BIOINFORMATICS ORIGINAL PAPER Vol. 24 no. 18 2008, pages 2057–2063
doi:10.1093/bioinformatics/btn365

Systems biology

Apparently low reproducibility of true differential expression discoveries in microarray studies

Min Zhang[1,†], Chen Yao[2,†], Zheng Guo[1,2,*], Jinfeng Zou[1,‡], Lin Zhang[2,‡], Hui Xiao[1], Dong Wang[1], Da Yang[1], Xue Gong[1], Jing Zhu[2], Yanhui Li[2] and Xia Li[1,*]

[1]School of Bioinformatics Science and Technology, Harbin Medical University, Harbin 150086 and [2]Bioinformatics Centre and School of Life Science, University of Electronic Science and Technology of China, Chengdu 610054, China

Received on April 14, 2008; revised and accepted on July 14, 2008
Advance Access publication July 16, 2008
Associate Editor: Olga Troyanskaya

ABSTRACT

Motivation: Differentially expressed gene (DEG) lists detected from different microarray studies for a same disease are often highly inconsistent. Even in technical replicate tests using identical samples, DEG detection still shows very low reproducibility. It is often believed that current small microarray studies will largely introduce false discoveries.

Results: Based on a statistical model, we show that even in technical replicate tests using identical samples, it is highly likely that the selected DEG lists will be very inconsistent in the presence of small measurement variations. Therefore, the apparently low reproducibility of DEG detection from current technical replicate tests does not indicate low quality of microarray technology. We also demonstrate that heterogeneous biological variations existing in real cancer data will further reduce the overall reproducibility of DEG detection. Nevertheless, in small subsamples from both simulated and real data, the actual false discovery rate (FDR) for each DEG list tends to be low, suggesting that each separately determined list may comprise mostly true DEGs. Rather than simply counting the overlaps of the discovery lists from different studies for a complex disease, novel metrics are needed for evaluating the reproducibility of discoveries characterized with correlated molecular changes.

Contact: guoz@ems.hrbmu.edu.cn; lixia@ems.hrbmu.edu.cn

Supplementaty information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

As an ‘Array of Hope’ (Lander, 1999), microarray technology has enormous influence on modern biological researches. However, as ‘An Array of Problems’ (Frantz, 2005), microarray technology has been challenged by many criticisms about its reliability (Frantz, 2005; Marshall, 2004; Tan *et al*., 2003). Often, it is the low reproducibility of the differentially expressed genes (DEGs) lists for a disease that raises doubts about the reliability of microarrys (Ein-Dor *et al*., 2005; Miklos and Maleszka, 2004). Impressively, even using technical replicated samples for intra- or inter-platform comparisons, DEG detection still shows very low reproducibility (Tan *et al*., 2003). On the other hand, many studies (Shi *et al*., 2006; Tong *et al*., 2006) suggested that most microarray platforms can generate rather reliable and reproducible measurements. Specifically, the MAQC (MicroArray Quality Control Consortium) (Shi *et al*., 2006) studies suggested that the lack of reproducibility of DEG lists may come from the common practice of using stringent P-value cutoffs to determine DEGs. Thus, they suggested choosing genes with large changes combining with a less stringent P-value cutoff to increase the reproducibility of DEG lists, which was criticized for being short of statistical control (Klebanov *et al*., 2007).

The reproducibility of gene lists is often measured by the percentage of overlapping genes (POG) (Ein-Dor *et al*., 2006; Irizarry *et al*., 2005; Shi *et al*., 2006) between gene lists from different microarray datasets. (Ein-Dor *et al*., 2006) analyzed the POG of gene lists selected according to the correlation of gene expressions with sample labels and concluded that, because of large biological variations, it might need thousands of samples to reach a high POG score. However, they did not use a proper statistical control to guarantee that the lists comprised mostly true discoveries, which might be misleading because the POG score can be large for two gene lists sharing mostly false discoveries. Here, by a statistical model treating the POG between gene lists from different datasets as outcomes of a random experiment, we show that even when using identical samples in technical replicate tests, with small technical variations (Tan *et al*., 2003), it is still highly possible that the DEG lists obtained with statistical control of false discoveries are very inconsistent. Therefore, the low reproducibility of DEG lists from current technical replicate tests does not directly indicate low quality of microarray technology.

By resampling subsamples from three large cancer datasets as well as simulated data, we show that the number of the DEGs detected in each subsample by using SAM (significance analysis of Microarray) (Tusher *et al*., 2001) method under FDR (false discovery rate) control (Benjamini and Hochberg, 1995) increases greatly as the sample size increases. The wide and complex expression changes in a complex disease are separately detectable at different sample size levels, further reducing the overall

[*]To whom correspondence should be addressed.
[†]M.Zhang and C.Yao contributed equally to this work.
[‡]J.Zou and L.Zhang contributed equally to this work.

reproducibility. On the other hand, in contrast to the common belief that current small microarray studies will introduce many false discoveries (Klebanov *et al*., 2007), we show that it is entirely possible that each separately determined list comprises mostly true discoveries.

In many other high-throughput postgenomic areas such as proteomics (Ransohoff, 2005b) and metabolomics (Broadhurst and Kell, 2006), the irreproducibility problem in finding molecular markers of complex diseases also exists and often leads to disappointments and disputes among investigators. However, we will show in this article that the low apparent reproducibility of discoveries generated in postgenomic areas does not prove lack of reliability of the high-throughput technology platforms used. This problem might reflect a kind of 'culture clash' (Frantz, 2005) between the systems biology and the traditional biology. For a complex disease characterized with many coordinated changes of disease markers, we need novel concepts and metrics to evaluate the reproducibility of discovery lists at the systems biology level by considering the correlation of molecular changes, rather than simply counting the overlaps of discoveries from different studies.

2 METHODS

2.1 Data preprocessing and normalization

Three large cancer datasets are analyzed. The prostate cancer cDNA microarray data (Lapointe *et al*., 2004) consists of 62 primary prostate tumors and 41 normal prostate specimens measured for 46 205 clones. The liver cancer cDNA microarray data (Chen *et al*., 2002) contains 82 primary hepato-cellular carcinoma (HCC) and 74 non-tumor liver tissues measured for 23 093 clones. The overall missing rate with respect to the whole data in each dataset is 10% and 5% for prostate and liver cancer data, respectively, and a lower missing rate may reflect higher data quality. The leukemia data (Yeoh *et al*., 2002) consists of 79 TEL-AML1 and 64 Hyperdiploid samples measured for 12 600 probe sets by Affymetrix U95A GeneChip (Affymetrix Incorporated, Santa Clara, CA). The original authors Yeoh *et al*. found a high reproducibility of measurement signals between replicate samples. Additionally, we analyze a subset of the MAQC dataset (AFX_1) for technical replicated samples (Shi *et al*., 2006).

The cDNA data is log2-transformed and then normalized as median 0 and SD 1 per array, as adopted in Oncomine database (Rhodes *et al*., 2007). The CloneIDs with missing rates above 20% are deleted. The remaining missing values are replaced by using the *k*NN imputation algorithm ($k=15$) (Troyanskaya *et al*., 2001). The Affymetrix GeneChip data is preprocessed by the robust multi-array analysis (RMA) and then between-array median normalized (Irizarry *et al*., 2005). The most recent (July, 2007) SOURCE database (Diehn *et al*., 2003) is used for annotating CloneID to GeneID. Because all the current normalization procedures are debatable (Do and Choi, 2006), we additionally try LOWESS (Yang *et al*., 2002) and median global (Quackenbush, 2002) normalizations for cDNA data. For the Affymetrix GeneChip data, we additionally apply the commonly used software MAS5.0 (Gautier *et al*., 2004) which performs background correction using neighboring probe sets. For the gene selection problem, global normalizations as adopted in this study are proper choices because local normalizations usually require selecting non-DEGs beforehand.

2.2 Selection of DEGs

In the real datasets, we use the most popular SAM (samr_1.25 R package) method (Tusher *et al*., 2001) to select DEGs. In the statistical model, we use *t*-test to select DEGs because the simulated data is ideally normally distributed. While, the multiple statistical tests are controlled by FDR defined as the expected percentage of false positives among the claimed DEGs (Benjamini and Hochberg, 1995). Because the FDR estimation of SAM might be overly conservative (Xie *et al*., 2005; Zhang, 2007), we also apply the FDR estimation method suggested by Zhang (2007) following the idea of Xie *et al*. (2005), and refer it as the modified SAM method.

2.3 Evaluation of the apparent reproducibility

The reproducibility of gene lists is often measured by the POG metric (Ein-Dor *et al*., 2006; Irizarry *et al*., 2005; Shi *et al*., 2006). However, because the POG metric depends on the lengths of gene lists (Chen *et al*., 2007; Shi *et al*., 2005), it cannot be used to compare the reproducibility of gene lists with different lengths. Therefore, we refer to the POG score as apparent reproducibility.

To study some major factors affecting the POG score, we first analyze a simple statistical model: all the DEGs are supposed to have the same expected fold change (FC) and coefficient of variance (CV) at the original measurement (intensity or ratio) level and the data is log-normally distributed in both groups of samples. Then, we can reason that the log-expression follows normal distribution with equal variance in two groups of samples. Thus, *t*-test can be ideally used to detect DEGs. When using *n* samples per group and FDR control level *fdr* to detect DEGs with FC = *fc* and CV = *cv*, the expected power β and POG of the DEG lists can be calculated as below (see details in Supplementary Methods):

$$\beta = t'_{df,\lambda}(-c) + 1 - t'_{df,\lambda}(c) \tag{1}$$

$$E(POG) = \beta \times \left(\frac{fdr^2 \times \pi}{(1-\pi)(1-fdr)} + 1 - fdr \right) \tag{2}$$

where π is the proportion of DEGs. c is determined by *fdr* (Pawitan *et al*., 2005a). $\lambda = \sqrt{n/2} \times \left(\log fc / \sqrt{\log(cv^2+1)} \right)$ is the parameter of the non-central *t*-distribution, and $df = 2n-2$.

When selecting two DEG lists with length l_1 and l_2 from N genes, the probability that they share at least k genes by random chance can be calculated by the hypergeometric probability model.

For DEGs in real data with heterogeneous expression changes, we use a mixture model (Pawitan *et al*., 2005b) to estimate the pattern of the FC and CV distributions. Then simulated data are created using the estimated parameters from the real data to illustrate some more complex changes of the POG (see detail in Supplementary Methods). Additionally, we also simulate the heterogeneous differential expressions of DEGs by a model proposed by Perelman *et al*. (2007). Briefly, variance σ_i^2 varies randomly, following the scaled inverse of a χ^2-distribution $d_0 s_0^2 / x_{d_0}^2$ with d_0 degree of freedom. Fold difference is zero for non-DEGs and follows normal distribution $N(0, v_0\sigma_i^2)$ for DEGs. Here, the tuning parameters s_0, v_0 and d_0 are set as 0.5, 1 and 12, respectively, to balance the variance differing moderately among genes.

3 RESULTS

3.1 Low apparent reproducibility of DEG selection

In general, the relationship between the POG expectation and some variables such as FDR and sample size is complicated, as shown in Figure 1. However, some trends can be observed.

(1) The expected POG increases as the FC increases (or CV decreases), when fixing the other parameters. In Equation (2), a larger FC (or a smaller CV) will produce a larger λ and smaller c, leading to a higher power and POG. Figure 1A and D, respectively, demonstrates the POG changing with the increased FC (or CV) for the selected DEGs with two CV (or FC) values. Here, the FDR control level is 1%, the proportion of DEGs is given as $\pi = 10\%$ and the sample size is five per group.

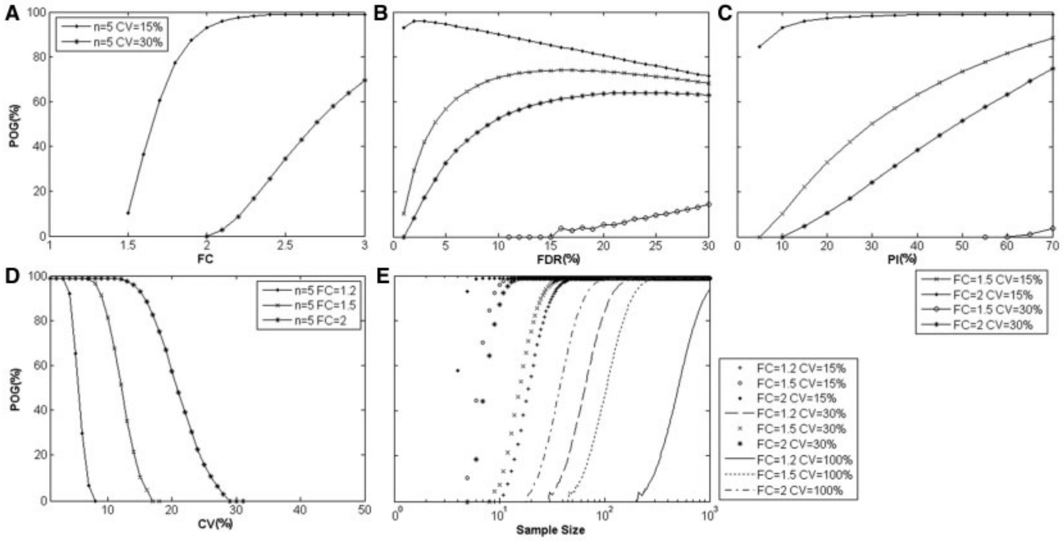

Fig. 1. Distributions of the POG with some parameters. The changes of the POG are represented in line plots against each variable in *x*-axis when fixing the other parameters in the expected POG model. (**A**) FC. (**B**) FDR control level. (**C**) PI (π) and (**D**) CV. (**E**) Sample size. Same legend for (B) and (C).

(2) When *fdr* is very small (≈0), Equation (2) can be approximately simplified as $E(POG) \approx \beta \times (1 - fdr) \approx \beta$. Thus, if FDR is stringently controlled, the POG is approximately equal to the power. Fixing the other parameters, when the FDR changes in an acceptable range, the POG goes up with the increasing FDR level. Figure 1B shows that as the FDR increases from 1% to 30%, most POGs of the selected DEGs with different FC and CV values increase. However, when FC is 2 and CV is 15%, the POG achieves the highest value at a small FDR control level (about 3%) and then decreases as the FDR keeps increasing because of the increased false positives. Here, the proportion of DEGs is given as $\pi = 10\%$ and the sample size is five per group.

(3) Generally, when fixing the other parameters, a larger π will increase the length of DEGs and thus a higher POG. As shown in Figure 1C, as π becomes larger, the POGs of the selected DEGs with different FC and CV values increase. Here, the FDR control level is 1% and the sample size is five per group.

(4) When fixing the other parameters, using more samples can increase the power and thus the POG. Figure 1E shows the changed POGs of different kinds of the selected DEGs as the sample size increases. Here, the FDR control level is 1%, and the proportion of DEGs is given as $\pi = 10\%$.

In technical replicate tests with no biological variation, the CV of the original signals could be <15%, indicating acceptable technical quality (Shi *et al*., 2006). However, at this CV level, when using five samples per group to detect DEGs with 1.5 FC, the expected POG is only about 10% and it will decrease as the FC decreases (Fig. 1). Notably, when FC increases from 1.5 to 2, the POG jumps from 10% to 93% because the power increases from 10% to 94%. Note that the POG is approximately equal to the power at a stringent FDR control level. The above analysis suggests that it is highly likely, rather than 'surprisingly' (Marshall, 2004), that DEG lists from small-scaled technical replicated tests (Shi *et al*., 2006; Tan *et al*., 2003) are very inconsistent.

In the presence of large biological variations (Klebanov and Yakovlev, 2007), lower POG scores are expected. If the total CV increases to 30%, the expected POG approaches to zero while using five samples per group to detect DEGs with 1.5 FC. When the CV is 100%, using 200 samples can only achieve 46% POG for DEGs with 1.5 FC, and thousands of samples are required to achieve 50% POG for DEGs with smaller changes (FC = 1.2).

3.2 Heterogeneous expression changes in real data

Here, we analyze three large cancer datasets. To mimic small experiments from each dataset, we randomly produce subsets at different sample size levels, ranging from five samples per group (cancer or normal) to the highest sample size level of the dataset.

Figure 2A shows the results for the cDNA microarry data normalized by the method adopted in Oncomine database and for the Affymetrix GeneChip data normalized by RMA. In each dataset, by using SAM with 1% FDR control, the average number of the DEGs identified across 100 resampling subsets increases dramatically as the sample size increases. For example, even when the sample size increases from 35 to the highest level of 40 samples per group in the prostate cancer data, the median number of DEGs increases from 1612 to 1889. Figure 2A also shows the trend of the POG increasing with the increased sample size. However, the POG may be overestimated since larger subsamples from a dataset will have larger sample overlaps. Similar results are observed when using 10% FDR control, the other normalization methods (Supplementary Figs S1 and S2) and the modified SAM method (Supplementary Figs S3 and S4).

It is known that, at a FDR control level, increasing sample size will increase the power of an appropriate statistical test in finding true DEGs while decreasing the probability of declaring non-DEGs as DEGs (Pawitan *et al*., 2005a). The observation that the number of the detected DEGs increases with sample size suggests that the three

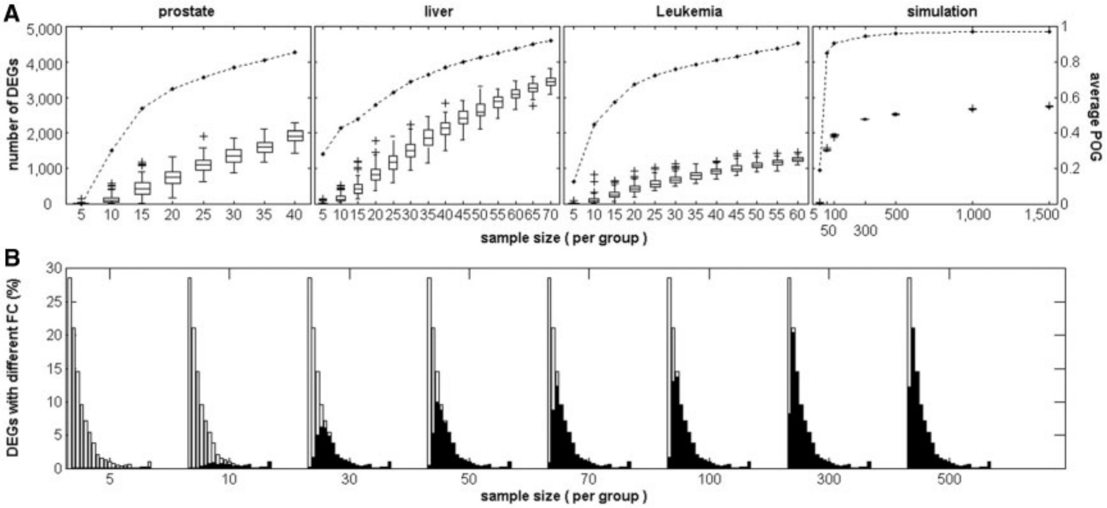

Fig. 2. The numbers of the DEGs detected and the corresponding average POG under different sample size conditions. (**A**) The DEGs identified with 1% FDR control. For each dataset, a box-and-whisker plot shows the DEG number in the left *y*-axis across 100 resampling subsets. The symbol '+' represents outlier data. The corresponding average POG is represented in the right *y*-axis as line plot. (**B**) The heterogeneous expressions of the predefined DEGs in the simulated data. The white bars represent the proportions of the DEGs with FC ranging from 1 to 5 stepped by 0.2, while the DEGs with FC above 5 are grouped together. The black bars represent the fractions of the DEGs detected under different sample size conditions. Genes with smaller expression changes could be detected when more samples are used.

relatively 'large' datasets might still be inadequate for studying the heterogeneous expression changes in cancers.

Then, as a proof of principle, we analyze the heterogeneous differential expressions by a simulation model (Perelman *et al.*, 2007), which is free of any hidden systemic bias possibly existing in real data. As described briefly in Section 2, the simulated data is normally distributed, with variance and FC varying across genes. We produce samples each consisting of 10 000 genes among which 30% is defined as DEGs. Now, as the sample size increases, the number of the detected DEGs gradually converges to the predefined number of DEGs, allowing false positives controlled by 1% FDR (Fig. 2A). For example, using 1500 samples per group, we can detect about 91% of all the predefined DEGs, while excluding most non-DEGs. Figure 2B shows smaller expression changes will be detected with larger samples.

Considering the heterogeneous differential expressions in the real data, we use the mixture model (Pawitan *et al.*, 2005b) to estimate the pattern of the FC and CV distributions in the three cancer data and the MAQC data, respectively. Then simulated data with 10% DEGs are created using the estimated parameters to mimic the heterogeneous differential expressions in each dataset. Using five samples per group, the POGs of the DEGs selected with 1% FDR control from the simulated data for these cancer data are all near zero, while that for the MAQC data is relatively large as 62%. Due to the nature of the MAQC's experimental design for technical replicate tests, most DEGs from the MAQC data have large FC (>3-fold) and low CV which show a strikingly different profile from the heterogeneous cancer datasets (Supplementary Fig. S5).

3.3 DEG lists comprising mostly true discoveries

Current FDR control procedures, including the one adopted in SAM, may be unstable in small samples, especially in the presence of correlated expression changes. We thus evaluate the actual FDR of a DEG list detected in simulated small samples, according to the predefined DEGs. Using 100 resampling subsamples with five samples per group, at 1% FDR control level, the median actual FDR is near zero and the median number of the detected genes is only 16. At 10% FDR control level, the median actual FDR is 3% and the median number of the detected genes increases to 139.

In each real dataset, by using SAM with 1% FDR control, we empirically define the DEGs obtained from the full samples as a nominal gold standard set (Pavlidis *et al.*, 2003). However, because such a gold standard set may include only a small fraction of the true positives, it is very likely that many DEGs from small samples will be wrongly judged as false discoveries, leading to enlarged nominal actual FDR estimates. Nevertheless, when using 100 subsamples with 5 samples per group, at 1% FDR control level, the median nominal actual FDR is also near zero for each dataset (Table 1), while the median number of the detected DEGs is only 6, 17 and 9 for prostate, liver and leukemia data, respectively. As the control FDR level goes up to 10%, these numbers increase to 45, 92 and 37, while the nominal median actual FDR is <3% for each dataset. Similar results are observed in the data normalized differently and using the modified FDR method (Supplementary Tables S1 and S2).

The low actual FDR levels in both simulated and real data indicate that it is entirely possible that each list from small samples comprises mostly true discoveries, though the DEG lists separately detected in different studies tend to be inconsistent (Fig. 2A). Notably, the genes subtly altered have a strong impact on FDR via increasing the size of the proportion of DEGs. Most current FDR control approaches often underestimate this proportion and thus overestimate the FDR value (Pawitan *et al.*, 2005b), especially when the multiple tests are not independent with each other.

Table 1. The power and actual FDR (using 1% and 10% FDR control, five samples per group)

FDR	Dataset (normalization)	Number[a]	Tp number[b]	Power[c]	actual FDR[d]
1%	Simulation	16 (4)	16 (4)	0.005 (0.001)	0(0)
	Prostate (Oncomine)	6 (5)	6 (4)	0.003 (0.002)	0(0)
	Liver (Oncomine)	17 (8)	17 (7)	0.005 (0.002)	0(0)
	Leukemia (RMA)	9 (6)	9 (6)	0.006 (0.004)	0(0)
10%	Simulation	139 (16)	134 (16)	0.045 (0.005)	0.03 (0.01)
	Prostate (Oncomine)	45 (88)	44 (83)	0.01 (0.019)	0.01 (0.02)
	Liver (Oncomine)	92 (73)	88 (70)	0.015 (0.012)	0.02 (0.02)
	Leukemia (RMA)	37 (62)	35 (59)	0.016 (0.027)	0.03 (0.04)

This table shows when using 100 resampling subsamples with five samples per group in each dataset, under 1% and 10% FDR control, the median and the quartile deviation (in the parentheses) of:

[a] The total number of the DEGs detected.

[b] The number of the detected DEGs appearing in the nominal gold standard.

[c] The power of the resampling experiments for DEG detection.

[d] The nominal actual FDR.

Despite an overall low POG level, some genes are frequently selected from different subsamples, which are often associated with the cancer under study (Supplementary Table S3).

4 DISCUSSION

Here in this article, we compare the DEG lists from different subsamples in a same study to avoid any platform and site differences. According to our analysis, even in some ideal situations like technical replicate tests with small technical variations, the DEG lists can still be very inconsistent. Therefore, the apparently low reproducibility of DEG lists from current technical replicate tests does not indicate low quality of microarray technology. In different studies for a complex disease such as cancer, it will be more difficult to obtain a consistent DEG list, though each separately determined list may comprise mostly true DEGs. Obviously, genes with modest or small changes will decrease the overall POG. The MAQC suggested a FC-based approach with a less stringent P-value to balance statistical significance and reproducibility (Shi *et al.*, 2006). In our simple model, when the CV is given, larger FC can lead to larger POG expectation. This result partially supports the MAQC suggestion. Therefore, for reaching a higher (apparent) reproducibility, we do not recommend using very stringent FDR control level in determining DEGs. However, the relationship between the statistical significance (FDR) and the reproducibility (POG) is complex. Note that the POG from the MAQC data that tends to be large is specific to their technical replicate data characterized with small CV and inherently large FCs (brain samples versus cell culture). In most biological datasets, there is a more intricate interplay between CV, FC and POG. For example, in the three cancer datasets, genes with larger FC tend to have larger CV (Supplementary Fig. S5), which might decrease the POG of DEG lists even if they only include genes with large FC.

It has been suggested that using thousands of samples for a disease could finally produce a reproducible DEG list (Ein-Dor *et al.*, 2006), which, however, would still be hardly reproducible in small samples. It has also been suggested that we may find consistent DEGs by extracting the genes frequently detected over many samplings (Li *et al.*, 2004; Qiu *et al.*, 2006). However, as suggested by this study, this practice tends to miss most of the significant genes because of the low reproducibility of DEG lists from different subsamples. Thus, only accepting reproducible data would infer misleading biological conclusions (Hakes *et al.*, 2008).

It is believed that the reproducibility of scientific discoveries is of fundamental importance (Marshall, 2004) and 'a study that cannot be reliably reproduced has little value' (Klebanov *et al.*, 2007). However, currently, both the concepts and metrics of reproducibility of DEG selection are often loosely defined, using intuitive terms such as consistency (Ein-Dor *et al.*, 2006), concordance (Shi *et al.*, 2006), agreement (Shi *et al.*, 2006), stability (Ein-Dor *et al.*, 2006; Qiu *et al.*, 2006), commonality (Klebanov *et al.*, 2007) or rediscovery rate (Xu and Li, 2003). As mentioned in Section 3, the most frequently used POG metric depends on the lengths of gene lists (Chen *et al.*, 2007; Shi *et al.*, 2005) and cannot be used to compare gene lists with different lengths. Here in this article, we just use POG to demonstrate that apparently low reproducibility can be produced from technically reliable measurements, without concluding how high the reproducibility is or which list is more stable. To provide a better statistical basis, as described in Section 2, we may calculate the probability P of a POG score observed by random chance. Excluding the cases of no common genes, the low POGs in Figure 1E from the model study are actually highly statistically significant, with P-values ranging from 10^{-12} to 10^{-11}. As for the studies by MAQC, the reported lower POG on the National Cancer Institute (NCI) platform (Shi *et al.*, 2006) does not necessarily mean that its measurement quality is lower than the others, because the DEG lists produced by this platform are much shorter. Actually, all the POGs in the MAQC data from different platforms, including the NCI platform, achieve very small P-values ranging from 10^{-11} to 10^{-13}. Alternatively, adjusting POG metric by list lengths could be considered but it would still be difficult to interpret the magnitude of scores corrected by different chance overlaps. Till now, no single metric to properly evaluate the reproducibility of discovery lists can be universally recommended. We note that the reproducibility of microarray studies can be evaluated at raw intensity or log ratio level by using metrics such as coefficient variance (Shi *et al.*, 2006), Pearson correlation coefficient (Guo *et al.*, 2006) and intra-class correlation coefficient (Dobbin *et al.*, 2005). However, high reproducibility at the measurement level does not guarantee high reproducibility of a consecutive analysis since statistical decision-making procedures do matter.

On the other hand, as a sign that two microarray studies have detected a same result for a disease, it is not necessary that the DEG lists themselves are consistent (Subramanian *et al.*, 2005). For example, for a same experiment, although gene lists generated by different statistical methods can be strikingly different, they could be rather consistent according to the functional modules they overrepresented (Guo *et al.*, 2006; Hosack *et al.*, 2003; Zhu *et al.*, 2007). Recently, we also showed that the DEG lists with very different lengths detected under varied statistical thresholds and from different studies could be functionally consistent according to their semantic similarity. For example, for the two prostate cancer datasets produced by different platforms, although the DEG lists selected using SAM at 10% FDR control level had only ∼20% overlaps, their semantic similarity was still high and statistically significant ($P < 0.05$) (Yang *et al*., 2008). Thus, some functional aspects of the expression changes in a disease could be captured by only a fraction of DEGs. Similarly in large-scale screens for cancer mutations, the inconsistent candidate lists across studies tend to be functionally consistent (Chen *et al*., 2007). Besides cancer heterogeneity, the inconsistency of cancer genes detected from current mutation screens might also be a statistical outcome of using inadequate samples.

In general, low apparent irreproducibility of discoveries (e.g. disease markers) is a common problem in many other high-throughput postgenomic areas such as proteomics (Ransohoff, 2005b) and metabolomics (Broadhurst and Kell, 2006). Biologically, a complex disease is often characterized with many coordinated molecular changes (e.g. gene or protein expressions) and their statistical rankings also fluctuate, which alone can result in low discovery consistency. Specifically in microarray data, studying correlated differential expressions in a continuous spectrum might be a reasonable attempt (Klebanov *et al*., 2006).

Notably, most high-throughput technologies can generate huge data from a few patients, which might introduce high-dimensional or overfitting problem in data anayses (Broadhurst and Kell, 2006; Frantz, 2005; Ransohoff, 2004). However, the high-dimensional problem might not be so serious as it looks like, because molecular changes are often correlated under a disease condition (Guo *et al*., 2007; Klebanov *et al*., 2006; Subramanian *et al*., 2005). Only a few independent (non-redundant) components may exist behind the huge data (Guo *et al*., 2005; Xu *et al*., 2006). It is true that using small samples can hinder the discovery of important disease markers (Marshall, 2004), especially when we improperly treat each molecular change as an independent event. However, whether a study is 'small' depends on the correlation structure of the data and the type of the discoveries we are searching for, rather than only on the naive number of the samples. Finally, we note that the reproducibility problem is only one of the fundamental issues in validation of high-throughput discoveries. Being equally important, systematic biases in experimental designs and other problems remain to be resolved in high-throughput systems biology studies (Ransohoff, 2004, 2005a).

ACKNOWLEDGEMENTS

Funding: This work was supported in part by the National Natural Science Foundation of China (Grant Nos. 30170515, 30370388, 30571034).

Conflict of Interest: none declared.

REFERENCES

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B Met.*, **57**, 289–300.

Broadhurst,D.I. and Kell,D.B. (2006) Statistical strategies for avoiding false discoveries in metabolomics and related experiments. *Metabolomics*, **2**, 171–196.

Chen,J. *et al*. (2007) Reproducibility of microarray data: a further analysis of microarray quality control (MAQC) data. *BMC Bioinformatics*, **8**, 412.

Chen,X. *et al*. (2002) Gene expression patterns in human liver cancers. *Mol. Biol. Cell*, **13**, 1929–1939.

Diehn,M. *et al*. (2003) SOURCE: a unified genomic resource of functional annotations, ontologies, and gene expression data. *Nucleic Acids Res.*, **31**, 219–223.

Do,J.H. and Choi,D.K. (2006) Normalization of microarray data: single-labeled and dual-labeled arrays. *Mol. Cells*, **22**, 254–261.

Dobbin,K.K. *et al*. (2005) Interlaboratory comparability study of cancer gene expression analysis using oligonucleotide microarrays. *Clin. Cancer Res.*, **11**, 565–572.

Ein-Dor,L. *et al*. (2005) Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics*, **21**, 171–178.

Ein-Dor,L. *et al*. (2006) Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proc. Natl Acad. Sci. USA*, **103**, 5923–5928.

Frantz,S. (2005) An array of problems. *Nat. Rev. Drug Discov.*, **4**, 362–363.

Gautier,L. *et al*. (2004) affy–analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*, **20**, 307–315.

Guo,L. *et al*. (2006) Rat toxicogenomic study reveals analytical consistency across microarray platforms. *Nat. Biotechnol.*, **24**, 1162–1169.

Guo,Z. *et al*. (2007) Edge-based scoring and searching method for identifying condition-responsive protein-protein interaction sub-network. *Bioinformatics*, **23**, 2121–2128.

Guo,Z. *et al*. (2005) Towards precise classification of cancers based on robust gene functional expression profiles. *BMC Bioinformatics*, **6**, 58.

Hakes,L. *et al*. (2008) Protein-protein interaction networks and biology-what's the connection? *Nat. Biotechnol.*, **26**, 69–72.

Hosack,D.A. *et al*. (2003) Identifying biological themes within lists of genes with EASE. *Genome Biol.*, **4**, R70.

Irizarry,R.A. *et al*. (2005) Multiple-laboratory comparison of microarray platforms. *Nat. Methods*, **2**, 345–350.

Klebanov,L. *et al*. (2006) A new type of stochastic dependence revealed in gene expression data. *Stat. Appl. Genet. Mol. Biol.*, **5**, Article7.

Klebanov,L. *et al*. (2007) Statistical methods and microarray data. *Nat. Biotechnol.*, **25**, 25–26; author reply 26–27.

Klebanov,L. and Yakovlev,A. (2007) How high is the level of technical noise in microarray data? *Biol. Direct*, **2**, 9.

Lander,E.S. (1999) Array of hope. *Nat. Genet.*, **21**, 3–4.

Lapointe,J. *et al*. (2004) Gene expression profiling identifies clinically relevant subtypes of prostate cancer. *Proc. Natl Acad. Sci. USA*, **101**, 811–816.

Li,X. *et al*. (2004) Gene mining: a novel and powerful ensemble decision approach to hunting for disease genes using microarray expression profiling. *Nucleic Acids Res.*, **32**, 2685–2694.

Marshall,E. (2004) Getting the noise out of gene arrays. *Science*, **306**, 630–631.

Miklos,G.L. and Maleszka,R. (2004) Microarray reality checks in the context of a complex disease. *Nat. Biotechnol.*, **22**, 615–621.

Pavlidis,P. *et al*. (2003) The effect of replication on gene expression microarray experiments. *Bioinformatics*, **19**, 1620–1627.

Pawitan,Y. *et al*. (2005a) False discovery rate, sensitivity and sample size for microarray studies. *Bioinformatics*, **21**, 3017–3024.

Pawitan,Y. *et al*. (2005b) Bias in the estimation of false discovery rate in microarray studies. *Bioinformatics*, **21**, 3865–3872.

Perelman,E. *et al*. (2007) Detecting differential expression in microarray data: comparison of optimal procedures. *BMC Bioinformatics*, **8**, 28.

Qiu,X. *et al*. (2006) Assessing stability of gene selection in microarray data analysis. *BMC Bioinformatics*, **7**, 50.

Quackenbush,J. (2002) Microarray data normalization and transformation. *Nat. Genet.*, **32** (Suppl.), 496–501.

Ransohoff,D.F. (2004) Rules of evidence for cancer molecular-marker discovery and validation. *Nat. Rev. Cancer*, **4**, 309–314.

Ransohoff,D.F. (2005a) Bias as a threat to the validity of cancer molecular-marker research. *Nat. Rev. Cancer*, **5**, 142–149.

Ransohoff,D.F. (2005b) Lessons from controversy: ovarian cancer screening and serum proteomics. *J. Natl Cancer Inst.*, **97**, 315–319.

Rhodes,D.R. *et al*. (2007) Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles. *Neoplasia*, **9**, 166–180.

Shi,L. *et al*. (2005) Cross-platform comparability of microarray technology: intra-platform consistency and appropriate data analysis procedures are essential. *BMC Bioinformatics*, **6** (Suppl. 2), S12.

Shi,L. *et al*. (2006) The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat. Biotechnol.*, **24**, 1151–1161.

Subramanian,A. *et al*. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA*, **102**, 15545–15550.

Tan,P.K. *et al*. (2003) Evaluation of gene expression measurements from commercial microarray platforms. *Nucleic Acids Res.*, **31**, 5676–5684.

Tong,W. *et al*. (2006) Evaluation of external RNA controls for the assessment of microarray performance. *Nat. Biotechnol.*, **24**, 1132–1139.

Troyanskaya,O. *et al*. (2001) Missing value estimation methods for DNA microarrays. *Bioinformatics*, **17**, 520–525.

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

Xie,Y. *et al*. (2005) A note on using permutation-based false discovery rate estimates to compare different analysis methods for microarray data. *Bioinformatics*, **21**, 4280–4288.

Xu,J.Z. *et al*. (2006) Peeling off the hidden genetic heterogeneities of cancers based on disease-relevant functional modules. *Mol. Med.*, **12**, 25–33.

Xu,R. and Li,X. (2003) A comparison of parametric versus permutation methods with applications to general and temporal microarray gene expression data. *Bioinformatics*, **19**, 1284–1289.

Yang,D. *et al*. (2008) Gaining confidence in biological interpretation of the microarray data: the functional consistence of the significant GO categories. *Bioinformatics*, **24**, 265–271.

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.

Yeoh,E.J. *et al*. (2002) Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell*, **1**, 133–143.

Zhang,S. (2007) A comprehensive evaluation of SAM, the SAM R-package and a simple modification to improve its performance. *BMC Bioinformatics*, **8**, 230.

Zhu,J. *et al*. (2007) GO-2D: identifying 2-dimensional cellular-localized functional modules in Gene Ontology. *BMC Genomics*, **8**, 30.