Normalization strategies for mRNA expression data in cartilage research

K. Fundel Ph.D.,† J. Haag Ph.D.,‡, P. M. Gebhard M.Sc.,‡, R. Zimmer Ph.D., Professor†
and T. Aigner M.D., D.Sc., Professor†*
† Institute for Informatics, Ludwig-Maximilians-Universität München, Germany
‡ Institute of Pathology, University of Leipzig, Germany

Summary

Objective: Normalization of mRNA data, i.e., the calculation of mRNA expression values comparable in between different experiments, is a major issue in biomedical and orthopaedic/rheumatology research, both for single-gene technologies [Northern blotting, conventional and quantitative polymerase chain reaction (qPCR)] and large-scale gene expression experiments. In this study, we tested several established normalization methods for their effects on gene expression measurements.

Method: Five standard normalization strategies were applied on a previously published data set comparing peripheral and central late stage osteoarthritic cartilage samples.

Results: The different normalization procedures had profound effects on the distribution as well as the significance values of the gene expression levels. All applied normalization procedures, except the median absolute deviation scaling, showed a bias towards up- or down-regulation of genes as visualized in volcano plots. Of interest, the P-values were much more depending on the normalization procedure than the fold changes. Ten commonly used housekeeping genes showed a significant variability in between the different specimens investigated. The gene expression analysis by cDNA arrays was confirmed for these genes by qPCR.

Conclusion: This study documents how much normalization strategies influence the outcome of gene expression profiling analysis (i.e., the detection of regulated genes). Different normalization approaches can significantly change the P-values and fold changes of a large number of genes. Thus, it is of vital importance to check every individual step of gene expression data analysis for its appropriateness. The use of global robustness and quality measures for analyzing individual outcomes can help in estimating the reliability of final microarray study results.

Key words: Gene expression, Chondrocytes, Housekeeping genes, Bioinformatics, Biostatistics.

Introduction

Normalization of mRNA data, i.e., the calculation of mRNA expression values comparable in between different experiments, is a major issue in biomedical and orthopaedic research. This is related to single-gene technologies such as Northern blotting, conventional and quantitative polymerase chain reaction (qPCR) as well as large-scale gene expression screens (i.e., cDNA and oligo arrays etc.). In all instances, the question arises how measured mRNA levels correlate from one experiment, cell or tissue type to another. Obviously, a rather accurate answer to this question is essential in order to be able to draw any reasonable conclusions on the gene regulation in the context of interest. Sources for potential errors include experimental variabilities on all levels of the processing procedures (i.e., usually starting from the RNA-isolation, cDNA-synthesis and PCR-reaction) (see Table I).

In order to tackle this problem, many different approaches have been adapted over the years depending on the experiment performed: the total RNA or ribosomal RNA approach, the housekeeping gene approach for single-gene analyses and the globalisation approach for multi-gene analyses in particular. Each of these approaches is based on a basic biological assumption about cellular gene expression. In particular, in each case some population of RNA molecules is assumed to be present at a constant level in all investigated cells. Therefore, this population can serve as a biological internal standard (an overview over the discussed normalization methods is given in Table II).

Previously, many papers appeared reporting on the comparison of different normalization methods in different experimental set-ups (i.e., cell and tissue types as well as different physiological and pathological conditions). In this study, we tested several established normalization methods on a previously published data set relevant for the orthopaedic/rheumatology field and demonstrate how variable gene expression measurements are depending on the method used. We also suggest a strategy to follow up in order to normalize large gene expression data sets.

Materials and methods

TISSUE ASSERVATION AND ARRAY PERFORMANCE

The tissue asservation and array performance as well as the basic data set were published previously. In brief, expression levels of more than...
Overview of contributions to gene expression measurement errors. Errors that depend on the sample or the experimental protocol are called array- or sample-dependent, since they affect all spots/genes of an array equally. Errors depending on sequence (either the expressed mRNA or the spotted cDNA/oligonucleotide) are called gene-dependent (RT: reverse transcription).

<table>
<thead>
<tr>
<th>Variable</th>
<th>Spotted gene sequence</th>
<th>Depends on sample</th>
<th>Experimental protocol</th>
<th>Remedy possible</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of cells in sample</td>
<td>-</td>
<td>++</td>
<td>-</td>
<td>If cells can be counted</td>
</tr>
<tr>
<td>RNA-isolation efficiency</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>Internal standards</td>
</tr>
<tr>
<td>RT/labelling efficiency</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>Partly by internal standards</td>
</tr>
<tr>
<td>Hybridization/washing efficiency</td>
<td>++</td>
<td>-</td>
<td>++</td>
<td>Yes (two-channel labelling technologies)</td>
</tr>
<tr>
<td>Exposure time</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>Trivial</td>
</tr>
</tbody>
</table>

4000 genes were evaluated by cDNA array analysis in cartilage from human femoral condyles: normal articular cartilage \( (n = 18; \text{45–88 yrs}) \) and early degenerated cartilage \( (n = 20; \text{43–91 yrs}) \) were obtained from autopsies, within 48 h of death. Osteoarthritic cartilage was obtained from total knee replacements (low grade: \( n = 21; \text{61–84 yrs}, \text{moderate/high grade: } n = 19; \text{61–84 yrs}) \). Cartilage was considered to be normal if it showed no significant macroscopic softening or surface fibrillation and a Mankin’s grade\(^5\) of less than 3. Early degenerative cartilage was defined as cartilage which showed a moderate fibrillation and softening, but no advanced erosion of the articular cartilage corresponding to a Mankin’s grade 3–6. Cases of rheumatoid arthritis were excluded from the study. Only primary degenerated and not regenerative cartilage (osteoophytic tissue) was used.

### Table II

<table>
<thead>
<tr>
<th>Method</th>
<th>Biological assumption</th>
<th>Pros</th>
<th>Cons</th>
<th>Scaling approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>Housekeeping</td>
<td>The housekeeping genes are constitutively expressed in the cells (i.e., the housekeeping genes are expressed at the same level and are only weakly regulated in the cells investigated)</td>
<td>- Fast and easy to compute</td>
<td>- Biological assumption is often not fulfilled (i.e., housekeeping genes are in many circumstances regulated)</td>
<td>Divide expression values of genes of interest by the intensity of the pre-selected housekeeping genes; overall: application of a multiplicative factor per array</td>
</tr>
<tr>
<td>Total RNA – globalization</td>
<td>The total number of mRNA molecules per cell is constant (i.e., the hybridization intensities for all samples investigated, summed over all genes, should be equal)</td>
<td>- Fast and easy to compute</td>
<td>- The biological assumption is often not fulfilled (i.e., different cells and different status of cells have different amounts of total RNA per cell ranging from 1 μg/1 mio cells up to 100 μg/1 mio cells)</td>
<td>Divide expression values of genes of interest by the sum of all signals; overall: application of a multiplicative factor per array</td>
</tr>
<tr>
<td>Centralization</td>
<td>The regulation of genes is well behaved (e.g., most genes are not significantly regulated or about equal numbers of genes are up- and down-regulated in different cells/samples investigated)</td>
<td>- Parameter estimation procedure is based on all expression values (and not on arbitrarily selected genes)</td>
<td>- Returns good results only if shape and spread of signal distribution is similar in the samples investigated</td>
<td>Find most probable consistent pairwise scaling based on central tendency of expression ratios; overall: application of a multiplicative factor per array</td>
</tr>
<tr>
<td>MAD</td>
<td>The median expression level and the spread of expression values should be the same in all samples/cells investigated</td>
<td>- Fast and easy to compute</td>
<td>- Returns good results only if shape of signal distribution is similar in the samples investigated</td>
<td>Scale spread of expression value distribution (MAD) to the same value and set median expression level to same level; overall: application of an additive and a multiplicative factor per array</td>
</tr>
<tr>
<td>Percentile normalization</td>
<td>The median expression level or any other percentile, is constant for all samples/cells investigated</td>
<td>- Fast and easy to compute</td>
<td>- Does not retain any differences in signal value distribution</td>
<td>Divide by the median value of the respective array or set median expression level (or any other percentile, e.g., 75% percentile) to a given value; overall: application of a multiplicative factor per array</td>
</tr>
</tbody>
</table>
CHONDROCYTE CULTURE AND GENE EXPRESSION PROFILING WITH REAL-TIME qPCR (TAGMAN)

Adult human articular chondrocytes were isolated and cultured (with and without stimulation with interleukin (IL)-1, insulin-like growth factor (IGF)-1, and tumor necrosis factor (TNF)-alpha) as described previously5. For the evaluation of gene expression levels of a panel of often used housekeeping genes, the Applied Biosystems “Human endogenous control plate” was used in an experimental series of three culture conditions (two independent donors). The assay was performed exactly according to the instructions of the manufacturer.

GENE EXPRESSION PROFILING WITH cDNA arrays

The data analyzed was obtained from custom designed cDNA microarrays produced and measured by GPC-Biotech AG ( Martinsried, Germany). A part of the spotted cDNA had been pre-selected for osteoarthritis (OA)-relevant genes. Each one of the analyzed microarrays contained 7467 spots representing 3648 genes.

DATA PROCESSING — NORMALIZATION

For normalization of cDNA array data five different standard approaches were used (for details see Causton et al., 20037). All microarray normalization methods rely on the assumption that a specific criterion should remain constant within an experiment; if necessary, normalization adapts data so that the criteria are met. Individual normalization methods vary by the criteria they consider and by the way data are adapted for fitting the required criteria (for an overview see e.g., Draghici 20038, Parmigiani et al., 20039, and Speed 200225). Here, we chose several normalization methods which are based on different assumptions and vary in the rigour of data modification.

Housekeeping genes normalizations assumes that a set of housekeeping genes show constant expression levels over all experimental conditions. If required, a multiplicative factor per array is applied to all expression values from the respective array so that the signal of the housekeeping gene(s) are the same for all arrays.

Total intensity normalization (globalization)2 assumes that the total amount of RNA in a cell remains constant and thus the total intensities should be equal for all arrays. Normalization is achieved by dividing expression values by the total intensity (i.e., the sum of all expression values) of the given array. Centralization1 assumes that regulation is well behaved, i.e., most genes are not significantly regulated or about equal numbers of genes are up- and down-regulated. This method estimates for all pairs of arrays the factor of proportionality. From the resulting matrix of pairwise factors, the method determines a multiplicative factor for each array that is used to adjust the array expression values to the overall distribution.

Percentile normalization assumes that a certain level of expression values should be equal for all arrays. Typically, the 50% percentile (eq, median) or 75% percentile is used, assuming that the expression level, below which 50% or 75% of the expression values are found, should be the same for all samples. A multiplicative factor is applied to the data so that the chosen percentile is at the same expression value for all arrays.

Median absolute deviation (MAD) scale normalization12 assumes that the median and the spread of the expression value distribution should be equal for all arrays. The spread of the expression value distribution is measured by the MAD, which is the median of the distances of all expression values to the median expression value. This method applies a multiplicative factor for adapting the spread of the expression value distribution and an additive factor to adapt the location (median) of the distribution.

DIFFERENTIAL GENE EXPRESSION

The fold change was determined as follows: individual fold changes for all pairs of samples derived from the two groups to be compared were calculated and the median of these individual fold changes represents the overall fold change for the given gene. The P-value was calculated by the Wilcoxon’s rank sum test. The gene P-values were converted into q-values by use of the R-library ‘qvalue’10. The q-value quantifies the false discovery rate, i.e., a q-value of 0.01 indicates that when selecting significant genes as the subset of all genes having a q-value < 0.01, 1% of the selected genes have to be expected to be false positives.

CRITERIA FOR METHOD SELECTION

For the given data, it was experimentally confirmed that the mRNA content was the same for all sample preparations, and thus expression intensities are expected to be similar for all measurements. Furthermore, the number of up- and down-regulated genes is expected to be balanced, as no specific activation events are investigated. Ideally, the known housekeeping genes should not show important variability between the samples.

RESULTS

COMPARISON OF DIFFERENT NORMALIZATION METHODS IN RELATION TO GENE EXPRESSION ANALYSIS

The different normalization procedures had profound effects on the gene expression levels in the investigated data sets [Fig. 1(A)]. In the raw data, gene expression levels of the low grade osteoarthritic cartilage samples were systematically lower than those of the moderate/high grade samples. Globalization and centralization do not alter the general pattern of expression levels. Percentile (50) normalization leads to a common median for all samples; yet, expression intensities are spread over a wider range for low grade osteoarthritic cartilage samples than for moderate/high grade osteoarthritic cartilage samples. Percentile (75) normalization results in similar expression values in the higher intensity range of all arrays; yet, the median and lower intensity values are systematically shifted. After MAD scale normalization the median is the same for all arrays, and the 25 and 75 percentiles are very similar; thus, after this normalization, no systematic shift can be observed between the two disease stages.

Similarly, the significance values for differential expression [Fig. 1(B)] in the investigated data sets depend on the normalization procedure. In the raw data, more genes appear to be up-regulated than down-regulated, which can be deduced from the volcano plot being biased to the right. The same bias can be observed after globalization, centralization and percentile (75) normalization. After percentile (50) normalization, more genes appear to be down-regulated than up-regulated. Only after MAD scale normalization, the volcano plot is approximately symmetric and thus approximately the same numbers of genes are determined to be up-regulated and down-regulated, respectively.

The choice of a normalization method directly affects the expression values as well as the calculated fold change and associated P-value of many genes of biological interest. Depending on the normalization procedure, genes may appear clearly regulated with a fold change greater than two or not regulated with a fold change close to one (e.g., matrix metalloproteinase (MMP)1 and MMP3). Other genes may appear slightly up-regulated or slightly down-regulated depending on the normalization procedure (e.g.,Col3A1 and Col6A1). Thus, the choice of a normalization procedure directly affects the biological interpretation of the data.

Of interest, the P-values (i.e., the significance level of changes detected) were much more depending on the normalization procedure than the fold changes (Fig. 3). Thus, genes may have a very significant P-value after one normalization but not after another (cf. Col6A1 in Fig. 2, the P-values of this gene vary between 10−5 and 1). Consequently, when selecting a set of differentially expressed genes based on their P-values, the application of one normalization method can lead to a set of genes that differ significantly from the set that would have been selected after application of a different normalization method. Fold changes obtained after application of different normalization methods show better correlation, i.e., genes that are determined to be up-regulated to a certain extent after one normalization, are generally also up-regulated at a similar level after a different normalization.

ANALYSIS OF HOUSEKEEPING GENE EXPRESSION VARIATIONS IN NORMALIZED GENE EXPRESSION DATA SETS

Ten commonly used housekeeping genes were analyzed (the list of investigated genes is given in legend of Fig. 4) in
Fig. 1. Effects of normalization. (A) Group-level plots for peripheral (samples 1–21) and central (samples 22–40) late stage osteoarthritic samples, this plot shows the 25%, 50% and 75% percentile for each sample as scatter and the median over these values for each sample group as solid line. (B) Volcano plots for group comparison P vs c for raw data and four different normalizations (for details see Data processing — normalization).
order to check them for constant expression levels across specimens investigated, because this would be the pre-condition to use them for standardizing data. All of these genes showed a significant variability in between the different specimens investigated [Fig. 4(A)]. The gene expression analysis by cDNA arrays was confirmed for these genes by quantitative PCR using stimulated vs non-stimulated adult human chondrocytes. Also, this showed considerable variation of expression levels depending on the stimulation used [Fig. 4(B)].

Discussion

This study mostly provides two important insights for biomedical researchers including the field of orthopaedics and rheumatology: it impressively documents how much normalization strategies influence the outcome of gene expression profiling analyses (i.e., the detection of regulated genes)\(^1\). Differing normalization approaches can change the calculation of \(P\)-values and fold changes of a large number of genes depending on the normalization method applied\(^1\). This results in an artificially high or low number of differentially regulated genes. Thus, more genes may seem to be regulated in one direction than in the other even though this might not be the case after adequate normalization. Also, the so-called “housekeeping” genes taken for normalization of single-gene and oligo-gene analyses are hardly to be considered to be constant even in rather homogenous sample groups (e.g., normal cartilage samples). Normalization is an important prerequisite for any quantitative data analysis in multi-gene and single-gene analyses. All available approaches in this respect have pros and cons and none of them appeared to perform optimally in all circumstances, e.g., none of them yielded similar expression levels for the sample groups, symmetric volcano plots, and constant housekeeping genes. This is in principle in line with previous similar studies based on array data derived from neuronal samples\(^1\) and others. Of note, in terms of biological understanding none of the approaches is fully covered: the assumption of the globalization method that the amount of mRNA per cell is constant is theoretically questionable for several reasons. Often the sum of all expression signals is dominated by the strongest signals (even after log-transformation)\(^11,15\). But strongly expressed genes are most likely to be regulated as they represent the major expression products of specialized cells (e.g., immunoglobulin chains for plasma cells, haemoglobin for erythroblasts, etc.). The total RNA approach rests on the assumption that, at every time point, each cell carries the same amount of total RNA, but it is well known that different cell types and cells in different conditions produce different amounts of total (and ribosomal) RNA, ranging from less than 2 mg to more than 100 mg total RNA per \(10^9\) cells. The housekeeping gene approach assumes that the expression of housekeeping genes, e.g., glyceraldehyde-3-phosphate dehydrogenase (GAPDH) or beta-actin (ACTB), is not significantly regulated. However, it becomes more and more clear that this assumption is wrong\(^15,18\), although regulation of these genes appears to be low compared to other genes. In fact, for defined cell types analyzed in rather comparable cell states it might still be a suitable method (in particular for techniques which do not allow to determine gene expression levels for a high number of different genes in parallel, such as qPCR, Northern blotting and RNase-protection assays).

One problem for all normalization methods remains the possibility of cell activation, implying the amplification of most cellular gene products (including housekeeping genes and ribosomal RNA). A proportional increase of the expression of all genes cannot be distinguished from an upscaling of all intensities due to any multiplicative error that takes effect on all spots of an array. However, though such activation is likely to occur to a certain degree, this appears to be
limited within the body. It is known that cells requiring high expression capacity in the body form polykaryons (e.g., syncytiotrophoblast, osteoclasts, giant cells, and striated muscle cells) or gain polyploidy (hepatocytes and karyocytes). In fact, there seems to be a biological limitation of single karyons with diploid genomes to support the expression machinery, limiting the systematic error introduced by activation of cells into any normalization procedure. On the other hand, if a cell doubles the expression of all its genes, it should still behave similar to two cells at the original expression level, rendering complete activation irrelevant for most biological questions (e.g., the response to drug treatment). Similarly, if all genes but one are up-regulated, it may be equally or even more useful to say that the one gene is down-regulated.

Overall, we believe that this study illustrates that it is of vital importance to check every individual step of gene expression data analysis for its suitability. Certainly, gene expression data need to follow statistical requirements. However, additionally it has to take experimental and biological background knowledge into account. For most single processing steps several alternatives exist. Therefore, it is...
important to check various possibilities and to look at their effects on the data. The use of global robustness and quality measures for analyzing individual outcomes is helpful for estimating the reliability of gene expression profiling results. Also, volcano plots can be used to estimate the general distribution of differences in expression levels. Generally, it appears desirable to apply a normalization method that modifies the original data as slightly as possible as to retain the original signal; yet, the normalization method should cover for any kind of inherent characteristics caused by undesired effects such as systematic shifts in the data. Thus, the normalization method needs to be selected in accordance with the characteristics of the investigated data set. For the analyzed data, we suggest to use the MAD algorithm based on our analysis. Globalization, the most commonly used technology, did certainly not fulfil the required criteria. This might be different of other data sets, though MAD normalization was shown for independent data sets to form optimally (own unpublished data). For housekeeping genes, all of these genes behaved largely in a relatively behaved manner: thus, in general no severe up- or down-regulation was observed, though none of these genes was constant in the data set investigated [as well as others (own unpublished results)]. Still, housekeeping genes might well be used for rough calculation of well related biological samples, which is usually the need. Specialized programs can help to select the most appropriate ones given a certain data set. Last not least, for gene expression analysis the same cautiousness is warranted as for all other technologies: what is artificial and what is relevant for understanding — this needs to be asked and answered with approaches beyond the one used in the specific study.

Fig. 3. (continued)
Fig. 4. (A) Effect of normalization on expression profiles of various housekeeping genes. Samples classified in four disease stages (normal cartilage 1–18, early osteoarthritic cartilage 19–38, peripheral late stage 39–59 and central late stage 60–78) are listed on the X-axes. Red dots indicate the expression value (on the Y-axes) for a given sample (median expression value over spots representing one gene); the blue line indicates the median expression level over all samples belonging to one disease class. The individual diagrams show distinct housekeeping genes (columns) and normalizations (rows). The plot shows that most so-called housekeeping genes show significant variations in their expression levels between different disease stages, and normalization has an important effect on the expression profiles of these genes (*the values completely outside of the usual range presumably are due to measurement errors). (B) Demonstration of the variation in expression levels of a panel of housekeeping genes (Applied Biosystems ‘Human endogenous control plate’) in three stimulation experiments as measured with the high-precision Taqman device. Y-axis: difference in number of PCR rounds, corresponding to log2 ratio of expression. RPS18: ribosomal protein S18; RPLP0: ribosomal protein, large, P0; ACTB; PPIA: peptidylprolyl isomerase A (cyclophilin A); GAPDH; PGK1: phosphoglycerate kinase 1; B2M: beta-2-microglobulin; GUSB: glucuronidase, beta; TFRC: transferrin receptor.
Conflict of interest

The authors declare no conflict of interest.

Acknowledgement

The authors wish to thank Sanofi-Aventis Pharma, Frankfurt, especially Eckart Bartnik and Joachim Saas for expert discussions, and the German ministry for research and education, which supported this work by funds to RZ and TA (projects BOA and BEX, grant 01GG9824) as well as to the DFG (grant 20/7-1) to TA.

References