ICN: a normalization method for gene expression data considering the over-expression of informative genes†

Lixin Cheng,‡*a Xuan Wang,‡b Pak-Kan Wong,a Kwan-Yeung Lee,a Le Li,a Bin Xu,c Dong Wang*gc and Kwong-Sak Leung*†

The global increase of gene expression has been frequently established in cancer microarray studies. However, many genes may not deliver informative signals for a given experiment, due to insufficient expression or even non-expression, despite the DNA microarrays massively measuring genes in parallel. Hence the informative gene set, rather than the whole genome, should be more reasonable to represent the genome expression level. We observed that the trend of over-expression for informative genes is more obvious in human cancers, which is to some extent masked using the whole genome without any filtering. Accordingly we proposed a novel normalization method, Informative CrossNorm (ICN), which performs the cross normalization (CrossNorm) on the expression matrix merely containing the informative genes. ICN outperforms other methods with a consistently high precision, F-score, and Matthews correlation coefficient as well as an acceptable recall based on three available spiked-in datasets with ground truth. In addition, nine potential therapeutic target genes for esophageal squamous cell carcinoma (ESCC) were identified using ICN integrated with a protein–protein interaction network, which biologically demonstrates that ICN shows superior performance. Consequently, it is expected that ICN could be applied routinely in cancer microarray studies.

1 Introduction

The global increase in gene expression has been established in numerous cancer microarray studies.1–6 Cells can produce more RNA when they have a high level of c-Myc which occurs frequently in human cancers.7 Also, Wang et al.1 found that genes tend to be over-expressed in various cancers from a large number of microarray datasets. However, microarrays simultaneously measure tens of thousands of genes, of which only a small subset is informative for biological interpretation.8–10 Others may not be informative signals because of insufficient expression or even non-expression. Furthermore, due to the intrinsic random noise in microarray experiments, a large number of genes are non-informative which substantially affects the conventional normalization as well as methods based on subset reference, such as data-driven invariant-set normalization.11–17 Moreover, selecting informative genes prior to subsequent differential analysis is reportedly effective in avoiding false positives.8,18 Consequently, the informative gene set, rather than the whole genome, is expected to be more reasonable for representing the overall levels of gene expression.

In this study, we have found the same trend for the informative genes in cancer microarray datasets as the aforementioned trend in the whole genome.1,4 When normalization procedures were performed, the basic assumptions were (1) only a small proportion of genes vary between samples (e.g. differentially expressed genes (DEGs)), and (2) the numbers of up- and down-regulated genes are similar.10–14 Conventional normalization methods are limited by the alleged basic assumptions, which are not universally valid and may risk over-normalization and thus a high false discovery rate.16,17 A common example of this is that the assumptions are invalid for data with an obvious upward shift, as widely exists in cancer datasets.1,4 It is oversimplified to assume that the regulation directions are symmetrical for all genes. For instance, if more than half of genes vary or are expressed higher in cancer samples, it is no longer credible to assume that the expression distribution for each sample is similar across conditions. Our findings, however, highlight the power of the informative gene set in representing the genome expression levels, which are biased when using the whole genome without any gene filtering.
CrossNorm, a normalization method proposed for microarray data with skewed expression patterns, outperforms the conventional methods when the DEG proportion is high (e.g. more than 30%). Nevertheless, although CrossNorm can detect more true up-regulated DEGs, it also blindly identifies a number of false discoveries. The method may mix up signals with noise and is not robust enough for data with a large fraction of unexpressed genes. Fortunately, I/NI-calls, a highly effective tool for gene filtering, can exclude uninformative genes by modeling gene expression values at the probe level. Because multiple probes of a probe set are designed to measure the same target mRNA, these probes should be expression correlated if the target mRNA concentration varies across the samples. A probe set or a gene is informative when most of its probes have similar expression patterns across samples, otherwise it is non-informative. A novel preprocessing method called Informative CrossNorm (ICN, a modified version of CrossNorm) is proposed here to overcome the difficulties in understanding gene expression data. In ICN, the informative genes are first selected from the raw expression data using I/NI-calls, and CrossNorm is then performed. In theory, CrossNorm is fully adapted to the informative gene expression data, among which genes show a global shift between conditions, and few noisy signals affect the performance of normalization. Based on several real-life and spiked-in datasets, ICN is shown to be highly effective and has consistently high precision, \( F_1 \) score and Matthews correlation coefficient (MCC).

In this paper, the datasets used, the methods employed and detailed studies of ICN are reported. We first analyzed 25 Affymetrix microarray datasets with paired normal and cancer samples covering 11 tissue types to comprehensively characterize the landscape of mRNA expression in cancer. Surprisingly, up to four fifths of these cancer datasets were found to be globally up-regulated for the informative gene sets. We then performed differential analysis on three spiked-in datasets with ground truth, demonstrating that ICN is consistently superior to the other five compared normalization methods, some of which are strongly based on standard assumptions (RMA, MAS5.0 and LOESS) and some are less dependent (LVS and CrossNorm).

Finally, nine potential therapeutic target genes for esophageal squamous cell carcinoma (ESCC) were identified using ICN integrated with a protein–protein interaction network. Overall, we found that using CrossNorm based on an informative gene set can help improve the precision of DEG identification in cancer samples with global expression increase. To our knowledge, this is the first study using only an informative gene set for the large-scale analysis of cancer genomic data. Because the global increase of gene expression is frequent in cancers and considering the high performance of ICN, it is expected that ICN could be applied routinely in cancer microarray studies.

2 Results

2.1 Properties of informative genes in cancers

Previous studies have indicated that genes tend to be extensively up-regulated in cancer tissues compared to the normal counterpart in cancer microarray datasets. These findings reveal that the conventional normalization methods are not likely to produce trustworthy results, because they are based on standard assumptions, but the gene expression distributions are different in normal and cancer samples. However, a considerable proportion of genes embedded in the microarray are not informative in interpreting the experiment, and revisiting this trend solely restricting attention to the informative genes is desirable. In this study, we first compared the expression distributions between the normal and cancer samples for the entire gene set and the informative gene set, respectively, among the 25 microarray datasets. Genes in the entire gene set and the informative gene set were both detected to be expressed higher in the cancer samples than in the normal samples in most datasets (Fig. 2).

Moreover, the change is even more obvious when solely analyzing the informative gene set. In other words, an increase in the expression level from the normal to the cancer samples is even enhanced in the informative gene set when compared with the entire gene set. In particular, in the ESCC106 dataset (Fig. 2A), genes in the raw data are expressed significantly higher in the cancer group than in the normal group \((P < 0.0004)\), whereas the statistical significance is improved \((P < 0.0003)\) when focusing only on informative genes. This phenomenon, an improvement in the statistical significance, is also observed in other 10 datasets (Fig. 2F). In addition, for ESCC34 (Fig. 2B), no improvement in statistical significance is observed after gene filtering, but the trend of over-expression remains obvious for the informative gene set with the \(P\)-value equal to 0.0439. Nine datasets exhibited a similar phenomenon, three of which have statistical significance and the remaining six, such as OTSCC24, are not significant in the expression level increase (Fig. 2D).

![ICN flow-chart](image-url)
Moreover, for CRC64 (Fig. 2C), the average expression values of the cancer and normal states for the entire gene set is 4.90 and 5.06, respectively, which indicates that the genes are unlikely to express higher in cancer tissues. By contrast, the overall expression level of cancer is higher than the normal state (7.55 vs. 7.44) when only using the refined informative genes. The detailed data for all datasets are presented in Table S1 (ESI†). Consequently, up to four-fifths of the cancer datasets were observed to be globally up-regulated for the informative gene set (Fig. 2F), which indicates that an unbalanced expression pattern for the informative genes extensively exists in the cancer datasets, and further confirms that genes tend to express higher in cancer tissues than in normal tissues.

Furthermore, we observed that the fold-change (FC) distributions of the informative gene set and the entire gene set were distinct for the raw data. The gene-wise density plots of FC for the 25 datasets frequently skew to the right side slightly for the genome, whereas the plots skew greatly to the right when they have a single peak or two peaks with the right peak far from zero for the informative genes (Fig. S2, ESI†). Specifically, the FC distribution of the informative genes for Pancreatic32 apparently skews to the right side with the mean value approximately equal to 0.8, whereas the distribution has two peaks for Pancreatic78 with the right peak (mean value is approximately 1) far from the left one with the mean value around zero. Both these findings reveal that the informative genes are more likely to be up-regulated in cancer tissues and a large fraction of them are differentially expressed.

2.2 Method of evaluation of the golden spiked-in dataset

To comprehensively illustrate the large effect of Informative CrossNorm (ICN) in differential analysis, we compared ICN with five other well-known preprocessing methods, Quantile,11 LOESS,13 MAS5.0,14 CrossNorm,4 and LVS,16 respectively. Quantile, LOESS, and MAS5.0 are widely applied conventional methods on the basis of standard assumptions, while CrossNorm and LVS are less so and are usually used for data with unbalanced expression patterns.

2.2.1 Properties of the informative gene set in the spiked-in dataset

Since true positives and true negatives are known for the spiked-in datasets, they are usually used to assess the abilities of preprocessing methods for quantifying the FC in an expression, using the correlation between the nominal and the observed FCs. Another advantage of the spiked-in dataset is that, unlike in silico simulation data, all measurements are carried out through the same procedures as the ordinary experiment; hence, the dataset experiences the same technical variation.20 For the golden spiked-in dataset, a total of 14 040 probe sets are detected in the DrosGenome1 platform. This dataset consists of three types of probe sets, spiked probe sets with a series
of concentration-folds greater than 1 (1309), spiked probe sets with unchanged concentration (2551) and empty probe sets without spiked RNAs (10 144). When filtering features using I/NI-calls, 98% of the empty probe sets are identified as noisy genes and are, thus, eliminated. In total, 75% of the spiked probe sets are selected as informative genes for further analyses, 70% for changed, and 77% for unchanged probe sets (Fig. S3, ESI†). The proportion of the identified DEGs increases dramatically from 9% to 42% before and after gene filtering. In other words, the DEGs account for a considerable percentage for the informative gene set and again the standard assumptions do not suit these informative genes for normalization.

2.2.2 Performance evaluation of the golden spiked-in dataset. To comprehensively compare the performance of ICN with the other normalization methods, we evaluated these methods on the golden spiked-in datasets using FC and the \( P \) value of the \( t \)-test separately as the DEG identification procedure. For the FC criteria, it is clear that ICN performs better than all others in the MA-plot and shape analysis figure (Fig. 3A and B). The true DEGs (bright points), the probe sets with assigned high concentration-folds, are far from the baseline, whereas the non-DEGs (black points) are close to the baseline for ICN. For the other methods, however, the difference between the nominal DEGs and non-DEGs is ambiguous in the MA-plot in terms of the expression magnitude. For instance, although the true DEGs are detected with high and positive FCs, meanwhile, a high proportion of non-DEGs are blindly selected as down-regulated DEGs with negative FCs, for LOESS and Quantile. Numerous empty probe sets

Fig. 3 Performance evaluation of the six normalization methods for identifying DEGs using the fold-change and statistical \( P \)-value, respectively. (A) The MA-plot for the six methods and (C) the PA-plot. (B and D) Bar plots of the detected DEG proportion for each concentration-fold corresponding to (A) and (C), respectively. The color depth represents the assigned concentration-fold. Darker colors indicate higher concentration-folds. A grey point/bar represents an empty probe set, and a black one represents a probe set with unchanged concentration.
were incorrectly identified as DEGs for MAS5.0, CrossNorm, and LVS. In Fig. 3B, accordingly, the true positives (bright bars) of ICN are comparable with others whereas the false discoveries (the gray and black bars) are extremely low. LOESS and Quantile select several unchanged probe sets as DEGs, while LVS, MAS5.0 and CrossNorm are very sensitive to noise.

By contrast, ICN outperforms the comparisons when assessed by the criterion of the \( P \)-value in the \( t \)-test. As described in Fig. 3C, the probe sets with high concentration-folds are usually distinct from the unchanged and empty probe sets for data preprocessed by ICN. For CrossNorm and the conventional methods, identifying the differences among true DEGs, empty probe sets, and unchanged probe sets is difficult. LVS succeeded ICN and was better than the other four methods, because very few unchanged and empty probe sets were identified as DEGs. Furthermore, it is obvious in Fig. 3D that ICN can achieve an extremely low false discovery rate (FDR) with a comparable statistical power of DEG detection, whereas the FDRs are rather high for the others although a few can identify more true DEGs. In this situation, LOESS, Quantile, and LVS tend to select nominal non-DEGs as DEGs (black bars) while MAS5.0 and CrossNorm are not powerful in identifying adequate true DEGs.

For comprehensive and accurate ICN measurements, four general measures, evaluations, precision, recall, \( F_1 \) score, and MCC, were used for further evaluations. It is apparent in Fig. 4 that ICN has the highest precision, \( F_1 \)-score, and MCC for all DEG selection criteria, FC, \( P \)-value, and \( Q \)-value. The precision maintains a value of 0.8 for ICN, and the \( F_1 \)-score and MCC are approximately 0.6. For measuring the recall, ICN can give about 0.5 for all these criteria, which is not as high as the other methods (nearly 0.6). In addition, ICN consistently gives the lowest FPR across all methods. Table 2 describes all these measures, and the top-two highest scores are marked in bold for each of them. All these results indicate that ICN can effectively control the FPR and it is superior to all the well-known normalization methods when all measures are comprehensively considered. Moreover, our findings confirm the previous result that the conventional normalization methods may dramatically reduce the variance of true expression signals.\(^{11-13}\)

### 2.3 Performance evaluation of the affycomp spiked-in datasets

To evaluate the performance of ICN in a dataset with a low proportion of DEGs, which is a more common case for non-cancerous studies, we analyzed two other spiked-in datasets, HGU95A and HGU133A, from the ‘affycomp’ project.\(^{21}\)

For the HGU95A dataset, we used two sample groups with seven duplicates that were spiked-in with assigned concentrations. We first tested the differential expression between these two sample groups by using the FC and the \( t \)-test on the informative gene set (ICN). The assigned DEGs are apparently separated from the genes with unchanged concentration (Fig. S4a and b, ESI†), although a few are not identified as informative genes. For other methods without gene filtering, however, the detected DEG set contains several false positives (black points in Fig. S4, ESI†), and thus results in a high FDR. A similar result was observed in the HGU133A dataset (Fig. S4c and d, ESI†), in which we employed two groups of arrays with triplicate samples spiked-in with known concentrations, although the two categories of genes are not separated adequately clearly, like in the HGU95A dataset.

Fig. S5a (ESI†) illustrates that for the HGU95A dataset, ICN maintains the highest scores in all measures except recall for the DEG identification criteria of the FC, \( P \)-value, and \( Q \)-value. Furthermore, a similar result can be observed for HGU133A, in which ICN gains the highest precision, \( F_1 \)-score, and MCC when the FC and \( P \)-value are used for selecting DEGs (Fig. S5b, ESI†). The precision is approximately 1 for the three conventional methods, but less than 10% nominal DEGs are identified when setting the threshold of the \( Q \)-value equal to 0.1. A comprehensive consideration of all measures indicates that ICN is superior to the comparisons for the two spiked-in datasets regardless of the DEG detection criteria. ICN essentially enhances the precision in selecting DEGs with an acceptable recall.

#### 2.4 Application of expression and interaction analysis to the ESCC data

Finally, we integrated the ICN-detected DEGs with the protein–protein interaction (PPI) network to obtain novel insights into...
In the beginning, 2361 up-regulated and 918 down-regulated DEGs were identified from the ESCC106 dataset using two sample t-tests with the threshold of the FDR adjusted P-value of 0.01. In a recent study, Lin et al. discovered 31 genes as druggable candidates that can be potentially targeted for therapy in treating ESCC. Among those therapeutic targets, 24 were involved in the integrated study of both gene expression and interaction. 10 of the 24 candidate targets were selected as DEGs; this cannot occur accidently in the context that overall 13436 genes were tested in parallel (hypergeometric test, P = 0.0096). Furthermore, similar results were obtained for the ESCC34 dataset (Fig. 5A), which indicates that the genes acting distinct in expression between the cancer and normal states are highly relevant to the therapeutic target genes.

We then generated the PPI subnetworks for proteins encoded by DEGs in the ESCC106 and ESCC34 datasets, and executed a targeted study of the subnetwork concentrating on the proteins that interact with the known therapeutic target proteins to determine the potential regulators of pathogenicity for ESCC. As shown in Fig. 5B and C, the final subnetwork only contains the DEG targets and their interactors that are highly connected with them (e.g., gene products interacted with at least 3 DEG targets). Overall, nine genes were predicted as therapeutic candidates, which are LYN, MAPK1, SHC1, STAT1, STAT3, PDGFRB, PTPN11, SOCS1, and PLCG1. Five genes among them, STAT3, PDGFRB, PTPN11, SOCS1, and PLCG1, are already confirmed for their functional relevance to cancers. For the other four candidate targets, SHC1, MAPK1, LYN and STAT1, all of them participate in the pathway of chemokine signaling (hypergeometric test, P = 2.72 × 10⁻⁹), which plays a pivotal role in tumor progression and metastasis. Dysregulated expression of several chemokine signaling pathways has been implicated in the progression of several cancers, including breast cancer, gastric cancer, pancreatic cancer, etc. Besides, Wong et al. have illustrated that esophageal cells with mutant p53 and POSTN reveal upregulation of the STAT1 network and STAT1-dependent target genes. Overall, the integrated analysis of the PPI network and large-scale gene expression data reveals its ability to indicate cancer-related signaling networks, although additional experiments are required to further establish the functions of these putative therapeutic targets. This finding further demonstrates that ICN-based differential analysis is highly biologically interpretable.

3 Discussion

In this paper, we have provided a novel normalization procedure, informative CrossNorm (ICN), which can be used to pre-process microarray datasets with a global shift between conditions relying on the informative gene set and CrossNorm. Normalization is a critical step in eliminating biases in microarray data for accurate analysis. However, conventional normalization methods are usually not powerful in dealing with samples with

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Accession number</th>
<th>All genes</th>
<th>Informative gene</th>
<th>Informative gene ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>pancreatic32</td>
<td>GSE16515</td>
<td>5467</td>
<td>14291</td>
<td>0.2614</td>
</tr>
<tr>
<td>pancreatic78</td>
<td>GSE15471</td>
<td>5467</td>
<td>22747</td>
<td>0.416</td>
</tr>
<tr>
<td>pancreatic90</td>
<td>GSE28735</td>
<td>32321</td>
<td>4956</td>
<td>0.1533</td>
</tr>
<tr>
<td>gastric62</td>
<td>GSE13911</td>
<td>5467</td>
<td>13477</td>
<td>0.2465</td>
</tr>
<tr>
<td>gastric24</td>
<td>GSE19826</td>
<td>5467</td>
<td>16032</td>
<td>0.3932</td>
</tr>
<tr>
<td>RCC20a</td>
<td>GSE6344</td>
<td>22863</td>
<td>6390</td>
<td>0.2868</td>
</tr>
<tr>
<td>RCC20b</td>
<td>GSE6334</td>
<td>22645</td>
<td>4648</td>
<td>0.2053</td>
</tr>
<tr>
<td>ESCC106</td>
<td>GSE20347</td>
<td>22277</td>
<td>10488</td>
<td>0.4708</td>
</tr>
<tr>
<td>ESCC 24</td>
<td>GSE29001</td>
<td>22277</td>
<td>10818</td>
<td>0.4856</td>
</tr>
<tr>
<td>breast26</td>
<td>GSE10780</td>
<td>5467</td>
<td>15063</td>
<td>0.2755</td>
</tr>
<tr>
<td>breast24</td>
<td>GSE16873</td>
<td>22863</td>
<td>7779</td>
<td>0.3491</td>
</tr>
<tr>
<td>otscc40</td>
<td>GSE13601</td>
<td>12625</td>
<td>4524</td>
<td>0.3583</td>
</tr>
<tr>
<td>otscc24</td>
<td>GSE9844</td>
<td>5467</td>
<td>12492</td>
<td>0.2285</td>
</tr>
<tr>
<td>crc34a</td>
<td>GSE18205</td>
<td>5467</td>
<td>20349</td>
<td>0.3722</td>
</tr>
<tr>
<td>crc34b</td>
<td>GSE32323</td>
<td>5467</td>
<td>20349</td>
<td>0.3722</td>
</tr>
<tr>
<td>crc64</td>
<td>GSE8671</td>
<td>5467</td>
<td>10669</td>
<td>0.1951</td>
</tr>
<tr>
<td>hcc20</td>
<td>GSE29721</td>
<td>5467</td>
<td>18018</td>
<td>0.3295</td>
</tr>
<tr>
<td>hnscc44</td>
<td>GSE66631</td>
<td>12625</td>
<td>3772</td>
<td>0.2988</td>
</tr>
<tr>
<td>lung34</td>
<td>GSE31552</td>
<td>32321</td>
<td>3093</td>
<td>0.9957</td>
</tr>
<tr>
<td>lung88</td>
<td>GSE18842</td>
<td>5467</td>
<td>16028</td>
<td>0.2932</td>
</tr>
<tr>
<td>lung54</td>
<td>GSE7670</td>
<td>22283</td>
<td>7999</td>
<td>0.3365</td>
</tr>
<tr>
<td>lung66</td>
<td>GSE10072</td>
<td>22283</td>
<td>5544</td>
<td>0.2488</td>
</tr>
<tr>
<td>lung120</td>
<td>GSE19804</td>
<td>5467</td>
<td>21795</td>
<td>0.3986</td>
</tr>
<tr>
<td>ptc40</td>
<td>GSE29265</td>
<td>5467</td>
<td>15397</td>
<td>0.2816</td>
</tr>
</tbody>
</table>

Each dataset is denoted using the cancer type followed by the sample size.
global variation, such as a dominant scenario in cancer datasets, and may even aggravate technical artifacts in measurements because of the violation of the normalization assumptions. ICN is effective in processing cancer datasets for downstream differential analysis and furthermore, it stands second to none for the conventional methods even when normalizing datasets satisfying the standard assumptions.

I/NI-calls is applied in ICN for informative gene selection. It is an effective tool for eliminating potentially uninformative features to reduce noise and increase precision in detecting DEGs. In principle, according to the user specifications, the ICN model can be extended naturally using any other gene filtering procedure, such as FLUSH. Similarly, the step of CrossNorm can employ other normalization methods, and is not only limited to Quantile. Applying these methods through the ICN strategy, which solely uses the informative signals, considerably enhances the robustness and accuracy of estimating the transcript level in cancer microarrays.

The downside is that traditional expression microarrays are designed for measuring the transcript abundance of all nucleic acids in a sample, but cannot make a distinction between different splice isoforms. Also, the I/NI method does not take alternative splicing into consideration and it is a rigorous way to identify the information probe-set or genes. A fraction of alternatively spliced genes might be falsely identified as non-informative genes and excluded for subsequent analyses. More generally, filtering genes before normalization or differential analysis may omit some true DEGs, which cannot be further tested, but it is an obvious trade-off between precision and recall; otherwise, a higher FDR will be generated as long as no perfect gene filtering methods are launched.

Additionally, gene filtering and DEG selection should be independent under the null hypothesis for DEG selection. In ICN, the former step eliminates the uninformative genes on the basis of the correlation between probes, whereas the latter step utilizes the t-test for two-class comparison. Type I error rate control is not affected by the step of gene filtering.

The results have crucial implications for past and future cancer studies based on microarray samples with non-negligible differences. Hence, it is highly recommended to routinely apply ICN for cancer microarray studies. In future work, we expect that ICN could be used with success on RNA-seq expression data, as the ICN strategy can be theoretically applied to other profile-like data where experiments have global expression variations between samples.

4 Datasets and methods

4.1 Datasets

4.1.1 Cancer datasets. This study has used 25 real-life cancer datasets collected from the open source database Gene Expression Omnibus with pairwise cancer and normal samples. To derive rigorous inference from the results obtained from our analyses, each experiment consists of no less than 20 samples in order to be statistically relevant. In our paper, each dataset is denoted using the cancer type followed by the sample size, for example, the dataset with accession number GSE23400 is
4.1.2 Spiked-in datasets. Three spiked-in datasets have been used for evaluation, which were produced by controlled experiments with assigned concentration-folds. The first dataset is known as the ‘golden spiked-in’ dataset proposed by Choe et al., whereas the other two datasets are from the open challenge ‘Affycomp III’ (http://affycomp.biostat.jhsph.edu/). For simplicity, each dataset is denoted by its platform type, namely DrosGenome1, HGU95A, and HGU133A, respectively.

Spiked-in dataset DrosGenome1. This dataset is designed to identify differentially expressed genes (DEGs) using Affymetrix DrosGenome1 chips. It consists of three case arrays and three control arrays. 3860 probe sets were assigned with known concentration-folds, of which 2551 probe sets have the same concentration between two groups and 1309 probe sets have a series of predefined concentration-folds {1.2, 1.5, 1.7, 2.0, 2.5, 3.0, 3.5, 4.0} between arrays. Other 10 144 probe sets with no spiked concentration were considered as technique noise.

Spiked-in dataset HGU95A. This dataset is the first part of the ‘affycomp’ website and consists of 59 Affymetrix HGU95A arrays grouped into 20 experiments. The RNA concentration of 14 genes was predefined with particular concentrations of a ‘Latin square’ design mixture ranging from 0.25 to 1024 pM. We employed two groups, each consisting of seven arrays with the same spiked concentrations. Eventually, profiles with 12 606 probe sets and 14 arrays were laid out.

Spiked-in dataset HGU133A. This dataset, the second part of the ‘affycomp’ website, is also based on a ‘Latin square’ experimental design. It consists of 42 HGU133A arrays with 42 spiked-in genes spiked RNA concentrations ranging from 0.125 to 512 pM. The concentration-folds are 2 for all of these 42 spiked-in probe sets and unknown for the other 22 258 probe sets. Each experiment contains three technical replicate arrays and thereby a total of 42 arrays are stored in the dataset. In the study, we compared experiments 6 and 7, which were separated by one shift in the ‘Latin square’ design with a fold-change of 2.

4.2 Informative gene selection
Applying the expression data for all genes without filtering may result in high false negative rates. In ICN, informative genes are identified using I/NI-calls. Although several feature filtering techniques are available, they often have shortcomings such as overfitting and selection bias; therefore, these techniques are not able to produce effective results. For I/NI-calls, it can efficiently remove false positives, namely genes biologically irrelevant to the experiments that can produce erroneous output in the subsequent analyses. Moreover, it can enhance the statistical power of identifying significant features, because the number of features is reduced as some of those which would otherwise be identified as significant are eliminated. In general, the informative genes refined by I/NI-calls are biologically more interpretable than the entire genome.

In particular, the core algorithm of I/NI-calls is a factor analysis—a multivariate method to determine common structures in data with multiple probes measuring the same gene or probe set. Across arrays, multiple probes in an identical probe set measure the same target transcript; therefore, they should be positively correlated if meaningful variations exist in the mRNA concentration across the arrays. A probe set is informative when several of its probes exhibit the same trend in mRNA concentration across the arrays; otherwise, a dissimilar expression pattern indicates that these probes are masked by noise and thereby cannot represent a probe set, and this type of probe sets is non-informative. I/NI calls uses an R package, FARMS, to quantify the signal-to-noise ratio of each probe set. Probe sets with a signal-to-noise ratio smaller than 1 are considered as not informative and hence are eliminated. Please refer to www.bioinf.jku.at/software/farms/farms.html for more detailed information.

4.3 Cross normalization
In general, the proportion of DEGs is expected to be small enough so that the transcript intensity distributions among arrays are assumed to be similar or even identical for the conventional normalization methods, such as Quantile, MASS.50, and LOESS. Quantile and MASS.50 are two widely applied methods for single-channel microarray data while LOESS was originally designed for two-channel data. When processing single-channel data, the LOESS normalization fits LOESS curves to MA-plot and adjusts the arrays with respect to each other by forcing log-ratios (M) to be scattered around a constant. The LVS method uses a non-linear model to fit each array to a given reference array merely focusing on the pre-identified invariant genes. For uniformity, prior to normalization, all the raw data were firstly background-corrected using the RMA algorithm and then probes in the same probe-set were averaged to a summarized expression value.

On the basis of the quantile normalization, CrossNorm expands upon the conventional assumptions and assumes that the merged disease and normal transcript intensities have the same distribution. Hence, it is reasonable to perform quantile on the translocated expression matrix to acquire a comparable expression profile. CrossNorm is effective in differential analysis for datasets with a large fraction of DEGs; nevertheless, it does not have a satisfactory performance in controlling false positives for datasets with numerous noisy signals or non-informative genes.

4.4 ICN procedure
The ICN work-flow is illustrated in Fig. 1. The raw data are background-corrected and summarized to obtain an expression matrix for all genes. Simultaneously, these raw data are processed using I/NI-calls to identify informative genes required for subsequent analyses. CrossNorm is then performed for the expression matrix only containing informative genes. Ultimately, the ICN procedure returns an expression matrix solely for informative genes whose expression values are biologically interpretable. The ICN procedure can be summarized as follows:
(1) Apply I/NI-calls to the raw data to identify a set of informative genes for further analyses.

(2) Background-correct and summarize the raw data to obtain an expression matrix.

(3) Perform CrossNorm to the expression matrix merely containing informative genes in step (1).

(4) Identify DEGs among the informative gene set using an appropriate method, (e.g., fold-change or t-test).

4.5 Quantification of differential expression

An initial and fundamental objective of a microarray is to identify genes whose expression level changes between conditions. In particular, in cancer studies, the aim is to compare gene expression levels between normal and cancer states. In this study, if not stated otherwise, we consider a gene to be significantly differentially expressed based on the following criteria: (i) it has a log fold-change higher than 1.3 and (ii) it has a Q value lower than 0.1, which is the adjusted P value based on the false discovery rate (FDR, Benjamini & Hochberg). The P value is simply obtained using a two sample t-test.

4.6 Evaluation of differential expression

In this paper, we measure the performance of each normalization method using precision, recall, $F_1$-score and Matthews correlation coefficient (MCC). Precision is the fraction of truly detected DEGs and recall is the fraction of correctly detected DEGs over all known DEGs. The $F_1$-score is defined by the following equation:

$$F_1 = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} = \frac{2 \times \text{TP}}{2 \times \text{TP} + \text{FP} + \text{FN}}$$

The second phase of the MAQC project (MAQC-II) indicates that the MCC value is a better performance metric than other measurements, such as AUC-ROC, because by definition a ROC curve constructs the performance over all possible cutoffs whereas only one or a few cutoff(s) are used to detect gene differential expression. MCC is estimated as follows:

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(	ext{TP} + \text{FP}) \times (	ext{TP} + \text{FN}) \times (	ext{FP} + \text{TN}) \times (\text{TN} + \text{FN})}}$$

For both of the two above equations, TP, TN, FP, and FN represent the true positive, true negative, false positive and false negative, respectively.

4.7 Enrichment analysis

A hypergeometric probability model was used to test whether the interesting genes are enriched within a special category.

$$P = 1 - \sum_{i=0}^{k} \frac{C_j \times C_N^{M-j}}{C_N^M}$$

where $N$ is the number of entire background genes, $M$ is the number of genes within one category, $n$ is the number of interesting genes selected from the background genes, and $k$ is the number of genes selected from the category. The resultant $P$ value reflects the probability of extracting up to $k$ possible $n$ genes in $N$ drawings.

Author contributions

L. C. and D. W. conceived and designed the experiments. L. C. analyzed the data and performed the experiments. L. C., K. L., D. W., P. W., K. L., L. L., and B. X. wrote the manuscript. All authors reviewed and approved the final manuscript.

Conflict of interest

None of the authors have any competing interests in the manuscript.

Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CrossNorm</td>
<td>Cross normalization</td>
</tr>
<tr>
<td>ICN</td>
<td>Informative CrossNorm</td>
</tr>
<tr>
<td>ESCC</td>
<td>Esophageal squamous cell carcinoma</td>
</tr>
<tr>
<td>DEG</td>
<td>Differentially expressed gene</td>
</tr>
<tr>
<td>MCC</td>
<td>Matthews correlation coefficient</td>
</tr>
<tr>
<td>FC</td>
<td>Fold-change</td>
</tr>
<tr>
<td>FDR</td>
<td>False discovery rate</td>
</tr>
<tr>
<td>PPI</td>
<td>Protein–protein interaction</td>
</tr>
<tr>
<td>MAQC</td>
<td>Microarray quality control</td>
</tr>
</tbody>
</table>

Acknowledgements

This work was supported by a Direct Grant from the Chinese University of Hong Kong and a GRF Grant (Project Reference 414413) from the Research Grants Council of Hong Kong SAR, China. It was also supported by the Natural Science Foundation of Heilongjiang Province of China (C2015027) and the Scientific Research Fund of Heilongjiang Provincial Education Department (12541426).

References

19 S. E. Choe, M. Boutros, A. M. Michelson, G. M. Church and M. S. Halfon, Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset, Genome Biol., 2005, 6(2), R16.