Genomics 103 (2014) 21–30

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Avoiding pitfalls in gene (co)expression meta-analysis

Gabriel Östlund [a,b], Erik L.L. Sonnhammer [a,b,c,*]

[a] Stockholm Bioinformatics Centre, Science for Life Laboratory, Box 1031, SE-17121 Solna, Sweden
[b] Department of Biochemistry and Biophysics, Stockholm University, Sweden
[c] Swedish eScience Research Center, Sweden

ARTICLE INFO

Article history:
Received 9 May 2013
Accepted 22 October 2013
Available online 30 October 2013

Keywords:
Differential expression
Differential coexpression
Cancer gene expression
Microarray data processing

ABSTRACT

Differential gene expression analysis between healthy and diseased groups is a widely used approach to understand the molecular underpinnings of disease. A wide variety of experimental and bioinformatics techniques are available for this type of analysis, yet their impact on the reliability of the results has not been systematically studied.
We performed a large scale comparative analysis of clinical expression data, using several background corrections and differential expression metrics. The agreement between studies was analyzed for study pairs of same cancer type, of different cancer types, and between cancer and non-cancer studies. We also replicated the analysis using differential *coexpression*.
We found that agreement of differential expression is primarily dictated by the microarray platform, while differential coexpression requires large sample sizes. Two studies using different differential expression metrics may show no agreement, even if they agree strongly using the same metric. Our analysis provides practical recommendations for gene (co)expression analysis.

1. Introduction

The use of microarray or RNA-seq technologies makes it possible to measure the expression level of all human transcripts. By measuring expression for patient cohorts and comparing to healthy controls one can identify genes that show an altered level of expression — differential expression between healthy and diseased. This approach has been used to identify genes involved in many diseases, as well as to create prognostic gene signatures capable of classifying patients into high/low risk groups.

Despite the promising potential and results with high statistical significance, doubts have been raised about their reliability. Surveys of published top lists of differentially expressed genes have shown a low degree of overlap between microarray expression studies for ovarian cancer [11] as well as between microarray expression studies and non-microarray-based clinical and biological data for schizophrenias [20]. One might however question the statistical backing of the reported overlaps being low. While numerically low, they are often significantly larger than would be expected by chance.

Also the reliability of prognostic gene signatures has been questioned. Michiels et al. [19] studied the stability of gene signatures for disease classification by examining the consistency of signatures generated from multiple randomly resampled sets and found that there was a large dependence on the set of patients of the study. They suggested that larger sample sizes are needed for accurate results, the need for which has been corroborated by other studies [7,28].

If variation of expression across samples for a gene is large, detecting differential expression becomes difficult due to overlapping expression between healthy and diseased samples. It could also have the effect that studies with biased sampling would be very inconsistent regarding such a gene. Further, a change to a regulatory gene might affect the regulated genes while the expression of the regulatory gene remains non-differential [15]. However, important genes without detectable differential expression can potentially be discovered by examining the coexpression of gene pairs and the change of coexpression between healthy and diseased individuals (for review see [6]). Differential coexpression (*DC*) will likely be less error-prone for genes with a large expression variation or biased sampling, and may therefore be more consistent across studies than differential expression.

While differences in experimental procedures, biological differences and differences in composition of tissue samples or groups of patients can affect the results, there is also evidence that procedures chosen for data analysis and calculation of metrics can have a profound effect. Studies using homogeneous data with controlled RNA abundances have been used to compare different preprocessing methods in combination with different metrics of differential expression (*DE*) [4,17,31]. Even in controlled conditions that lack biological variation, a strong dependency on data processing methods was observed. The optimal combination of methods however varied between data sets.

Cross-platform consistency, i.e. the reproducibility between different microarray chips or other techniques, is an important measure of reliability. This has been examined for platforms from different microarray providers in mouse [10] and human [2,26], showing a far from perfect consistency of differential expression across platforms. In both studies fold change (FC) performed better than significance analysis of

* Corresponding author.
E-mail address: Erik.Sonnhammer@scilifelab.se (E.L.L. Sonnhammer).

0888-7543/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ygeno.2013.10.006

microarrays (SAM) [27] and *t*-test. While consistency between different platforms can be high, consistency across different Affymetrix microarrays has been found to be dependent on the similarity of probe sets and the expression level of the transcript [22]. For example, Hwang et al. [16] showed that Affymetrix microarray results cluster according to platform rather than biological sample. They also noted that expression patterns were often similar for different platforms, but at different absolute levels, suggesting that gene coexpression would be less affected. The extent to which coexpression is sensitive to data processing, sample sizes and platform has however not been tested systematically.

Evaluating consistency in a statistically sound manner is of utmost importance. The expected result of comparing two top lists is highly dependent on the chosen number of genes to include. An alternative approach, which avoids the arbitrary choice of how many top DE genes to consider, is to use the entire list of DE-ranked genes and calculate a global overlap between two such lists. [33] provide a good overview of different methods for sound comparisons of ranked lists, which are all incorporated in the GeneSelector R package.

If several effective methods exist for detecting DE, why is the consistency of results not higher? The problem may be related to e.g. the poor generalizability of prognostic gene signatures, which is likely caused by putting too much focus on achieving high accuracy while issues of consistency are ignored [12]. This is exemplified by e.g. the work of Zervakis et al. [30] who showed the importance of complementing cross validation with independent validation in order not to overestimate accuracy. That two independent sets of differentially expressed genes, representing alternative gene signatures for the same disease, exhibit a low consistency between studies could merely be an indication that multiple valid alternative sets exist. For cancer, good outcome prognosis is easily obtained even for gene signatures based on random gene sets [21]. This indicates that because cancer exhibits vast phenotypical changes, with rewiring of the cellular machinery and uncontrolled growth, statistically significant results from differential (co)expression analysis can readily be obtained. In general one would expect to see large changes of gene expression for cancer compared to normal.

However, what if the expression changes in one cancer are observed in multiple cancer types, or even in non-cancer diseases? If that were the case, can we be certain that a significant differential expression, or coexpression, between normal and diseased states, is informative about our disease of interest? Or will the same genes/gene pairs be identified as being differentially expressed/coexpressed in related diseases, or even in diseases in general? In order to test this hypothesis we extend previous work by performing a systematic large scale analysis of clinical expression data sets, taking into account both microarray platform generation, sample size, background correction, as well as metric. We do this with a focus on cancer on account of the heterogeneity of the disease and the high availability of data, and perform the analysis both for DE and DC (see Fig. 1 for an overview of the workflow and concepts).

2. Methods

2.1. Literature mining

GEO [3] was searched for cancer studies including both normal and diseased samples using Affymetrix chips with at least 5 samples per sample group. The initial list was restricted to studies with at least 3 studies per cancer type, same tissue for disease/normal, and with all samples analyzed using the same Affymetrix platform. Five studies were removed in this step. Annotations were retrieved using GEOmetadb [32] and were manually annotated for diseased/normal states to obtain roughly coherent sample groups across studies. Where possible, samples from whole tissues were chosen over samples from single cell types, and samples from cell lines were always excluded. Samples were annotated to maximize the number of samples of matching tissue. This resulted in a set of 4 studies of lung cancer, 7 of colon cancer and 11 of breast cancer. For the tissues of these 3 cancer types, GEO was searched for non-cancer studies but otherwise with the same criteria as before. For breast no study could be found with normal/disease groups; instead a study of parous/nulliparous women (having given birth/not given birth) was selected. The 3 non-cancer studies were processed in the same manner as the cancer studies. For details of the included studies, see Table S1.

2.2. Data retrieval and processing

Data retrieval and processing was carried out in an automated fashion using R. Briefly, raw.cel files were retrieved using GEOquery [5] and background corrected with both MAS5 and GCRMA using the affy [8] package with default settings. Probes were mapped to entrezid using annotate [9], when multiple probes mapped to the same gene they were averaged.

3. Metrics

3.1. DE

t-Test and SAM were calculated on log2 transformed expression values while FC was calculated on non-transformed expression values.

3.1.1. Fold change (FC)

For each gene the average was calculated for both normal and diseased states, initial fold change was calculated by dividing the diseased state average with the normal state average. Prior to ranking, initial fold changes below 1 were transformed by dividing 1 with these fold changes; this was in order to treat downregulation equally to upregulation.

3.1.2. t-Test

A two sided *t*-test was performed using the t.test function and subsequently BH (Benjamini–Hochberg) corrected using padjust.

3.1.3. SAM

SAM was performed using the sam function from siggenes [25] with default settings.

3.2. DC

3.2.1. dSpearman (dS)

Spearman correlations were calculated for each gene pair, separately for the diseased and normal states. dSpearman was subsequently calculated as the absolute of the difference in correlation between the states.

4. Measuring the agreement between studies with *overlapScore*

In previous work, usually the set of shared differentially (co) expressed genes has been used to assess the agreement between studies. However, this approach depends on the choice of arbitrary cutoffs. An approach that does not suffer from this drawback is to instead measure the intersections between entire gene lists, ranked according to differential (co)expression, for all possible depths [29]. This raw overlap score is calculated by taking the cumulative sum of intersections over all depths. Because the intersection increases towards the end, the contribution of each depth is weighted by a decay in order to prioritize overlap at the top of the list.

It is recommended to determine the optimal decay through resampling, but due to the large number of comparisons, this was not feasible. Instead we opted to compare different decays by simulation. Random lists were generated from a template by either adding general Gaussian noise, adding Gaussian noise proportional to rank or adding both general and proportional noise. A linear penalty was found to give robust scores with good signal to noise ratio whereas a quadratic penalty gave worse robustness/signal to noise and an exponential penalty required list length specific tuning in order to give comparable quality to a linear penalty (data not shown).

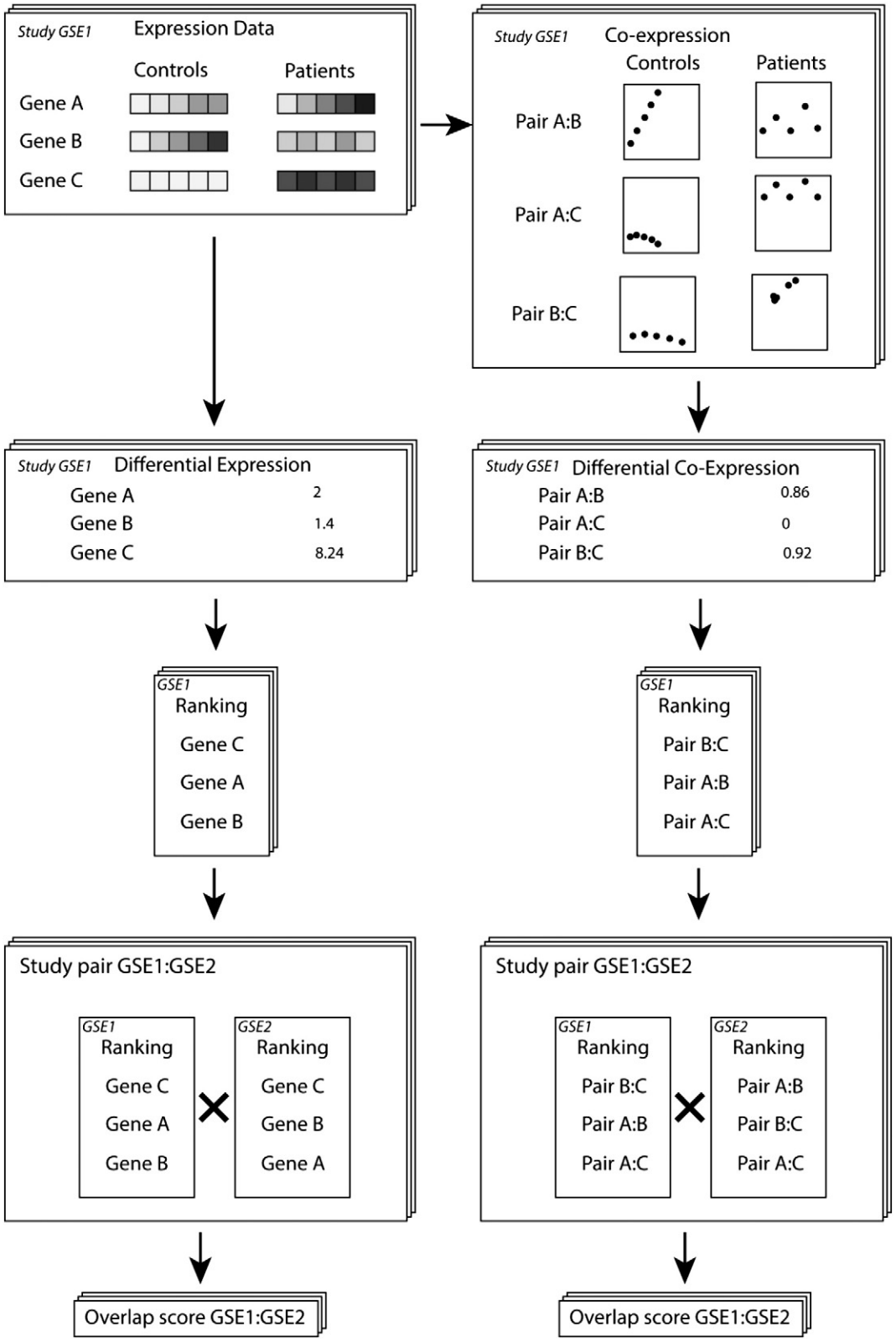


Fig. 1. Workflow and concepts. Starting from a study containing expression data for both patients and controls, differential expression is calculated and used to rank genes. The ranking of differential expression for pairs of studies is compared, resulting in an overlap score measuring the agreement. The same source data is used to calculate the correlation between all pairs of genes, separately for patients and controls. The difference in correlation for the two groups is calculated and used to rank gene pairs. Same as for differential expression, the ranks are then used to compare pairs of studies in order to gain a measure of their agreement.

The overlap score was calculated for pairs of studies using an optimized version of getStabilityOverlap from geneSelector (Boulesteix and Slawski, 2009). However, the expected overlap score depends on the list length, and it does not by itself provide a significance estimate. Getting a high intersection at an early point is more likely with a shorter list, hence one would expect a higher score for shorter lists.

To resolve this, we model the expected score for a comparison between two random lists using the hypergeometric distribution, and use this to normalize for list length. Calculating the probability of observing an overlap score equal to or higher than a given limit is computationally infeasible as the number of possible random lists is equal to the factorial of the list length. It is however possible to calculate the probability of observing an intersection equal to or greater than a given limit, given a list length and depth. Similarly, given a p-value, list length and depth, it is possible to calculate a limit of intersection (***I***) for which it is less likely than p to observe an equal or higher intersection.

$$P(\text{intersection} >= \boldsymbol{I} | \text{list length}, \quad \text{depth}) < p$$
$$\boldsymbol{I} = f(\text{list length}, \quad \text{depth}, \quad p).$$

Such a limit of intersection can be calculated for all depths in a fairly efficient manner (see Supplementary material for details) for any given p-value and list length. We then calculate the overlap score for ***I*** across all depths to associate each p-value with an equal or higher overlap score. This enables determining significant deviance from the expected as the p-value used to generate ***I*** is a conservative approximation for the probability of observing an overlap score equal or higher than the overlap score for ***I***.

In order to obtain a boundary above for which observed overlap scores are significant, ***I*** was calculated for all pairs of studies using cIterHyper (Supplementary material) with a multiple testing corrected p-value corresponding to a family wise error rate (FWER) below 0.01, to obtain ***I***. The multiple testing corrected p-value was obtained by dividing the desired FWER (0.01) with the number of pairwise comparisons (2 background corrections $*$ 4 metrics $*$ 300 pairs of studies), see Fig. 2 for details. Significance boundaries were then obtained by calculating the overlap score for the generated ***I*** for each pair.

As mentioned, overlap score depends on list length. For example, there is an approximately 5% difference in the overlap score at FWER < 0.01, for the smallest intersection of DE studies compared with the largest. This effect was adjusted for by normalizing the overlap score for all pairs to a common size of intersection (list length). Briefly, overlap scores were generated for ***I*** based on p-values ranging between −5 and 5 standard deviations of the normal distribution, separately for each intersection size. The overlap scores and corresponding p-values for each list length (the intersection of two gene lists) were then used to fit beta distributions. These beta distributions were first used to convert the overlap scores of study pairs to p-values, using the beta distribution for the intersection size of the two studies in each pair. Then the p-value of each study pair was converted back to an *overlapScore*, using the beta distribution for the normalization intersection size, which was set to the number of genes in the GPL570 platform, i.e. 19701. See Suppl. Fig. 1 for an illustration of this process.

4.1. Significance testing

Significance testing for differences between pair types was performed using wilcox_test from the coin R package [13].

5. Results

We address a number of important questions in differential expression (DE) and differential coexpression (DC) analysis. For instance, how consistent are results across studies of the same or different diseases? How does the choice of experimental platform affect the results? How does the choice of data processing and DE measure affect the results? DC can be seen as a high-dimensional complement to DE; what strengths and weaknesses does it have?

To answer such questions, we collected 25 gene expression studies with normal and cancer states for lung, breast and colon. As the choice of background correction and metric can have an impact, we processed the data using both MAS5 and GCRMA before proceeding to calculating DE by fold change, *t*-test, and SAM p-values, as well as DC by differential Spearman correlation. The studies were then compared using the ranks of metrics, rather than raw metrics, of genes and pairs. This was done for each metric separately for both DE and DC. While using ranks may result in a loss of precision it is more robust to systematic differences between the studies. See Fig. 2 for an overview of the analysis pipeline. This resulted in a total of 1800 pairs of studies for DE and 600 pairs for DC.

Choosing a set (small) cutoff of the top K genes/pairs for determining overlap is inherently volatile due to the risk of obtaining a high, or low, overlap purely by chance. It would likely introduce bias in that any cutoff would be more suitable for some pairs of studies than for others. Any results or conclusions drawn from the comparison would thus be highly dependent on the chosen cutoff. We instead opted to use the overlap score which includes all data contained in both studies, and thus does not require choosing an arbitrary cutoff. It is conceptually quite simple; it is a weighted average of intersection over all depths, with decreasing weights for increasing depth. This means that an overlap of highly differentially (co)expressed genes/pairs contributes more to the score than an overlap of lowly (co)expressed genes/pairs. The way it was used in this study corresponds to the average intersection over all list depths and thus allows for easy interpretation. Since the likelihood of observing a given overlap score is dependent on list length, all overlap scores were normalized by list length to *overlapScores* for a length corresponding to the number of genes in the GPL570 platform (Suppl. Fig. 1).

To determine significant agreement between pairs of studies, the expected overlap score for pairs of random lists can be modeled using the hypergeometric distribution. However, this is not fully appropriate for gene lists. The hypergeometric distribution assumes independence between genes and since genes are often functionally dependent of each other, one would not expect the model to be fully accurate. Furthermore, some genes (such as stress response genes), might be more prone to be altered compared to others (such as housekeeping genes). Also, experimental procedures, measurements and data processing could also introduce biases. In an extreme case one could picture a method that gives genes a score according to their alphabetical ordering would show a perfect agreement for any two studies compared. Because of this it is important to contrast the results of the comparison against cases where one would not expect to find any/as high agreement, as a baseline.

We analyzed the agreement between studies in three different categories of pairs: (1) pairs of same cancer type (*same cancer*), (2) pairs of different cancer type (*different cancer*), and (3) pairs between cancer and non-cancer studies (*cancer non-cancer*). In principle one would expect same cancer types to overlap more than different cancer types, and both of these to overlap more than non-cancers. If this is not observed, it would raise questions about the reliability and biological meaningfulness of the data and/or methodology. If the agreement between cancer types is as large as within the same type, one would not expect to be able to draw conclusions regarding which genes are important for a specific cancer type, but only about the importance for cancer in general. If the aim is to construct prognostic gene signatures, one would not expect to be able to discern between different cancer types. Similarly, if agreement between cancer studies and non-cancer studies is significantly high, it would be uncertain if conclusions based on DE or DC would be applicable to cancer rather than (related) diseases in general.

With the resulting collection of pairwise comparisons of studies (Table S2), we analyzed the data in an attempt to determine whether the different processing procedures would give consistent results for studies of the same cancer type, and if it would be more consistent between cancers of the same type than between different types or between cancer and non-cancer.

5.1. DE agreement depends mainly on microarray platform

Clustering pairs of studies based on DE showed a strong preference to group together studies that used the same Affymetrix platform

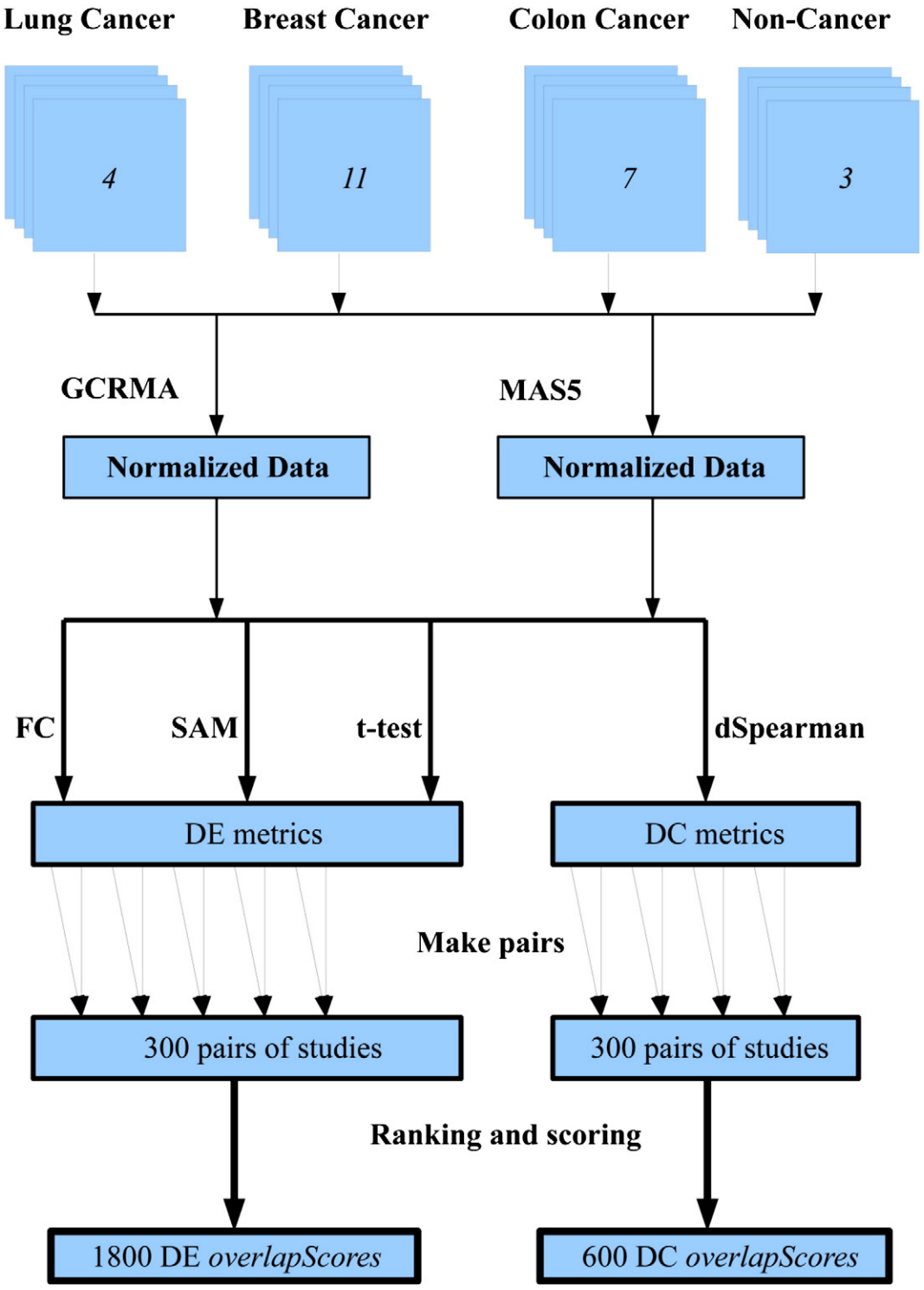
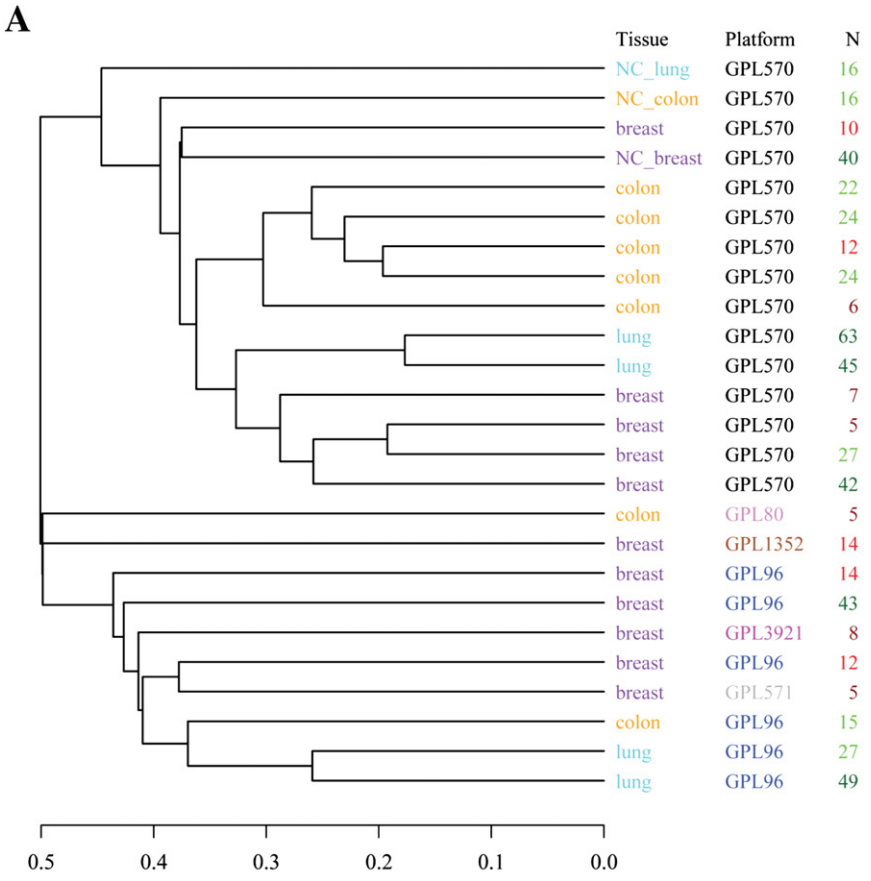
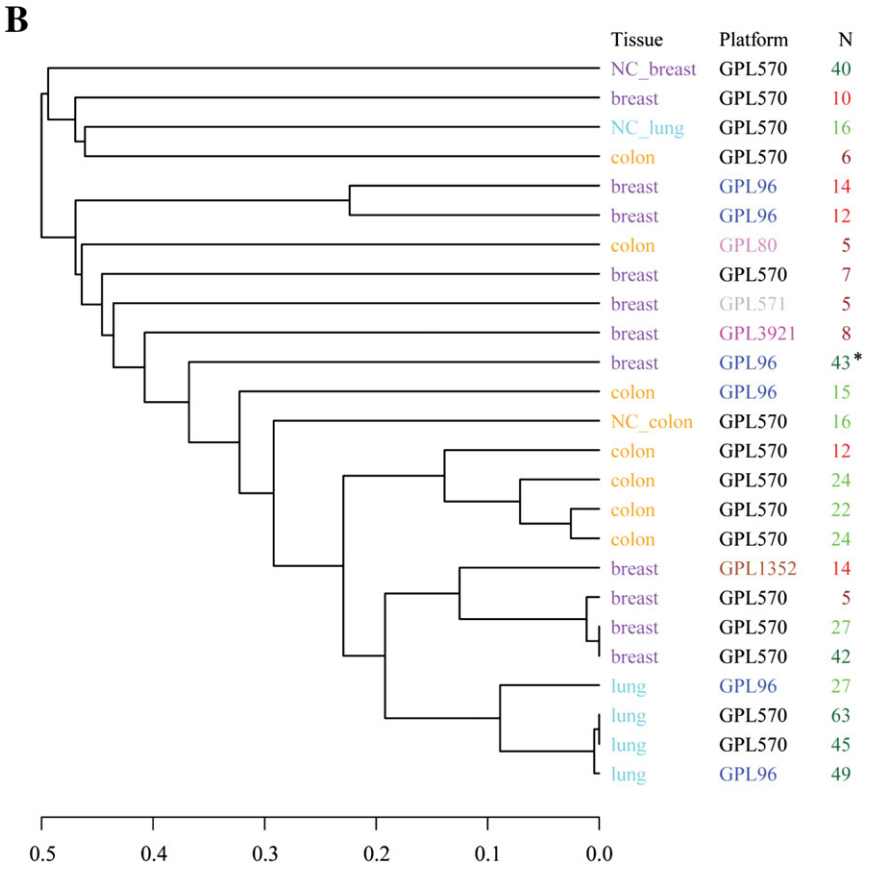


Fig. 2. Analysis pipeline for making pairwise comparisons (overlap scores) between clinical gene expression data sets using a range of normalizations and metrics. Briefly, 25 clinical studies were normalized with both GCRMA and MAS5. Differential expression (DE) and differential coexpression (DC) were calculated using three DE metrics (FC, SAM, and *t*-test) and one DC metric (dS). This resulted in 1800 ($2 * 3 * (25 * 24) / 2$) pairs of studies for DE and 600 ($2 * (25 * 24) / 2$) for DC.

(Fig. 3A). For instance, all studies using GPL570 are in one group which includes all non-cancer studies, whereas studies of the same cancer type using other platforms fall outside of the group. Within a cluster of studies using the same platform however, the studies are correctly clustered according to cancer type.

By plotting the *overlapScore* distributions for all study pairs, separated into same platform (Fig. 4A) and different platform (Fig. 4B) it is clear that hardly any (5%) of the pairs from different platforms obtain a significant *overlapScore*. For study pairs using the same platform the *overlapScore* is generally significant and correlated with the minimum sample size (in either study), but this trend is not observed for pairs using different platforms. In light of this, further analysis for DE was performed only with pairs where both studies use the same platform. The effect of the average sample size was also examined but was found to have a lower impact than the minimum sample size (data not shown).


A
Tissue Platform N
NC_lung GPL570 16
NC_colon GPL570 16
breast GPL570 10
NC_breast GPL570 40
colon GPL570 22
colon GPL570 24
colon GPL570 12
colon GPL570 24
colon GPL570 6
lung GPL570 63
lung GPL570 45
breast GPL570 7
breast GPL570 5
breast GPL570 27
breast GPL570 42
colon GPL80 5
breast GPL1352 14
breast GPL96 14
breast GPL96 43
breast GPL3921 8
breast GPL96 12
breast GPL571 5
colon GPL96 15
lung GPL96 27
lung GPL96 49
0.5 0.4 0.3 0.2 0.1 0.0


B
Tissue Platform N
NC_breast GPL570 40
breast GPL570 10
NC_lung GPL570 16
colon GPL570 6
breast GPL96 14
breast GPL96 12
colon GPL80 5
breast GPL570 7
breast GPL571 5
breast GPL3921 8
breast GPL96 43*
colon GPL96 15
NC_colon GPL570 16
colon GPL570 12
colon GPL570 24
colon GPL570 22
colon GPL570 24
breast GPL1352 14
breast GPL570 5
breast GPL570 27
breast GPL570 42
lung GPL96 27
lung GPL570 63
lung GPL570 45
lung GPL96 49
0.5 0.4 0.3 0.2 0.1 0.0

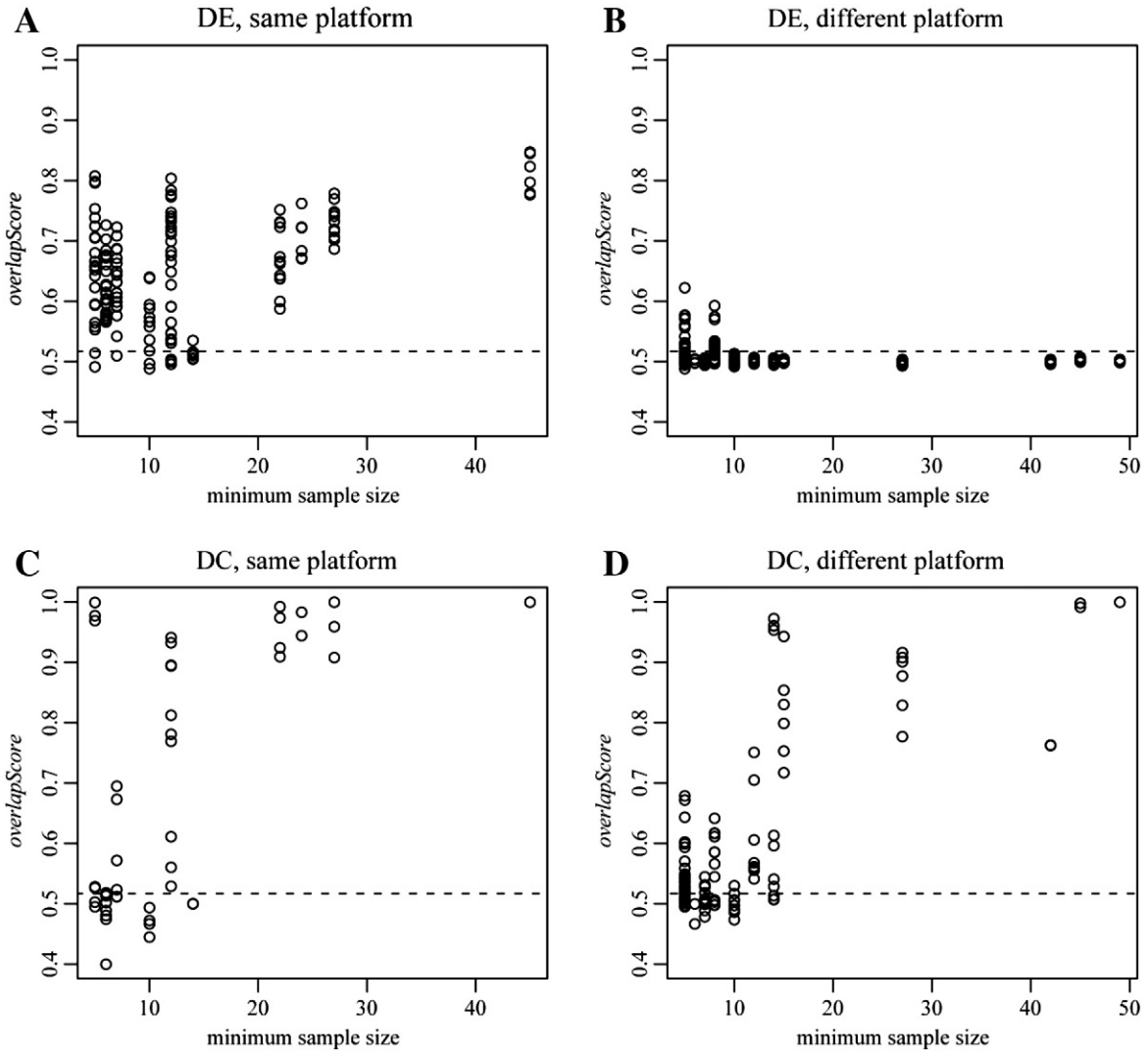


Fig. 4. Microarray platform and sample size dependence when comparing two studies. Distributions of *overlapScores* between studies using the same microarray platform (A for differential expression (DE); C for differential coexpression (DC)) and pairs using different platform (B for DE; D for DC) are shown and all categories of pairs are included. Each * *overlapScore* is plotted at the minimum sample size of any sample group in the pair. The horizontal dashed line denotes the limit of significance (FWER < 0.01) based on the hypergeometric distribution. For DE, hardly any study pairs using different platforms reach a significant *overlapScore* (B), but for DC this platform dependence is not observed. Instead, DC has a more pronounced dependence on the sample size.

5.2. DC agreement depends on sample size but not on microarray platform

As shown in Fig. 3B, DC does not suffer from the strong platform dependence as DE. Studies of the same cancer type using different platforms cluster together, where they do not with DE. However, studies with few samples tend to be clustered near the root in biologically not meaningful ways. This is the Achilles heel of DC — if the number of samples is too low, the Spearman correlation becomes unreliable. Surprisingly, one breast cancer study GSE15852 with 43 or more samples in both groups did not cluster with other high-sample size breast cancers (marked with a star in Fig. 3B). The study was conducted by Pau Ni et al. [23] and was prompted by a prior study by Hsiao et al. [14], which showed that cancer predisposition varies between western and Asian women. Like Pau Ni et al. [23] we see no notable difference for DE from western studies that use the same microarray platform. However, the divergent placement when using DC indicates that the difference in predisposition might be linked to changes in regulation.

In Figs. 4C and D, no clear difference can be seen between the *overlapScore* distributions for pairs using the same or different platforms. There is however a more pronounced tendency than for DE to get a lower *overlapScore* with smaller number of samples, although some exceptions exist. Pairs between GSE21422 and two other breast cancer studies have an *overlapScore* close to 1 despite GSE21422 having only 5 samples for the normal group. Agreement between GSE21422 and other breast cancer studies is generally low though, confirming the impact on reliability from the low number of samples.

Fig. 3. UPGMA (unweighted pair group method with arithmetic mean) trees based on *overlapScores*, showing the relatedness of different clinical gene expression studies measured by differential expression (DE) (A) and differential coexpression (DC) (B). The cancer type (tissue), microarray platform, and minimum number of samples are color coded for readability. *Non-cancer* studies are named NC_*. In the DE tree, the platform is the dominant clustering factor, while for DC the studies are generally correctly clustered according to cancer type, except for studies with few samples. DE was calculated with FC and GCRMA normalization; DC was calculated as difference in Spearman correlation (dS) with MAS5 normalization. Trees constructed for other combinations of metric and normalization show a similar general structure.

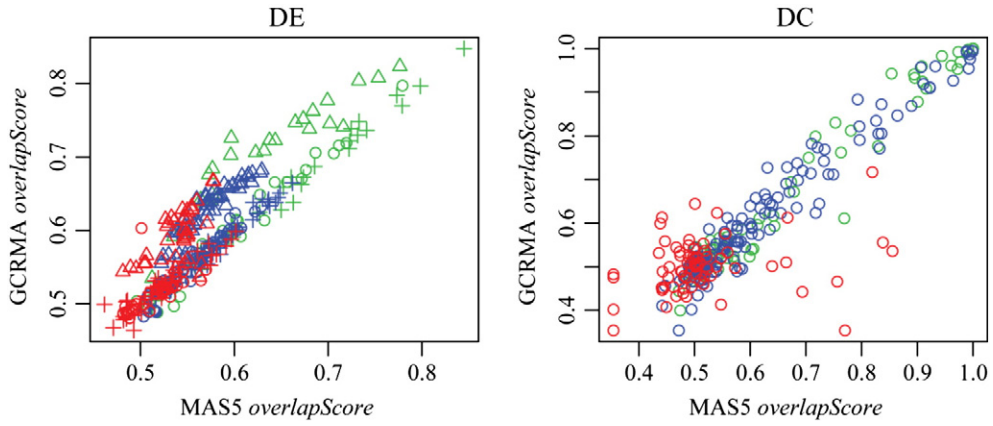


Fig. 5. Scatter plot of *overlapScores* when normalizing with MAS5 versus with GCRMA. The left plot shows differential expression (DE, same platform pairs), using the averaged *overlapScores* for all three DE metrics, and the right plot the differential coexpression (DC, all pairs). In the left plot the DE metrics are denoted by symbols: circle for *t*-test, triangle for FC and cross for SAM. The colors correspond to the three pair categories: green for *same cancer* pairs, blue for *different cancer* pairs, and red for *cancer non-cancer* pairs. There is a strong correlation between the normalizations for both DE and DC, with a larger uncertainty for *cancer non-cancer* pairs for DC. Notably, there is an increase in *overlapScore* when using FC in combination with GCRMA; this is observed for all pair types.

5.3. The choice of background correction has minor impact

Prior studies have shown that the background correction can have a noticeable impact on DE accuracy for a single study [17,28,31]. Does it affect the agreement between studies too? To examine the impact of background correction, *overlapScores* using MAS5 and GCRMA were compared for the three pair types. As seen in Fig. 5, there is a high agreement between the two background corrections. The Pearson correlation for DC is 0.93, and 0.89 for DE when restricted to only same platform pairs but including all three DE metrics. For individual DE metrics the correlation is higher. For FC it is 0.93, with GCRMA giving systematically higher *overlapScores* than MAS5. For SAM and *t*-test the correlation is 0.99 and 0.97 respectively, with no bias towards either background correction. Based on this, subsequent analysis was performed using MAS5 and GCRMA pooled for SAM, *t*-test, and dSpearman, while the two background corrections were generally handled separately for FC.

5.4. Cancer and non-cancer studies agree significantly

We examined the *overlapScores* that each DE metric yields on average for the three categories of pairs, and tested if there were significant differences between the categories. All metrics gave the highest mean *overlapScore* for *same cancer* pairs; lower for *different cancer* pairs; and yet lower for *cancer non-cancer* pairs (Fig. 6). FC corrected with GCRMA resulted in a substantially higher mean *overlapScore* than the other metrics, for all categories. In fact, FC with GCRMA resulted in a higher mean *overlapScore* between *cancer non-cancer* pairs than all the other metrics yielded between *different cancer* pairs. SAM produced a relatively clearer separation of pair types than other metrics, although all metrics showed a significant difference between the pair categories. It is noteworthy that the mean *overlapScore* for *cancer non-cancer* pairs was above the significance (FWER < 0.01) level using FC or SAM, and just below it for *t*-test. If the changes in cancers and the non-cancer diseases were truly independent one would not expect to observe a single *cancer non-cancer* pair with an *overlapScore* above the significance level. However, significant scores for *cancer non-cancer* pairs were observed for all metrics. This indicates that even though a given cancer type has its unique characteristics, basic processes exist that are activated in very different diseases, which produces a gene expression agreement higher than expected by chance. For example, GSE18842 (lung cancer) and GSE30010 (breast non-cancer) when compared using FC with GCRMA normalization has an *overlapScore* of 0.61, p < 1.19e–210. Such general processes may e.g. be connected to inflammation, which can be caused by a multitude of disease. When comparing pair categories the dependence of DC agreement on the number of samples is also clear. As seen in Fig. 6, dS *overlapScores* for pairs with a small sample size are quite low; there is no significant difference between *same-* and *different cancer* pairs, and *cancer non-cancer* pairs generally have *overlapScores* below the significance level. There is a vast increase in *overlapScores* for *same-* and *different cancer* pairs with a higher number of samples. The separation between

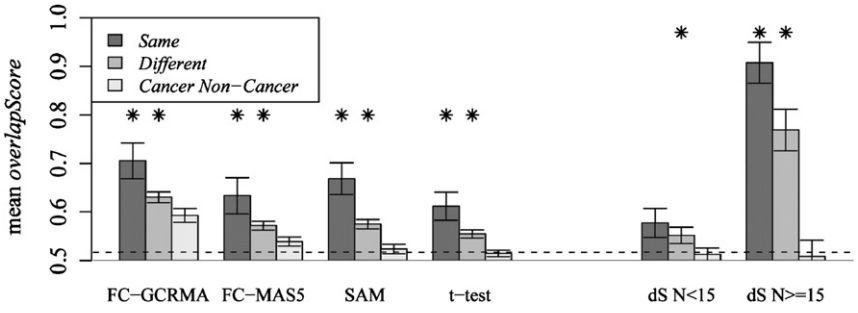


Fig. 6. Mean *overlapScores* for pairs of three categories: (1) *same cancer type*; (2) *different cancer type*; and (3) *cancer non-cancer*. Asterisks denote a significant (p < 0.01) difference (Wilcoxon rank sum test) compared to the category to the right for the same metric, and the error bars correspond to a 99% CI. The horizontal dashed line denotes the limit of significance (FWER < 0.01). There is a significant difference between all categories for all metrics, except for between *same-* and *different cancer* pairs for dS with a low number of samples. SAM shows the most pronounced separation for DE. It is even more pronounced for DC, but only when the number of samples is high. Notably, except for *t*-test and dS, *cancer non-cancer* pairs have a significant mean *overlapScore*, and *different cancer* pairs have a highly significant mean *overlapScore* for all metrics. SAM, *t*-test and dS used the average from GCRMA and MAS5 normalizations.

pair categories is good, and *cancer non-cancer* pairs remain below the significance level. For DE the number of samples had no effect in this analysis (data not shown).

5.5. Using different metrics decreases agreement

In order to study the impact of comparing studies that use different metrics, we selected a subset of breast studies. We then proceeded to compare study pairs in all possible combinations of metrics and background corrections. Surprisingly, as seen in Fig. 7, while pairs where one study uses FC and the other *t*-test reach a fairly high overlap, the agreement is near the limit for significance (FWER < 0.01) when one study uses FC and the other uses SAM. In fact, the *overlapScore* is similar for *cancer non-cancer* pairs using the same metric as for *same cancer* pairs when SAM is used for one study and FC for the other. Using the same background correction compared to using different background corrections had a negligible impact (data not shown).

6. Discussion

We have developed a generalized method to compare ranked gene lists and have applied it to assess agreement between clinical gene expression data. Comparing ranked lists is a very general and common task in many types of research, particularly in functional genomics. This methodology avoids choosing an arbitrary cutoff parameter, which is how such comparisons are normally made. It is a generalized framework and should find useful applications in many other situations when comparing ranked lists. The biological analysis that we have performed gives many new insights into what artifacts that can be expected from choosing among standard techniques for normalization and measuring DE. Among the most striking results are the huge dependence on microarray platform for DE, and that using different DE metrics for two studies can eradicate an agreement that is strong when using the same metric. Both of these problems are non-issues for DC, but this approach requires sufficiently high sample sizes.

While this study was performed only using older Affymetrix microarrays, the results can be contrasted against other studies using newer Affymetrix microarrays or platforms from other providers. In a comparison by Affymetrix [1], the Affymetrix GeneChip Human Genome U133 Plus 2.0 (GPL570) platform had a 0.77 correlation of FC with both the newer human exon and human gene platforms. It was commented that discrepancies in part could be due to the fact that GPL570 does not discriminate between different transcript isoforms. In a study by Robinson and Speed [24] using the same data, the GPL570 platform had an overlap for the top 2000 differentially expressed genes of 65% with the other platforms. As a comparison, the two independent lung cancer studies used here that employ GPL570 have a 0.86 correlation of FC and an overlap greater than 70% for the top 2000 differentially expressed genes for all DE metrics. While this is a single example, it is quite noteworthy that two independent studies using the same platform but based on completely different biological material have a higher agreement than the same platform has with newer platforms based on the exact same biological material. These results can also be contrasted against comparisons between platforms using different technologies, as done by Guo et al. [10]. All of the 5 platforms they tested overlapped the other ones by less than 65% for the top 2000 differentially expressed genes, even without adjusting for the shorter list length in this comparison. In other words, platforms of different technologies appear to be less coherent than different platforms of the same technology/manufacturer.

The fact that different metrics agree poorly even when each metric is informative indicates a lack of comparability. A way to improve this could be to combine different metrics. For example, an ad hoc combination of two metrics by applying a cutoff based on one and ranking by the other has been used [10,26]. A more prudent approach would be to systematically evaluate ways of combining metrics.

This study was performed without taking differences between studies into account, and instead assuming that published studies should be of sufficient quality to provide generalizable results. As has been shown in our analysis, this assumption is sometimes compromised; for example some studies have too few samples in at least one group to be reliable with *t*-test or dSpearman. Another confounding factor is that different studies are performed using different compositions of cancer- and tissue sub types. Correcting for this would be desirable but is often difficult due to poor annotations; it would also have the negative effect to reduce the number of samples.

While expression agreement is high for *same cancer* pairs of studies, agreement is also significant between studies of different cancer types. Generally there is also agreement between cancer and non-cancer studies that is significantly higher than expected by chance. The reason is probably that some genes are more likely to show differential (co)expression, regardless of cancer or not. This is not unexpected as any type of disorder in an organ may trigger the same types of stress responses, such as apoptosis, proliferation, or repair mechanisms. Indeed, based on gene expression analysis, cancer can be seen as a wound that does not heal [18].

What are the implications of such cross-disease commonalities for differential gene expression analysis? Even if significant DE/DC is observed for a disease, one cannot safely conclude that these genes are specifically affected in this disease, rather than coming from general responses. To get disease-specific information, for instance to develop biomarkers for a disease, one would need to compare the obtained DE p-values to an empirical background. This way it might be possible to identify e.g. genes that are differentially expressed to a significant degree in breast cancer when compared to other cancer types. Extending this reasoning, it would be prudent to construct a map of DE/DC across studies and diseases to identify genes/gene pairs that unspecifically show DE/DC. These could either be removed from consideration, or their DE/DC could

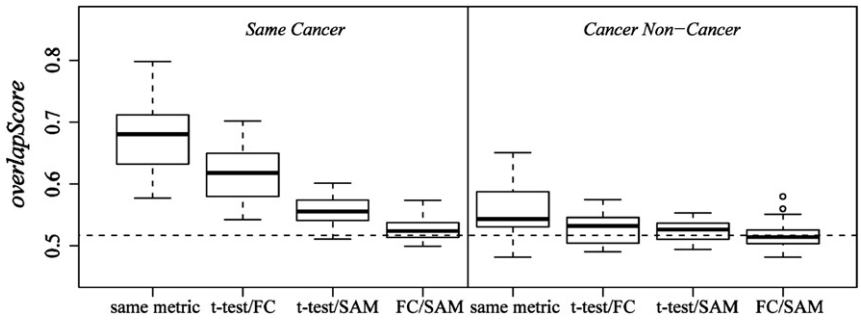


Fig. 7. Comparing studies using different metrics. *OverlapScore* distributions were generated for breast cancer studies with either *same cancer* pairs or *cancer non-cancer* pairs. Each box corresponds either to pairs using different metrics or all pairs using the same metric pooled. The horizontal dashed line denotes the limit of significance (FWER < 0.01). When DE for the two studies is calculated using different metrics the *overlapScore* is clearly lower than when the same metric is used. SAM compared to *t*-test or FC has as low, or lower, *overlapScore* as *cancer non-cancer* pairs using the same metric. Boxes mark the median and quartiles, while the whiskers mark the most extreme data point within 1.5 times the interquartile range from the box.

be given some penalty. Although such a map would be limited to the expression data currently available, it would enable contrasting results of single studies, in order to come closer towards drawing disease-specific conclusions.

R scripts for automatic downloading and processing of GEO gene expression data sets are available upon request.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.ygeno.2013.10.006.

References

[1] Affymetrix, Human Gene 1.0 ST Array Performance, http://media.affymetrix.com/support/technical/whitepapers/hugene_perf_whitepaper.pdf2007.
[2] M. Barnes, J. Freudenberg, S. Thompson, B. Aronow, P. Pavlidis, Experimental comparison and cross-validation of the Affymetrix and Illumina gene expression analysis platforms, Nucleic Acids Res. 33 (18) (2005) 5914–5923.
[3] T. Barrett, D.B. Troup, S.E. Wilhite, P. Ledoux, C. Evangelista, I.F. Kim, M. Tomashevsky, K.A. Marshall, K.H. Phillippy, P.M. Sherman, et al., NCBI GEO: archive for functional genomics data sets — 10 years on, Nucleic Acids Res. 39 (2011) D1005–D1010.
[4] S.E. Choe, M. Boutros, A.M. Michelson, G.M. Church, M.S. Halfon, Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset, Genome Biol. 6 (2005) R16.
[5] S. Davis, P.S. Meltzer, GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor, Bioinformatics 23 (2007) 1846–1847.
[6] A. de la Fuente, From 'differential expression' to 'differential networking' — identification of dysfunctional regulatory networks in diseases, Trends Genet. 26 (2010) 326–333.
[7] B. Di Camillo, T. Sanavia, M. Martini, G. Jurman, F. Sambo, A. Barla, M. Squillario, C. Furlanello, G. Toffolo, C. Cobelli, Effect of size and heterogeneity of samples on biomarker discovery: synthetic and real data assessment, PLoS One 7 (2012) e32200.
[8] L. Gautier, L. Cope, B.M. Bolstad, R.A. Irizarry, affy — Analysis of Affymetrix GeneChip data at the probe level, Bioinformatics 20 (2004) 307–315.
[9] R. Gentleman, annotate: Annotation for microarrays, R package version 1.28.12013.
[10] L. Guo, E.K. Lobenhofer, C. Wang, R. Shippy, S.C. Harris, L. Zhang, N. Mei, T. Chen, D. Herman, F.M. Goodsaid, et al., Rat toxicogenomic study reveals analytical consistency across microarray platforms, Nat. Biotechnol. 24 (2006) 1162–1169.
[11] B. Gyorffy, M. Dietel, T. Fekete, H. Lage, A snapshot of microarray-generated gene expression signatures associated with ovarian carcinoma, Int J. Gynecol. Cancer 18 (2008) 1215–1233.
[12] Z. He, W. Yu, Stable feature selection for biomarker discovery, Comput Biol. Chem. 34 (2010) 215–225.
[13] T. Hothorn, K. Hornik, M.A.v. d Wiel, A. Zeileis, Implementing a class of permutation tests: the coin package, J. Stat. Softw. 28 (2008) 1–23.
[14] W.C. Hsiao, K.C. Young, S.L. Lin, P.W. Lin, Estrogen receptor-alpha polymorphism in a Taiwanese clinical breast cancer population: a case–control study, Breast Cancer Res. 6 (2004) R180–R186.
[15] N.J. Hudson, A. Reverter, B.P. Dalrymple, A differential wiring analysis of expression data correctly identifies the gene containing the causal mutation, PLoS Comput. Biol. 5 (2009) e1000382.
[16] K.B. Hwang, S.W. Kong, S.A. Greenberg, P.J. Park, Combining gene expression data from different generations of oligonucleotide arrays, BMC Bioinformatics 5 (2004) 159.
[17] R.A. Irizarry, Z. Wu, H.A. Jaffee, Comparison of Affymetrix GeneChip expression measures, Bioinformatics 22 (2006) 789–794.
[18] R. Kalluri, ZeisbergM. , Fibroblasts in cancer, Nat. Rev. Cancer 6 (2006) 392–401.
[19] S. Michiels, S. Koscielny, C. Hill, Prediction of cancer outcome with microarrays: a multiple random validation strategy, Lancet 365 (2005) 488–492.
[20] G.L. Miklos, R. Maleszka, Microarray reality checks in the context of a complex disease, Nat. Biotechnol. 22 (2004) 615–621.
[21] J.D. Mosley, R.A. Keri, Intrinsic bias in breast cancer gene expression data sets, BMC Cancer 9 (2009) 214.
[22] A. Nimgaonkar, D. Sanoudou, A.J. Butte, J.N. Haslett, L.M. Kunkel, A.H. Beggs, I.S. Kohane, Reproducibility of gene expression across generations of Affymetrix microarrays, BMC Bioinformatics 4 (2003) 27.
[23] I.B. Pau Ni, Z. Zakaria, R. Muhammad, N. Abdullah, N. Ibrahim, N. Aina Emran, N. Hisham Abdullah, S.N. Syed Hussain, Gene expression patterns distinguish breast carcinomas from normal breast tissues: the Malaysian context, Pathol. Res. Pract. 206 (2010) 223–228.
[24] M.D. Robinson, T.P. Speed, A comparison of Affymetrix gene expression arrays, BMC Bioinformatics 8 (1) (2007) 449.
[25] H. Schwender, siggenes: Multiple testing using SAM and Efron's empirical Bayes approaches, 2009.
[26] L. Shi, W.D. Jones, R.V. Jensen, S.C. Harris, R.G. Perkins, F.M. Goodsaid, L. Guo, L.J. Croner, C. Boysen, H. Fang, et al., The balance of reproducibility, sensitivity, and specificity of lists of differentially expressed genes in microarray studies, BMC Bioinformatics 9 (Suppl. 9) (2008) S10.
[27] V.G. Tusher, R. Tibshirani, G. Chu, Significance analysis of microarrays applied to the ionizing radiation response, Proc. Natl. Acad. Sci. U. S. A. 98 (2001) 5116–5121.
[28] R.G. Verhaak, F.J. Staal, P.J. Valk, B. Lowenberg, M.J. Reinders, D. de Ridder, The effect of oligonucleotide microarray data pre-processing on the analysis of patient–cohort studies, BMC Bioinformatics 7 (2006) 105.
[29] X. Yang, S. Bentink, S. Scheid, R. Spang, Similarities of ordered gene lists, J. Bioinforma. Comput. Biol. 4 (3) (2006) 693–708.
[30] M. Zervakis, M.E. Blazadonakis, G. Tsiliki, V. Danilatou, M. Tsiknakis, D. Kafetzopoulos, Outcome prediction based on microarray analysis: a critical perspective on methods, BMC Bioinformatics 10 (2009) 53.
[31] Q. Zhu, J.C. Miecznikowski, M.S. Halfon, Preferred analysis methods for Affymetrix GeneChips. II. An expanded, balanced, wholly-defined spike-in dataset, BMC Bioinformatics 11 (2010) 285.
[32] Y. Zhu, S. Davis, R. Stephens, P.S. Meltzer, Y. Chen, GEOmetadb: powerful alternative search engine for the Gene Expression Omnibus, Bioinformatics 24 (2008) 2798–2800.
[33] A.L. Boulesteix, M. Slawski, Stability and aggregation of ranked gene lists, Brief Bioinform. 10 (5) (Sep. 2009) 556–568, http://dx.doi.org/10.1093/bib/bbp034.