EPIG-Seq: Extracting Patterns and Identifying Co-expressed Genes from RNA-Seq Data

Jianying Li1,2,4 and Pierre R. Bushel2,3,\*

1Integrative Bioinformatics Group, 2Microarray and Genome Informatics Group, 3Biostatistics and Computational Biology Branch, National Institute of Environmental Health Sciences, Research Triangle Park, NC 27709

4Kelly Government Solutions, Research Triangle Park, NC 27709

\*Corresponding author

Pierre R. Bushel

111 T.W. Alexander Drive,

RTP, NC 27709

Phone: 919-316-4564

Email: bushel@niehs.nih.gov

Abstract

RNA-sequencing (RNA-seq) provides genome-wide representation of gene expression. RNA-Seq data is count-based rendering many normal distribution models inappropriate. We adapted our extracting patterns and identifying co-expressed genes methodology for RNA-seq (EPIG-Seq). To identify patterns, a count-based correlation measures similarity between expression profiles, a quasipoisson model estimates dispersion and a location parameter indicates the magnitude of differential expression. EPIG-Seq categorizes genes to the patterns that they correlate with. Using simulated data, we show that EPIG-Seq is superior to EPIG when the data is transformed by popular RNA-Seq normalization methods and performs equally as well as other pattern analysis methods when count data is used. We applied EPIG-Seq to toxicogenomics and cancer datasets to identify co-expressed genes related to the underlying biology of the conditions. EPIG-Seq is available at: www.niehs.nih.gov/research/resources/software/biostatistics/epig-seq.

Key words: gene expression, RNA-seq, clustering, pattern analysis, toxicogenomics, EPIG-Seq

Background

The advantages of RNA-sequencing (RNA-Seq) over microarray to measure gene expression have been widely reported [[1](#_ENREF_1)]. Recently, methods have been developed to analyze RNA-Seq data based on normalization of the read counts or using the raw count data [[2-4](#_ENREF_2)]. The former has advantages in that it adjusts the data according to sequencing library size (read depth), accounts for the length of transcripts and allows the use of analysis tool designed specifically for microarray data. However, normalized RNA-Seq data has limitations which can adversely impact the analysis (inflated variation among replicates, zeros in the data, low and highly expressed genes squelching, etc.). On the other hand, using the raw read counts avoids the shortcomings of normalization but requires modeling of the data to estimate overdispersion, accounting for the read depth and filtering to avoid cases of inflated zeros. In addition, statistical models of count data based on Poisson, beta- or negative-binomial distributions can be severely impacted by outliers in the data although recently developed approaches overcomes this pitfall by analyzing the data with a log-linear model or Wicoxon statistic and resampling in a non-parametric manner [[5-7](#_ENREF_5)]. Unfortunately, there is a paucity of methodologies that can identify correlated gene expression patterns from RNA-Seq count data across biological conditions (i.e., time course, dose response, multiclass study designs).

Extracting patterns and identifying co-expressed genes (EPIG) was designed for microarray data to detect genes with similar expression profiles across biological conditions [[8](#_ENREF_8)]. EPIG uses the Pearson correlation to group genes by similarity of expression across treatment groups, a signal-to-noise ratio (SNR) to compare the level of a genes expression to background noise and magnitude of fold change to assess the level of differential gene expression between test samples and controls/baseline. Normalizing RNA-Seq data to transform it into microarray gene expression space permits it to be analyzed by EPIG but comes with the aforementioned caveats.

We adapted the EPIG methodology for the identification of co-expressed genes from RNA-Seq data (EPIG-Seq). Patterns of gene expression across experimental groups are determined using a similarity measure for count data [[9](#_ENREF_9)]to ascertain similarity between expression profiles, a quasipoisson model [[7](#_ENREF_7)] to estimate overdispersion in the data and a location parameter as a measure of the magnitude of difference between test samples and control/baseline. EPIG-Seq then categorizes each gene expression profile to the pattern for which it has the highest correlation. The EPIG-Seq approach is impervious to variations in read depths, inflated zeros, outliers in the data and overdispersion. Using simulated data, we show that EPIG-Seq is superior to EPIG when the data is transformed by the popular RNA-Seq reads per million (RPM) normalization method and performs equally as well as other pattern analysis methods when count data is used. We also apply EPIG-Seq to a publicly available RNA-Seq data sets from the SEquence Quality Control (SEQC) toxicogenomics [[1](#_ENREF_1)] arm of the MicroArray Quality Control (MAQC) consortium and from The Cancer Genome Atlas (TCGA) [[10](#_ENREF_10)] breast cancer data portal. We identify several co-expressed genes related to modes of action of the chemical agents in the toxicogenomics data set and we also determine genes that are commonly known to be associated with the breast cancers subtypes and some putatively involved in oncogenic pathways.

**Materials and Methods**

**Data**

**TCGA breast cancer RNA-Seq data**

Count-level breast cancer RNA-Seq data from The Cancer Genome Altas (TCGA) [[10](#_ENREF_10)] was downloaded from NCI portal: <https://tcga-data.nci.nih.gov/tcga/dataAccessMatrix.htm>. We focused on the mRNAseq data produced on the Illumina GAII sequencer. Raw sequences were analyzed by the TCGA team with SeqWare <https://seqware.github.io/> (Version=0.7.0) using RNASeqAlignmentBWAWorfklow (Version=0.7.10) and RNASeqQuantificationWorfklow (Version=0.7.2). Tumor samples were classified by the mRNA subtypes [[11-13](#_ENREF_11)]. In our study, we only used data of the following four subtypes: luminal A, luminal B, Her2-enriched, and basal-like. To generate 4 sample data sets of reasonable size (n=50), for each one, we randomly selected 10 lanes from each tumor subtypes plus 10 lanes from normal tissues.

**Simulated RNA-Seq data**

To simulate RNA-Seq count data for each *k*th pattern where k{1,…,K}, we adopted and extended a method published previously [[14](#_ENREF_14)]. Briefly, the count was simulated based on a negative binomial distribution with dispersion parameter estimated from a known data set. Let denote the count of a gene *g* {1,…,*G*}, belonging to a group *m* {1,…,*M*} of a sample *S*{S1, S2,…,SN} such that:



~NB( mean = , var= ( 1 + ))

where, is the parameter measuring the dispersion in the data and is the true mean of the data. We set the dispersion parameter to be the same for all samples. Thus, = . Here,



where, is the sequencing depth for the *m*th group in sample *S*.  = 107for ~Unif[0.7,1.4]. The estimation of the sample mean and  are obtained from a DESeq [[15](#_ENREF_15)] analysis of RNA-Seq raw count data from the 10 randomly chosen human normal breast tissue samples obtained from TCGA (see previous section).. Then, to generate simulated patterns of genes across groups of samples, we defined, where S1 is the control group of samples, Sm is a group of samples with a particular phenotype,  and

.

The parameter  denotes the lower bound of differential expression for the *g*th gene between Sm and S1; it was set at different levels for the respective groups. We let  denote the set of genes that are up-regulated and  denote the set of genes that are down-regulated.

**Toxicogenomics RNA-Seq data**

RNA-Seq data from the MicroArray Quality Control phase III (Sequence Quality Control (SEQC)) crowd source toxicogenomics (TGxSEQC) effort [[1](#_ENREF_1)] is available in the National Center for Biotechnology Information Sequence Read Archive (SRA) [[16](#_ENREF_16)] under accession number SRP024314. We used the training set data containing RNA-seq data from the livers of male Sprague-Dawley rats exposed to one of 15 chemicals or vehicle and route matched controls. Sets of three chemicals share one of five modes of action (MOA). Three MOAs are associated with well-defined receptor-mediated processes—peroxisome proliferator-activated receptor alpha (PPARA), orphan nuclear hormone receptors (CAR/PXR) and aryl hydrocarbon receptor (AhR). The other two are non-receptor-mediated—DNA damage (DNA\_Damage) or cytotoxicity (Cytotoxic). The data is comprised of paired-end 100 bp reads from Illumina HiScanSQ or HiSeq2000 systems with depths between 23 – 25 million reads. Specific details of the study design and sample collection are available in the TGxSEQC publication. For each sample, a two-step alignment was performed using the fastq files. In the first step, raw reads were aligned with Novoalign v2.08.01 (www.novocraft.com) against rat genome rn4 downloaded from the UCSC ftp server (ftp://hgdownload.cse.ucsc.edu/goldenPath/rn4). The intermediate bam files generated with Novoalign were then parsed through customized scripts to summarize mapping results. Unmapped reads were passed to Novoalign again and were mapped to the rat RefSeq transcriptome gene model (release version 52, March 5, 2012) downloaded from the NCBI ftp server (ftp://ftp.ncbi.nih.gov/refseq). The alignment results were parsed through customized scripts to summarize mapping results and then merged with the results generated in step one.

**EPIG-Seq**

Pattern extraction

**Correlation**

A compiled RNA-seq gene expression dataset consists of a 2-dimensional matrix, in which each row represents a gene expression profile and each column represents a sample. We denote x*ij* as the count of reads from sample *j* mapped to a gene *i* and *xkj* as the count of reads from sample *j* mapped to a gene *k*. To measure the count level correlation between two gene profiles, we adopted the similarity measure for count data as defined by Cao et al. {Cao, 1997 #20} is:



where



and *a* is the total number of samples with read counts mapped to either profile. *CYd* is maximized by assigning *xij* = 1 and *xkj* = *Dk* for *a*/2 times and likewise *xkj* = 1 and *xij* = *Di* *a*/2 times. Here

,  for *xij* ≥ 1 and *xkj* ≥ 1. Thus, maximum *CYd* = *D1* + *D2* + *D3*, where

,

,

,

*b* is the total number of samples with read counts present in profile *i* only and *c* is the total number of samples with read counts present in profile *k* only. As such,= 0 when two profiles are totally difference and = 1 when the two are identical.

**Magnitude of change**

We define the strength of a gene expression profile’s signal according to the value of the test-statistic location parameter obtained from a Wilcoxon rank sum non-parametric test [[17](#_ENREF_17)] measuring the difference between the ranks of the expression of the genes in sample *X* vs those in sample *Y*. Here, sample *X* is the biological replicates from the treated, perturbed or diseased group and sample *Y* is the biological replicates from the controls. The *g*th gene expression profile’s signal is therefore:



When the sample size for each group is small (i.e., ≤ 30), the estimation of the Z-statistic from the Wilcoxon rank sum test can be spurious. In such a case, we default to measure the strength of the *g*th gene’s differential expression according to the value of the Hodges-Lehmann location parameter estimator  for the difference between two groups[[17](#_ENREF_17)]. Briefly,  is the median of the Walsh averages, the average of all possible pairs of differences between the ranks of the *g*th gene RNA-Seq counts in sample X vs the ranks of the RNA-Seq counts in sample Y. Thus, for a set of N paired differences observations, there will be N(N+1)/2 Walsh averages. The median of all the Walsh averages is equal to. Hence,

.

**Dispersion**

Count data is known to be dispersed. The variance-to-mean ratio (*VMR*) is a measure of dispersion () and is the inverse of signal to noise (SNR). If is larger than 1, a dataset is said to be overdispersed. For each *g*th gene expression profile, we estimate using a quasi-Poisson regression to model the data. For Poisson distributed data, the variance is equal to the mean, i.e., V(Yg) = E(Yg) = *g*. The quasi-Poisson likelihood model is commonly used for overdispersed count data as it incorporates the dispersion parameter  into the Poisson model [[18](#_ENREF_18)]. In doing so, the variance of the response (Yg) is a linear function of the mean, V(Yg) =  and dispersion estimated as



where *n* is the sample size, *c* is the number of estimated parameters andis an inverse function of the linear predictors. Here, the inverse function is a “log” link in the form of a generalized linear model:

 , where for the *g*th gene expression profile in the *j*th sample, *Y* is the read count, *X* is the independent variable and ** is the random error term. In our study, one set of dispersion parameters were estimated from the TCGA data and was used in simulating the research data. In general modeling, dispersion was often estimated from the aforementioned Quasi-Poisson likelihood model for individual profile.

Categorization of gene expression profiles to patterns

Once the patterns have been extracted, the measure is used to correlate the *i*th profile to the *k*th pattern, which is often represented by the profile initialized as the seed. The profile is assigned to the pattern to which it has the highest similarity to. Once all the profiles are assigned, a representative profile for each of the patterns is determined by the highest median correlation to the other profiles in the pattern. Briefly, for the *i*th gene expression profile and for the *k*th pattern it is assigned to, a Pattern Correlation Score



is computed as the median of the correlations among the *i*th profile (*xi*) to all other profiles (*xj*) assigned to pattern *k* (*Pk*). Until no more profiles are reassigned, the  measure is used to correlate the *i*th profile to the *k*th pattern and assign it to the pattern with which it has the highest correlation.

Since the *CYs* does not denote the directionality of the correlation, we restrict the assignment of a profile if and only if, corr(ind(Zi),ind(Zj)) for all groups = 1. Here corr() is the Pearson correlation and ind() is the indication of + or – of the location parameter Z.

Searching the parameter space

To extract patterns (step 1) and categorize gene profiles to the candidate patterns (step 2), user-defined parameters are set. To search for optimal parameter settings, in the first step, we fixed the parameters for profile categorization in step 2 with the *CYs* similarity at 0.8 and location parameter St2 at 2. Then, we searched the parameter space for step 1. These included: (1) *CYs* similarity [0.5 – 0.9] at interval 0.1, (2) location parameter St1 at [1 – 5] at interval 0.5, and dispersion tail (on both side) [1 – 5%] at interval 1%. With the parameter set for step 1 fixed, we searched the parameter space for step 2. These included: (1) *CYs* similarity [0.7 – 0.9] at interval 0.1 and (2) St2 at [1 – 5] at interval 0.5. To evaluated the impact of the parameters on the EPIG-seq performance, we used the Adjusted Rand Index [[19](#_ENREF_19)] together with true number of real patterns from the simulated data.

Assessing the significance of extracted pattern

To assess the significance of the extracted patterns, we performed *B* number of bootstrapped assignments of *P* random gene profiles to a pattern and compute the PCS each time to compare to the observed PCS for that pattern. Briefly, for *B* times and for a given pattern containing *P* gene profiles, we randomly select *P* number of gene profiles from the dataset. Then, for the selected *P* random profiles, we compute the bootstrapped PCS. The *p*-value for a pattern is computed as the number (*n*) of times one of these bootstrapped scores is greater than the observed score. Thus, p-value = *n*/*B*.

Comparing two clustering outcomes

To assess the reproducibility of pattern extraction and subsequent categorization of genes, we adopted a method for accessing the clustering reproducibility [[20](#_ENREF_20)]. To compare two clustering results during the recursion and reallocation portion of the gene categorization to patterns part of EPIG-Seq (step 2), the Mutual Information (MI) between the two clustering results was used.  MI, when log base 2 is used, is interpreted in units of bits and is the reduction in uncertainty about the samples groupings in one clustering outcome given the other clustering outcome (i.e., clustering A vs clustering B).

Let: 

where, *p*(*a,b*) is the joint probability distribution of *A* and *B*, and Ent(*X*) is the entropy of clustering outcome X. Given *k* clusters,



where *pi* is the probability of the *i*th cluster membership in clustering X.  To adjust the MI for chance occurrence,



where, EMI is the expectation of the mutual information.  AMI ranges between 0 and 1.  An AMI value of zero means that the two clustering outcomes are independent whereas an AMI of 1 indicates that the two clustering are highly associated.

The EPIG-Seq algorithm pseudo code

**Step 1: Extract candidate profiles as seeds to patterns**

1. Define the parameter set θ for pattern extracting;
2. **for** i = 1..n **do**

Calculate all pairwise CYs correlation

**end**

1. Delete the *i*th profile if the number of profiles with CYs >= Rt1 is < Mt
2. For remaining profiles, estimate
   1. the location parameter
   2. the dispersion
3. Delete profile if location parameter < St1 or [dispersion < 95th percentile or > 5th percentile]
4. Remove profile with max(dispersion) if the top 5 correlated profiles overlap with another profile
5. Remove profile with correlation >=0.9 to another profile

Remaining profiles are defined as candidate seeds for patterns **Step 2: Categorize profiles to patterns**

1. Define the parameter set θ for categorization;
2. **for** i = 1..n **do**

Initially assign the ith profile to the kth pattern if CYs >= Rt2 and location parameter > St2

**end**

1. Until no more moves (when 1-AMI < 0.0001)

**for** i = 1..m **do**

Update patterns with profiles with the highest median PCS

Assign the ith profile to the kth pattern if CYs >= Rt2

**end**

1. Report final assignment of profiles to patterns

**Results**

**Components of EPIG-Seq**

We patterned the development of EPIG-Seq off of the steps and components that make up EPIG [[8](#_ENREF_8)] for analyzing gene expression patterns from microarray data. As shown in table 1, whereas EPIG uses Pearson correlation, magnitude of change and signal-to-noise for similarity, response and dynamic range, EPIG-Seq uses a CYs measure, a the magnitude of a Wilcoxon statistic and dispersion for count RNA-Seq count data. There are several advantages of EPIG-Seq on the analysis of RNA-Seq data. First, it supports cases where the read count is “zero”. Second, it doesn’t require normalization due to differences in total read count per sample/lane of RNA-Seq. Third, it handles the discrete Poisson distribution typical of RNA-Seq count data and uses a quasi-Poisson model to account for dispersion in the data. Finally, when within group sample sizes are small, it uses the robust and non-parametric Hodges-Lehmann estimator as magnitude location parameter. As shown in Figure 1, the 1st step in EPIG-Seq is to find all patterns in the data. Once the patterns are identified, in step 2, the expression profiles are categorized to the patterns to cluster them. Clustering is performed iteratively until the patterns with the categorized gene profiles stabilize. The profile is assigned to the pattern to which it has the highest similarity to. Once all the profiles are assigned, a representative profile for each of the patterns is determined by the Pattern Correlation Score (PCS) which is the highest median correlation to the other profiles in the pattern. Since the *CYs* correlation measure for count data does not denote directionality, we restrict the assignment of a profile if and only if, it is in the values of the location parameter for each sample are in the same direction for the pattern and the profile. To ascertain significance of an extracted pattern with categorized gene profiles, we randomly assign gene profiles to the patterns and compute p-values from a non-parametric distribution of PCSs.

**Evaluation of EPIG-seq**

To test the robustness EPIG-seq, we simulated RNA-seq count data across four groups. Six patterns were simulated (Figure 2) with parameters shown in Table 2. Patterns 1-5 (n = 200 profiles in each for a total of 1000 real pseudo genes) are considered real biological responses (i.e., the mean varies) across the samples whereas pattern # 6 (n = 19000 profiles) is considered unresponsive (i.e., noisy pseudo genes) and as such, no profiles from this pattern should be extracted. The heat map representation of the data from a comparison the samples in all groups to the average of the baseline (1st group) illustrates the distinct differences of the patterns (Figure 3A). As shown in Figure 3B, the distribution of the samples in 3D space by principal component analysis of the CYs measures of the 20,000 simulated genes from the six patterns separate reveals that the four groups of data are separated fairly well, although there is some overlapping which makes the extraction a challenge. Noticeably, the first three PCs explained almost 90% of the total variability in the data.

**Determination of the optimal parameters for EPIG-Seq**

To find the best parameter setting to analyze the simulated data, we performed a search space scan by fixing the CYs and location parameter St2 for step 2 at 0.8 and 5 respectively while varying the CYs measure between 0.5 and 0.9, the location parameter (St1) from 1-5 and dispersion percentile (PCT) 1-5% (both sides) for step 1 of EPIG-Seq. Figure 4A depicts the plotting of the adjusted Rand index (ADI) for 5 – 10 patterns extracted indicating the validity of the clustering of the profiles to the patterns. A higher ADI denotes a better clustering. There should be only 5 real patterns extracted. As can be seen, a CYs measure of 0.9 and PCT of 5 yielded the most valid clustering of the data into 5 patterns. Of those parameters together, St1 = 5 gave the best results (data not shown). Fixing step 1 parameters at these values and varying step 2 CYs from 0.5 to 0.9 and St2 from 1 - 5 revealed that a CYs of 0.8 and St2 = 3.5 resulted in the best results (Figure 4B). Although a CYs of 0.9 yielded a higher ARI, the cardinality of the gene profiles in the patterns were extremely low (only 76 of 1000 extracted compared to 441 of 1000).

**EPIG-Seq extraction of profiles from simulated data**

As seen in Figure 5A, through the EPIG-seq processing of the simulated data with the optimal parameter settings, all five real patterns were successfully extracted without extracting the “noise pattern” pattern6. Shown is the RPM normalized count data of the profile with the highest PCS compared to the average of the baseline samples. In the figure showed, samples belonging to the different groups are color coded with: group1 (labeled as “Baseline”) in red, group2 in green; group3 in blue and group4 in purple. Within each group, although the expected mean under a condition in a given pattern was defined at the simulation, we did observe clear oscillation of replicated sample data points owing to the impact imposed by dispersion in the model. Regardless, neither the artificially introduced noises nor the dispersion in data prevented the EPIG-Seq from recovering all the simulated real patterns with p-values < 0.06. As show in Table 3, patterns 1 – 5 had 316, 189, 129, 207 and 130 gene profiles categorized to each respectively with an average sensitivity of 68% and average specificity of 93%. Heat map representation of the normalized data shows good overall categorization of the profiles to their respective patterns (left color bar in Figure 5 B). However, the right color bar revelas that for all patterns except B, there are some assignments of profiles to other pattrns that the one originally simulated to. This is expected, and reflected in the 68% sensitivity, since variation is introduced into the data by simulating the dispersion parameter from the normal breast tissue samples. Despite that, the distribution of the samples by PCA on the categorized genes from the five patterns indicates very good separation. EPIG extracted the patterns with sensitivity = XXX and specificity = YYYY. This is due to the fact that it extracted some noisy genes from Group F.

**EPIG-Seq analysis of real data**

To evaluate how EPIG-Seq performed on real but samples data, we applied the analyzed some of the TCGA breast cancer RNA-seq data. We randomly selected 10 lanes for each “subtype” [[11](#_ENREF_11)] and constructed balanced data of the four breast cancer subtypes plus normal breast tissue as a control. Using XXXX parameter settings, EPIG-seq extracted 4 patterns 4 or six patterns from either of the TGCA sampled data sets with between 192 and 344 genes in total per sampled set (Table 4 and Supplemental Figures 1-4). The general silhouette of the genes in the patterns reveal that they are relatively tight and separated well with a pattern extracted from sample 4 having the maximum silhouette (Table 4 and Supplemental Figures 5-8). Since the datasets were randomly selected from the pool, there were known and unknown factors that could introduce noises into the results. We adopted the adjusted mutual information (AMI) criterion to assess the agreement between the clustering results. From the clustering comparison (Table 5), good AMI agreement was observed, although comparisons between dataset 1 and dataset 3, and between dataset 2 and dataset 3 yielded low scores of 0.464 and 0.461 respectively. This points to possibly sampled data set 3 being somewhat of an outlier.

**Comparing against available methods EPIG ~~and ORIGEN~~ (combine with table 1 & 3, results 1) The rest should go to discussion, also need confusion matrix for both EPIG and EPIG-Seq**

To show the validity of the newly proposed EPIG-Seq method, we compared two highly regarded count level RNA-Seq analysis method: DEseq and SAMseq; we also compared the analysis with the EPIG on the RPM normalized data and ORIGEN on the squared rooted count data. To better address the advantages and disadvantages, all the comparison was performed on the aforementioned simulated dataset with known expected expression values. In the DEseq comparison, we followed author’s suggestion (ref) using “pooled” method to estimate the dispersion while using the locfit package to fit a dispersion-mean relation as described in the paper (ref). To get the “differentially expressed genes” (DEGs), we used a negative binomial test between each “experimental” group and the “basal” line. In other words, groups 2 – 4 were each compared to “group 1”; and DEGs were extracted with adjusted p-value < 0.05 and fold change requirement was at 1.5, 2.5, and 4 for the corresponding comparison respectively. In the end, we got 1130, 1006, and 758 DEGS from those three comparisons. Not surprisingly, all 758 DEGs from the comparison between group 4 and group 1 were the subset of 1006, which was got from the comparison between group 3 and group 1. Owing the fold change requirement set for comparison between group 2 and group1, the test results were a little inflated with approximate 10% as the falsely discovered. In addition to those falsely discovered DEGs, there were 96 DEGs pulled from the comparison between group3 and group1 were not included in the set between group2 and group1. On the other hand, about another 100 DEGs from comparison between group2 and group1 were not recovered in comparison between group3 and group1. Although, DEseq performed reasonably well in term of recovering the known DEGs, obviously, the co-expressed genes that primarily make the pattern 3 & 4 were easily missed with the pair-wise comparisons.

SAMseq (ref) is another efficient non-parametric testing methods extended from its predecessor SAM (ref) to directly apply to count level RNA-seq data. We compared the analysis with SAMseq using the mutli-class comparison with default parameter setting at FDR = 0.20. As expected, SAMseq worked very efficiently and reported 1256 DEGs including all those 1000 simulated genes that made up the first five informative patterns. Not surprisingly, it included 256 false discovered DEGs, which is a little more than the claimed false discovery rate (20%). When we lowered the FDR to 0.05, SAMseq reported 1035 DEGs (data now shown), and it again included all those 1000 simulated genes that made up the first five informative patterns. When we re-visited the method comparison metrics (table xx), SAMseq used similar approaches in its main statistical assumption, except the CYs as the correlation measurement in. It also is also proved very efficient on our simulated RNA-seq data.

EPIG performed reasonably well on the simulated data (figure xx) with RPM normalization. It was able to recovered the five main simulated patterns, however the patterns tended to contain noisy gene profiles, which were known not belonging to the patterns. i.e. the last pattern had 96 noise genes categorized, which was an FDR of 23.4%. ORIGEN was also able to recover the five known patterns (figure xx), but it also reported much more sub-patterns, which were not included in the simulation. (enough **Comparing to EPIG or ORIGEN??)**

**EPIG-Seq on TCGA data, figure 7 and table 6)**

**Discussion**

In contrast to common analysis strategies for genome level monitoring of responses from experimental specimens at different treatment conditions or perturbation stresses, we explored a route to reveal the hidden mechanism through assessing the systematic behavior crossing across multiple conditions. Our hypothesis is that a significant profile/pattern supported by a group of co-expressed genes could serve as a key signature of the responses to outside exposures. As shown in table 1, our newly proposed EPIG-Seq method, although shares some similarities with other highly-remarked methods, stands out by itself with its special advantages, not only provides researcher a new approach for monitoring the systematic responses across multiple conditions, it also avoids unjustified data pre-processing and normalization procedures.

**Briefly mention the comparison to SAMseq and DEseq**

To show the validity of the newly proposed EPIG-Seq method, we compared two highly regarded count level RNA-Seq analysis method: DEseq and SAMseq; we also compared the analysis with the EPIG on the RPM normalized data and ORIGEN on the squared rooted count data. To better address the advantages and disadvantages, all the comparison was performed on the aforementioned simulated dataset with known expected expression values. In the DEseq comparison, we followed author’s suggestion (ref) using “pooled” method to estimate the dispersion while using the locfit package to fit a dispersion-mean relation as described in the paper (ref). To get the “differentially expressed genes” (DEGs), we used a negative binomial test between each “experimental” group and the “basal” line. In other words, groups 2 – 4 were each compared to “group 1”; and DEGs were extracted with adjusted p-value < 0.05 and fold change requirement was at 1.5, 2.5, and 4 for the corresponding comparison respectively. In the end, we got 1130, 1006, and 758 DEGS from those three comparisons. Not surprisingly, all 758 DEGs from the comparison between group 4 and group 1 were the subset of 1006, which was got from the comparison between group 3 and group 1. Owing the fold change requirement set for comparison between group 2 and group1, the test results were a little inflated with approximate 10% as the falsely discovered. In addition to those falsely discovered DEGs, there were 96 DEGs pulled from the comparison between group3 and group1 were not included in the set between group2 and group1. On the other hand, about another 100 DEGs from comparison between group2 and group1 were not recovered in comparison between group3 and group1. Although, DEseq performed reasonably well in term of recovering the known DEGs, obviously, the co-expressed genes that primarily make the pattern 3 & 4 were easily missed with the pair-wise comparisons.

SAMseq (ref) is anther efficient non-parametric testing methods extended from its predecessor SAM (ref) to directly apply to count level RNA-seq data. We compared the analysis with SAMseq using the mutli-class comparison with default parameter setting at FDR = 0.20. As expected, SAMseq worked very efficiently and reported 1256 DEGs including all those 1000 simulated genes that made up the first five informative patterns. Not surprisingly, it included 256 false discovered DEGs, which is a little more than the claimed false discovery rate (20%). When we lowered the FDR to 0.05, SAMseq reported 1035 DEGs (data now shown), and it again included all those 1000 simulated genes that made up the first five informative patterns. When we re-visited the method comparison metrics (table xx), SAMseq used similar approaches in its main statistical assumption, except the CYs as the correlation measurement in. It also is also proved very efficient on our simulated RNA-seq data.

**Acknowledgements**

We thank the MicroArray Quality Control phase III (MAQC-III) Sequence Quality Control (SEQC) toxicogenomics working group members (TGxSEQC) Weida Tong, Charles Wang, Joshua Xu, Binsheng Gong and Zhenqiang Su for contributing the RNA-Seq mapped reads from the raw data. We appreciate the ORIGEN clustering results provided by Shyamal Peddada. This research was supported, in part, by the Intramural Research Program of the National Institutes of Health (NIH), National Institute of Environmental Health Sciences (NIEHS).

References:

1. Wang C, Gong B, Bushel PR, Thierry-Mieg J, Thierry-Mieg D, Xu J, Fang H, Hong H, Shen J, Su Z, et al: **The concordance between RNA-seq and microarray data depends on chemical treatment and transcript abundance.** *Nat Biotechnol* 2014, **32:**926-932.

2. Anders S, Huber W: **Differential expression analysis for sequence count data.** *Genome Biol* 2010, **11:**R106.

3. Robinson MD, McCarthy DJ, Smyth GK: **edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.** *Bioinformatics* 2010, **26:**139-140.

4. Zhou YH, Xia K, Wright FA: **A powerful and flexible approach to the analysis of RNA sequence count data.** *Bioinformatics* 2011, **27:**2672-2678.

5. Li J, Tibshirani R: **Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-Seq data.** *Stat Methods Med Res* 2013, **22:**519-536.

6. Li J, Witten DM, Johnstone IM, Tibshirani R: **Normalization, testing, and false discovery rate estimation for RNA-sequencing data.** *Biostatistics* 2012, **13:**523-538.

7. Lund SP, Nettleton D, McCarthy DJ, Smyth GK: **Detecting differential expression in RNA-sequence data using quasi-likelihood with shrunken dispersion estimates.** *Stat Appl Genet Mol Biol* 2012, **11**.

8. Chou JW, Zhou T, Kaufmann WK, Paules RS, Bushel PR: **Extracting gene expression patterns and identifying co-expressed genes from microarray data reveals biologically responsive processes.** *BMC bioinformatics* 2007, **8:**427.

9. Cao Y, Williams WP, Bark AW: **Similarity Measure Bias in River Benthic Aufwuchs Community Analysis.** *Water Environment Research* 1997, **69:**95-106.

10. Cancer Genome Atlas N: **Comprehensive molecular portraits of human breast tumours.** *Nature* 2012, **490:**61-70.

11. Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, et al: **Molecular portraits of human breast tumours.** *Nature* 2000, **406:**747-752.

12. Sorlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, Hastie T, Eisen MB, van de Rijn M, Jeffrey SS, et al: **Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications.** *Proceedings of the National Academy of Sciences of the United States of America* 2001, **98:**10869-10874.

13. Parker JS, Mullins M, Cheang MC, Leung S, Voduc D, Vickery T, Davies S, Fauron C, He X, Hu Z, et al: **Supervised risk predictor of breast cancer based on intrinsic subtypes.** *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 2009, **27:**1160-1167.

14. Soneson C, Delorenzi M: **A comparison of methods for differential expression analysis of RNA-seq data.** *BMC bioinformatics* 2013, **14:**91.

15. Anders S, Huber W: **Differential expression analysis for sequence count data.** *Genome biology* 2010, **11:**R106.

16. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, Dicuccio M, Edgar R, Federhen S, et al: **Database resources of the National Center for Biotechnology Information.** *Nucleic Acids Res* 2008, **36:**D13-21.

17. Hollander MW, Douglas A. ;Chicken, Eric: *Nonparametric Statistical Methods.* John Wiley & Sons; 2013.

18. McCullagh P, Nelder JA: *Generalized linear models.* 2nd edn. Boca Raton: Chapman & Hall/CRC; 1998.

19. Rand WM: **Objective criteria for the evaluation of clustering methods.** *Journal of the American Statistical Association* 1971, **66:**846-850.

20. Romano SB, James; Vinh, Nguyen xuan; Verspoorr, Karin: **Standardized Mutual Information for Clustering Comparisons: One Step Further in Adjustment for Chance.** *Proceedings of the 31st International Conference on Machine Learning, Beijing, China JMLR: W&CP* 2014, **32**.

Table 1. Comparison of EPIG-Seq vs. EPIG

|  |  |  |
| --- | --- | --- |
| Algorithm | EPIG | EPIG-Seq |
| Data type | Continuous | Count level |
| Distribution assumption | Gaussian | Poisson |
| Correlation measurement | Pearson's | CYs |
| Spread of the data | Variance | Dispersion |
| Magnitude | LogRatio | Wilcoxon test |
| Dynamic range | SignalToNoise | Variance-to-mean ratio (VMR) |

Table 2. Simulated RNA-seq data with mean fold chang values.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Pattern | # of genes | Group 1 | Group 2 | Group 3 | Group 4 |
| Group A | 200 | 1 | 1.5 | 2.5 | 4 |
| Group B | 200 | 1 | -1.5 | -2.5 | -4 |
| Group C | 200 | 1 | -1.5 | 4 | 2.5 |
| Group D | 200 | 1 | 4 | 2.5 | 1.5 |
| Group E | 200 | 1 | 4 | 4 | 4 |
| Group F | 19000 | 1 | 1 | 1 | 1 |

Table 3 Sensitivity and specificity of EPIG-seq extraction of simulated data.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Pattern** | **1** | **2** | **3** | **4** | **5** | **Total** | **Sensitivity (%)** | **Specificity (%)** |
| Group A | 7 | 132 | 11 | 0 | 0 | 150 | 88.00 | 66.00 |
| Group B | 0 | 0 | 169 | 0 | 0 | 169 | 100.00 | 84.50 |
| Group C | 135 | 26 | 5 | 0 | 0 | 166 | 81.33 | 67.50 |
| Group D | 0 | 0 | 0 | 15 | 166 | 181 | 91.71 | 83.00 |
| Group E | 9 | 23 | 0 | 111 | 43 | 186 | 59.68 | 55.50 |

Table 4 EPIG-Seq clustering efficiency of patterns extracted from the TCGA sampled data.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Sample #** | **GS** | **MS** | **# of Patterns** | **# of Genes** |
| 1 | 0.31 | 0.54 | 6 | 192 |
| 2 | 0.37 | 0.51 | 4 | 169 |
| 3 | 0.21 | 0.52 | 6 | 344 |
| 4 | 0.41 | 0.59 | 4 | 197 |

Table 5. Agreement of cluster extracted from the TGCA sampled data.

|  |  |
| --- | --- |
| **Samples Compared** | **Agreement** |
| 1 vs 2 | 0.770 |
| 1 vs 3 | 0.524 |
| 1 vs 4 | 0.452 |
| 2 vs 3 | 0.691 |
| 2 vs 4 | 0.751 |
| 3 vs 4 | 0.500 |

All comparisons based on AMI except for those with sample 2 where concordance was used.

Table 6 GO biological processes of MOA categorized genes.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Pattern #** | **# of Genes** | **Top GOBP** | **p-value** | **FDR** |
| 1 | 9 | GO:0006631 - Fatty acid metabolic process | 3.8E-06 | 4.4E-03 |
| 2 | 10 | GO:0055114 - Oxidation reduction process | 2.3E-02 | 2.1E+01 |
| 3 | 10 | GO:0042592 - Homeostatic process | 6.0E-02 | 5.5E+01 |
| 4 | 4 | **-** | **-** | **-** |

Table 7 KEGG enrichment of breast cancer categorized genes.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Pattern # | # of Genes | Breast Cancer Related Biology | p-value | FDR |
| 1 | 29 | GO:0042803 - Protein homodimerization activity (S100A16, CENPF, APOE, PLOD1, TOP2A) | 2.20E-03 | 2.50E+00 |
| 2 | 114 | KEGG:04512 - Extra cellular matrix-receptor interaction | 2.8E-04 | 3.0E-01 |
| 3 | 46 | KEGG:03320 - Peroxisome proliferator-activated receptor signaling pathway | 6.0E-02 | 5.5E+01 |
| 4 | 8 | CD59, ITGB1 and 5 ribosomal protein genes | **-** | **-** |

**Figure lengends**

Figure 1a EPIG-Seq workflow

Figure 1b CY example

Figure 2 Simulated RNA-Seq data with five patterns (pattern 1-5) carrying biological meaning and pattern 6 as background noise

Figure 3 Unsupervised analyses on gene profiles from simulated data with five patterns and background noise. 3A. A conventional PCA analysis was performed on pair-wise CYs as the covariance matrix. 3B. A hierarchical clustering reveals prominent patterns in simulated data

Figure 4 Parameter-space searching for the optimized default choice for end users. 4A. Parameter optimization in step one: there were four panels with labels on the top indicating the choice of location parameter St1 at [5-10]; within each panel, the x-axis shows the CYs as the similarity measure [0.5 – 0.9]; the color code was for the two-tailed dispersion cut off at 1 – 5 %. The y-axis in each panel was the Adjusted Rand Index [0-1]. 4B. Parameter optimization in step two, x-axis shows the CYs as the similarity measure [0.5 – 0.9]; the y-axis was the Adjusted Rand Index [0-1]; the color code was for indicating the choice of location parameter St2 at [2-5].

Figure 5 EPIG-Seq results running from the simulated dataset (aforementioned). 5A. The thumbnails of the 5 extracted simulated patterns extract by EPIG-Seq. Group1 (labeled as “Baseline”) in red, group2 in green; group3 in blue and group4 in purple. 5B. The hierarchical clustering reveals prominent patterns in simulated data. 5C. conventional PCA analysis was performed on pair-wise CYs as the covariance matrix.

Figure 6 The application of EPIG-Seq on the SEQC toxicity data. 6A. Thumbnails of the four significant extracted patterns extract by EPIG-Seq (the MOA was color coded) 6B MOA data extracted patterns PCA 6C. The hierarchical clustering (heatmap) that reveals prominent patterns in the SEQC data

Figure 7. Thumbnails of the four significant extracted patterns extract by EPIG-Seq on TCGA breast cancer data. (four the breast cancer subtypes and one normal group were color coded)