Coex-Rank: An approach incorporating co-expression information for combined analysis of microarray data

Jinlu Cai[1,*], Henry L. Keen[2], Curt D. Sigmund[2,3,7] and Thomas L. Casavant[4,5,6]

[1]Genetics Department, Albert Einstein College of Medicine, New York, New York, 10461, USA

[2]Department of Pharmacology, The University of Iowa, Iowa City, IA, 52242, USA

[3]Department of Internal Medicine, The University of Iowa, Iowa City, IA, 52242, USA

[4]Department of Electrical and Computer Engineering, The University of Iowa, Iowa City, IA, 52242, USA

[5]Department of Biomedical Engineering, The University of Iowa, Iowa City, IA, 52242, USA

[6]Center for Bioinformatics and Computational Biology, The University of Iowa, Iowa City, IA, 52242, USA

[7]Center for Functional Genomics of Hypertension, The University of Iowa, Iowa City, IA, 52242, USA

Summary

Microarrays have been widely used to study differential gene expression at the genomic level. They can also provide genome-wide co-expression information. Biologically related datasets from independent studies are publicly available, which requires robust combined approaches for integration and validation. Previously, meta-analysis has been adopted to solve this problem.

As an alternative to meta-analysis, for microarray data with high similarity in biological experimental design, a more direct combined approach is possible. Gene-level normalization across datasets is motivated by the different scale and distribution of data due to separate origins. However, there has been limited discussion about this point in the past. Here we describe a combined approach for microarray analysis, including gene-level normalization and Coex-Rank approach. After normalization, a linear modeling process is used to identify lists of differentially expressed genes. The Coex-Rank approach incorporates co-expression information into a rank-aggregation procedure. We applied this computational approach to our data, which illustrated an improvement in statistical power and a complementary advantage of the Coex-Rank approach from a biological perspective.

Our combined approach for microarray data analysis (Coex-rank) is based on normalization, which is naturally driven. The Coex-rank process not only takes advantage of merging the power of multiple methods regarding normalization but also assists in the discovery of functional clusters of genes.

1 Introduction

High-throughput microarray technologies have become popular for genome-wide investigation of gene expression profiles. Careful experimental design followed by a variety

* To whom correspondence should be addressed. Email: caijinlu117@gmail.com

doi:10.2390/biecoll-jib-2012-208

of proper computational analyses can reveal interaction of genes and related biological pathways [1]. For the data analysis, a common goal is to detect differentially expressed genes (DEGs) between controls and cases or in response to specific factors, such as time and dose effects. Different laboratories may carry out microarray experiments with related biological experimental design, but using different types of platforms. Due to the high cost of microarrays, many studies suffer from the problem of small sample size, which may lead to a high false discovery rate (FDR) in determination of DEGs [2]. Combining related but independent microarray datasets increases sample size and may result in higher reliability of novel gene candidate discovery from a statistical view [3]. For example, a combined approach may be able to detect small but consistent changes. In fact, this is one of the motivating factors for the construction of public microarray databases, such as Gene Expression Omnibus (GEO) [4]. In another way, successful combined analysis demonstrates the reproducibility of these studies [5], which is a fundamental issue in validation of biological experiments.

However, rarely is a direct combined analysis suitable for microarray studies, as complications arise from biological variations and technical differences. Meta-analysis, which has been well-studied in statistics, is a practical way to solve this problem. The application of meta-analysis to microarray data has been demonstrated by different groups, yet no consensus has been reached as to the best method. Hong,F. et al. evaluated the performance of different microarray meta-analysis methods and recommended approaches derived from two different philosophies. One is the t-based modeling approach, and the other is a rank-product approach, which has the advantage of robustness in ranking genes over the t-based method, but only provides relative prioritization of genes [6].

As an alternative to meta-analysis, a more direct combined approach is also possible for datasets with highly similar biological design. With the development of microarray technology, more comprehensive arrays become available for researchers in biological fields. For example, exon arrays are designed to focus on exon level analysis, but also provide accurate assessments for gene expression. Thus, there exists a series of microarray datasets with similar biological samples but from different array platforms. Obviously, there are scale and distribution differences among those datasets. To solve this issue, gene-level normalization across datasets has typically been performed, but the details of this have not been widely discussed in combined analysis of microarray datasets.

Gene level normalization is generally the preferred option for microarray analysis in a single study, and this has been revealed by an application of the M-A based loess normalization to a wholly defined control dataset from a “spike-in” experiment [7]. A previous study regarding the comparison of probe level normalization methods suggested that complete data methods including the M-A based loess normalization and the quantile normalization have better performance compared to other methods making use of a baseline array [8]. Therefore, we adopt both M-A based loess normalization and quantile normalization, and then mix them with scale normalization for gene level implementation.

After gene-level normalization, a liner model is set up, which helps to identify lists of differentially expressed genes. Different normalization methods lead to lists of relevant genes, and rank-aggregation approach is used to merge the power of different normalization methods.

To further complement the rank-aggregation approach, we have incorporated co-expression information to prioritize DEGs. The co-expression pattern of genes at the mRNA level can be recognized from a large set of microarray data. The rich body of data in GEO serves to provide this added dimension to our method. Genes with similar mRNA expression profiles are likely to be regulated via the same mechanism or share common functions [9]. This correlated information is useful for detecting or prioritizing genes with weak differential

expression, since these genes are expected to co-express with other highly DEGs. A statistical method of predicting genes with differential expressions based on co-expression patterns has already been proposed [10]. Moreover, rank-aggregation for similar items has been investigated as well [11]. Thus, we modified the rank-aggregation approach using genome-wide co-expression information, which we term as Co-expression-Rank-aggregation (Coex-Rank).

In this article, we describe an approach for Coex-Rank featured analysis to combine microarray data via normalization. We applied this to our own S-PPAR (a mutant PPARγ) dataset. A simulation study was also conducted to demonstrate that the strength of this method is not limited to our specific datasets.

2 Methods

2.1 Motivating datasets

Before introducing our method, we first provide two motivating datasets, with similar, but non-identical experimental designs or platforms. A combined analysis to explore genes with significant expression-level changes will illustrate the potential power of the Coex-Rank featured approach from both statistical and biological perspectives.

Our laboratory has generated transgenic mice with dominant negative PPARγ (P467L) targeted to vascular smooth muscle cells (VSMCs) and these mice (called S-PPAR mice) have been shown to exhibit severe aortic dysfunction [12]. PPARγ has effects in vascular smooth muscle cells (VSMCs), with impact to cardiovascular diseases [12]. Two independent microarray experiments were carried out using mRNA from the aortas of these mice, compared to wild-type littermate controls (denoted as S-PPAR datasets). The first experiment was performed using the Affymetrix mouse genome 430 2.0 array (expression array), with only 2 control and 3 transgenic samples. The second set of samples from the same mice took advantage of the Affymetrix mouse exon 1.0 ST array (exon array); this time with 5 control and 7 transgenic samples. More details are available in supplementary files.

To generate gene-level expression values, we used the Robust Multi-chip Average (RMA) algorithm [13, 14]. For expression array data, the implementation was carried out using the *affy* package of R [15] and resulted in 45,101 probe-sets. The Affymetrix Expression Console software (http://www.affymetrix.com/) was applied to data from the exon arrays and 101,176 gene-level probe-set records were generated. Next, we attempted to remove redundant and ambiguous probe-sets so that comparisons across platforms could be performed. First of all, probe-sets without annotations such as gene symbols or mRNA accession information were removed. In the case of multiple probe-sets matching the same gene, we selected the probe-set with the most significant p-value. Student's T test with equal variance was used to calculate the p-value, comparing control vs. transgenic samples. Through the above steps, 26,599 probe-sets on the expression array and 33,797 probe-sets on exon array were retained. Then, we combined probe-sets from two datasets if they had any annotations in common. For example, there is one record with annotations "NM_015781 /// Nap1l1" from the expression array data and another record from the exon array data annotated as "D12618 /// Nap1l1", therefore, they can be merged into a new record as they share the same gene symbol "Nap1l1". Following this rule, we finally generated a combined dataset with 18,307 records.

2.2 Normalization

Normalization is naturally driven by the relative scale or differences in the distribution of expression levels among arrays from multiple studies. In the case of S-PPAR data for example

(see Figure 1), the distributions of gene expression intensities are dissimilar between two platforms. In our implementation, we applied scale normalization first, which is capable of correcting linear variations, followed by either quantile or M-A based loess normalization.

Scale normalization is sometimes referred as global normalization, which enforces an equal median or mean intensity criteria for all arrays [15]. In our implementation, we selected a method based on median, which is less sensitive to extreme data points. Quantile normalization enforces an equal distribution of intensity values across all the arrays [15].

M-A based Loess normalization is a classical method for cDNA array normalization and can also be applied to two one-channel arrays. First, Y and X denote the $\log_2$-scaled expression values from two arrays, and M denotes the difference of Y and X, while A represents the average of Y and X. That is, M=Y–X and A=(Y+X)/2. The M'-A' plot after loess regression should show a cloud of points scattered about the M'=0 axis and Y', X' are generated afterwards [16].

Loess normalization can be realized via two different approaches -- either a median-base method or a trim-mean method. For the median-base method, consider the S-PPAR combined data mentioned above. In each iteration, Y proceeds from array X_1 to array X_{17}, while X is the array storing the median of the median intensities of all arrays (termed as X_{base}), therefore there are 17 rounds of loess regressions. For each loess regression, X is selected dynamically based on the current expression values of all arrays, and both Y' and X' are used to update Y and X. The pseudo code of this algorithm is as follows:

```
for ( i in 1:#iteration ) /*the number of iterations*/
{
    for (j in 1:#sample)  /* sample size=17 in our S-PPAR dataset*/
    {
        Y=X_j ; X=X_base ;
        Loess normalization using Y and X ;
    }
}
```

For the trim-mean method, in each iteration, Y proceeds from array X_1 to array X_{17}, while X is the reference array, dynamically generated consisting of the 0.05 trim mean of all 17 arrays. As X is only a series of reference arrays, only Y is updated using Y'.

For loess normalization, the regression can also be performed using only rank-invariant genes. The size of the rank-invariant gene set is data dependent. Genes are defined as rank-invariant as described in a previous study [17].

2.3 Linear model

After normalizing using different methods, we generated lists of significantly changed genes for further comparison or validation by a simple linear model. A variety of complex methods have been proposed, but they do not necessarily perform better than a simple one. Further, complex methods may add background noise and even induce bias if all assumptions are not satisfied [18]. For example, consider our S-PPAR data, a linear model can be constructed for each gene by the following formula:

$$Y = b + a_1 \times X_1 + a_2 \times X_2 ,$$

where Y is the observed value of gene expression and b is the baseline level of gene expression. Data from expression array and wild type are considered as the baselines. The exon array effect is indicated by a_1 and X_1 = (0 or 1). The S-PPAR mutant effect is measured by a_2 and X_2 = (0 or 1) as well. The regression is carried out using the *lm()* function of R and then ANOVA is used to test the statistical significance of a_2. The +/- sign of a_2 indicates up or down regulation and the absolute value of a_2 indicates fold-change, which is different from the original scale but can still be used to rank genes or indicate relative changes.

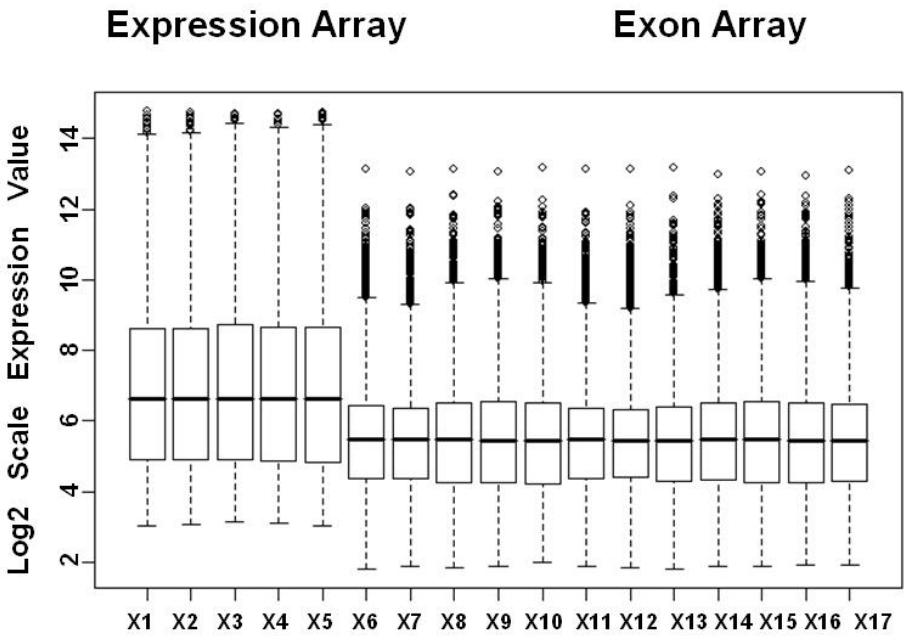

Figure 1: The boxplots of all 17 arrays from S-PPAR datasets. X1-X5 refer to data from the expression arrays and they show different distributions from X6-X17 plots of the exon arrays. This plot is generated by the *boxplot()* function of R.

2.4 Co-Expression-Rank-aggregation (Coex-Rank)

Multiple lists of up and down regulated genes can be generated from different normalization methods. To take advantage of the power from merging all these lists, we investigated the rank-aggregation method, which focuses on finding a robust list with minimum distance among all available ordered lists of genes. The *RankAggreg* package of R is publicly available [19]. For choices of distance function, this package concentrates on the two most popular ones: Spearman foot distance and Kendall's tau distance. The realization of rank-aggregation is provided with two different algorithms: a Cross-entropy Monte Carlo algorithm (CE) and a Genetic algorithm (GA).

For Coex-Rank, we modified the R implementation of rank-aggregation by incorporating co-expression information into the approach. The goal of Coex-Rank is to prioritize genes that are highly correlated with already-top-ranked genes. For instance (see Figure 2), Gene_a and Gene_a' are highly correlated in expression. Gene_a is a top-ranked gene on all input lists for Coex-Rank, but Gene_a' is present at the bottom of some of the input lists. Through our Coex-Rank process, Gene_a' will be pulled up onto top of the output list.

For our implementation, the co-expression information is included in the distance calculation step. The co-expression information is obtained from a combination of microarray datasets with samples from similar tissues of the same species to avoid bias. To be consistent with our case study, mouse S-PPAR data, we added four more microarray datasets using blood vessels of mice and the total sample size increased to 59 (more details are available in supplementary

files). The co-expression co-efficients calculation was based on the probe-sets matching with the final combined S-PPAR dataset as described in section 2.1.2. Then, for any two genes, the Pearson's correlation co-efficient was calculated from 59 pairs of records.

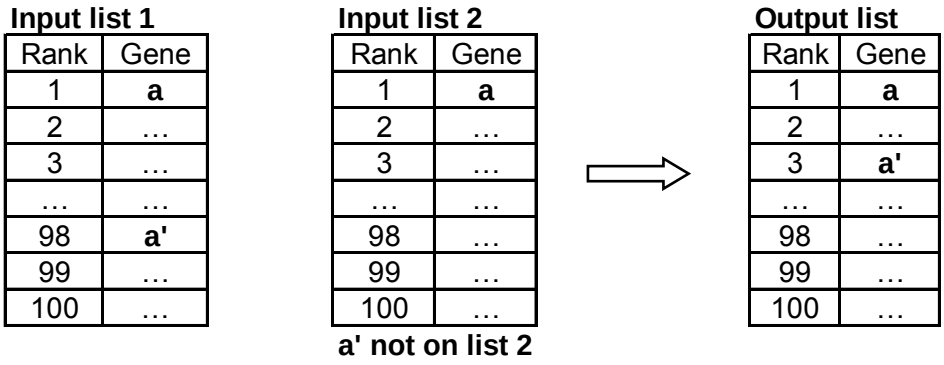

Figure 2: Demonstration of Coex-Rank approach. Gene_a and Gene_a' are two assumed genes. Gene_a is a top-ranked gene on all input lists for Coex-Rank processing, but Gene_a' is only present at the bottom of some of the input lists or even absent from some fo the input lists. Coex-Rank approach prioritizes Gene_a' because it is highly correlated with already-top-ranked Gene_a.

Distance calculation with co-expression information is the core part of Coex-Rank algorithm. The distance D( ) between two ranked gene lists L_1 and L_2, given the co-expression co-efficients, is defined as follows:

$$D(L_1,L_2)= \frac{1}{2} \times (F(L_1,L_1\text{-co}) + F(L_2,L_2\text{-co})),$$

where F( ) is either the Spearman footrule or the Kendall's tau distance of two lists [11]. List L_1-co contains all the genes from list L_1 but the rank information is obtained from list L_2. For genes also present on list L_2, their ranks remain the same, while for genes only present on list L_1 but not on list L_2, the ranks of their highly correlated genes from list L_2 are used instead. There is a cut-off value for co-expression co-efficients for consideration. For example, Gene_a is only present in list L_1, and it has n highly correlated genes on list L_2. The rank of Gene_a on list L_1-co is defined as follows:

$$L_1\text{-co-rank (Gene\_a)} = \frac{1}{n} \times \sum_{i=1}^{n} [\frac{L_2\text{-Rank (Gene\_i)}}{Co(Gene\_a, Gene\_i)}], i = 1,2, \ldots n.$$

Co(Gene_a,Gene_i) denotes the co-expression co-efficient between Gene_a and Gene_i (we used Pearson correlation co-efficient in implementation) and L_2-Rank (Gene_i) is the rank of Gene_i on list L_2. For genes only present on list L_1 but not on list L_2, if they do not have any highly correlated genes from list L_2, their ranks are assigned as Length(L_2)+1, where Length( ) is the length of the gene list.

The R program is freely available for download from http://code.google.com/p/coex-rank/downloads/list with simple data as an example.

3 Results

3.1 Similar effect of different normalization methods

For our mouse S-PPAR data, 10 different normalization methods were implemented. They were quantile, loess-median-base, loess-median-base-invariant, loess-trim-mean, loess-trim-mean-invariant and the same 5 methods utilizing scale normalization first.

For the loess-median-base approach, 10 iterations were chosen for normalization. More iterations should result in more similar distributions of intensities from different arrays. As many as 50 iterations were tried, but no significant improvement in the results was observed (see Supple. Figures 1-3 and Supple. Table 1).

The loess-trim-mean approach could narrow the distribution of intensities after running through a large number of iterations. In an extreme example of 50 iterations, the boxplots of intensities degenerate into many repeated data points. Thus, we selected 5 iterations for the S-PPAR data, which produced similar distributions of intensities as other approaches.

For loess regression based on rank-invariant genes, a separate analysis (data shown in section 3.2) showed that no more than 1,000 genes significantly changed between control and transgenic groups. So we used 17,000 as the size of our rank-invariant gene-set.

After normalization using each method, linear models were created for each gene. An ANOVA test was applied to generate lists of up/down regulated genes due to the S-PPAR effect. Next, a comparison of 10 up-regulated lists was performed, each with the top 100 genes ranked by p-value (see Table 1). In the table showing the size of the union of any two gene lists, the largest set contains 129 genes, which indicates that lists from any two normalization methods have about 70% overlapping genes at least. For down-regulated genes, the results are similar (see Supple. Table 2). Though different normalization methods were applied, similar gene lists were generated, which motivated us to apply the rank-aggregation approach to utilize information from all the normalization methods.

Table 1: Size table of union of any two lists from different normalizations. The first row and the first column show the index of normalization Methods. The numbers in the table are the size of union of any two list from different normalization method. The maximum union size is 129, shown in bold.

	1	2	3	4	5	6	7	8	9	10
1	Quantile	100	113	112	110	108	**129**	123	127	122
2	scale-quantile		113	112	110	108	**129**	123	127	122
3	loess-trim-mean			102	111	109	128	**129**	125	126
4	scale-loess-trim-mean				110	108	127	127	125	124
5	loess-trim-mean-invariant					103	127	123	126	119
6	scale-loess-trim-mean-invariant						126	123	126	120
7	loess-median-base							128	115	127
8	scale-loess-median-base								127	122
9	loess-median-base-invariant									127
10	scale-loess-median-base-invariant									

3.2 Combined analysis increases statistical power

By increasing sample size, statistical power of an analysis will tend to increase. For our S-PPAR data, the combined analysis has a sample size of 17, while the separate datasets have sample sizes of 5 or 12. Comparison of the two different analyses demonstrated the benefit of the larger sample size.

For the separate analysis, student's T test with equal variance was used to compare control vs. transgenic samples. This statistical test is mathematically equivalent to a one-way ANOVA test. When we selected p-value<0.005 as cut-off value, we could achieve roughly twice the

number of genes via combined analysis, compared to the separate approach (see Table 2). The statistics of the combined analysis were based on the scale-loess-trim-mean-invariant normalization method; other normalization methods resulted in similar numbers.

Table 2: Comparison of combined and separate analyses of S-PPAR data using cut-off value: p-value < 0.005 and FDR < 0.05. Numbers of total DEGs and up/down-regulated genes are shown separately for expression/exon array data and combined analysis. Both the p-value and FDR are used as cutoff criteria. The combined analysis demonstrates a better statistical power.

	Expression Array		**Exon Array**		**Combined Analysis**	
	p-value	**FDR**	p-value	**FDR**	p-value	**FDR**
#Total	288	**5**	218	**23**	583	**286**
#Up	200	**2**	115	**9**	283	**140**
#Down	88	**3**	103	**14**	300	**146**

We also corrected for multiple comparisons effect using the the *qvalue* package of R [20]. When we set FDR (also called q-value) < 0.05, we could see a dramatic improvement with the combined analysis, from 5 genes from the expression arrays, 23 genes from the exon arrays to 286 genes from the combined analysis (see Table 2). Two different cut-off values were set and more genes were selected as DEGs in the combined analysis, indicating the increasing statistical power of this approach.

3.3 Complementary advantage of Coex-Rank

Gene lists from 5 normalization methods starting with scale normalization were used as the input for rank-aggregation and Coex-Rank approaches. For example, for the up-regulated genes, considering p-value < 0.005 as the cut-off, 5 gene lists were generated and then the genes were ranked either by p-value or fold change, which resulted in 10 different lists. The top 100 genes were selected from each list and then served as the input for both rank-aggregation and Coex-Rank approaches. The reason we chose 100 genes from each list was that these genes were significantly up-regulated according to the FDR < 0.05 cut-off value.

The parameter settings for the rank-aggregation step were the default values (spearman footrule distance and cross-entropy algorithm), except that the maximum-iteration was increased from 1000 to 1500 for our S-PPAR data. For Coex-Rank approach, one more parameter for the cut-off value of co-expression co-efficients was set as 0.95 for our S-PPAR data. The output of both rank-aggregation and Coex-Rank approaches were lists, each with 100 genes.

We note that the rank-aggregation and the Coex-Rank methods, both generated different lists of genes, but that they shared about 70% genes in common (73 for up-regulated genes and 67 for down-regulated genes). To investigate the biological significance of these genes, we focused on the enrichment of annotations. We compared the gene lists from two approaches by clusters generated by DAVID [21] (the default medium and low classification stringencies were used). Coex-Rank approach led to slightly more enrichments (see Table 3) due to the incorporation of co-expression information. If we focused on individual genes, for instance, Calpain 13 – CAPN13 , was only reported to be up-regulated via Coex-Rank approach. Increased calpain activity has been associated with vascular dysfunction and enhanced angiotensin II signaling. Thus inhibition of calpains has several beneficial actions, preventing

cardiovascular remodeling in angiotensin II-induced hypertension [22]. Moreover, another gene TIMP4 - tissue inhibitor of metalloproteinase 4 – was only found to be down-regulated in the Coex-Rank analysis. TIMPs normally inhibit MMPs (metalloproteinases). This inhibitory action is mostly thought to be a good thing. So in the S-PPAR mice, a potential hypothesis is that decreased TIMP4 leads to increased MMPs which in turn leads to vascular remodeling or damage or at least a predisposition to those consequences. Linking back to PPARγ, it has also been shown to have binding sites near TIMP4 gene in a Chip-Seq experiment [23].

Table 3: Comparison of gene annotations enrichment for both rank-aggregation and Coex-Rank approaches. Bothe medium and low stringencies are used to generate clusters of up/down-regulated for each approaches and Coex_Rank method achieves more clusters compared to Rank-aggregation only method.

	#Clusters of up-regulated genes		#Clusters of down-regulated genes	
	Medium Stringency	**Low Stringency**	Medium Stringency	**Low Stringency**
Rank-aggregation	1	**5**	3	**5**
Coex-Rank	3	**7**	3	**6**

However, Coex-Rank approach prioritizes genes highly correlated with already-top-ranked genes on the input lists at the cost of excluding some already-top-ranked genes. Therefore, we decided to add non-overlapping genes from the Coex-Rank approach to the top 100 genes from rank-aggregation and in total we promoted 127 up-regulated genes and 133 down-regulated genes to the final reported lists (see Supple. Files Sppar_up.xls and Sppar_down.xls). These up-regulated genes generate 7 clusters according to DAVID (with low classification stringency). For down-regulated genes, 9 clusters were generated.

4 Discussion

To confirm that our result shown in section 3.2 regarding the advantage of combined analysis over separate analysis was not dataset dependent, we conducted a simulation study consisting of one dataset from the exon arrays and one dataset from the expression arrays. Each dataset had six samples, three controls v.s. three treatments and each sample covered 18,307 genes. Consider the exon array dataset for example, generated as follows:

(1) The sample means μ_i (i = 1, 2, 3…18,307) were from a real dataset. Four arrays using mammary gland were exacted from GSE10246 and the same probe-sets were selected as in our S-PPAR case study. Sample means were calculated for 18,307 genes separately.

(2) Background variations were added according to the following formulas:

$$Y_{ij}=\mu_i+Z_{ij} \quad (i=1,2,3\ldots18307, j=1,2,3,4,5,6),$$

$$Z_{ij} \sim N(0,\sigma^2),$$

$$\sigma=\alpha\times(0.3-0.02\times\mu_i)\times G_i,\ G_i\sim Gamma(5).$$

Y_{ij} refers to the expression value of the i^{th} gene from the j^{th} sample and α is a parameter controlling the scale of variation [2]. We evaluated α = 0.1, 0.2, and 0.3 to demonstrate different levels of background noise. Here, we also made the assumption that the

amount of variation is μ_i dependent. As it is often seen in real data, genes with smaller expression values have larger proportional variations [2].

(3) The first 200 genes from treated samples were added with differential expression values as follows [2]:

$$Y_{ij}=\mu_i+Z_{ij}+\delta_{ij} \ (i = 1,2,3 \ldots\ldots 200, j = 4,5,6),$$

$$\delta_{ij} = 0.2\times(2\times B_{ij} - 1) \times G_i ,$$

$$B_{ij}\sim \text{Bernoulli}(0.5),\ G_i \sim \text{Gamma}(5).$$

The simulation data from expression arrays were generated in a similar way. At step 1, the four arrays using mammary gland were extracted from GSE15998 and at step 3, the differential expression value for a specific gene was scaled by the ratio of sample means from two platforms.

We then generated 10 datasets for each platform. We applied both separate analysis and combined analysis including normalization and linear regression followed by an ANOVA test as described in our Methods section. We used a p-value cutoff of 0.001 to select significantly changed genes. The number of differentially expressed genes was averaged for calculation of sensitivity and specificity and FDR respectively for expression array data, exon array data and a combined dataset. As shown in Table 4, the combined analysis increases the Sensitivity and reduces the FDR compared to the separate analysis, with Specificity remaining consistent (around 0.99) at different levels of background noise. The consistency of specificity is due to the nature of microarray data, as the expression levels of most genes are unchanged.

Table 4: Comparison of combined and separate analyses based on simulation. Combined analysis has advantage in increasing of sensitivity and decreasing of FDR compared with expression array or exon array only analysis. Different background variation has been evaluated via α = 0.1, 0.2 and 0.3.

	Expression Array		**Exon Array**		**Combined Analysis**	
	Sensitivity	**FDR**	Sensitivity	**FDR**	Sensitivity	**FDR**
α = 0.1	0.89	**0.09**	0.69	**0.11**	0.94	**0.07**
α = 0.2	0.67	**0.12**	0.40	**0.17**	0.82	**0.07**
α = 0.3	0.47	**0.16**	0.24	**0.28**	0.64	**0.09**

5 Conclusion

In this article, we describe an approach for combined analysis of microarray data, starting from normalization, and proceeding to rank-aggregation / Coex-Rank procedures. Different normalization methods were discussed and compared. Rank-aggregation / Coex-Rank approach was employed to generate final robust lists of genes with differential expression. This approach is flexible regarding normalization procedures and takes advantage of merging the power of multiple methods. Moreover, incorporating the co-expression information in the rank-aggregation approach helps to discover functional clusters of genes.

In this paper, Coex-Rank was applied to generate robust results from different normalizations. It can be applicable to merge gene lists from potentially incompatible methods arising from statistical tests as well. For example, multiple significance testing methods have been

proposed for microarray experiments with small sample size, and thus our Coex-Rank solution also provides an alternative to a seemingly arbitrary choice among many good methods. Coex-Rank is not limited to microarray data, and it is open to prioritize any lists of genes from other high-throughput technology, such as deep sequencing.

Acknowledgements

The work was supported by the National Institutes of Health [1T32GM082729-01, HL062984]; and the National Heart, Lung and Blood Institute [NS024621]. We also would like to thank Dr. Jian Huang and Dr. Kai Wang for providing comments on the manuscript.

References

[1] J. Quackenbush. Microarray data normalization and transformation. *Nat Genet*, 32:496-501, 2002.

[2] C. Kooperberg, A. Aragaki, A. D. Strand and J. M. Olso. Significance testing for small microarray experiments. *Stat Med*, 24(15):2281-2298, 2005.

[3] D. Ghosh, T. R. Barette, D. Rhodes and A. M. Chinnaiyan. Statistical issues and methods for meta-analysis of microarray data: a case study in prostate cancer. *Funct Integr Genomics*, 3(4):180-188, 2003.

[4] R. Edgar, M. Domrachev and A. E. Lash. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*, 30(1):207-210, 2002.

[5] M. Zhang, L. Zhang, J. Zou, C. Yao, H. Xiao, Q. Liu, J. Wang, D. Wang, C. Wang and Z. Guo. Evaluating reproducibility of differential expression discoveries in microarray studies by considering correlated molecular changes. *Bioinformatics*, 25(13):1662-1668, 2009.

[6] F. Hong and R. Breitling. A comparison of meta-analysis methods for detecting differentially expressed genes in microarray experiments. *Bioinformatics*, 24(3):374-382, 2008.

[7] 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*, 6(2):R16, 2005.

[8] B. M. Bolstad, R. A. Irizarry, M. Astrand and T. P. Speed T. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185-193, 2003.

[9] D. J. Allocco, I. S. Kohane and A. J. Butte. Quantifying the relationship between co-expression, co-regulation and gene function. *BMC Bioinformatics*, 5:18, 2004.

[10] Y. Lai. Genome-wide co-expression based prediction of differential expressions. *Bioinformatics*, 24(5):666-673, 2008.

[11] D. Sculley. Rank Aggregation for Similar Items. In *Proceedings of 2007 SIAM International Conference on Data Mining*, 2007.

[12] C. M. Halabi, A. M. Beyer, W. J. de Lange, H. L. Keen, G. L. Baumbach, F. M. Faraci and C. D. Sigmund. Interference with PPAR gamma function in smooth muscle causes vascular dysfunction and hypertension. *Cell Metab*, 7(3):215-226, 2008.

[13] R. A. Irizarry, B. M. Bolstad, F. Collin, L. M. Cope, B. Hobbs and T. P. Speed. Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Re*s, 31(4):e15, 2003.

[14] R. A. Irizarry, B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf and T. P. Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249-264, 2003.

[15] B. M. Bolstad, R. A. Irizarry, M. Astrand and T. P. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185-193, 2003.

[16] S. Dudoit, Y. H. Yang, M. J. Callow and T. P. Speed. Statistical methods for identifying genes with differential expression in replicated cDNA microarray experiments. *Statistica Sinica*, 12(1):111-139, 2002.

[17] C. R. Pelz, M. Kulesz-Martin, G. Bagby and R. C. Sears. Global rank-invariant set normalization (GRSN) to reduce systematic distortions in microarray data. *BMC Bioinformatics*, 9:520, 2008.

[18] D. B. Allison, X. Cui, G. P. Page and M. Sabripour. Microarray data analysis: from disarray to consolidation and consensus. *Nat Rev Genet*, 7(1):55-65, 2006.

[19] V. Pihur and S. Datta. RankAggreg, an R package for weighted rank aggregation. *BMC Bioinformatics*, 10:62, 2009.

[20] J. D. Storey and R. Tibshirani. Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A*, 100(16):9440-9445, 2003.

[21] D. W. Huang, B. T. Sherman, Q. Tan, J. R. Collins, W. G. Alvord, J. Roayaei, R. Stephens, M. W. Baseler, H. C. Lane and R. A. Lempicki. The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. *Genome Biol*, 8(9):R183, 2007.

[22] E. Letavernier, J. Perez, A. Bellocq, L. Mesnard, A. de Castro Keller, J.-P. Haymann and L. Baud. Targeting the calpain/calpastatin system as a new strategy to prevent cardiovascular remodeling in angiotensin II-induced hypertension. *Circ Res*, 102:720-728, 2008.

[23] T. Adhikary, K. Kaddatz, F. Finkernagel, A. Schönbauer, W. Meissner, M. Scharfe, M. Jarek, H. Blöcker, S. Müller-Brüsselbach and R. Müller. Genomewide analyses define different modes of transcriptional regulation by peroxisome proliferator-activated receptor-β/δ (PPARβ/δ). *PLoS One*, 6(1):e16344, 2011.