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

Jianying Li, Jeff W. Chou and Pierre R. Bushel

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 quasibinomial model estimates dispersion and the Wilcoxon rank sum Z-statistic indicates the magnitude of differential expression. EPIG-Seq categorizes genes to the patterns that they correlation with. Using simulated data, we show that EPIG-Seq is superior to EPIG and the CLICK clustering algorithm 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

Introduction

The advantages of RNA-sequencing (RNA-Seq) over microarray to measure gene expression have been widely reported (refs). Recently, methods have been developed to analyze RNA-Seq data based on normalization of the read counts or using the raw count data (refs). 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.) and there is no clear consensus of which normalization method is best (ref). On the other hand, using the raw read counts avoids the shortcomings of normalization but requires modeling of the data to estimate overdisperion, accounting for the read depth and filtering to avoid cases of inflated zeros. In addition, statistical models of count data based on Poisson 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 (Biostatistics. 2012 Jul;13(3):523-38. doi: 10.1093/biostatistics/kxr031, SAMSeq: Stat Methods Med Res. 2013 Oct;22(5):519-36. doi: 10.1177/0962280211428386). 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 detecting genes with similar expression profiles across biological conditions (ref). EPIG uses 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 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 (ref) to ascertain similarity between expression profiles, a quasibinomial model (ref) to estimate overdispersion in the data and the Wilcoxon rank sum Z-statistic (ref) 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 and the CLICK clustering algorithm 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. W also apply EPIG-Seq to a publicly available toxicogenomics data set and identify several co-expressed genes related to modes of action of the chemical agents.

Methods and Materials

**Data**

Simulated RNA-Seq data

To simulate RNA-Seq count data for each *k*th pattern where k{1,…,K}, we used a negative binomial distribution to model counts from a publicly available dataset of normal tissue as previously described (Soneson C, Delorenzi M. A comparison of methods for differential expression analysis of RNA-seq data. BMC Bioinformatics. 2013 Mar 9;14:91. doi:10.1186/1471-2105-14-91. PubMed PMID: 23497356). Briefly, let  denote the count of the *g*=1,…,*G* genes in the *m*=1,…,*M* groups of sample *s*{S1, S2,…,SN} such that:

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

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]. For each *g*th gene, an estimation of the sample mean and are obtained from a DESeq (ref) analysis of RNA-Seq raw count data from 10 randomly chosen human samples obtained from The Cancer Genome Atlas (TCGA) data portal (ref). 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. We let  denote the get of genes that are up-regulated and  denote the set of genes that are down-regulated. For all genes in the simulations, we let =

1.5 for m = 1 and 2, 2.5 for m = 3 and 4.0 for m = 4 in pattern 1;

4.0 for m = 1, 2.5 for m = 2 and 1.5 for m = 3 and 4 in pattern 2;

1.5 for m = 1 and 2, 4.0 for m = 3 and 2.5 for m = 4 in pattern 3;

1.5 for m = 1, 4.0 for m = 2, 2.5 for m = 3 and 1.5 for m = 4 in pattern 4;

4.0 for m = 1 to 4 in pattern 5; and

1.5 for m = 1 to 4 in pattern 6.

Finally, we set *G* = 200 for each *k*th pattern generating a total of 1200 simulated genes.

Sampled Cancer RNA-Seq data (Jianying add here)

Toxicogenomics RNA-Seq data

RNA-Seq data from the MicroArray Quality Control phase III (Sequence Quality Control (SEQC)) crowd source toxicogenomics (TGxSEQC) effort (ref) is available in the Sequence Read Archive (SRA) 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 (ref). 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 number of reads from sample *j* in profile *i* and *xkj* as the number of reads from sample *j* in profile *k*. As such, the similarity measure for count data as defined by Cao et al. (ref) is:



where



and *a* is the total number of samples with read counts in both profiles. *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. = 0 when two profiles are totally difference and = 1 when the two are identical.

\*\*\*\*\*\*\* Give an example with table and computation in the results to orient the reader \*\*\*\*\*\*\*\*

**Magnitude of change**

We define the strength of a gene expression profile’s signal according to the value of the z-statistic obtained from a Wilcoxon rank sum non-parametric test (ref) 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 estimator  for the difference between two groups (ref). Briefly,  is the median of the Walsh averages (ref), 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 including each difference with itself. Thus, for a set of N = (nx + ny) observations, there will be N(N+1)/2 Walsh averages. The median of all the Walsh averages is equal to. Hence,

.

\*\*\*\*\*\*\*\*\*\* The Hodges-Lehmann estimator needs to be implemented in EPIG-Seq \*\*\*\*\*\*\*\*\*\*\*\*

**Dispersion**

Count data is known to be dispersed (ref). 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 (McCullagh and Nelder 1989 generalized linear models). 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.

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. 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 average 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 sum of the Spearman rank correlations (**) among the *i*th profile (*xi*) to all other profiles (*xj*) assigned to pattern *k* (*Pk*) divided by the size of the pattern (i.e., the number of profiles in the pattern minus 1). 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 unless the profile exhibits low or high dispersion or a small magnitude of change.

Mutual Information (MI) between two random variables is dimensionless and with log base 2, interpreted in units of bits (ref). MI in terms of comparing two clustering outcomes (i.e., A and B), is the reduction in uncertainty about the cluster groupings in one outcome given knowledge of another (ref). A high MI value indicates that the two clusterings are similar where zero MI between the two cluster outcomes means that they are independent. Normalized MI = MI/(Entropy\_A+EntropyB)/2. NMI ranges between 0 and 1.

The EPIG-Seq algorithm pseudo code

Extract patterns with gene profile cluster size *Mt* > 5 and thresholds *CYst*, *St* and *VMRt* set at 0.95, 0.5 and the 5th percentile of log10 VMR respectively as follows:

**Extract patterns**

Calculate pairwise CYs correlations

Tally candidate patterns with >= Mt profile with Rt1

Delete patterns with low Mt

Delete profiles with low R to patterns

Model data to estimate dispersion and perform Wilcoxon Rank sum test

Delete pattern profiles with (low or high dispersion) or small magnitude (Wilcoxon rank sum Z-statistic < St1

Delete patterns with overlapping profiles to make them mutually exclusive

Remove patterns with similar correlation (>=0.8)

**Categorize profiles to patterns**

Use extracted patterns as seeds

Until no more moves (i.e., if 1-NMI < 0.001):

Correlate profiles to patterns using CYs similarity

Update patterns with profile with the highest average Spearman rank of correlation with other pattern profiles

Add profile to pattern with the highest correlation (>= Rt2)

Remove pattern profiles with (low or high dispersion) or small magnitude (Wilcoxon rank sum Z-statistic < St2)

Return final assignment.

Add brief description of methods for EPIG, ORIGEN and SAMSeq

Results

Discussion

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).