Normalization Using Weighted Negative Second Order Exponential Error Functions (NeONORM) Provides Robustness Against Asymmetries in Comparative Transcriptome Profiles and Avoids False Calls

Sebastian Noth, Guillaume Brysbaert, and Arndt Benecke*

Systems Epigenomics Group, Institut des Hautes Etudes Scientifiques/Institut de Recherches Interdisciplinaires, CNRS/INSERM, 91440 Bures sur Yvette, France.

Studies on high-throughput global gene expression using microarray technology have generated ever larger amounts of systematic transcriptome data. A major challenge in exploiting these heterogeneous datasets is how to normalize the expression profiles by inter-assay methods. Different non-linear and linear normalization methods have been developed, which essentially rely on the hypothesis that the true or perceived logarithmic fold-change distributions between two different assays are symmetric in nature. However, asymmetric gene expression changes are frequently observed, leading to suboptimal normalization results and in consequence potentially to thousands of false calls. Therefore, we have specifically investigated asymmetric comparative transcriptome profiles and developed the normalization using weighted negative second order exponential error functions (NeONORM) for robust and global inter-assay normalization. NeONORM efficiently damps true gene regulatory events in order to minimize their misleading impact on the normalization process. We evaluated NeONORM's applicability using artificial and true experimental datasets, both of which demonstrated that NeONORM could be systematically applied to inter-assay and inter-condition comparisons.

Key words: microarray, statistical analysis, normalization, asymmetry, robustness

Introduction

Studying the cellular transcriptome and its dynamics using microarray technology has become a common application in modern biomedical research (*1*), and has spurred over the past two decades many novel insights into the mechanisms of gene regulation and physiopathology (*2–5*). Microarray technology has seen considerable advances in the number of transcripts that can be detected simultaneously as well as the precision of the measurement. Several all-in-one commercial solutions and a multitude of array scanner systems in the public domain have emerged. However, these technologic solutions often diverge in the design of the probes and the method of transcript detection, making the statistical analysis of transcriptome data challenging and non-uniform (*6*, *7*). Microarray technology cannot be used for absolute but only relative quantification of the abundance of individual transcripts (*8*). Obviously, this fact greatly complicates the successful analysis of microarray experiments. As for inter-assay comparisons, which hence need to be carried out for relative quantification between two different biologic samples, efficient and accurate normalization methods have to be developed (*9–11*).

In the first step, intra-assay normalization methods account for variations in print-tip quality, irregular sample distribution over the array surface, irregularities in the surface itself, camera aperture related distortions, non-linear detection-dye dynamic ranges, and (depending on the technology used) many other phenomena that contribute to technical variations for individual probes (*10*, *12*, *13*). Then, inter-assay normalization techniques in the optimal case capture technical variations (which are due to sample preparation, extraction, amount, or labeling), quality variations, dye related differences (2 dye setup) and vari-

***Corresponding author.**
E-mail: arndt@ihes.fr

ations, chemical batch variations, array batch variations, and so on. If intra-array normalization methods have been successfully applied before, inter-assay normalization techniques usually require a global rescaling of the entire datasets relative to each other and thus are of homogenous nature (*11*, *14*).

Many of the non-linear and basically all linear normalization methods thereby make two assumptions about the nature of the data: (1) invariance in the absolute quantity of a majority of transcripts, and (2) symmetry of the probe variation distribution (*9–14*). Note that normalization methods are based solely on control probes for which synthetic transcripts are added at different moments during the experimentation, assuming analogously invariance and symmetry of the control signals. Since solely internal control based normalization fails to capture sample related technical variation and is not very robust, it is rarely used by itself. The first invariance assumption for the majority of probes, in the limit of large probe sets, seems to hold and can be biologically justified. Furthermore, this assumption is necessarily required for the principle of inter-assay normalization based on probe signals to be meaningful. The second assumption about symmetry of the probe variation distribution, which is exploited in non-discriminant averaging normalizations, cannot be justified from a biological point of view. This assumption only holds true for technical replicates generated from a single biologic sample.

Biological questions frequently pose concern about the dynamics in the transcriptome profile when comparing two different states of cell differentiation (*15*), or the differences and/or similarities between a physiologic and a pathologic state of a cell (*16*). When comparing distinct physiologic situations, the symmetry assumption seems to be poorly reflecting biological reality. A multitude of biologic processes can be cited that will consequently lead to asymmetric changes in the expression profile of a cell. For instance, inherently asymmetric processes such as apoptosis or mitotic repression will lead to a majority of genes being down-regulated and only a comparatively small fraction being induced (*17*, *18*). Furthermore, due to their very different mechanistic nature, transcription activation and repression mechanisms follow distinct dynamics. Consequently, such asymmetries are frequently observed in "real world" data (*19*, *20*). Since these asymmetries are not corrected in technical and/or biological replicates, they reflect true biological variations. Most inter-assay normalization methods do not account for such asymmetries and use non-discriminant averaging (mean, median) for normalization of these data. It is obvious that averaging methods, provided the asymmetry is of sufficient significance to exceed numerical precision, will lead to suboptimal normalization factor estimation. Such considerations, especially in the case of the so-called boutique arrays where the invariance assumption cannot be made, have lead to the development of discriminant normalization methods (*21*, *22*). Here, only either pre-defined or conditionally determined subsets of probes are used for the normalization process. For the former, either the so-called house-keeping genes that are thought to be invariant in their expression across a large number of cellular conditions, or external and synthetic probe/target pairs are added to the experimental pipeline at different points in time, and then are considered for normalization. However, such methods have significant disadvantages. The notion of a house-keeping gene is at best empiric and often circumstantial (*23*). Dynamic determination (no *a priori* assumptions made) of the probes to be considered for normalization could be a solution to this problem. Unfortunately, few methods exist for such a type of inter-assay normalization (*22*, *24*). Those methods, however, are either expression rank-based (*22*, *25–27*), or depend on statistical tests that are performed after inter-assay comparison/subtraction profiles have been calculated (*14*). Especially in view of the increasing sensitivity of some microarray solutions, analysis methods that do not or only partially consider information on individual signal variance, will lead to suboptimal statistical interpretation of the transcriptome data (*9*).

While analyzing high-density kinetic transcriptome data for myeloid cell differentiation, we have realized that the asymmetries in the probe variance distributions did significantly compromise the median-based inter-assay normalization. We employed the novel AB1700 platform (Product Info: http://www.appliedbiosystems.com) for transcriptome analysis. As detailed in Materials and Methods, several non-linear and linear intra-assay normalization techniques were systematically and automatically applied during the primary analysis of raw image data. The resulting probe signal estimates were considered sufficiently normalized to directly proceed with inter-assay comparisons (http://www.appliedbiosystems.com). However, when carefully analyzing the phenomenon of asymmetric probe variance distributions, we realized

that additional means of inter-assay normalization were required. Furthermore, after carefully reviewing the literature as well as the data generated by a different technology, we realized that this problem seems to be not limited to a particular technology or biologic model (*19*, *20*). Unsatisfied with existing solutions to the normalization problem, we developed a novel method, namely NeONORM, for inter-assay normalization that is insensitive to asymmetries in probe variation distributions. This study summarizes the development of NeONORM and its evaluation using synthetic test and "real world" data.

Results and Discussion

Derivation of the NeONORM error function

We specifically thought to develop an inter-assay normalization method for transcriptome or similar data that overcomes the problems associated with asymmetric heavy tails in fold-change distributions (Figure 1). This method would supposedly be applied once (generally non-linear) intra-assay normalizations (such as print-tip correction) have already successfully been applied. Our central assumption in the reasoning leading up to the NeONORM method is that except for technical replicates, the common hypothesis of symmetry of fold-change distributions is not well founded for inter-assay comparisons. In order to avoid implementing several different methods specifically tailored to and distinguishing explicitly between technical replicates, biological replicates, and comparisons of truly different conditions, ideally, this novel method would also intrinsically adapt according to the nature of the inter-assay comparison. Therefore, the method would behave in the limit of technical replicates and highly symmetric fold-change distributions, which is closely similar to the existing, averaging, and linear methods such as Median normalization.

In addition to a quadratic error function (quadratic in the signal difference x), we construct a damping function such that it would restrict the maximum contribution of a large logQ (defined as the binary logarithm of the signal quotient) on the local quadratic error (small variation Δx of x, see Figure 2A). In the limit of Δx approaching zero, the influence of the damping function should disappear. For illustration, we have sketched the product of two

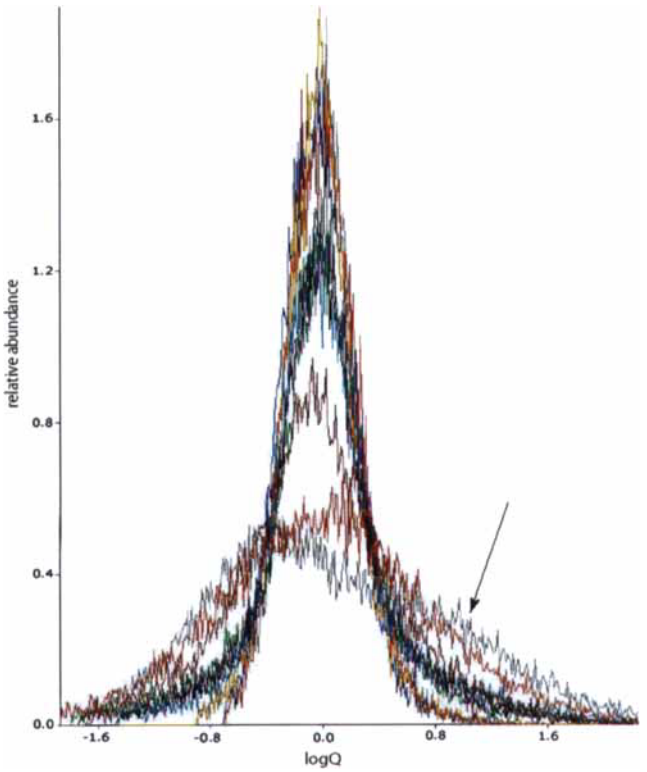

Fig. 1 Asymmetric heavy tails in fold-change distributions. A histogram view of nine different "real-world" superimposed fold-change distributions is shown as calculated from recent experimental work in our laboratory. The arrow indicates asymmetric heavy tails that are observed to different degrees in such experimental data.

such functions in Figure 2A. A simple multiplication of both the quadratic error function with the damping function obviously would result in lower error contributions once $|x|$ exceeds the equilibrium point (EP) between the contributions of both functions, and can hence not be used. The NeONORM error function, however, approximates the multiplicative functions in the bounds of $|x|$ <EP, and for $|x|$ >EP, the error becomes essentially invariant (Figure 2A). Additionally, the individual contributed errors would be scaled as a function of the quality of measurement (cumulated variances over both individual signals). The formal derivation of this modified NeONORM error function can be found in the supporting online material ("NeONORMformalism.pdf"). In short:

Damping the derivative of the quadratic error function with a Gaussian:

$$damp\,(z) = e^{-z^2} \tag{1}$$

yields the overall error function after integration and introduction of weights:

$$err_{neoC}(a) = \sum_{i=1}^{n} w_i \cdot \left(1 - e^{-(x_i - a)^2}\right) \qquad (2)$$

Testing $err_{neoC}(a)$ on real data produced very good results. Nevertheless, a parameter k to control the sensitivity was introduced to yield the final NeONORM error function:

$$err_{neo}(a) = \sum_{i=1}^{n} w_i \cdot \left(1 - e^{-\frac{(x_i - a)^2}{2k^2}}\right) \qquad (3)$$

For any x_i, an individual minimum is obtained if k is sufficiently small, and the greater k is, the more err_{neo} approaches the quadratic error function:

$$\lim_{k \to 0} \frac{1 - err_{neo}(a,k)}{k} = \sum_i \delta(x_i - a) \qquad (4)$$

(Dirac generalized functions)

and

$$\lim_{k \to \infty} err_{neo}(a,k) = err_{sqr}(a) \qquad (5)$$

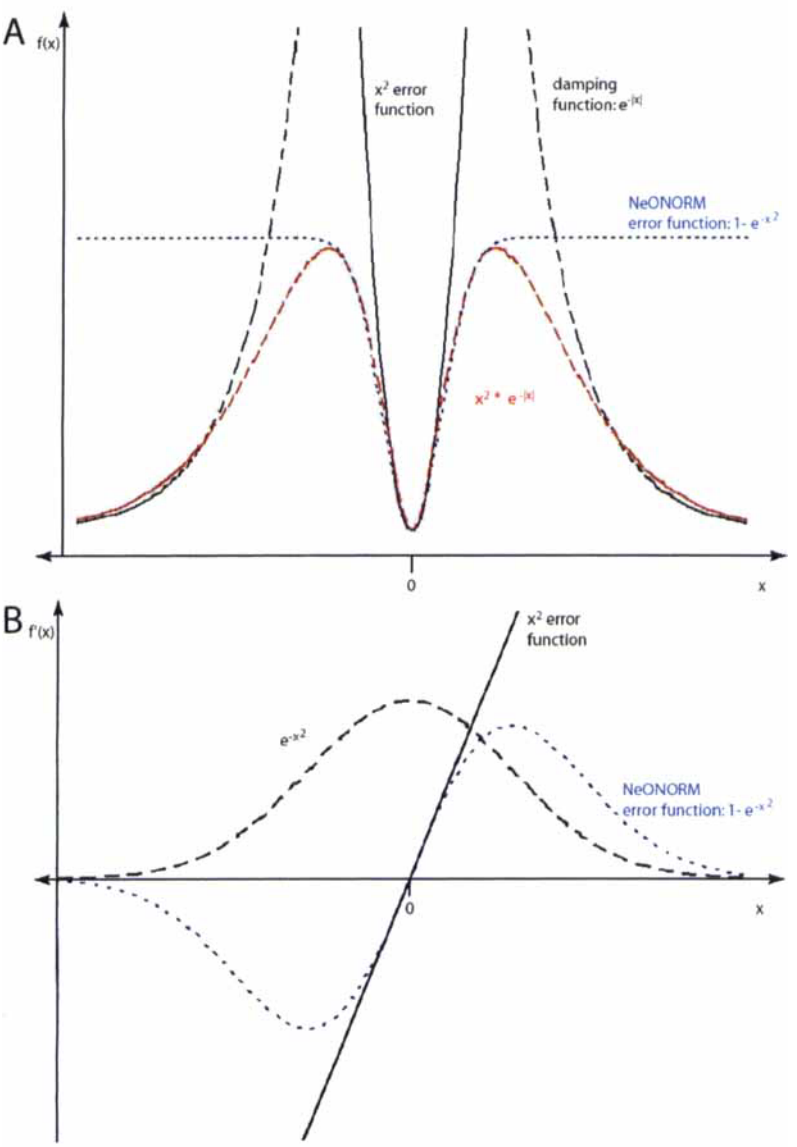

Fig. 2 The NeONORM error function. **A**. Schematic representation of the standard quadratic error function (solid black), as well as a damping function (dashed black), the product of both (dashed red), and the novel NeONORM error function (dashed blue) that we indirectly derived from the former. **B**. NeONORM damping function (dashed black) for the first derivative of the NeONORM error function, and the first derivate of the standard quadratic (solid black) and NeONORM (dashed blue) error functions.

Properties of the NeONORM error function

The NeONORM error function has very interesting properties with respect to the free parameter k. As can be seen in Equation (3), the entire error to be minimized is in fact a sum of individual contributing error *functions* $err_i(a)$, each symmetrically centered at x_i and scaled by w_i. Sums of these individual error functions have interesting properties that shall be discussed here, referring in particular to Figure 3.

The free parameter k, which we also refer to as the "sensitivity parameter", has crucial influence on the shape of the total error function. For two values x_i and x_k, differences in their summation depending on their distance are depicted in Figure 3, where $\tau = |x_i - x_k|/k$.

The point $p_m = (x_i + x_k)/2$ is crucial for distinguishing three different cases of summation by inspecting the sign of the second derivative of err_{neo} at p_m:

$$\begin{cases} \tau < 2 : err''_{neo}(p_m) > 0 : p_m \text{ is a minimum} \\ \tau = 2 : err''_{neo}(p_m) = 0 : p_m \text{ is a saddle point} \\ \tau > 2 : err''_{neo}(p_m) < 0 : p_m \text{ is a maximum} \end{cases}$$

In the second case, the saddle point is a "wide" minimum and indicates a bifurcation point. In the third case, two distinct minima exist, as can be easily seen in Figure 3. Note that the above holds true only for equal weights w_i and w_k.

The parameter k thus determines a border of discrimination between clusters of points: no two points that have a distance less than $2k$ will create distinct minima.

Unfortunately, at present, we are not able to derive a formal definition of k. The parameter therefore has to be empirically derived.

Sign plots of the first order derivatives of the NeONORM error function as visual indicators

In order to further illustrate the properties of the NeONORM error function at different values for k, we plotted the first derivative over the range of $-2 < a < 2$ and $10^{-3} < k < 10^1$. In Figure 4A, we show a sign plot (blue=negative, red=positive) and a 3D-surface plot of the function over the same parameter space. We compared two technical replicates generated from total RNA of HT29 cells [HT29(1) vs. HT29(2), see Materials and Methods]. As becomes evident from these plots, and as expected, the NeONORM error function at large k possesses a single minimum (in the sign plot: blue/red border) corresponding to the minimum of the quadratic error function (mean). With decreasing k, the sole minimum bifurcates into several

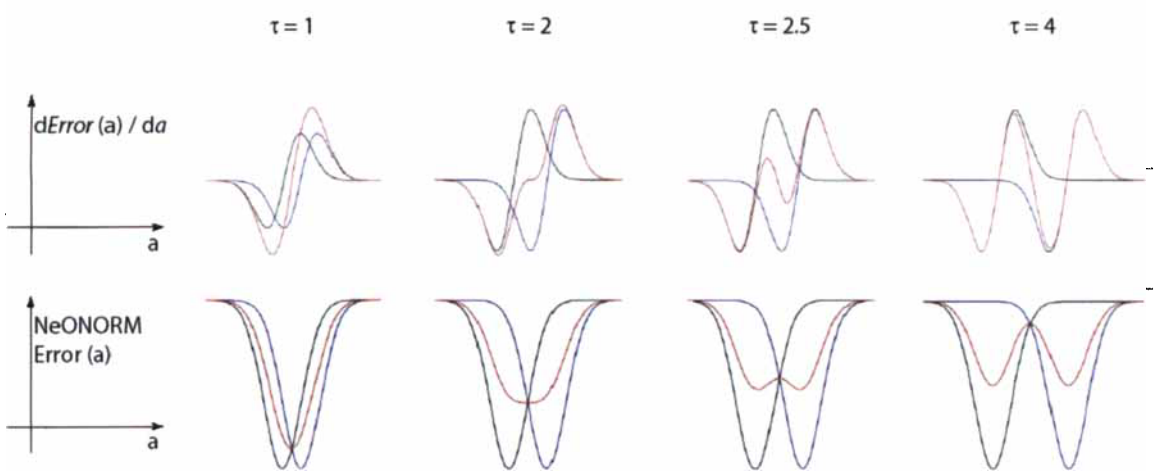

Fig. 3 Properties of the NeONORM error function. The NeONORM (composite) error function has exactly one minimum for $\tau \leq 2$ in the limit of identical weights for both individual error functions. Thereby τ denotes the absolute distance between the global minima of the individual error functions in units of k. Here k is the NeONORM sensitivity parameter identical to the absolute distance between the inclination points and the global minimum of each function. In the upper panel, we schematize the derivatives of the individual NeONORM contributing error functions (black: probe 1, blue: probe 2), and their composite (red) for different values of τ at constant k (shown are the first order derivates in the normalization factor a). The lower panel displays the original functions. Only when τ increases above 2, the NeONORM error function acquires two distinct minima.

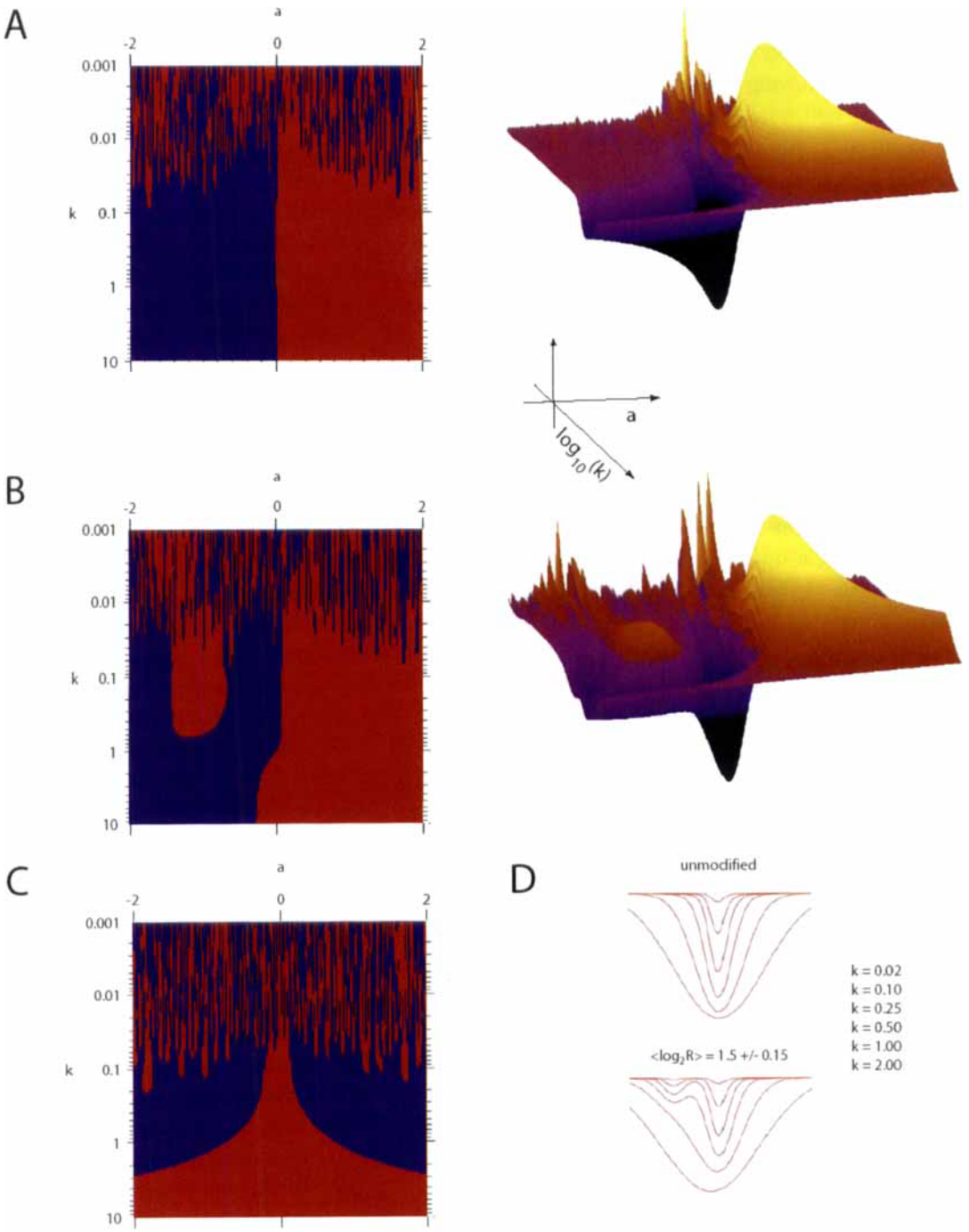

Fig. 4 The NeONORM sensitivity parameter k. **A**. Sign plot and 3D-surface plot for the first order derivative in a of the NeONORM error as a function of the sensitivity parameter k and the normalization factor a. Sign plots are used to illustrate the zero crossings of the function (blue-red and red-blue boundaries). The data represented are a subtraction from two HT29 technical replicates HT29(1) vs. HT29(2). **B**. As in panel A for a subtraction between the modified, artificial dataset HT29(1)mod ¼ 1.5±0.15 vs. HT29(2). In order to generate the modified dataset, a random chosen quarter of all probe signals of the HT29(1) dataset was individually multiplied with a different random value drawn from a normal distribution with $\mu=2^{1.5}$ and $\sigma=0.1*\mu$. **C**. As in panel A, but only a sign plot for the second order derivative in a is shown. **D**. The NeONORM error function for selected increasing values for k is shown for the HT29(1) vs. HT29(2) (upper) and the HT29(1)mod ¼ 1.5±0.15 vs. HT29(2) datasets. Note that k=0.02 is the flattest curve in both cases.

local minima. These bifurcations process onset at multiple a over a relatively short range of k. They hence reflect the growing (with decreasing k) influence of smaller and smaller clusters of co-varying probes generating local minima in the error landscape. At the limit of k approaching zero (the granularity or numerical resolution of the given data), individual probes will generate local minima in the error function, whereas in the limit of k approaching infinity, the data are "appraised" by the function in the same way as in the quadratic error function.

We then plotted the first derivative of the NeONORM error function for an inter-assay comparison where we used artificially distorted data (in order to generate asymmetry in the logQ distribution, see Materials and Methods). Briefly, the HT29(1)mod 1/4 1.5 ± 0.15 dataset was generated from HT29(1) by randomly choosing a quarter of all probe signals and multiplying each of them with a factor randomly drawn from a normal distribution with a mean of $2^{1.5}$ and a variance of $0.1*2^{1.5}$. As becomes evident from both plots in Figure 4B, the perturbation of the underlying data generates a massive distortion in the HT29(1)mod 1/4 1.5 ± 0.15 vs. HT29(2) comparison. The multiplication of a subset of probe signals with an almost constant factor essentially subdivides the probe set into two distinct clusters. This effect is noticeable at relative large k when compared to the bifurcations observed within very symmetric data (Figure 4A). The NeONORM error function hence captures strong clustering in the data. A second phenomenon can be observed in the sign plots of the artificial dataset. With decreasing k, the relative position of the global minimum of the error function shifts in a. Only at sufficiently small k, thus at sufficiently high sensitivity, the NeONORM error function returns the optimal normalization factor a_{min}. This obviously has major implications for the choice of k. At the same time, it illustrates the advantage of the NeONORM method in that averaging error functions (comparable to NeONORM in the limit of k approaching infinity) return suboptimal normalization factor a_{min}.

For further illustration purposes, we also plotted the second derivative of the NeONORM error function over the same parameter space with the unmodified HT29 data. Here, the commencement of the multiple bifurcation zone already appears at larger k (Figure 4C). In Figure 4D, we show the NeONORM error function at selected k (increasing from top to bottom) for both datasets, the unmodified (top) and modified (bottom) HT29 inter-assay comparison. In the case of the asymmetric data, the appearance of a second minimum at small k is clearly visible.

Choosing NeONORM parameter k

Since at present we are unable to derive a formal description of k, we had to empirically derive a suitable value for k. Thereby an equilibrium between two considerations had to be found. Whereas for k tending to zero the sensitivity of the NeONORM error function is maximal, the global minimum in the error function is the result of fewer and fewer probes, making the estimation of the normalization factor a less and less robust against variations and numerical inadequacy of the data. Using the artificial datasets, as well as some one hundred different experiments performed in our own laboratory, we have found that k=0.20 provides a stable estimate for a_{min}, and should be considered the lower bound for k. The derivation of a from $a_{k=0.20}$ at smaller k thereby is systematically inferior to the numerical precision imposed by the data. At the same time, k=0.20 should also be the upper bound for k since the maximal attainable sensitivity/precision of the NeONORM method is reached. Note that k=0.20 is found close to the commencement of the bifurcation zone in almost all the data we have analyzed. The algorithmic implementation of NeONORM is capable of correctly identifying the global minimum of the error function for arbitrarily small k, thus a_{min} is correctly estimated even within the bifurcation zone. Since we cannot rule out the possibility that k needs to be adjusted for different microarray platforms (for example, due to different precision of the numerical values), we have successfully applied the NeONORM method with k=0.20 to nine inter-assay comparisons of Affymetrix datasets (http://www.affymetrix.com). The evaluation of the NeONORM method presented below has been done at k=0.20.

"Asymmetric" test data

In order to evaluate the NeONORM method, several well defined test datasets had to be acquired. First, we were looking for "real world" inter-assay comparisons where the fold-change distributions were as close to symmetric as possible. To this end, we decided to use two technical replicates generated in our laboratory from the total RNA extracted from HT29 cells. These technical replicates, here denominated as HT29(1) and HT29(2), share a Pear-

son correlation of R=0.993, thus are indeed highly reproducible for the single biologic condition (Supporting online material: "HT29.txt"). In order to generate asymmetries in a controlled fashion in the fold-change distributions, the HT29(1) dataset was modified according to two distinct procedures to generate: (1) A dataset with ever increasing perturbations at ever larger distance to a=0, and with a constant and small variance over the perturbation (Figure 5; see Materials and Methods). This dataset was used to simultaneously test the performance and the robustness of the NeONORM method. Having up to a quarter of all probe signals artificially induced with a logQ average of up to $2^{1.5}$, and a constant <logR>/<variance> ratio of 0.1, this dataset resulted in highly significant perturbations easily detectable in the sign plots shown in Figure 5 (Supporting online material: "HT29(1)robustness.txt"). (2) A second dataset where exponentially increasing perturbations with a constant and large variance are introduced (Table 1; see Materials and Methods). For small ratios of modified probe signals f (n=~32,500 for our AB1700 data), the perturbations were very modest (Supporting online material: "HT29(1)sensitivity.txt"). Hence, this dataset can be used to evaluate the sensitivity of the NeONORM method at k=0.20.

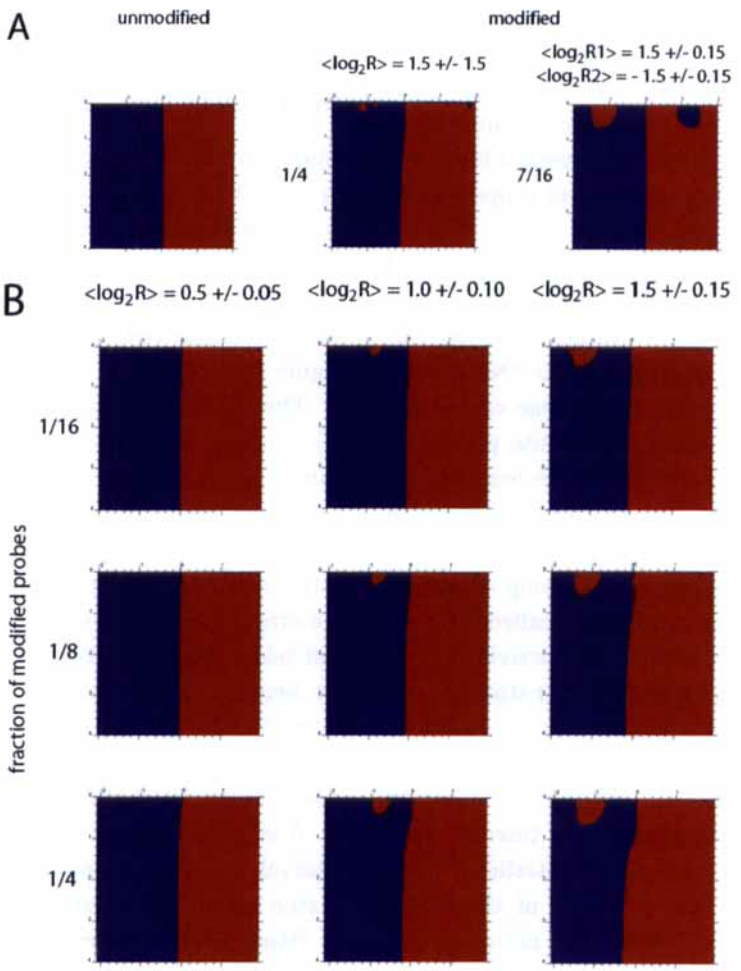

Fig. 5 Artificial "asymmetric" datasets. Sign plots of the NeONORM error function for the first series of artificial test data generated are shown (see Materials and Methods). **A**. From left to right: (1) Unmodified (original) dataset HT29(1) vs. HT29(2). (2) Asymmetrically modified dataset HT29(1) ¼ 1.5±1.5 (random chosen quarter, μ=$2^{1.5}$ and $\sigma = \mu$). (3) Double modified dataset HT29(1) ¼ 1.5±0.15 was once more modified by subsequently choosing another random quarter of probe signals and multiplying them with a random value drawn from a normal distribution with μ=$2^{-1.5}$ and σ=0.1*μ. This second operation generates almost symmetrically modified data, where the average total of modified probe signals is 7/16. **B**. As in panel A. To the right of each row the fraction of modified probe signals (1/16, 1/8, 1/4), and on the top of each column the average ratio change parameters ($\log_2$ of μ and the corresponding σ) are indicated.

Table 1 Pearson Correlations for Artificial Test Datasets

	Pearson correlations	
	vs. HT29(1)	vs. HT29(2)
HT29(1) unmodified	1.000	0.993
mod 1/256 1.5±1.5	0.984	0.977
mod 1/128 1.5±1.5	0.961	0.954
mod 1/64 1.5±1.5	0.931	0.927
mod 1/32 1.5±1.5	0.837	0.832
mod 1/16 1.5±1.5	0.794	0.790
mod 1/8 1.5±1.5	0.793	0.786
mod 1/4 1.5±1.5	0.718	0.713

In conclusion, the artificial datasets generated here to evaluate the absolute and relative performance of the NeONORM method are of high quality as the nature of the perturbations covers and exceeds the entire range of feasible naturally occurring asymmetries, and the perturbations, due to the process of how they were generated, have similar statistical properties to real experimental data.

Finally, we chose two datasets from retinoic acid (RA) treated and untreated NB4 cells obtained recently in our laboratory as an example of "real" biologic data (Supporting online material: "NB4.txt"). The effect of RA on changes in the gene expression profiles of myeloid cells and human acute promyelocytic leukemia [APL; the NB4 cell line was derived from an APL patient (*28*)] will be discussed in detail elsewhere (data not shown). Briefly, RA directly acts upon the transcriptional activity of a group of nuclear receptor transcription factors, the so-called retinoid receptors (*29*), thereby switching their activity from a repressor state to a very potent activator state of gene expression (*30*). The effects of RA on gene expression are very immediate and drastic (becoming apparent after as short as 15–30 min). Within the observed time scale of a couple of hours, RA thereby commits NB4 cells to reentry into differentiation, which is otherwise blocked by the presence of the PML-RARalpha oncogene (*31*). Given the nature of the activity switch, as well as the rapidness of the physiologic response, we rightly assumed that these datasets would display some degree of asymmetry in the fold-change distributions.

Note that all of these datasets had previously undergone identical multiple non-linear and linear normalization steps to account for systematic intra-assay variations before the artificial perturbations were introduced and before these data were used for evaluating NeONORM (see Materials and Methods).

Algorithmic implementation of the NeONORM method

Based on the above reasoning, we have developed an algorithm capable of generating and minimizing the negative second order exponential error functions used by the NeONORM method. For illustration of the algorithmic implementation, a pseudo-code description of the NeONORM algorithm is displayed in Figure 6.

The algorithm consists of the following three steps: (1) Finding intervals $I_1 \ldots I_n$ that each contains a minimum; (2) Iteratively approaching for each $j = 1 \ldots n$ the minimum contained in the interval I_j until a defined precision is reached, yielding a candidate value a_j for the minimum in I_j; (3) Calculating the error values $err_j = err_{neo}(a_j)$ for each candidate and selecting a_j that corresponds to the minimal error amongst $err_1 \ldots err_n$. The algorithm is implemented in Java in the framework of the *ace.map* suite for microarray statistical data analysis, which will be described in detail elsewhere.

A value for k needs to be defined as well as a search interval s and a maximum tolerated derivative *conv*. A step size d is calculated as a function of k.

Step 1: The interval s is sampled using $\lceil s/d \rceil$ steps. In every step i, the sign of the derivative of err_{neo} at x_i, namely s_i, is calculated. If s_i is positive and s_{i-1} is negative, a zero crossing that belongs to a minimum occurs between x_{i-1} and x_i. This interval is added to a vector of candidates. This first sampling step is only required for small k when one cannot assume that there is only a single minimum. Other methods to find multiple minima could be used here, but this method has the advantage to use a defined

```
NeONORM

parameters:     data sets X={X^m, X^v}, Y={Y^m, Y^v}, each of size n
                k
                a0, the initial guess for a
                s, the interval size the minimum is searched for inside
                d, the sample step width
                max_iter, the maximum number of iterations
                conv, the maximum allowed derivative error

let Q := {qi:= log2(Xmi / Ymi) | 0<i≤n}
let W := {wi:= 1/sqrt(Xvi^2 + Yvi^2) | 0<i≤n}

rem   calculates the cumulated neonorm error used for
rem   initial minimum detection
subroutine error(a, k) begin
    sum := 0
    for all i | 0<i≤n
        sum := sum + wi * (1-exp(-(qi-a)^2/(2*k^2)))
    endfor
    return sum
end subroutine

rem   calculates the derivative of the cumulated neonorm error
rem   used to iteratively determine a local minimum
subroutine errorDeriv(a, k) begin
    sum := 0
    for all i | 0<i≤n
        sum := sum + wi * exp(-(qi-a)^2/(2*k^2)) * (a-qi) / k^2
    endfor
    return sum
end subroutine

rem   iteratively decreases the distance to the true minimum starting
rem   with the given interval bounds a0-s, a0+s. returns when the absolute
rem   of the remaining derivative falls below conv
subroutine minimum(a0, s, k, conv) begin
    count := 0
    apos := a0-s
    aneg := a0+s
    dpos := error(apos, k)
    dneg := error(aneg, k)
    if sign(epos) = sign(eneg) then
        return with failure
    endif
    if sign(epos) ≠ 1 then
        swap(apos, aneg)
        swap(dpos, dneg)
    endif
    do
        at := (apos+aneg)/2
        dt := errorDeriv(at, k)
        if sign(dt) = 1
            apos := at
            dpos := dt
        else
            aneg := at
            dneg := dt
        endif
        count := count+1
    while count<max_iter AND |dt| > conv
    return at
end subroutine

rem   determines a set of intervals containing a minimum each using sign
rem   changes from negative to positive, iteratively finds the precise
rem   position inside of each of the interval, calculates the corresponding
rem   error values and returns the position with the smallest such error
subroutine absmin(k, s) begin
    varArray arr
    lastsign := sign( errorDeriv(a0-s, k) )
    for ac=a0-s+d to a0+s step d
        sign := sign( errorDeriv(ac, k) )
        if sign = 1 AND lastsign = -1 then
            add element (ac-d/2) to arr
        endif
        lastsign := sign
    endfor
    for all elements arri in arr
        ei := minimum(arri, s/2, k)
    endfor
    B := 1
    E := e1
    for all i | 1<i≤size_of(arr)
        if ei<E then
            E := ei
            B := i
        endif
    endfor
    return aB
end subroutine
```

Fig. 6 The pseudo-code for the implementation of the NeONORM algorithm. Additional details for the implementation can be made available upon written inquiry to the authors.

number of steps and is shown to be sufficient for the described purpose.

Step 2: For every such candidate c_i, the minimum m_i is further approached iteratively, halving the size of the interval $[x_{\text{low}}, x_{\text{high}}]$ in every iteration step. In every step, the function calculates the error derivative d_e for the point p_m on half distance between x_{low} and x_{high}. If d_e is negative, x_{low} is replaced by x_m, otherwise x_{high}. The function stops once $|d_e| < conv$.

Step 3: For every minimum m_i, the actual error is calculated. The errors for all minima are compared and the mininum that belongs to the smallest error is returned.

Functions for calculating the cumulated error and its derivative sum up the results of corresponding functions that calculate individual error and error derivative for each gene, respectively.

Assume there are n genes in a dataset. For every single error or derivative, n exponentials have hence to be calculated (plus n times some basic arithmetical operations that are not regarded here). The step size d is proportional to k. So the number of exponentials to be calculated for Step 1 is proportional to $n \cdot s/k$.

The iterative method in Step 2 will find exactly one minimum in the given interval if at least one exists. If there exist several minima, the method will find one of them ignoring the existence of others. The method is fast, stable, and rarely requires more than 30 iterations to converge under $conv{=}10^{-4}$ for all considered cases so far.

The algorithm was tested for k less than the numerical precision of data provided (10^{-4}), distinguishing more than 500 minima in the interval $[-0.1 : 0.1]$.

Comparative validation of the NeONORM method

We proceeded to evaluate the performance, robustness, and sensitivity of the NeONORM method on the different "real world" and artificial datasets by comparative validation of three methods, that is, inter-assay normalization by the Median, LOWESS (LOcally WEighted Scatterplot Smoothing), and NeONORM methods. Figure 7 summarizes the key comparison results of the Median and NeONORM methods for the artificial datasets from Figure 5. In all the three panels of Figure 7A, frequency plots of the logQ distributions between two input sample datasets are depicted. On the unmodified HT29(1) vs. HT29(2) inter-assay comparison (Figure 7A1), both methods perform highly similarly as expected. As we have shown above formally for the NeONORM error function in the limit of large k (≥ 10), the error estimate becomes virtually identical to the one obtained by Median normalization. We have furthermore contended that the NeONORM error function at k=0.20 still closely resembles the quadratic error function in case of symmetric fold-change distributions (Figure 5A). We find this assumption confirmed using the technical HT29 replicates since the normalization factors a are virtually identical for both methods (a_M=0.0000, a_N=−0.0045). Note that the small differences of both curves in the frequency plot are rather due to rounding differences in the binning procedure. When both NeONORM and Median normalization methods are applied to an assay comparison with an asymmetric heavy tail [Figure 7A2: HT29(1) mod 1/4 1.5 ± 1.5 vs. HT29(2)], a clear difference becomes apparent. Note that NeONORM correctly identifies the error minimum and normalizes the asymmetric data such that the maximum of the frequency distribution falls upon logQ=0, whereas Median normalization shifts the frequency distribution significantly to negative logQ values (a_M=−0.5472, a_N=0.0000). In case of the compensatory doubly modified dataset [HT29(1)mod 1/4 1.5 ± 0.15 and mod 1/4 -1.5 ± 0.15 vs. HT29(2); Supporting online material: "HT29(1)double.txt"], the difference between both methods is less pronounced; however, NeONORM clearly much better normalizes this dataset as well (Figure 7A3: a_M=−0.1672, a_N=−0.0038). The fact that Median normalization does not lead to identical results here is due to the successively random choosing of probe subsets for modification, which, as can be easily shown, highly unlikely leads to perfectly symmetric data. Together with the initial unmodified data shown in Figure 7A1, this experiment demonstrates that NeONORM is insensitive to the presence of asymmetric heavy tails in the fold-change distributions, whereas averaging normalization methods for obvious reasons fail to correctly normalize the datasets. In Figure 7B, we look at Median (Figure 7B1) vs. NeONORM (Figure 7B2) normalization performance on the remaining nine artificial datasets shown in Figure 5B. As before, NeONORM correctly normalizes the different datasets such that the maxima of the increasingly perturbed datasets all superimpose at logQ=0. Median normalization, in contrast, leads to maxima being gradually (as a function of the perturbation of the artificial dataset) shifted towards negative logQ values (Figure 7B1). Figure 7B3 summarizes these

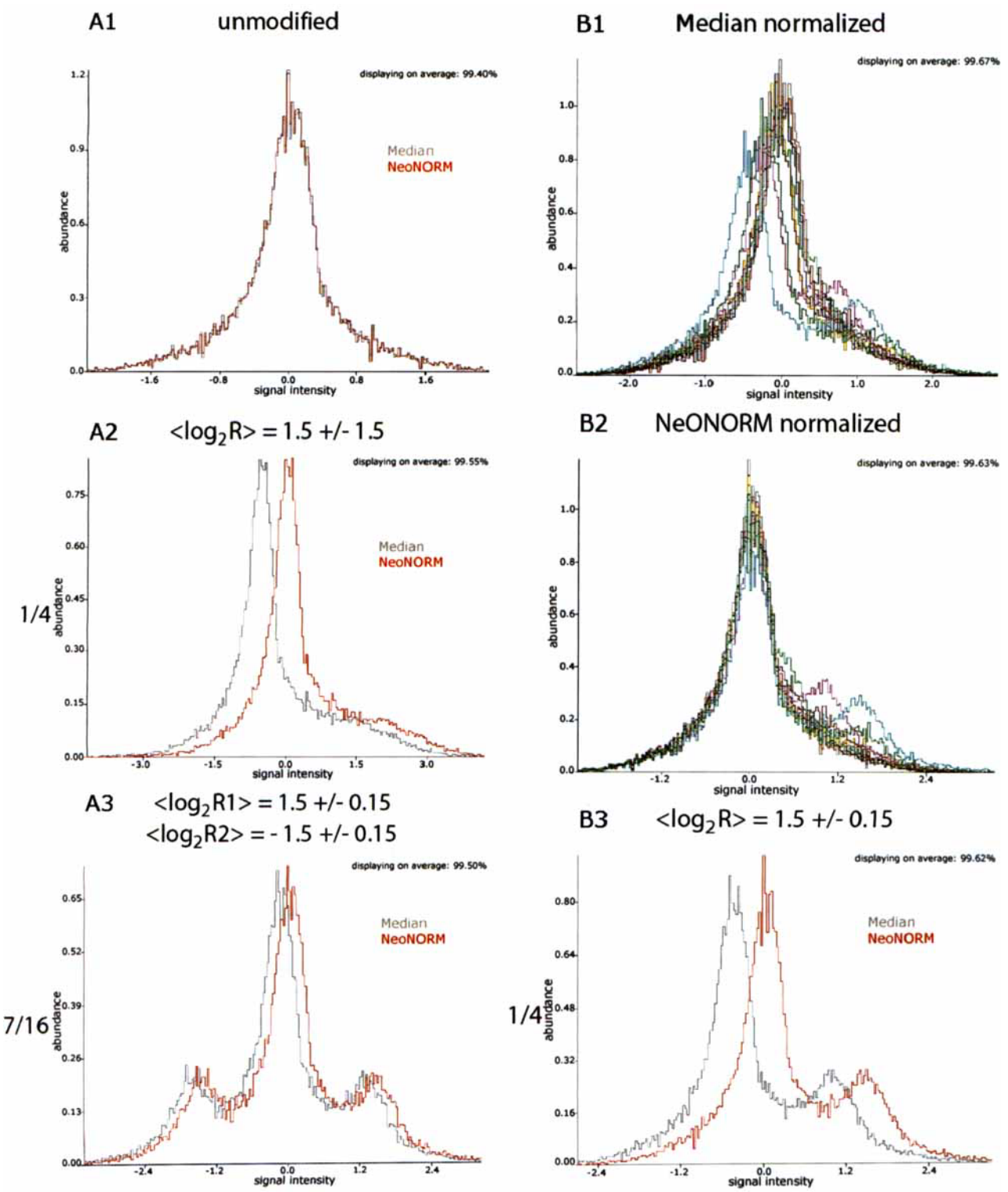

Fig. 7 Comparison of Median vs. NeONORM normalization. **A**. Direct comparison of Median vs. NeONORM on the datasets from Figure 5A. Frequency plots of logQ are shown simultaneously for both methods (Median in grey, NeONORM in red). **B**. Direct comparison of Median vs. NeONORM on the datasets from Figure 5B. Frequency plots of logQ are shown in the upper two panels simultaneously for all nine datasets (B1: Median normalized, B2: NeONORM normalized). B3: Frequency plots of logQ are shown simultaneously for both methods (Median in grey, NeONORM in red) on the HT29(1)mod ¼ 1.5±0.15 vs. HT29(2) dataset.

findings by directly comparing both methods on the most severe perturbed artificial dataset [HT29(1)mod 1/4 1.5 ± 0.15 vs. HT29(2); a_M=−0.4774, a_N= −0.0060]. In conclusion, the NeONORM method performs very well even on highly asymmetric data, and is robust with respect to the grade of the asymmetry [the range covered is from near zero (technical replicates shown in Figure 7A1) to a point that certainly exceeds any natural occurring gene regulatory event (Figure 7B3)].

In order to perform a sensitivity test of the NeONORM method, we used the second artificial test dataset. In this case, we additionally compared NeONORM performance relative to an implementation of the LOWESS algorithm (*32–34*). LOWESS is a very well performing non-linear normalization method widely used for intra- and inter-assay normalization (Ref. *12*, *32*; see Materials and Methods). The results of the direct comparisons are summarized in Figure 8A. The three panels show histogram plots for the logQ distributions of the second artificial dataset (Table 1; see Materials and Methods) as functions of the normalization methods. As was true for the previous dataset, Median normalization results in successive "migration" of the distribution maximum towards negative logQ values as a function of the severity of the perturbation (Figure 8A1). The LOWESS method seems to perform much better than Median normalization, as much as the maximum of the distribution is more stable around the logQ=0 point. However, when the perturbations of the original dataset become more and more severe (here especially evident for f =1/8 and f=1/4), the maximum of the distribution also shifts towards negative logQ values, indicating that the LOWESS method cannot correct for the asymmetries in the logQ distribution. Furthermore, due to the nature of the non-linear approach, the relative stability of LOWESS at low f values is counterbalanced by a significant distortion of the data (compare characteristic peaks at the maximum or the heavy tails between all distributions and with respect to the normalization methods). By contrast, NeONORM again performs equally well on all eight conditions, with the maxima of the distributions perfectly superimposed at logQ=0. Note that due to the linear nature of the normalization operation, no distortion of the distributions occurs. In order to demonstrate the significance of the differences in normalization factor a_i over this experimental series, we have generated a probe call table using arbitrary thresholds for logQ (as the nature and absolute value of the thresholds do not matter), and compared the differences in logQ calls (Table 2). It is apparent from this table that NeONORM, due to more accurate normalization, is the only method that shows constant logQ< −1 calls [see column with relative values, note that variation (negative values) is due to the normal distribution properties of the random drawing of factors during the perturbation of the data]. Furthermore, the logQ> 1 calls are proportional to the severity of the perturbation. Median normalized data, as expected by the nature of the operation, almost "grow" symmetrically over both thresholds, whereas LOWESS normalization generates satisfactory results for slightly biased test data and just starts to diverge significantly from NeONORM at $f > 1/32$. The distortion of the dataset also becomes apparent when comparing the total numbers of probe calls beyond the two thresholds.

In order to further quantify the relative performance of the different normalization techniques and demonstrate the robustness of NeONORM when dealing with asymmetric microarray data, we calculated the Type I and Type II errors resulting from NeONORM and Median normalization methods (the latter was found to be the least appropriate method in the above assays, see Table 2) on a selection of the synthetic datasets introduced above. Since those synthetic data were generated from the HT29(1) dataset (see Materials and Methods), we can calculate the number of probe calls that should occur in a comparison with HT29(1) using the correct normalization factor a_c=0. Any deviation from this calculated number is to be considered an error introduced through the normalization method that was applied. According to the nature of the false call (false positive or false negative), we can determine the type of error [incorrect rejection of a true null-hypothesis (I) or failure to reject a false zero hypothesis (II) of non-change]. Table 3 lists the results where the two types of errors were calculated as percent false calls when subtracting the four increasingly asymmetric synthetic datasets from the original HT29(1) dataset. We thereby investigated the 99% confidence interval for our zero-hypothesis and set a cut-off for determining the number of probe calls at |logQ| > 1.00. Note that due to the use of random number generators when creating the synthetic files, the calculated expected probe calls are associated with a variance. Thus, only errors superior to 0.038% are significant (bold-face). In agreement with the observations made before (Table 2), NeONORM consistently outperforms Median nor-

Table 2 Comparative Performance of Median, LOWESS, and NeONORM Methods on Artificial Microarray Data*

Profile [vs. HT29(2)]	Median					
	logQ< −1		|logQ| ≤ 1		logQ> 1	
	abs	rel	abs	rel	abs	rel
HT29(1) unmodified	2,139	0	28,300	0	2,193	0
mod 1/256 1.5±1.5	2,138	−1	28,213	−87	2,281	88
mod 1/128 1.5±1.5	2,171	32	28,144	−156	2,317	124
mod 1/64 1.5±1.5	2,218	79	27,990	−310	2,424	231
mod 1/32 1.5±1.5	2,424	285	27,673	−627	2,535	342
mod 1/16 1.5±1.5	2,640	501	27,013	−1,287	2,979	786
mod 1/8 1.5±1.5	3,493	1,354	25,477	−2,823	3,662	1,469
mod 1/4 1.5±1.5	5,305	3,166	22,387	−5,913	4,940	2,747
Profile [vs. HT29(2)]	LOWESS (f=0.3, iteration=2, δ=10E-5)					
	logQ< −1		|logQ| ≤ 1		logQ> 1	
	abs	rel	abs	rel	abs	rel
HT29(1) unmodified	2,043	0	28,334	0	2,255	0
mod 1/256 1.5±1.5	2,043	0	28,246	−88	2,343	88
mod 1/128 1.5±1.5	2,046	3	28,176	−158	2,410	155
mod 1/64 1.5±1.5	2,060	17	28,008	−326	2,564	309
mod 1/32 1.5±1.5	2,078	35	27,692	−642	2,862	607
mod 1/16 1.5±1.5	2,097	54	27,092	−1,242	3,443	1,188
mod 1/8 1.5±1.5	2,208	165	25,821	−2,513	4,603	2,348
mod 1/4 1.5±1.5	2,690	647	23,493	−4,841	6,449	4,194
Profile [vs. HT29(2)]	NeONORM (k=0.20, conv.=10E-4)					
	logQ< −1		|logQ| ≤ 1		logQ> 1	
	abs	rel	abs	rel	abs	rel
HT29(1) unmodified	2,144	0	28,301	0	2,187	0
mod 1/256 1.5±1.5	2,143	−1	28,214	−87	2,275	88
mod 1/128 1.5±1.5	2,144	0	28,144	−157	2,344	157
mod 1/64 1.5±1.5	2,143	−1	27,973	−328	2,516	329
mod 1/32 1.5±1.5	2,140	−4	27,678	−623	2,814	627
mod 1/16 1.5±1.5	2,129	−15	27,065	−1,236	3,438	1,251
mod 1/8 1.5±1.5	2,142	−2	25,751	−2,550	4,739	2,552
mod 1/4 1.5±1.5	2,144	0	23,267	−5,034	7,221	5,034

*Probe calls according to all three normalization methods are summarized for the second set of artificially generated data. Corresponding Pearson correlations are shown in Table 1. See Materials and Methods for a description of the modification procedure. abs = absolute number of probes, rel = relative (to the first row in each column) number of probes. Parameters for LOWESS and NeONORM methods are also indicated in the header row.

malization whether Type I or Type II error is considered. As a matter of fact, NeONORM is resistant against Type I errors even when Median normalization results in greater one percent of false positive calls, which corresponds here to 386 out of a total of 32,821 probes. When very significant asymmetries are introduced (up to a quarter of all probe signals being modified), NeONORM starts to generate some Type II errors; however, the values are still well below the Type II errors observed with Median normalization. NeONORM thus indeed significantly reduces the number of false calls due to inadequate or inefficient inter-assay normalization. Most importantly, NeONORM avoids Type I errors.

Finally, we also applied all the three normalization methods to a second "real world" dataset (NB4 RA 4h

Table 3 Type I and Type II Error Analysis for Median and NeONORM Normalization Using the Synthetic Datasets*

HT29(1) vs.: (|logQ| >1.00, p <0.01)	Absolute Error (%)			
	NeONORM		Median	
	Type I	Type II	Type I	Type II
HT29(1) unmodified	0.000	0.000	0.000	0.000
mod 1/256 1.5±1.5	0.030	0.000	0.012	0.000
mod 1/64 1.5±1.5	0.006	0.000	0.021	0.015
mod 1/16 1.5±1.5	0.000	**0.076**	**0.082**	**0.189**
mod 1/256 1.5±1.5	0.000	**0.207**	**1.176**	**1.222**

*Error is expressed as percent false probe calls at the given threshold and confidence interval for the zero-hypothesis of non-change in the subtraction profiles generated when comparing the modified datasets to the original HT29(1) dataset. Significant values are in bold type setting.

vs. NB4, see Materials and Methods, and also the paragraph on "Asymmetric test data"). As discussed above, given the nature of the physiologic response to RA, we had reasons to believe that such a dataset might be similar to some of the datasets presented in Figure 1 that show asymmetric heavy tails in the logQ distributions. Figure 8B summarizes the results we obtained by using the three normalization methods in direct comparison. In Figure 8B1, we show the first order derivative of the NeONORM error function at k=0.20 over the range from $-2 < a < 2$. The histogram plots of logQ distributions in Figure 8B2 recapitulate the results that were earlier obtained with the constructed test data. Median normalization leads to a very significantly shifted maximum (blue curve); LOWESS shows intermediate performance with respect to positioning of the maximum, but distorts the distribution (grey curve) when compared to NeONORM. Since these are non-controlled data (in the sense that we have not generated the asymmetry ourselves), we can only judge the relative performance. On the other hand, since the data are individually (intra-assay) median pre-normalized, the fact that the NeONORM normalization factor is close to zero [a_M=−0.4106, a_N=0.0000, note that a_L (given the non-linear nature) cannot be directly determined, and can only be estimated by renormalizing these data with NeONORM and determining the required normalization factor aa_{NL}, which repositions the LOWESS maximum back to superimpose exactly with the NeONORM value aa_{NL}=−0.0399] indicates that NeONORM here again performs close to optimum. As for the artificial data in Figure 8A, we also performed the probe call test (Table 4) to demonstrate the significant impact of choice of inter-assay normalization on potential results of comparative studies. Interestingly, due to the compression of the data during the LOWESS normalization, this method performs most poorly on the dataset (highest number of potentially false "down-regulated" calls, and lowest number of "up-regulated" calls) when compared to the other two methods. These results clearly can only be an indication of the relative performance of the three methods, as at present we can neither establish this dataset as representative for transcriptome studies, nor estimate the number of "down-" and "up-regulated" calls we should expect. However, given the detailed analysis we have performed above using artificially (in a well-controlled manner) modified data, NeONORM is the only method that shows robustness towards asymmetric logQ distributions. Given the fact that our two distinct strategies for generating these test datasets largely cover what could be expected to occur in real biologic samples as maximal asymmetry, and show that NeONORM scales very well with the degree of asymmetry and performs perfectly well in the limit of maximal symmetric data, we can draw the conclusion that NeONORM overcomes the problems associated with the standard symmetry hypothesis in linear inter-assay normalization.

Applicability of NeONORM

Since NeONORM is basically invariant to the level of asymmetry in the data, perfectly symmetric data are also processed correctly. Hence, NeONORM can be applied to any comparative inter-assay study without any *a priori* knowledge about, and regardless of the level of symmetry. Given their similitude in nature, NeONORM should also find its application in other

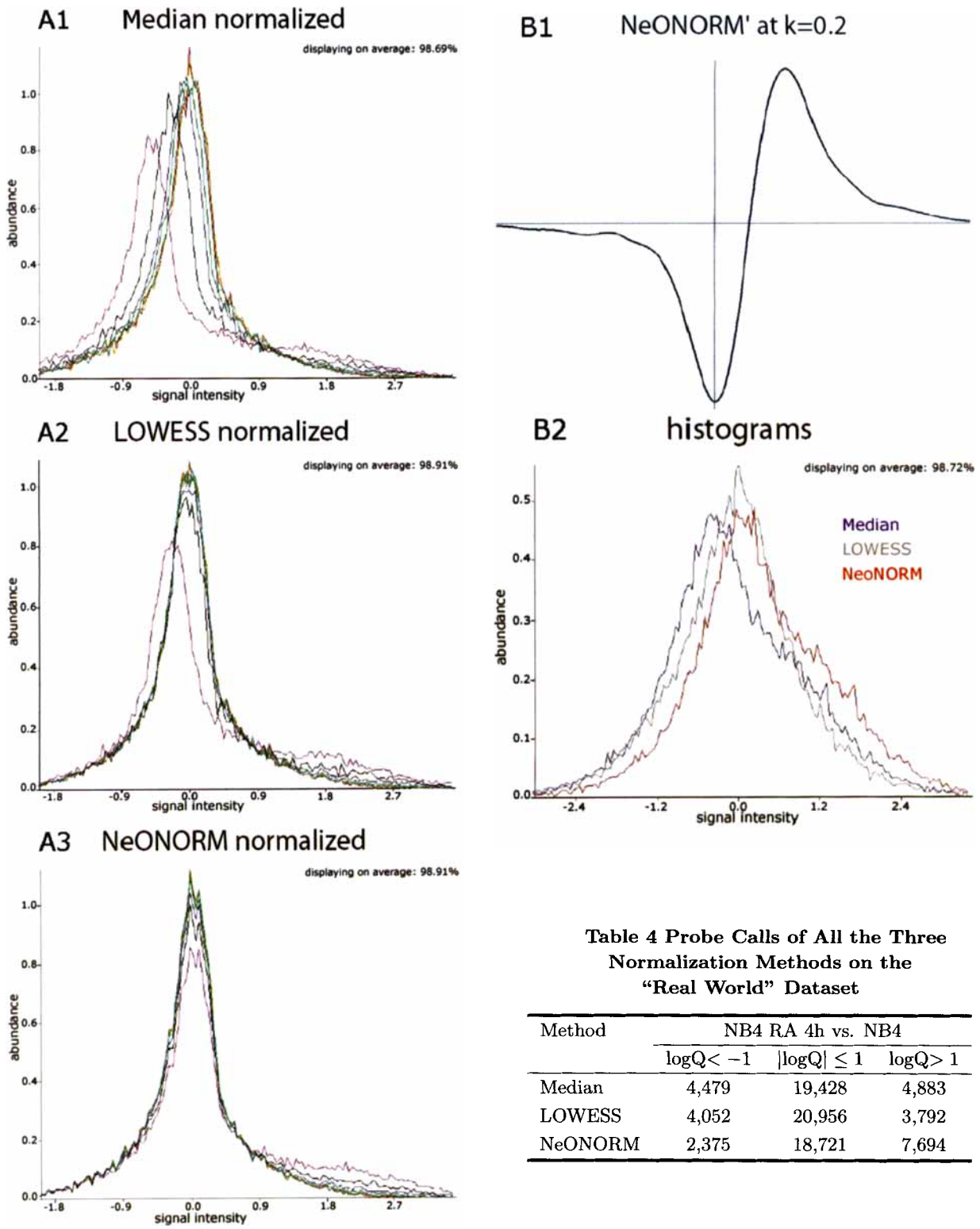

Table 4 Probe Calls of All the Three Normalization Methods on the "Real World" Dataset

Method	NB4 RA 4h vs. NB4		
	logQ< -1	$\|$logQ$\| \leq 1$	logQ> 1
Median	4,479	19,428	4,883
LOWESS	4,052	20,956	3,792
NeONORM	2,375	18,721	7,694

Fig. 8 Comparison of Median, LOWESS, and NeONORM normalization. **A**. Direct comparison of Median, LOWESS, and NeONORM normalization on the artificial datasets summarized in Table 1. Histogram plots of logQ are shown in the three panels simultaneously for all the six modified datasets and the original dataset. **B**. Direct comparison of Median, LOWESS, and NeONORM normalization on the "real world" NB4 RA 4h vs. NB4 dataset. B1: the first order derivative in a of the NeONORM error function at k=0.20, for the range $-2 < a < 2$. B2: Normalization of the NB4 RA 4h vs. NB4 according to all the three methods (Histogram plot is shown, blue = Median, grey = LOWESS, red = NeONORM).

functional genomics applications such as comparative genome hybridizations or ChIP-on-chip experiments, which seem even more prone to asymmetries given an even reduced dynamic range of the signal response. The same prerequisites as for microarray experiments will probably have to be met there, as non-linear intra-assay normalization methods should for similar reasons be applied first to the individual experiments.

Future challenges

Since the estimate of the over-all error is used during the iterative minimization procedure, and also the minimum of the error function is found at different values for a_{min} (as a non-trivial function of k), the choice of k should not be arbitrary. It seems obvious from the sign plots presented in Figures 4 and 5 that k for practical reasons should be sufficiently large to assure that the error function is not estimated within the multiple bifurcation region where individual or small group of co-varying probes attain significant influence. On the other hand, k should be chosen as small as possible in order to ascertain maximum sensitivity of the NeONORM method with respect to asymmetric logQ distributions. While other methods such as LOWESS normalization sometimes also possess even several "free" parameters (*32*, *33*), and much effort has been devoted to identify empirical optimal parameter ranges (*12*), we feel that in the case of NeONORM, a formal way of definition might be possible as parameter k directly reflects an inherent property of the analyzed data. To derive a formal way of selecting parameter k will thus represent a major challenge towards the perfection of the NeONORM method. We specifically invite suggestions or ideas towards such a formal derivation of k. At the same time, we have taken great care in the empiric derivation of a suitable k as demonstrated during the evaluation of the overall performance of the NeONORM method. The range of possible asymmetries covered through our artificial test datasets should assure that NeONORM performs in a robust fashion with respect to true physiologic datasets.

Finally, the integration of NeONORM with other methods of normalization deserves further interest. We clearly state that the NeONORM normalization can only be one of the several methods applied to current microarray data. However, the potential interplay of different non-linear intra-assay and linear inter-assay normalization methods does not seem to have been studied systematically yet. Potentially, a combination of particularly robust and situation/data adapted methods can be found and optimized to simplify, at a maximum of coherence, the microarray statistical analysis process.

Conclusion

The NeONORM method overcomes a current limitation of inter-assay normalization methods in that it is robust against asymmetries in the underlying fold change distributions. Such asymmetries reflect true changes in gene expression patterns rather than systematic experimental variation. NeONORM, a non-discriminant method, if combined with initial non-linear intra-assay normalization methods, could lead to better inter-assay normalization and thus better identification and estimation of gene regulatory phenomena in comparative transcriptome studies.

Materials and Methods

Cell culture and RNA extraction

Two different cell lines were used in the experiments. Human colon carcinoma HT29 cells (ATCC No: HTB-38) were cultivated under standard conditions in Dulbecco's Modified Eagle Medium (DMEM; Gibco, ProdNo: 41965-039) supplemented with 4.5 g/L glucose, 10% foetal bovine serum (FBS; PAA Laboratories GmbH, ProdNo: A15-649), antibiotics (a mix of 10 U/mL penicillin G and 10 mg/L streptomycine; Gibco, ProdNo: 15140-122), and 1.5 mM L-glutamine (Gibco, ProdNo: 25030-024). Cells were harvested during the exponential growth phase and subjected to RNA extraction. We have previously described cultivation and all-trans RA treatment of NB4 cells (*35*). Here, where appropriate, 10^{-6} M of all-trans RA (Sigma) was directly applied to the culture for 4 h prior to harvesting of the cells. RNA extraction was performed using the Qiagen RNeasy method according to the manufacture's recommendations (ProdNo: 75144). Quality and quantity of the isolated total RNA was determined using an Agilent 2100 Bioanalyzer as well as standard spectrophotometry.

Microarray technology

We used the novel Applied Biosystems AB1700 (ProdNo: 4338036) oligonucleotide-based microarray technology (http://www.appliedbiosystems.com).

For the present study, only Human Genome Survey Microarrays (V1.0, ProdNo: 4337467) were used, which contain probes for 29,918 validated human genes. Two assays with HT29 total RNA were performed as technical replicates [HT29(1) and HT29(2)]. One assay with NB4 cells that had been treated for 4 h with RA (NB4 RA 4h) and the other from untreated NB4 cells (NB4) have further been specifically used.

RNA labeling, hybridization, and detection

RNA labeling, hybridization, and detection were performed following the protocols supplied by Applied Biosystems together with the corresponding kits. 20 μg of total RNA sample was subjected to Chemiluminescence (CL) RT Labeling (Applied Biosystems, ProdNo: 4339628). Labeled cDNAs were then hybridized and detected (Applied Biosystems, ProdNo: 4346875).

Data preprocessing and primary analysis

Applied Biosystems Expression Array System Software (V1.1.1, ProdNo: 4364137) has been used to acquire the CL and fluorescence (FL) images and primary data analysis. Briefly, the primary analysis consists of the following individual operations: (1) Image correction; (2) Global and local background correction; (3) Feature normalization; (4) Spatial normalization; (5) Global normalization. Note that we renormalize the resulting data according to the median once more after having removed probes for which the Applied Biosystems Software has set flags greater than 2^{12}, indicating compromised or failed measurements (as recommended by Applied Biosystems). This secondary normalization is implemented in the *ace.map* suite.

Generation of artificially modified test datasets

Two artificially modified test datasets were generated by modifying the HT29(1) dataset. The modifications were carried out in order to generate asymmetries in a controlled fashion in the fold-change distributions. The general procedure for the modification can be summarized as follows:

A selected dataset X of size n is modified according to three parameters:

(1) f—the ratio of signals to be modified. Exactly $\lfloor f \cdot n \rfloor$ signals are modified;

(2) μ—the mean of the normal distribution from which the factor is drawn;

(3) σ—the variance of the normal distribution from which the factor is drawn.

The routine is implemented in Java using the supplied pseudo random number generator and the transformation function [java.util.Random.next Gaussian()], which implements the Polar Method by Box *et al* [http://java.sun.com/j2se/1.4.2/docs/api/java/util/Random.html#nextGaussian()].

The pseudo-code for the function can be given as:

```
additional variables a, b, c, r, n, frac

subroutine modify_dataset(X, f, μ, σ) begin
    let n := size_of(X)
    let frac := floor(f*n)

    rem unsort the data
    for c = 1 to 2*n begin
        let a := random_uniform()*n
        let b := random_uniform()*n
        exchange(x_a, x_b)
    endfor

    for c = 1 to frac begin
        let r := μ + σ * random_normal()
        let x_c := x_c * r
    endfor
    return X
end subroutine
```

The two generated artificial datasets differ by the ratio f, the mean μ, and the variance σ. The first dataset shown in Figure 5B was generated at constant σ=0.1*μ for all combinations of f=1/16, 1/8, 1/4 and μ=0.5, 1.0, 1.5. The second dataset discussed in Tables 1 and 2 was generated with constant $\mu=2^{1.5}$, constant $\sigma = \mu$, and varying f=1/256 to 1/4. Simply, the first dataset creates ever larger perturbations (with increasing f) at ever larger distance to a=0 (with increasing μ) with a constant and small variance of the perturbation. The second dataset creates exponentially increasing perturbations (with increasing f) at a constant distance to a=0 with a constant and large variance.

Inter-array normalization

(1) Median normalization, linear. Assuming symmetry in the fold change; after calculating the fold

changes, the median is subtracted from the $\log_2$-transformed signal quotients. (2) LOWESS normalization, non-linear. LOWESS is a method developed by Cleveland (*33*) in 1979 and since then has been frequently improved and modified. It was applied to microarray data analysis for the first time by Yang *et al* (*11*) in 2001. For the original method, four parameters have to be specified (which normally happens more or less arbitrarily), and changes in each of them lead to different results. A suggestion for optimized parameter selection was published by Berger *et al* (*12*) in 2004. LOWESS performs very well for poorly preprocessed data. The LOWESS implementation used to compare to NeONORM is a Java port of Cleveland's original FORTRAN code from 1985 freely available (http://netlib.bell-labs.com/netlib/go/lowess.f.gz), which was temporarily embedded in the *ace.map* platform for direct comparative testing.

For the Median and NeONORM normalization methods, profiles are biased by a single additive value identical for all probes. LOWESS normalization is performed on Bland-Altman-/MA-plots and hence additionally uses the average of the $\log_2$-transformed signal values (M) corresponding to one gene and, roughly speaking, generates a bias function, possibly different for every m∈M. LOWESS parameters used for all normalizations were: f=0.3; polynome order=1; iterations=2; delta=0.00001.

Data representation

Histograms and frequency plots appearing in the figures were generated using *ace.map*, however, are standard means of data representation. The 3D-surface plots in Figure 4 were rendered using *gnuplot 4.0*. Sign plots (blue = negative, red = positive) were rendered by an *ace.map* plug-in. Subtraction profiles consist of $\log_2$-transformed quotients (logQ) of signal intensities for the intersection set of the probe IDs, which were contained in the two input sample files used for the subtraction.

Acknowledgements

The authors thank Annick Lesne for very helpful comments on this work and critically reading the manuscript. All other members of the systems epigenomics group are thanked for stimulating discussions. Cécile Acquaviva is thanked for initial help with the cell culture of HT29 and NB4 cells. This work was supported by the European Hematology Association—José Carreras Foundation, the Institut des Hautes Etudes Scientifiques, the Centre National de la Recherche Scientifique (CNRS), the Institut National de la Santé Et de la Recherche Médicale (INSERM), and the French Ministry of Research through the "Complexité du Vivant—Action STICS-Santé" program (all to AB).

Authors' contributions

SN has significantly participated in the mathematical formulation of NeONORM, the generation of artificial data, the statistical data analysis, and manuscript preparation, as well as algorithmically implemented all of the three normalization methods. GB has contributed to the algorithmic implementation of NeONORM and manuscript preparation. AB has significantly participated in the mathematical formulation of NeONORM, the generation of artificial data, the statistical data analysis, and manuscript preparation. AB has designed and coordinated this study, and has acquired the experimental data. All authors read and approved the final manuscript.

Competing interests

The authors have declared that no competing interests exist.

References

1. Stoughton, R.B. 2005. Applications of DNA microarrays in biology. *Annu. Rev. Biochem.* 74: 53-82.
2. Stranger, B.E. and Dermitzakis, E.T. 2005. The genetics of regulatory variation in the human genome. *Hum. Genomics* 2: 126-131.
3. Chang, J.C., *et al.* 2005. The promise of microarrays in the management and treatment of breast cancer. *Breast Cancer Res.* 7: 100-104.
4. Raetz, E.A. and Moos, P.J. 2004. Impact of microarray technology in clinical oncology. *Cancer Invest.* 22: 312-320.
5. van Steensel, B. 2005. Mapping of genetic and epigenetic regulatory networks using microarrays. *Nat. Genet.* 37: S18-24.
6. Leung, Y.F. and Cavalieri, D. 2003. Fundamentals of cDNA microarray data analysis. *Trends Genet.* 19: 649-659.
7. Nguyen, D.V., *et al.* 2002. DNA microarray experiments: biological and technological aspects. *Biometrics* 58: 701-717.

8. Lipschutz, R.J., *et al.* 1999. High density synthetic oligonucleotides arrays. *Nat. Genet.* 21: 20-24.
9. Kerr, M.K., *et al.* 2000. Analysis of variance for gene expression microarray data. *J. Comput. Biol.* 7: 819-837.
10. Wilson, D.L., *et al.* 2003. New normalization methods for cDNA microarray data. *Bioinformatics* 19: 1325-1332.
11. Yang, Y.H., *et al.* 2001. Normalization for cDNA microarray data. In *Optical Technologies and Informatics* (eds. Bittner, M., *et al.*). International Society for Optical Engineering, San Jose, USA.
12. Berger, J.A., *et al.* 2004. Optimized LOWESS normalization parameter selection for DNA microarray data. *BMC Bioinformatics* 5: 194.
13. Workman, C., *et al.* 2002. A new non-linear normalization method for reducing variability in DNA microarray experiments. *Genome Biol.* 3: research0048.
14. Yang, Y.H., *et al.* 2002. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.* 30: e15.
15. Vaes, B.L., *et al.* 2005. Microarray analysis reveals expression regulation of Wnt antagonists in differentiating osteoblasts. *Bone* 36: 803-811.
16. Ein-Dor, L., *et al.* 2005. Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics* 21: 171-178.
17. Bell, B., *et al.* 2001. Identification of hTAF(II)80 delta links apoptotic signaling pathways to transcription factor TFIID function. *Mol. Cell* 8: 591-600.
18. Gottesfeld, J.M. and Forbes, D.J. 1997. Mitotic repression of the transcriptional machinery. *Trends Biochem. Sci.* 22: 197-202.
19. Sun, T., *et al.* 2005. Early asymmetry of gene transcription in embryonic human left and right cerebral cortex. *Science* 308: 1794-1798.
20. Zhang, M., *et al.* 2004. Foxj1 regulates asymmetric gene expression during left-right axis patterning in mice. *Biochem. Biophys. Res. Commun.* 324: 1413-1420.
21. De Smet, F., *et al.* 2002. Adaptive quality-based clustering of gene expression profiles. *Bioinformatics* 18: 735-746.
22. Martin, D.E., *et al.* 2004. Rank Difference Analysis of Microarrays (RDAM), a novel approach to statistical analysis of microarray expression profiling data. *BMC Bioinformatics* 5: 148.
23. Birnbaum, K., *et al.* 2003. A gene expression map of the *Arabidopsis* root. *Science* 302: 1956-1960.
24. Bolstad, B.M., *et al.* 2003. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 19: 185-193.
25. Pan, W. 2002. A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics* 18: 546-554.
26. Wang, H. and Huang, H. 2004. SED, a normalization free method for DNA microarray data analysis. *BMC Bioinformatics* 5: 121.
27. Khan, J., *et al.* 2001. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat. Med.* 7: 673-679.
28. Lanotte, M., *et al.* 1991. NB4, a maturation inducible cell line with t(15;17) marker isolated from a human acute promyelocytic leukemia. *Blood* 77: 1080.
29. Chambon, P. 1994. The retinoid signaling pathway: molecular and genetic analyses. *Semin. Cell Biol.* 5: 115-125.
30. Mangelsdorf, D.J., *et al.* 1995. The nuclear receptor superfamily: the second decade. *Cell* 83: 835.
31. Di Croce, L., *et al.* 2002. Methyltransferase recruitment and DNA hypermethylation of target promoters by an oncogenic transcription factor. *Science* 295: 1079-1082.
32. Dudoit, S., *et al.* 2002. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica* 12: 111-139.
33. Cleveland, W.S. 1979. Robust locally weighted regression and smoothing scatterplots. *J. Amer. Statist. Assoc.* 74: 829-836.
34. Cleveland, W.S. and Devlin, S. 1988. Locally weighted regression: an approach to regression analysis by local fitting. *J. Amer. Statist. Assoc.* 83: 596-610.
35. Heuze, M.L., *et al.* 2005. ASB2 is an Elongin BC-interacting protein that can assemble with Cullin 5 and Rbx1 to reconstitute an E3 ubiquitin ligase complex. *J. Biol. Chem.* 280: 5468-5474.

Supporting Online Material

http://www.iri.cnrs.fr/seg/NeONORMsuppData.zip