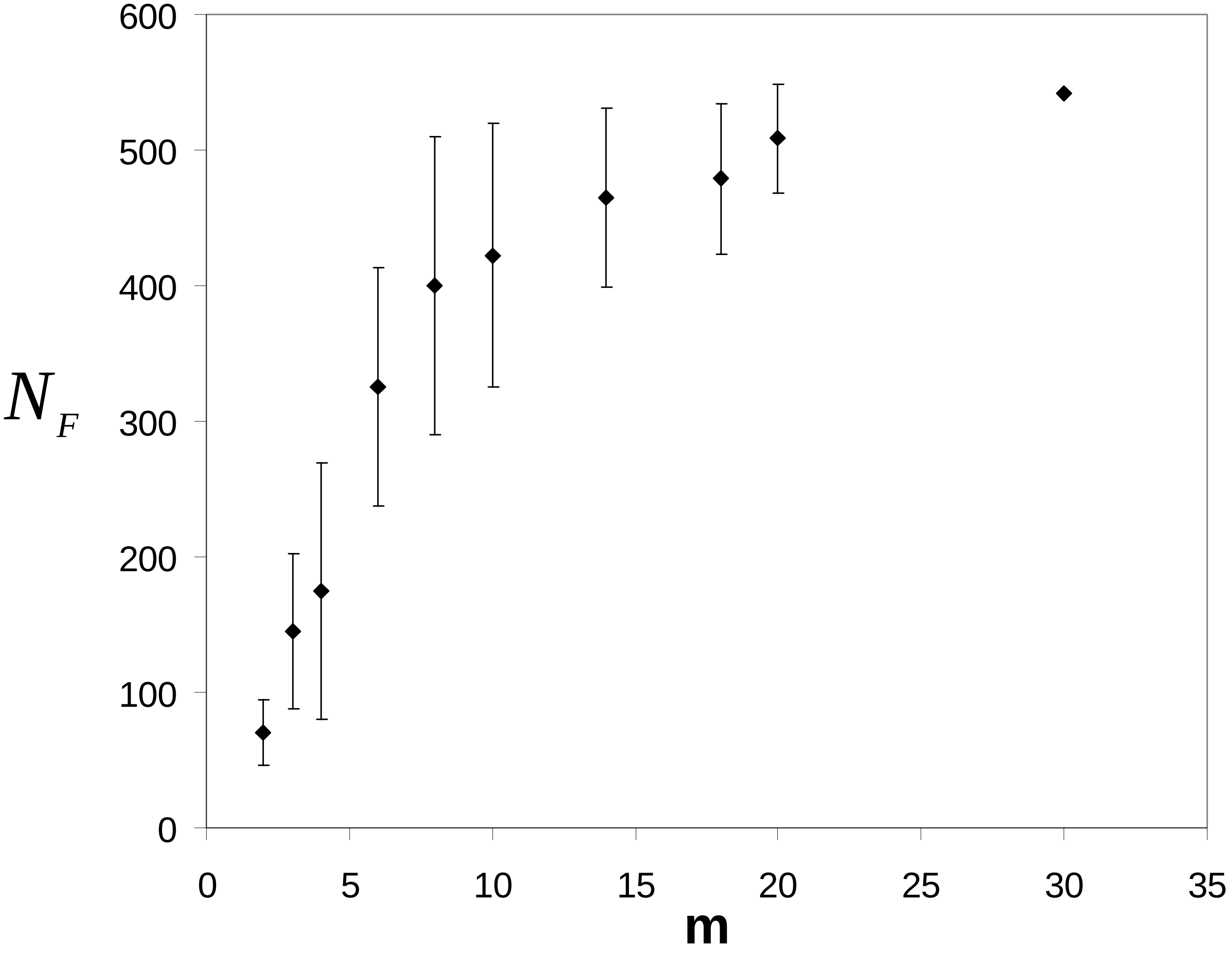


Fig. 6

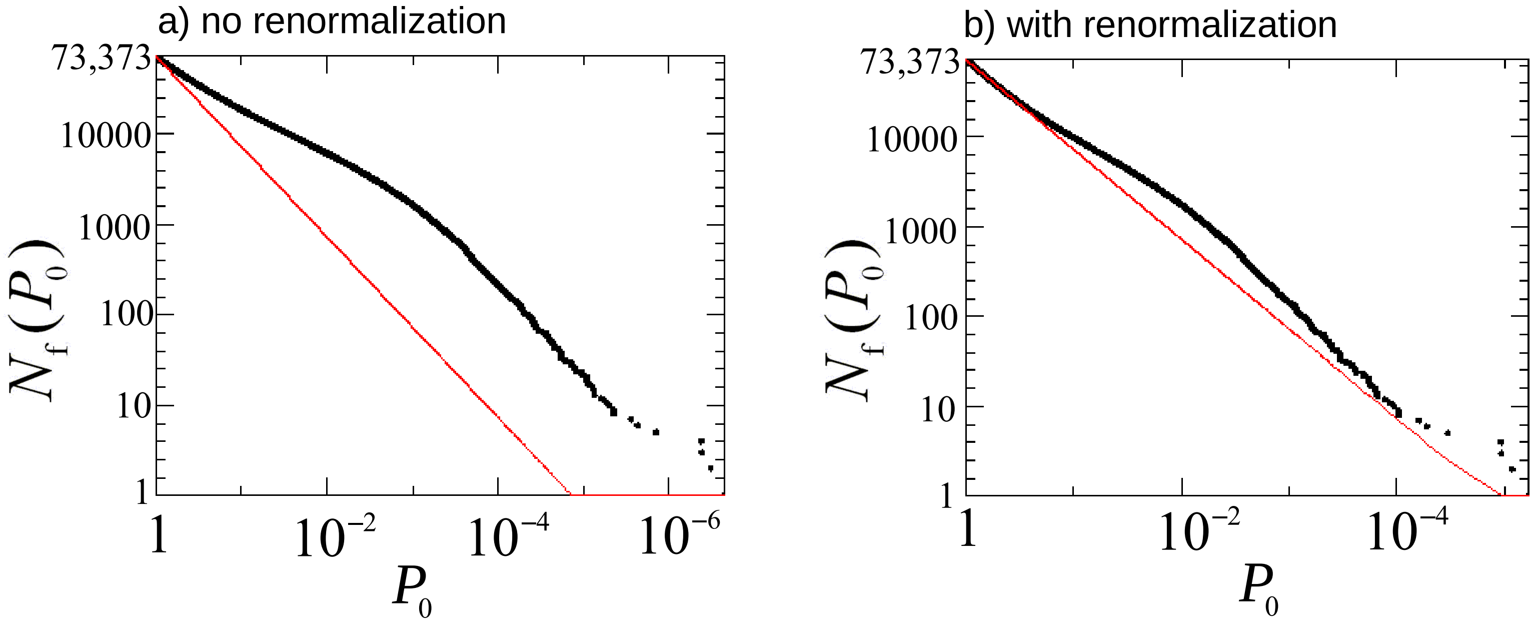

a) no renormalization
b) with renormalization
73,373
10000
1000
100
10
1
$N_f(P_0)$
1
10^{-2}
10^{-4}
10^{-6}
P_0

Fig. 7

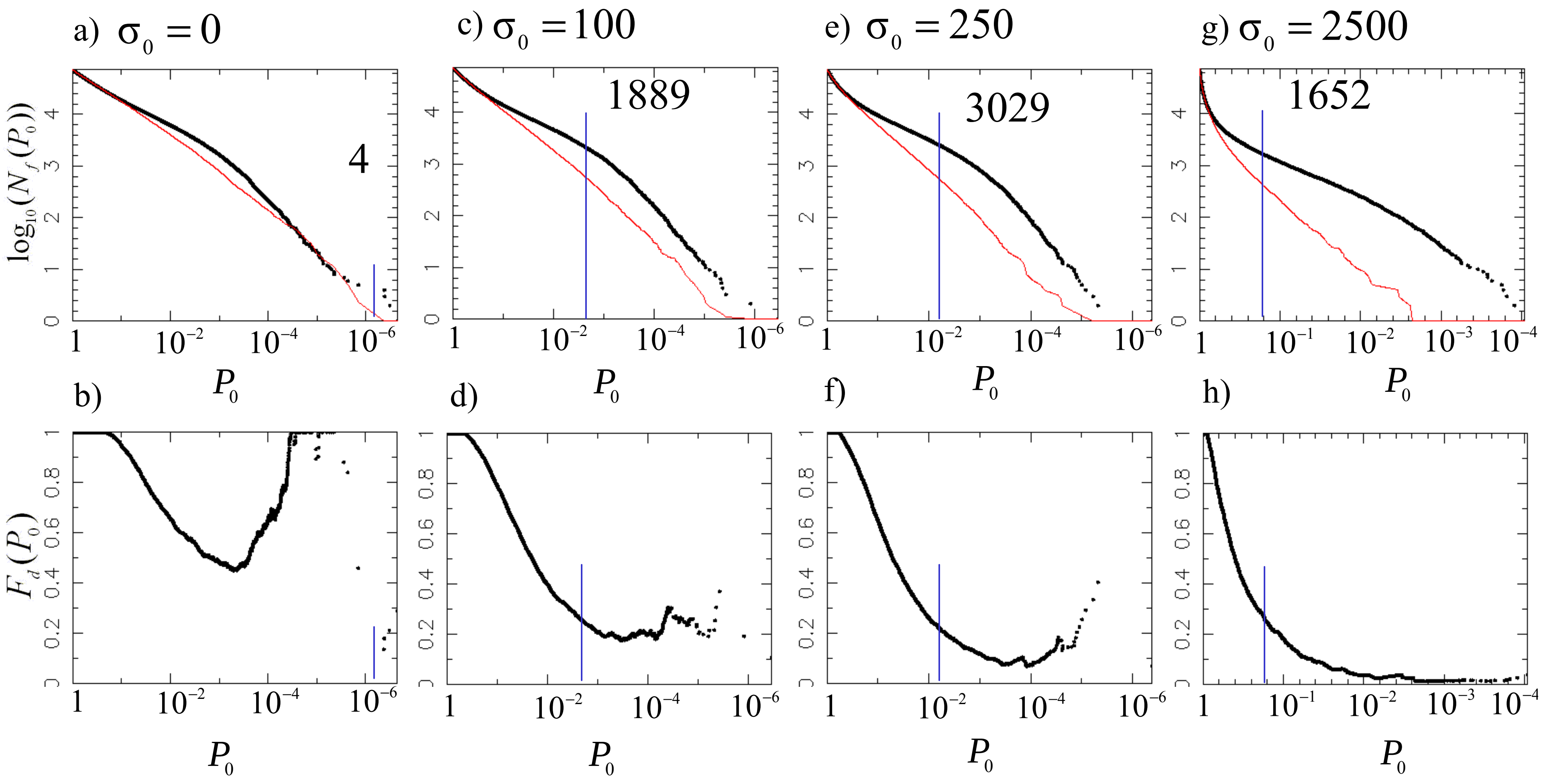

a) $\sigma_0 = 0$
c) $\sigma_0 = 100$
e) $\sigma_0 = 250$
g) $\sigma_0 = 2500$
$\log_{10}(N_f(P_0))$
4
1889
3029
1652
0
1
2
3
4
1
10^{-2}
10^{-4}
10^{-6}
10^{-1}
10^{-3}
P_0
b)
d)
f)
h)
$F_d(P_0)$
0
0.2
0.4
0.6
0.8
1

Fig. 8

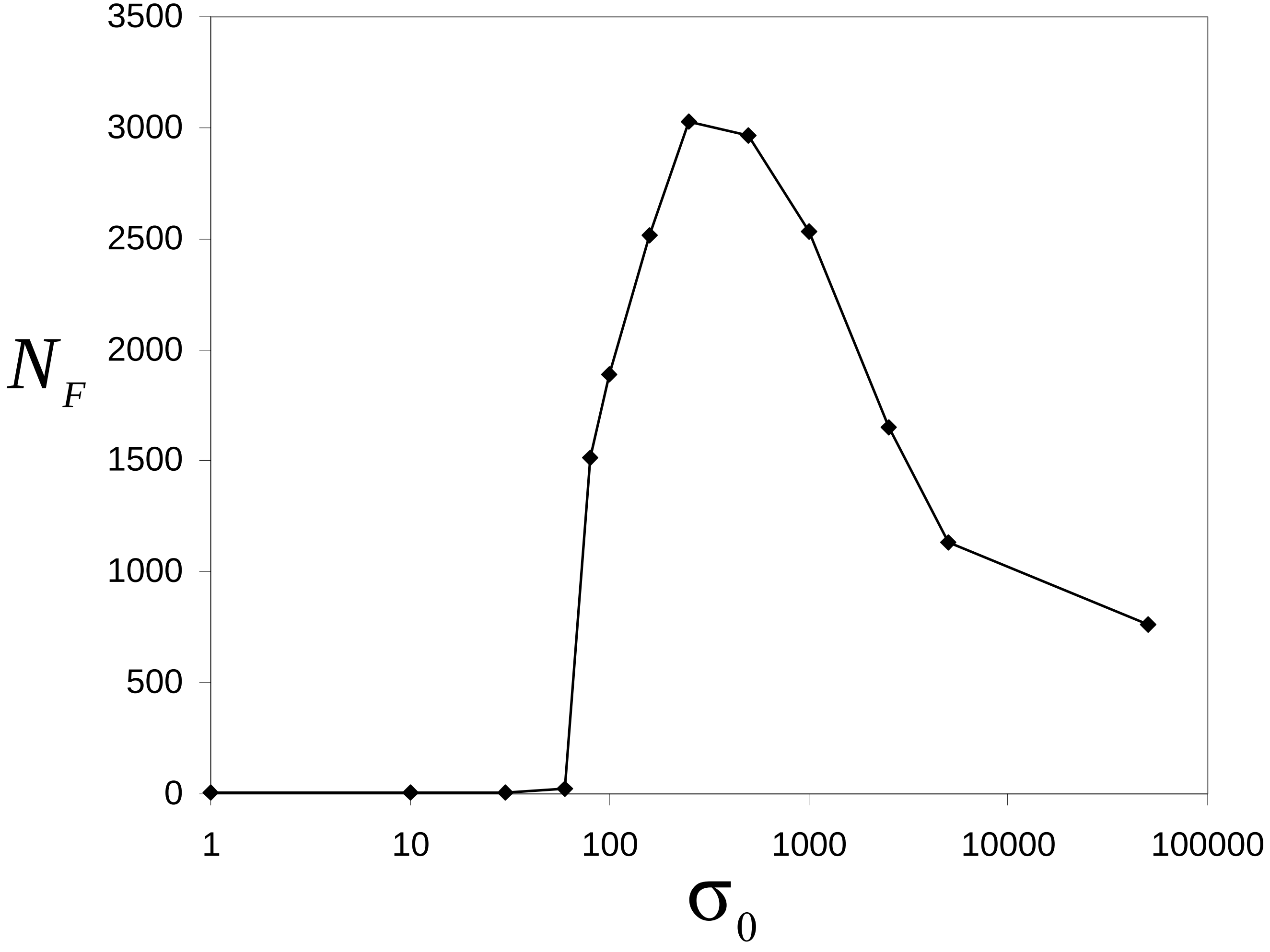

3500
3000
2500
2000
1500
1000
500
0
N_F
1
10
100
1000
10000
100000
σ_0

Fig. 9

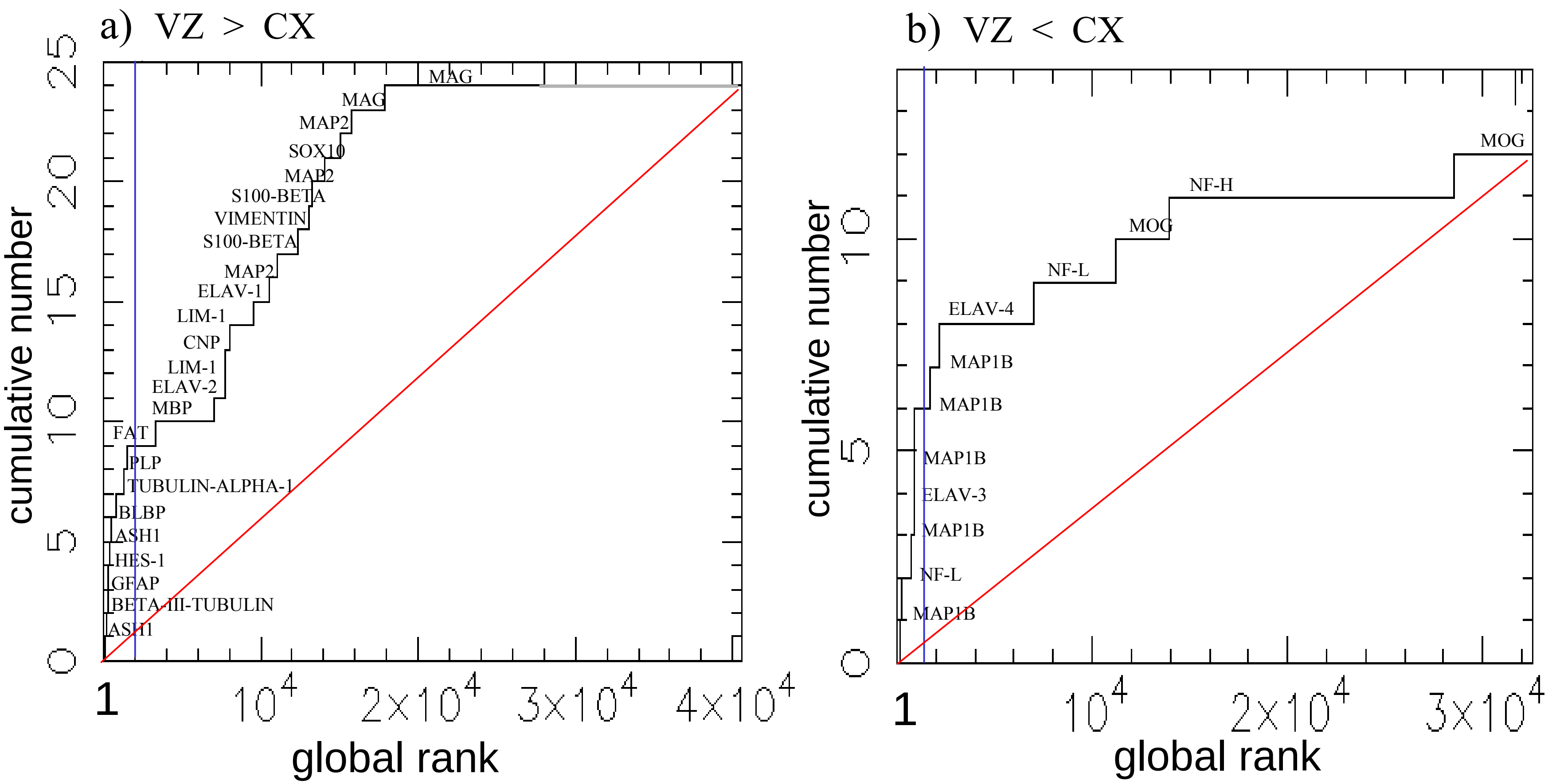

a) VZ > CX
cumulative number
0
5
10
15
20
25
1
10^4
2×10^4
3×10^4
4×10^4
global rank
MAG
MAG
MAP2
SOX10
MAP2
S100-BETA
VIMENTIN
S100-BETA
MAP2
ELAV-1
LIM-1
CNP
LIM-1
ELAV-2
MBP
FAT
PLP
TUBULIN-ALPHA-1
BLBP
ASH1
HES-1
GFAP
BETA-III-TUBULIN
ASH1
b) VZ < CX
cumulative number
0
5
10
1
10^4
2×10^4
3×10^4
global rank
MOG
NF-H
MOG
NF-L
ELAV-4
MAP1B
MAP1B
MAP1B
ELAV-3
MAP1B
NF-L
MAP1B

Fig. 10