NIH Public Access
Author Manuscript
ACM Conf Bioinform Comput Biol Biomed Inform (2013). Author manuscript; available in PMC 2015 January 26.

Published in final edited form as:
ACM Conf Bioinform Comput Biol Biomed Inform (2013). 2013 December 31; 2013: 419–429. doi: 10.1145/2506583.2506589.

An Island-Based Approach for Differential Expression Analysis

Abdallah M. Eteleeb,
Department of Computer, Engineering and Computer, Science, University of Louisville, Louisville, KY, USA, ametel01@louisville.edu

Robert M. Flight,
Department of Chemistry, University of Louisville, Louisville, KY, USA, robert.flight@louisville.edu

Benjamin J. Harrison,
Department of Anatomical, Sciences and Neurobiology, University of Louisville, Louisville, KY, USA, b.harrison@louisville.edu

Jeffrey C. Petruska, and
Department of Anatomical, Sciences and Neurobiology, University of Louisville, Louisville, KY, USA, j.petruska@louisville.edu

Eric C. Rouchka
Department of Computer, Engineering and Computer, Science, University of Louisville, Louisville, KY, USA, eric.rouchka@louisville.edu

Abstract

High-throughput mRNA sequencing (also known as RNA-Seq) promises to be the technique of choice for studying transcriptome profiles. This technique provides the ability to develop precise methodologies for transcript and gene expression quantification, novel transcript and exon discovery, and splice variant detection. One of the limitations of current RNA-Seq methods is the dependency on annotated biological features (e.g. exons, transcripts, genes) to detect expression differences across samples. This forces the identification of expression levels and the detection of significant changes to known genomic regions. Any significant changes that occur in unannotated regions will not be captured. To overcome this limitation, we developed a novel segmentation approach, Island-Based (IB), for analyzing differential expression in RNA-Seq and targeted sequencing (exome capture) data without specific knowledge of an isoform. The IB segmentation determines individual islands of expression based on windowed read counts that can be compared across experimental conditions to determine differential island expression. In order to detect differentially expressed genes, the significance of islands (*p*-values) are combined using *Fisher's* method. We tested and evaluated the performance of our approach by comparing it to the existing differentially expressed gene (DEG) methods: CuffDiff, DESeq, and edgeR using two benchmark MAQC RNA-Seq datasets. The IB algorithm outperforms all three methods in both datasets as illustrated by an increased auROC.

Correspondence to: Eric C. Rouchka.

Keywords

Differential Expression; RNA-Seq; Alternative Splicing

1. INTRODUCTION

Current next-generation sequencing technologies have afforded researchers the ability to sequence known and unknown mRNA transcripts that can be either coding or non-coding using RNA-Seq and captureSeq methodologies. Using the captureSeq approach, Mercer *et al.* [19] were able to expand by 12% the number of exonic structures that did not belong to known models. This indicates the power of next-generation sequencing approaches in providing novel information about the complexity of transcripts. Others have used RNA-Seq to expand the knowledge of transcribed regions [8, 28], including long noncoding RNAs (lncRNAs) and microRNAs (miRNAs) [33, 34]. Perhaps the most well-known of these studies was performed by the ENCODE Consortium [5], which focused on understanding encoded elements within the human genome. The GENCODE group relied heavily on RNA-Seq data to improve the accuracy of protein-coding regions, pseudogenes, and noncoding regions in the human genome [11, 14, 12].

The advent of RNA-Seq has enabled researchers and scientists to study the transcriptome at an unprecedented rate and has lately become the standard technology for transcriptome analysis. It is based on the direct sequencing of complementary DNA (cDNA) [20]. An RNA-Seq experiment starts with the extraction of total RNA or a portion such as polyadenylated RNA [32]. The extracted RNA is then converted to a library of double stranded cDNA and sheared into small fragments. In the next step, adapters are attached to one or both sides of each cDNA fragment. Using next-generation sequencing platforms, each cDNA fragment is sequenced and a short sequence (read) from one end of the fragment (single-end tag) or from both ends (paired-end tag) is obtained. The obtained reads are mapped to the reference genome or transcriptome to measure the abundance of each transcript.

Most RNA-Seq approaches developed for differential expression analysis follow a similar workflow where mapped reads are summarized according to known biological features such as exons, transcripts, or genes which restricts the mapping of read sequences to existing annotations. Thus, reads that map to regions outside annotated features will not be captured even in well annotated genomes (e.g. human and mouse) [22] and consequently changes in those regions will be missed. Additionally, previously undetected cassette-based isoforms will be ignored and summarized accordingly to known isoform annotations. While using known annotations allows for insightful analysis of how gene expression change in differing conditions, it also is limiting in understanding how the gene structure itself might also change.

To illustrate this problem, Pickrell *et al.* [23] found about 15% of mapped reads were located outside annotated exons in their Nigerian HapMap samples. Alicia *et al.* [22] showed an example of transcripts that fall outside annotated exons for the RNA binding

protein 39 gene in LNCaP prostate cancer cells. Our own work, highlighted in Figure 1, shows an expression level for microarray data that shows differential expression outside of a known rat gene. This differential transcription would be ignored by current analysis methods, even though it has been experimentally determined this region is part of the upstream gene. Looking at the annotated mouse homology, it can be inferred that the 3' UTR extends into this region, even if there is no support from the current rat annotation. Further analysis of this differentially expressed transcript shows an association with axonal localization [10].

In order to alleviate the issues resulting from dependence on annotations, we propose a novel segmentation approach, *Island-Based (IB)*, for analyzing RNA-Seq and targeted sequencing (exome capture) data. The segmentation methodology determines individual islands of expression based on windowed read counts for each individual sample that are then overlapped and compared across samples. To detect DEGs, the IB uses Fisher's method to combine the significance of sub-islands corresponding to a chromosomal region.

2. RELATED WORK

The detection of which genes have changed significantly between biological samples requires the use of statistical hypothesis tests to model count data from RNA-Seq experiments. Existing methods have implemented different statistical methods for this purpose. Currently, most statistical models are based on parametric assumptions for modeling RNA-Seq data. Discrete probability distributions such as *binomial, Poisson* and *negative binomial (NB)* distributions have been used to model RNA-Seq count data [15]. In RNA-Seq studies that use a single source of RNA, the distribution of counts across technical replicates for the majority of genes was indeed Poisson [22, 15]. Early methods [18, 31] were developed to detect DEGs based on this distribution. However, the Poisson distribution suffers from the inability to capture biological variability within RNA-Seq data [22, 15, 30] since the variance of the Poisson distribution is equal to the mean. Since the variance of many genes is likely to exceed the mean, this results in over-dispersion. Thus, Poisson-based analyses using biological replicates will be prone to high false positive rates. To address over-dispersion and account for biological variability, methods such as edgeR [26], DESeq [1], baySeq [9], and Cuffdiff [29] have been developed based on the negative binomial distribution (NB) to model read counts. To test for DEGs, both DESeq and edgeR use a variation of the Fisher's exact test adopted for the negative binomial distribution. Cuffdiff compares the log ratio of gene expression in two conditions against the log ratio of one and calculates the test statistics as $T=\frac{E[log(Y)]}{Var[log(Y)]}$, where Y is the log ratio of the normalized counts between the two conditions $(Y=\frac{FPKM_a}{FPKM_b})$. baySeq employs an empirical Bayesian approach to determine DE between conditions. For every gene, baySeq estimates two models: one assumes the expression pattern is the same and a second assumes the expression pattern is different across conditions. Thus, the posterior likelihood can be estimated using the prior estimates and the likelihood of the distribution of the data to decide if a gene is differentially expressed.

3. MATERIALS AND METHODS

3.1 Approach

We developed a novel approach, IB, in an attempt to overcome the limitation mentioned above and detect expression differences in any genomic region regardless of whether a genomic annotation is available. This approach (detailed in Sections 3.3–3.6) splits the genome into small fixed non-overlapping regions (windows) which are merged, based on their read count density, into larger regions called *islands*. Adjacent regions with similar densities are merged together constructing an island in a process called *segmentation*. First, reads are mapped to the reference genome to generate perbase abundances (Figure 2, step 1 and step 2). Region construction begins by first summarizing per-base read counts over a fixed window to minimize small variance in coverage due to noise (Figure 2, step 3). The size of the window is allowed to vary using smaller window sizes (10–30bp). Once the genome is split into windowed regions, each region is then classified as *high* or textitlow density (where the density is based on the number of reads) using an average threshold *t*. The calculation of this threshold is adapted from Zang *et al.* [35] where regions with read counts above or equal to the threshold *t* are classified as *high* density regions and regions with read counts below the threshold are classified as *low* density regions. The threshold *t* is sample specific and is defined based on a user-defined *p-value* and the *Poisson* distribution as an approximation for the expected number of reads:

$$\sum_{k=t}^{\infty} P(k, \lambda) \leq p - value$$

where *k* is the number of reads in a window and λ represents the average number of reads across all regions in the genome and is calculated as $\lambda = wS_j/G$, where *w* is the region size, S_j is the total number of reads in experiment *j*, and *G* is the effective genome length. To construct *islands*, contiguous high density regions are merged into larger high density islands and similarly contiguous low density regions are merged into larger low density islands (Figure 2, step 4). Low density regions denote the start and end points for individual high density islands. Each high density island is allowed to include a number of *low* density regions based on a pre-defined cost threshold *c*. Figure 3 shows an example where one *low* density region is allowed in a given island. To test for differentially expressed islands, constructed islands are overlapped between samples and split into smaller islands called sub-islands where the start and stop locations are different. Each sub-island then comprises an overlapping region between samples that can be subsequently tested for differential expression between conditions using statistical tests such as *t-test* or *Wilcoxon test* (Figure 2, step 5). Sub-islands constructed from *low* density islands across samples are removed and we keep only sub-islands constructed from *high* density islands in at least one sample.

3.2 Datasets

3.2.1 MAQC Datasets—To test the performance of the IB approach, two datasets related to the MicroArray Quality Control Project [17] were obtained. The experiments in the two

datasets analyze two biological samples: Ambion's human brain reference (Brain) and Stratagene's human universal reference RNA (UHR) [3]. In both datasets, the two samples were prepared using one library preparation and sequenced in seven lanes and two ow-cells using an Illumina Genome Analyzer II (GAIIx). The first dataset was sequenced with RNA-Seq reads of length 35bp with only one biological replicate [3]. This dataset was obtained from NCBI's Sequence Read Archive (SRA) with Accession IDs: SRX016359 and SRX016367 for Brain and UHR respectively. The second dataset was sequenced with 50bp RNA-Seq read length with one biological replicate [21]. This dataset was obtained from SRA with Accession IDs: SRX027129 and SRX027130 for Brain and UHR respectively.

3.2.2 qRT-PCR Datasets—As part of the MAQC project, 1044 genes were selected to be assayed by qRT-PCR. The expression of those genes were quantitatively measured for Brain and UHR samples using TaqMan Gene Expression Assay [3, 30]. This data is used as a "gold-standard" to evaluate the performance of our approach for detecting DEGs obtained from Gene Expression Omnibus (GEO) with series ID GSE5350. Four replicates were obtained for Brain (GSM129638-GSM129641) and four replicates for UHR (GSM129642-GSM129645). We removed genes whose identifiers are not present in RefSeq which results in a total of 1033 genes. We follow Bullard *et al.* [3] and Wan *et al.* [30] for processing this data and compute the expression level of each gene for each replicate. Thus, for gene *i* at replicate *j*, the expression is defined as:

$$Y_{i,j} \equiv \frac{log_2(\Delta C_{i,j})}{log_2(e)}$$

where $\Delta C_{i,j} = C_{i,POLR2A} - C_{i,j}$ denotes the original qRT-PCR expression (C is the normalized threshold cycle number and POLR2A is the reference gene). This was done to transform the original expressions, which are in log base-2, to the natural logarithmic scale. The log-fold change is then defined as the difference of average across the four replicates $\bar{Y}_{UHR,j} - \bar{Y}_{Brain,j}$. To define the DE genes (positive set) and non-DE genes (negative set), genes with absolute log-fold change > 2 are considered DE genes and genes with absolute log-fold change < 0.2 are considered as non-DE genes. Out of 1033 genes, 309 genes fall in the positive set and 174 genes in the negative set. Genes with absolute log-fold change > 0.2 and < 2 are discarded and not used in this study.

3.3 Mapping Read Sequences and Computing Per Base Counts

To generate per base abundances, short read sequences of the two samples (Brain and UHR) in both datasets are first mapped to the indexed reference genome (hg19) using Bowtie version 0.12.8 [16] with the default parameters allowing for two mismatches. The resulting SAM files are then converted via BAM into BED format (a tab-delimited text file that defines a feature track) and each file for each sample is split by chromosome and a per base count is computed for each chromosome separately using BEDTools [24]. Thus, if s_j represents sample j, then C_{s_j} represents the complete set of per base counts separately for each chromosome $C_{s_j} = \{C_{s_j,chr1}, C_{s_j,chr2}, \ldots, C_{s_jchr_k}\}$, where $C_{s_j,chr1}$ is the per base count for chromosome 1, $C_{s_j,chr2}$ is the per base count for chromosome 2, and so on for each of the

k chromosomes. The purpose of computing per base counts for each chromosome has two advantages: first the process is much faster than considering all chromosomes simultaneously. Second, the approach will have more exibility to work with specific chromosomes in case differential expression analysis needs to be performed for a particular chromosome.

3.4 Island Construction

In order to construct islands, a fixed window *w* with size 30bp is first applied to summarize read counts within a region for each sample. This can minimize small variances in coverage due to random uctuations in the per base counts. For each sample s_j, a set of regions R_{s_j} is constructed from the set of per base counts C_{s_j} for each chromosome. $R_{s_j} = \{R_{s_j,chr1}, R_{s_j,chr2}, \ldots, R_{s_jchr_k}\}$. Constructed regions are then classified as either a *high* (R_{high,s_j}) or *low* (R_{low,s_j}) density region using an average threshold *t* (see Section 3.1 for the computation of this threshold) with a *p*-value 0.05. Contiguous high density regions are merged to form high density islands and contiguous low density regions are merged to form low density islands. The high density islands are constructed from the set of the high density regions R_{high,s_j} and low density islands are constructed from the low density regions R_{low,s_j}. Thus, the complete set of islands in sample s_j is defined as $I_{s_j} = \{I_{s_j,chr1}, I_{s_j,chr2}, \ldots, I_{s_j,chr_k}\}$. In this study, we allow only one low density region to be included in each island constructed from high regions ($c = 1$) in order to prevent over-segmentation.

3.5 Testing Differential Sub-island Expression

The primary goal of sub-island DE testing is to test the null hypothesis H_0 that a sub-island has the same expression level between samples versus the alternative hypothesis H_1 that a sub-island has a significant difference between samples. To do that, the set of islands *I* of the two samples are first overlapped and compared between the two samples in both datasets. The comparison process starts with overlapping the islands constructed in the last step for each sample generating smaller regions called *sub-islands* which have different start and stop locations. Those sub-islands are the regions that will be tested for differential expression across samples. The sub-island constructed from *low* density islands in both samples were removed and only sub-islands constructed from *high* density islands in at least one sample were kept. To conduct an accurate comparison, read counts are first normalized based on the total number of mapped reads in each sample. We call this normalization method *Islands Per Million (IPM)* (which is an adaptation form of the well-known method *transcripts per million (TPM)*). The IPM method is defined as:

$$IPM = \frac{K_{ij}}{M_j} \times 10^6$$

where K_{ij} is the read counts of sub-island *i* in sample *j* and M_j is the total number of mapped reads for sample *j*. Since sub-islands tested for DE have the same length, it is not needed to include the sub-island length in the normalization computation. To test for DE sub-islands across the two samples, two statistical tests *Welch's t-test* and *Wilcoxon* test are used on the normalized *IPM* values. The Welch's t-test is an adaptation of the well-known *Student's t-*

test in which the test assumes that the two samples have unequal variances. The test statistic, *T*, and degree of freedom are defined as:

$$t=\frac{\bar{X}_1-\bar{X}_2}{\sqrt{\frac{s_1^2}{N_1}+\frac{s_2^2}{N_2}}}, df=\frac{\left(\frac{s_1^2}{N_1}+\frac{s_2^2}{N_2}\right)^2}{\frac{s_1^4}{N_1^2(N_1-1)}+\frac{s_2^4}{N_2^2(N_2-1)}}$$

where $\bar{X_i}$ represents the *i*th sub-island mean, s_i^2 represents the *i*th sub-island variance, and N_i represents the *i*th sub-island size. As an alternative to the parametric Welch's t-test, we also perform a nonparametric Wilcoxon Rank-Sum test (also known as Mann-Whitney U test). The Wilcoxon test is based on the ranks of the observations and not the raw data. The test-statistic, *T*, is calculated as the sum of ranks in the smaller group. To understand this test, suppose that $N_{1i}, \ldots, N_{ni}$ represents the read counts of islands *i* in *n* samples. If R_{ij} (*N*) is the rank of all counts N_{ij}, the Wilcoxon test statistic, *T* then is defined as:

$$T_i=\sum R_{ij}(N)-\frac{n_1(n_1+1)}{2}$$

where $N = n_1 + n_2$ and n_1 is the length of the sub-island in the first sample.

3.6 Combining the Significance of Sub-Islands

Since all the methods being compared report results at the gene level, an overall *p*-value for a gene needs to be generated from the sub-island *p*-values. Therefore, the *p*-values of the sub-islands that overlap with each gene in the gene set are combined using Fisher's method [6]. Fisher's method computes the overall *p*-value *p* by combining the significance of multiple tests using the formula:

$$-2\sum_{i=1}^{k} ln(p_i)=\chi^2_{2k,p}$$

where p_i is the *p*-value of the *i*th sub-island and *k* is the number of sub-islands tested. Thus, if none of the sub-islands are DE, the *p*-values p_i are independent and uniformly distributed on the unit interval $p_i \sim U(0, 1)$ which indicates the null hypothesis H_0 is true. Hence, $\chi^2_{2k,p}$ denotes the upper *p* point of the probability of a chi-squared distribution with 2*k* degrees of freedom [30, 4, 13].

3.7 Evaluation of Island-Based Approach for Detecting DEGs

To test the performance of the Island-based approach, the Receiver Operating Characteristic (ROC) is used to evaluate the relationship between sensitivity (TPR) and specificity (FPR). We evaluate the results of the IB approach for detecting DEGs by comparing it to three widely used methods: *Cuffdiff, DESeq*, and *edgeR*. For each method, the *p*-value is used to determine which genes are DE and which genes are not. Thus, for a given *p*-value threshold,

we consider genes with *p*-values smaller than or equal to the threshold as DE genes and genes with *p*-values greater than the threshold are non-DE genes. By using the qRT-PCR data as a "gold-standard", the predicted results are compared to the set of 483 genes (309 in the positive set and 174 in the negative set) and true positive rate (TPR) and false positive rate are calculated. These two measures are computed as follows:

$$TPR=\frac{TP}{TP+FN}, FPR=\frac{TN}{TN+FP}$$

where TP denotes the true positive, FN is the false negative, and TN is the true negative sets. Using this information, we generate ROC curves for all methods based on different *p*-value cutpoints using both datasets I and II. We use the area under the ROC curve (AUC) to measure the accuracy of each method and evaluate the performance for detecting DEGs. The area under the curve is calculated using the trapezoidal rule.

4. RESULTS AND DISCUSSION

Using the two MAQC benchmark datasets, we first present the results using the IB approach by applying two statistical tests *Welch's t-test* and *Wilcoxon test*. Then we compare the results of our approach to *Cuffdiff*, *DESeq*, and *edgeR*.

4.1 Construct Islands and Test for DE Sub-islands

To construct islands, short read sequences of the two samples Brain and UHR in each dataset were first mapped to reference genome (hg19) using Bowtie with the default parameters allowing for two mismatches. To construct regions, we applied a window of size 30bp. To classify regions, the average thresholds for the two samples were computed using a *p*-value of 0.05 resulting in $t = 3$ for both samples. Thus, regions with read count above or equal to 3 reads were classified as *high* density regions and regions with read counts below 3 reads were classified as *low* density regions. Using $t = 3$ and $c = 1$ (c is the cost threshold), islands were constructed for Brain and UHR samples for each dataset. Table 1 shows detailed information about the constructed islands for each sample in the two datasets. Table 1 indicates that the average length of *low* density islands is much larger than the average length of *high* density islands which agrees with the fact that a large portion of the human genome (about 98%) is non-coding and only about 2% is coding regions. Thus, the coding regions (or in this case, transcribed regions) should fall within high density islands. To test for DE sub-islands, two statistical tests *Welch's t-test* and *Wilcoxon* test were applied to compute the test statistics T and the *p*-values.

4.2 Evaluation and Comparison

The performance of IB was evaluated using the benchmark RNA-Seq datasets for the Brain and UHR samples. The portion of the qRT-PCR data that we selected in Section 3.2.2 with 309 genes in the positive set (true DE) and 174 genes in the negative set (true non-DE) was used to compare the results of the IB approach to the other DE methods. Since our approach

is based on combining the *p*-values of sub-islands that overlap with the genes, for all methods, the *p*-value was used as a measure of significance in this study.

When we computed the overall *p*-values for the two sets of genes, using dataset I, for the true DE set with 309 genes, nine genes were missing (none of the sub-islands overlapped with those genes) and 22 genes were missing from the true non-DE set with 174 genes. Per base counts for each of these missing genes were checked and it was determined they have low counts and consequently their corresponding sub-islands were classified as *low* density and were removed. To verify this conclusion, we compared the counts in Cuffdiff and in the count table used as an input to DESeq and edgeR. We found a strong agreement between our approach and the other methods in terms of low read counts. For instance, Cuffdiff reported that out of the nine missing genes, eight genes were not tested (NOTEST) indicating there were too few counts to perform a significance test and out of 22 missing genes, 20 were not tested for the same reason. Giving this strong evidence these genes are not DE between the two samples, they were treated as non-DE genes and counted as *false negatives (FN)* for the nine missing genes and as *true negatives (TN)* for the 22 genes. Similarly when we used our approach with dataset II, eight genes were missing in the positive set and 25 genes were missing in the negative set due to low counts. Results from Cuffdiff supported our approach in that Cuffdiff described all those genes as NOTEST which indicate the low counts. We performed the same filtering for dataset II and included those genes in the *false negative* set and *true negative* set in the calculation of true positive and false positive rates.

In order to compare our approach with other existing DE methods, we performed differential expression analysis for the same MAQC datasets (I and II) using Cuffdiff, DESeq, and edgeR and computed the *p*-values for the set of 483 (309+174) genes. With the exception of Cuffdiff, the differential expression analysis of DESeq and edgeR were performed using the same count table of all genes annotated in RefSeq. This count table was generated using `htseq-count` version 0.5.4p1 with the same RefSeq GTF file downloaded from the UCSC genome browser.

For the set of 483 genes, first we looked at the *p*-value distribution (Figure 4) generated by each method using dataset I and dataset II. Using a *p*-value cutoff ≤ 0.05 (5%), we could observe that our approach performs well in detecting the true DE genes whereas it performs slightly worse in detecting the true non-DE genes. This is illustrated in Figure 4 where the *p*-value histograms of the IB is highly skewed to 0 indicating that a large number of true DE genes will be detected (giving the fact that approximately 65% of the gene set falls in the positive set). Since this histogram is slightly skewed far from 0, there is a high possibility that the IB approach will not perform well in detecting true non-DE genes. In contrast, the *p*-value histograms of Cuffdiff, DESeq, and edgeR were not as highly skewed to 0 as the IB approach which indicates the likelihood of not performing well in detecting true DE genes. However, the histograms show a moderate shift toward 1 meaning those methods will perform well in detecting true non-DE genes.

Although Cuffdiff, DESeq, and edgeR did not perform well in detecting true DE genes, they were excellent in detecting almost the complete set of the true non-DE genes with 172, 173,

171, respectively out of 174. Table 2 shows the number of true positive (TP) and true negative (TN) genes detected by each method using a $p\text{-value} \leq 0.05$ and Figure 5 shows the bar graph of those numbers.

Table 2 indicates that the IB approach performs well in detecting TP genes whereas Cuffdiff, DESeq, and edgeR were much better in detecting the TN genes. As we see in Table 2 and Figure 5, the IB approach was not able to detect a high number of true non-DE genes like other methods. DE-Seq and edgeR performed similarly since both methods use similar statistical tests (a form of Fisher's exact test) and both model read counts by using a negative binomial distribution (NB). Thus, their results were close to each other. According to the DESeq documentation, DESeq is conservative in detecting DE genes. Thus, it is of no surprise we do not see a large number of true DE genes detected by DE-Seq. To plot the ROC curves for the four methods, we set different thresholds of the *p*-values and calculated the true positive rate (TPR) and false positive rate (FPR) for each method. Thus, a method that performs better will give a ROC curve with higher TPR than other methods with the same value of FPR. We computed the AUC and use it as a measure to compare the performance of each method. Figure 6 shows the ROC curves of the four methods on the two MAQC datasets. By looking at the AUC of each method in Figure 6, it is clear the two versions of our approach (*Welch's t-test* and *Wilcoxon*) outperform other methods in both datasets. The IB approach using the Wilcoxon test performed the best among the four methods with AUC = 0.897 for dataset I and AUC= 0.908 for dataset II. The IB approach using Welch's t-test also performs well similar to IB Wilcoxon in both datasets with AUC = 0.895 for dataset I and AUC = 0.871 for dataset II. Cuffdiff performed better than DESeq and edgeR but not as well as IB approach.

We further looked at the number of differentially expressed genes shared between each pair of methods (Table 3) for both datasets. This gives an indication on the level of agreement between each pair of methods in detecting the true DE genes. Table 3 indicates a strong agreement in detecting true DE genes between the two versions of IB approach. Compared to other methods, the two versions of our approach were able to detect almost all true DE genes detected by other methods for the two datasets. For instance, out of 190 true DE genes detected by CuffDiff, the IB approach was able to detect 183 and 178 respectively for the two versions in the first dataset. In the second dataset, the number is even higher as shown in Table 3. The same observation is applied for DESeq and edgeR where almost all true DE genes detected by those methods were also detected by our approach. This indicates the set of DE genes found by the IB approach contains a large number of DE genes found by other methods. To look at the overlap between all methods and determine the number of true DE genes and true non-DE gene shared between all methods, Figures 7 and 8 depicts the complete overlap between the number of TP and TN genes detected by each method.

One caveat with the choice of the MAQC datasets is the ratio of DE to non-DE genes is skewed in comparison to typical datasets where it might be expected that only 5–10% of the genes are differentially expressed. These datasets were chosen for comparative purposes since they contain experimental validation for differentially expressed genes. That being said, we have also applied the IB approach to whole transcriptome RNA-Seq data as well

(results not shown) for the datasets discussed in Figure 1. Initial results suggest a similar performance to the MAQC data with the majority of novel islands detected within or in close proximity to known transcribed regions.

5. CONCLUSION

In this paper, we proposed a novel approach for detecting differential expression in genome regions that does not rely on genomic annotations. The key idea of this approach is the segmentation methodology in which individual islands of expression are constructed based on windowed read counts and compared across experimental conditions to determine differential island expression. We illustrated how this approach is used to detect differences in expression without requiring any prior knowledge of isoforms where the only input to this approach is the raw data (short reads). To detect DEGs, Fisher's method for combining the significance of multiple tests was used. To evaluate the performance of the IB approach, we compared our results to three widely used methods for differential expression analysis using two benchmark MAQC RNA-Seq datasets. The IB approach was able to detect a high number of true DE genes using p-value ≤ 0.05 and performed the best among the four methods. However, the performance of detecting the true non-DE genes was not as good as we expected. Although the approach has detected a reasonable number of the true non-DE genes, it was not as high as the other methods considered. Considering the results we have, the performance of the IB approach can be considered on some level acceptable. However, it still leaves room for improvement in detecting true non-DE genes. There are several factors that affect the final result of the IB approach including but not limited to:

- The number of *low* density regions included in each *high* density island.
- The threshold t that used for classifying the regions into *high* and *low* density regions and its p-value.
- The window size used for splitting the genome.

Thus, several questions arise in this regard such as determining the best p-value to compute the threshold t, the optimal window size w, and the best cost threshold c for determining the number of *low* density regions to be included in *high* density islands. There are no clear answers and therefore these need to be investigated extensively by performing several experiments in order to determine the best values of those parameters.

The main focus of this research was to extend the knowledge of differentially expressed regions outside of known annotations. While this may be a fruitful approach for *de novo* transcriptome discovery, we have yet to compare it to *de novo* transcriptome assemblers such as Trans-Abyss [25], Oases [27], or Trinity [7]. This is due to the fact that our IB approach as currently constructed is a mapping-based methodology in contrast to these assembly-based methods. In the future, we will consider the applicability of the IB approach to *de novo* transcript assembly as well.

ACKNOWLEDGEMENTS

This work was partially funded by the Paralyzed Veterans of America Fellowship (2579 to BJH), Kentucky Spinal Cord and Head Injury Research Trust (9-12A and 10-10 to JCP) and National Institute of Health (NIH) grants P20GM103436, P20RR016481, 3P20RR016481-09S1, and R21NS080091. Its contents are solely the responsibility of the authors and do not represent the official views of the funding organizations.

REFERENCES

1. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. 2010 Oct.11(10):R106. 2010, DOI=http://dx.doi.org/10.1186/gb-2010-11-10-r106. [PubMed: 20979621]
2. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Statist. Soc. B. 1995 Oct; 57(1):289–300. 1995, DOI=http://dx.doi.org/10.2307/2346101.
3. Bullard JH, Purdom E, Hansen KD, Dudoit D. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC Bioinformatics. 2010 Feb.11:94. 2010, DOI=http://dx.doi.org/10.1186/1471-2105-11-94. [PubMed: 20167110]
4. Cousins RD. Annotated bibliography of some papers on combining significances or *p*-values. Available at arXiv: 0705.2209v2 [physics.data-an]. 2008 Dec. 2008.
5. ENCODE Project Consortium. A user's guide to the encyclopedia of DNA elements (ENCODE). PLoS Biol. 2011 Apr.9(4):e1001046. 2008. DOI=http://dx.doi.org/10.1371/journal.pbio.1001046. [PubMed: 21526222]
6. Fisher, RA. Statistical methods for research workers. Oliver and Boyd, Edinburgh, London: 1970.
7. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. 2011 May; 29(7):644–652. 2011, DOI=http://dx.doi.org/10.1038/nbt.1883. [PubMed: 21572440]
8. Halvardson J, Zaghlool A, Feuk L. Exome RNA sequencing reveals rare and novel alternative transcripts. Nucleic Acids Res. 2013 Aug; 41(1):e6–e6. 2013, DOI=http://dx.doi.org/10.1093/nar/gks816. [PubMed: 22941640]
9. Hardcastle TJ, Kelly KA. baySeq: Empirical Bayesian methods for identifying differential expression in sequence count data. BMC Bioinformatics. 2010 Aug.11(1):422. 2010, DOI=http://dx.doi.org/10.1186/1471-2105-11-422. [PubMed: 20698981]
10. Harrison BJ, Flight RM, Gomes C, Venkat G, Ellis SR, Sankar U, Twiss JL, Rouchka EC, Petruska JC. IB4-binding sensory neurons in the adult rat express a novel 3'UTR-extended isoform of CaMK4 that is associated with its localization to axons. J. Comp. Neurol. 2013 epub ahead of print. DOI=http://dx.doi.org/10.1002/cne.23398.
11. Harrow J, Denoeud F, Frankish A, Reymond A, Chen CK, Chrast J, Lagarde J, Gilbert JG, Storey R, Swarbreck D, Rossier C, Ucla C, Hubbard T, Antonarakis SE, Guigo R. GENCODE: producing a reference annotation for ENCODE. Genome Biol. 2010 Aug.7(Suppl 1):S4. 2006. DOI=http://dx.doi.org/10.1186/gb-2006-7-s1-s4. [PubMed: 16925838]
12. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, Barnes I, Bignell A, Boychenko V, Hunt T, Kay M, Mukherjee G, Rajan J, Despacio-Reyes G, Saunders G, Steward C, Harte R, Lin M, Howald C, Tanzer A, Derrien T, Chrast J, Walters N, Balasubramanian S, Pei B, Tress M, Rodriguez JM, Ezkurdia I, van Baren J, Brent M, Haussler D, Kellis M, Valencia A, Reymond A, Gerstein M, Guigo R, Hubbard TJ. GENCODE: The reference human genome annotation for The ENCODE Project. Genome Res. 2012 Sep; 22(9):1760–1774. 2012, DOI=http://dx.doi.org/10.1101/gr.135350.111. [PubMed: 22955987]
13. Hess A, Tyer H. Fisher's combined *p*-value for detecting differentially expressed genes using Affymetrix expression arrays. BMC Genomics. 2007 Apr.8:96. 2007, DOI=http://dx.doi.org/0.1186/1471-2164-8-96. [PubMed: 17419876]

14. Howald C, Tanzer A, Chrast J, Kokocinski F, Derrien T, Walters N, Gonzalez JM, Frankish A, Aken BL, Hourlier T, Vogel JH, White S, Searle S, Harrow J, Hubbard TJ, Guigo R, Reymond A. Combining RT-PCR-seq and RNA-seq to catalog all genic elements encoded in the human genome. Genome Res. 2012 Sep; 22(9):1698–1710. 2012, DOI=http://dx.doi.org/10.1101/gr.134478.111. [PubMed: 22955982]
15. Kvam VM, Liu P, Si Y. A comparison of statistical methods for detecting differentially expressed genes from RNA-seq data. Am. J. Bot. 2012 Feb; 99(2):248–256. 2012, DOI=http://dx.doi.org/10.3732/ajb.1100340. [PubMed: 22268221]
16. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009 Mar.10(3):R25. 2009, DOI=http://dx.doi.org/10.1186/gb-2009-10-3-r25. [PubMed: 19261174]
17. MAQC Consortium. The MicroArray Quality Control (MAQC) project shows inter- andintraplatform reproducibility of gene expression measurements. Nat. Biotechnol. 2006 Sep; 24(9):1151–1161. 2006, DOI=http://dx.doi.org/10.1038/nbt1239. [PubMed: 16964229]
18. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. Genome Res. 2008 Sep; 18(9):1509–1517. 2008, DOI=http://dx.doi.org/10.1101/gr.079558.108. [PubMed: 18550803]
19. Mercer TR, Gerhardt DJ, Dinger ME, Crawford J, Trapnell C, Jeddeloh JA, Mattick JS, Rinn JL. Targeted RNA sequencing reveals the deep complexity of the human transcriptome. Nat. Biotechnol. 2011 Nov; 30(1):99–104. 2011, DOI=http://dx.doi.org/10.1038/nbt.2024. [PubMed: 22081020]
20. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat. Methods. 2008 Jul; 5(7):621–628. 2008, DOI=http://dx.doi.org/10.1038/nmeth.1226. [PubMed: 18516045]
21. Nacu S, Yuan W, Kan Z, Bhatt D, Rivers CS, Stinson J, Peters BA, Modrusan Z, Jung K, Seshagiri S, Wu TD. Deep RNA sequencing analysis of readthrough gene fusions in human prostate adenocarcinoma and reference samples. BMC Med. Genomics. 2011 Jan.4:11. 2011, DOI=http://dx.doi.org/10.1186/1755-8794-4-11. [PubMed: 21261984]
22. Oshlack A, Robinson MD, Young MD. From RNA-seq reads to differential expression results. Genome Biol. 2010 Dec.11(12):220. 2010, DOI=http://dx.doi.org/10.1186/gb-2010-11-12-220. [PubMed: 21176179]
23. J. Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, Veyrieras J, Stephens M, Gilad Y, Pritchard JK. Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature. 2010 Apr; 464(7289):768–772. 2010, DOI=http://dx.doi.org/10.1038/nature08872. [PubMed: 20220758]
24. Quinlan AR, Hall IM. BEDTools: a exible suite of utilities for comparing genomic features. Bioinformatics. 2010 Mar; 26(6):841–842. 2010, DOI=http://dx.doi.org/10.1093/bioinformatics/btq033. [PubMed: 20110278]
25. Robertson G, Schein J, Chiu R, Corbett R, Field M, Jackman SD, Mungall K, Lee S, Okada HM, Qian JQ, Griffith M, Raymond A, Thiessen N, Cezard T, Butterfield YS, Newsome R, Chan SK, She R, Varhol R, Kamoh B, Prabhu A, Tam A, Zhao Y, Moore RA, Hirst M, Marra MA, Jones SJM, Hoodless PA, Birol I. De novo assembly and analysis of RNA-seq data. Nat. Methods. 2010 Nov; 7(11):909–912. 2010, DOI=http://dx.doi.org/. [PubMed: 20935650]
26. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010 Jan; 26(1):139–140. 2010, DOI=http://dx.doi.org/10.1093/bioinformatics/btp616. [PubMed: 19910308]
27. Schulz MH, Zerbino DR, Vingron M, Birney E. Oases: Robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics. 2012 Apr; 28(8):1086–1092. 2012, DOI=http://dx.doi.org/10.1093/bioinformatics/bts094. [PubMed: 22368243]
28. Sun K, Chen X, Jiang P, Song X, Wang H, Sun H. iSeeRNA: identification of long intergenic non-coding RNA transcripts from transcriptome sequencing data. BMC Genomics. 2013 Feb.14(Suppl 2):S7. 2013, DOI=http://dx.doi.org/10.1186/1471-2164-14-S2-S7. [PubMed: 23445546]
29. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts

and isoform switching during cell differentiation. Nat. Biotechnol. 2010 May; 28(5):511–515. 2010, DOI=http://dx.doi.org/10.1038/nbt.1621. [PubMed: 20436464]

30. Wan L, Sun F. CEDER: accurate detection of differentially expressed genes by combining significance of exons using RNA-Seq. IEEE/ACM Trans. Comput. Biol. Bioinform. 2012 Sep-Oct;9(5):1281–1292. 2012, DOI=http://dx.doi.org/10.1109/TCBB.2012.83. [PubMed: 22641709]
31. Wang L, Feng Z, Wang X, Wang X, Zhang X. DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. Bioinformatics. 2010 Jan; 26(1):136–138. 2010, DOI=http://dx.doi.org/10.1093/bioinformatics/btp612. [PubMed: 19855105]
32. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet. 2009 Jan; 10(1):57–63. 2009, DOI=http://dx.doi.org/10.1038/nrg2484. [PubMed: 19015660]
33. Yang JH, Li J, Jiang S, Zhou H, Qu LH. ChIPBase: a database for decoding the transcriptional regulation of long non-coding RNA and microRNA genes from ChIP-Seq data. Nucleic Acids Res. 2013 Nov; 41(D1):D177–D187. 2013, DOI=http://dx.doi.org/. [PubMed: 23161675]
34. Yang JH, Qu LH. deepBase: Annotation and discovery of microRNAs and other noncoding RNAs from deep-sequencing data. Methods Mol. Biol. 2012; 822:233–248. 2012, DOI=http://dx.doi.org/10.1007/978-1-61779-427-8_16. [PubMed: 22144203]
35. Zang C, Schones DE, Zeng C, Cui K, Zhao K, Peng W. A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. Bioinformatics. 2009 Aug; 25(15):1952–1958. 2009, DOI=http://dx.doi.org/. [PubMed: 19505939]

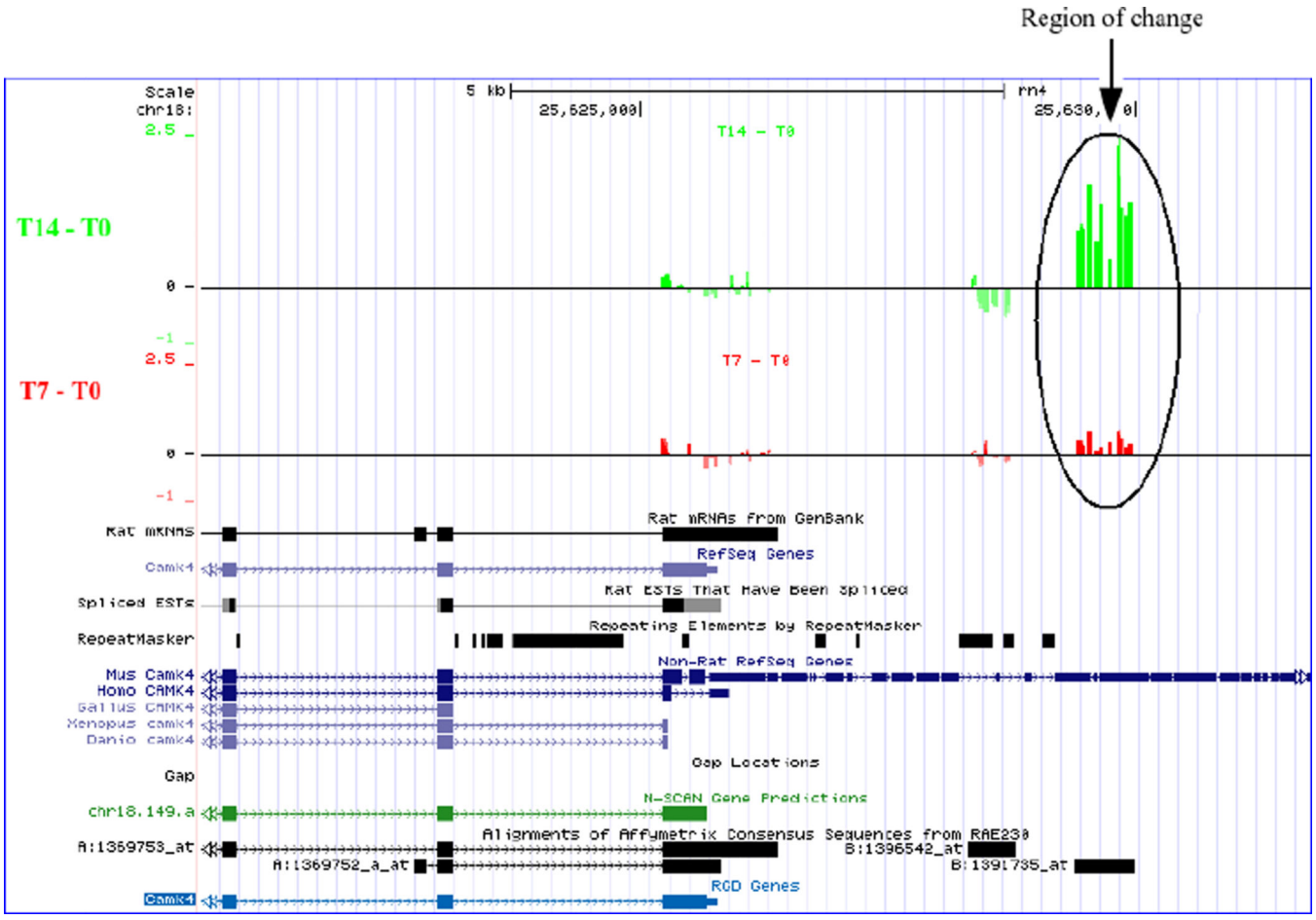


Figure 1.

Illustration of regions missed by current annotations. The 3' UTR region has a significant change between samples T14 vs T0 and samples T7 vs T0. The annotated mouse CaMK4 gene extends into this region. However, the corresponding rat CaMK4 gene annotation terminates prior to the differentially expressed region, which was subsequently verified to be part of the rat CaMK4 gene [10].

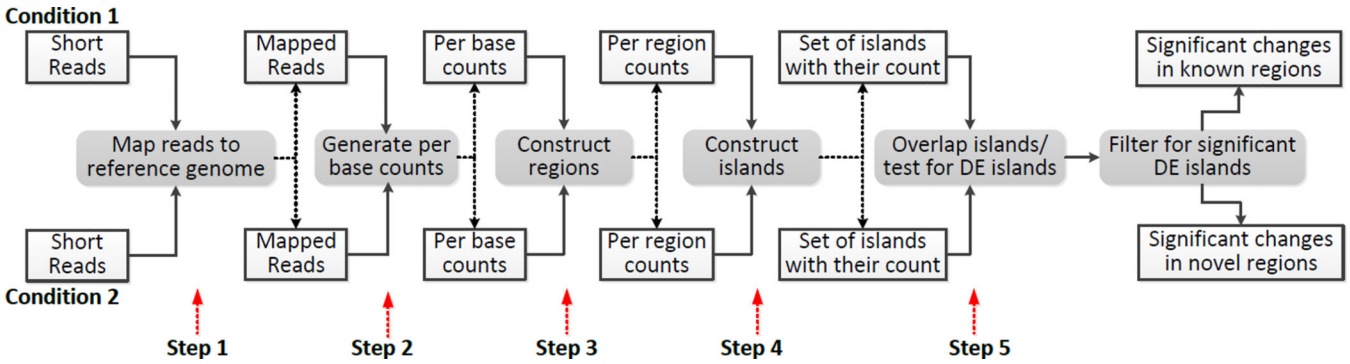


Figure 2.
Workow of the Island-Based approach.

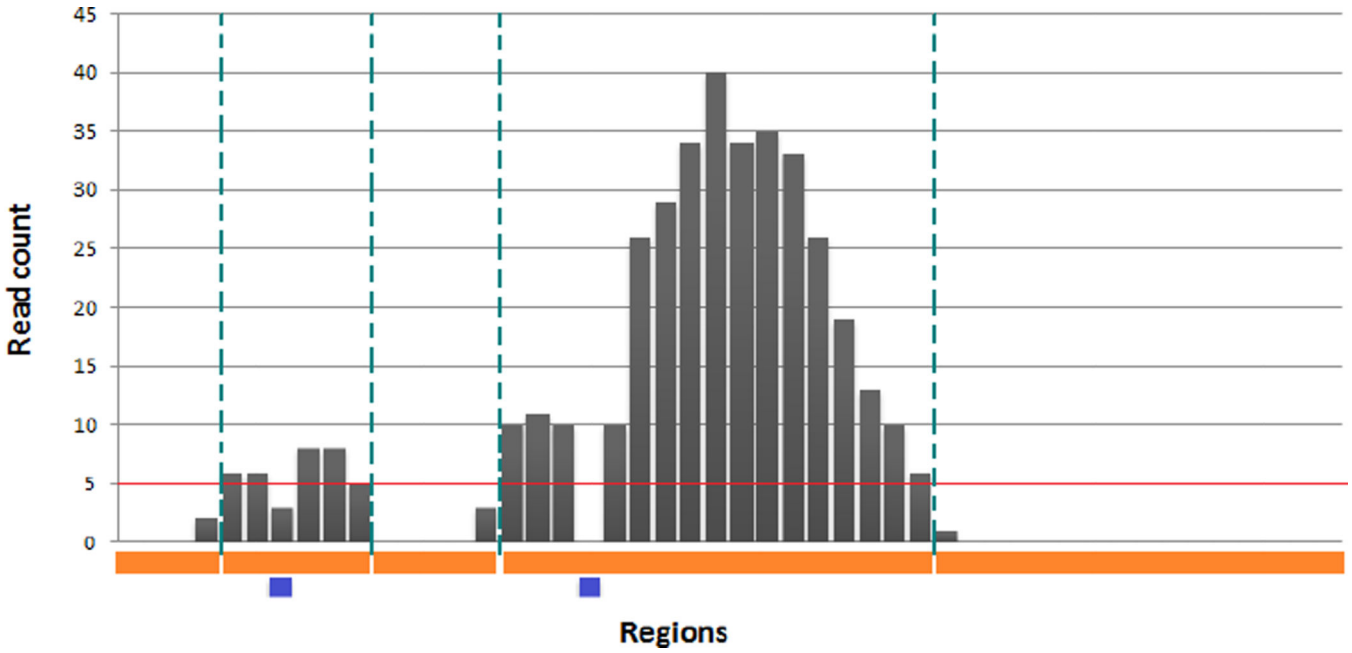


Figure 3.

Illustration of island definitions [35]. Regions are shown as genome coordinates along the x-axis with each bar representing one region. The y-axis denotes the read count for each region. The orange bar denotes the constructed islands using a threshold $t = 5$ (red line) and a cost threshold $c = 1$. The blue boxes show low density regions included in that island.

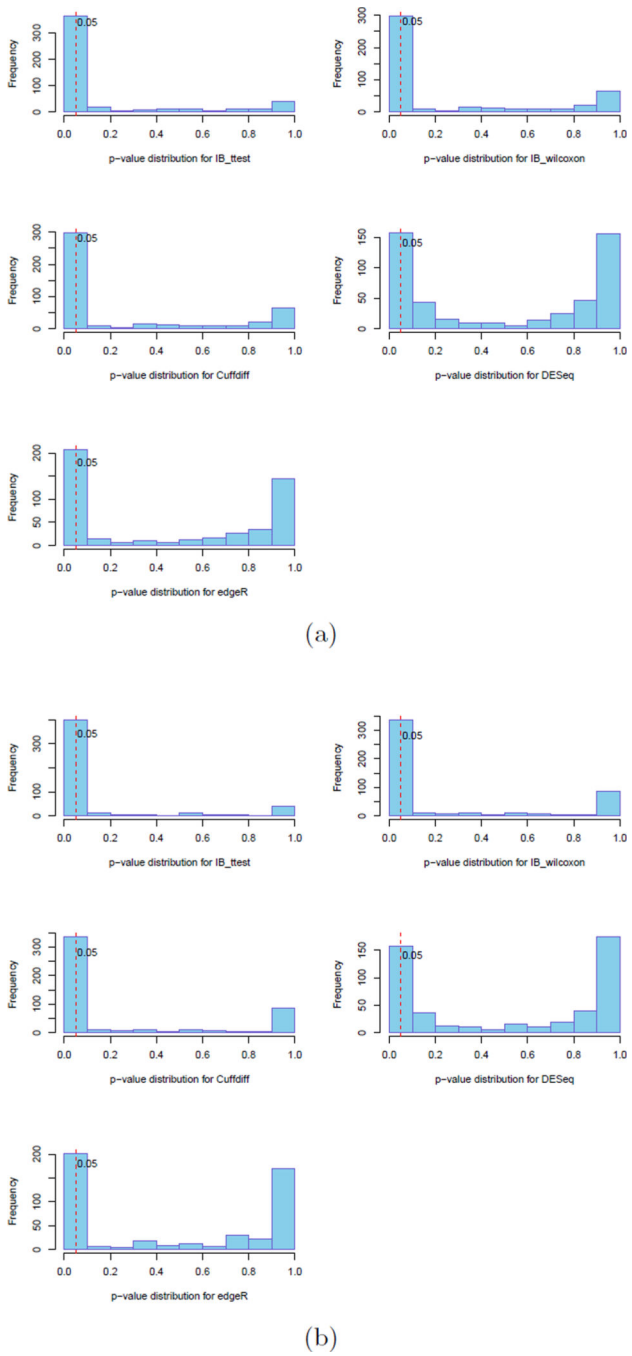


Figure 4.

The distribution of *p*-values for the four methods using qRT-PCR validated gene set: (a) The distribution of *p*-values for MAQC dataset I; (b) The distribution of *p*-values for MAQC dataset II.

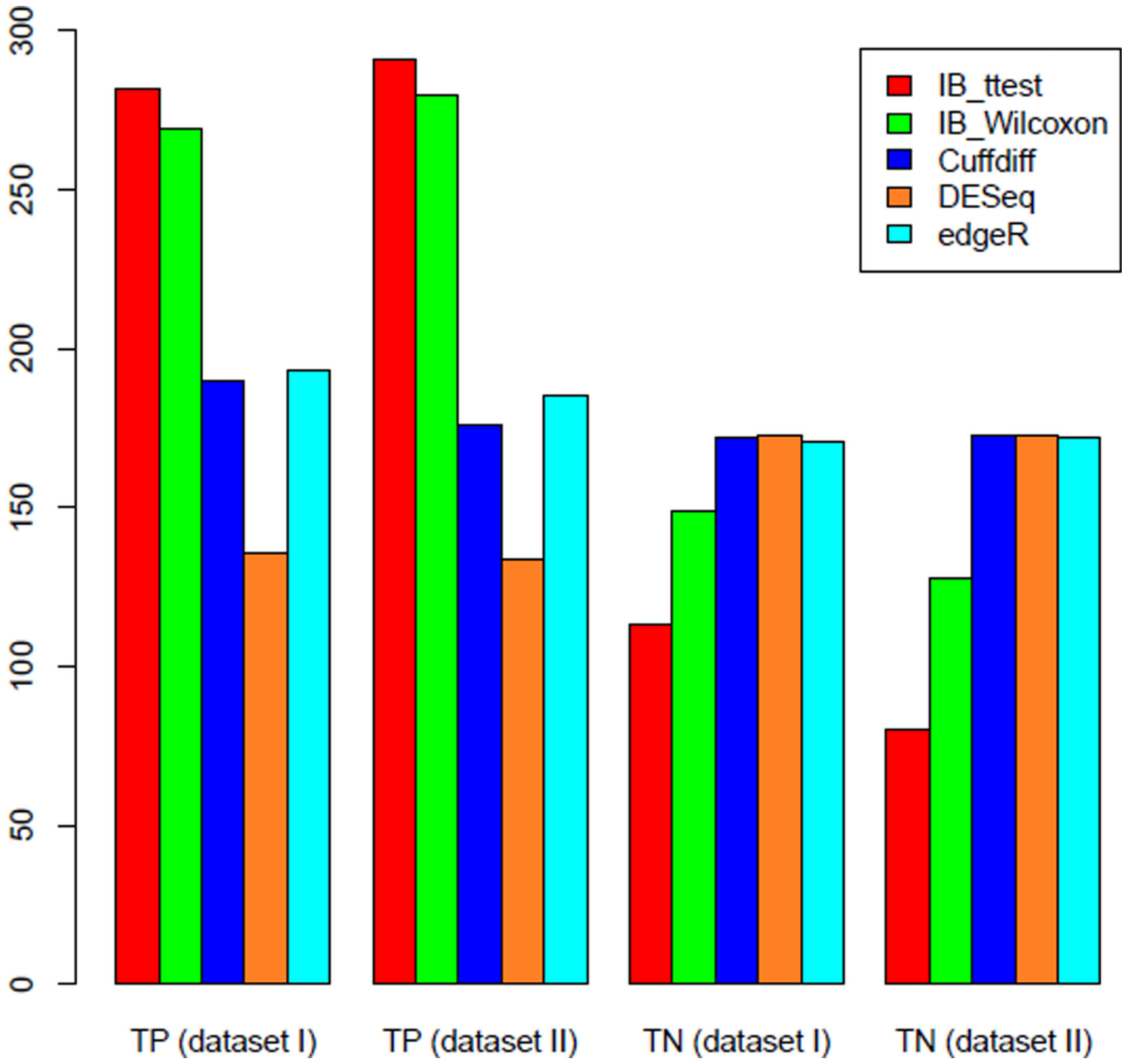


Figure 5.
The number of DE and non-DE genes detected by each method for the two datasets using *p*-value ≤ 0.05.

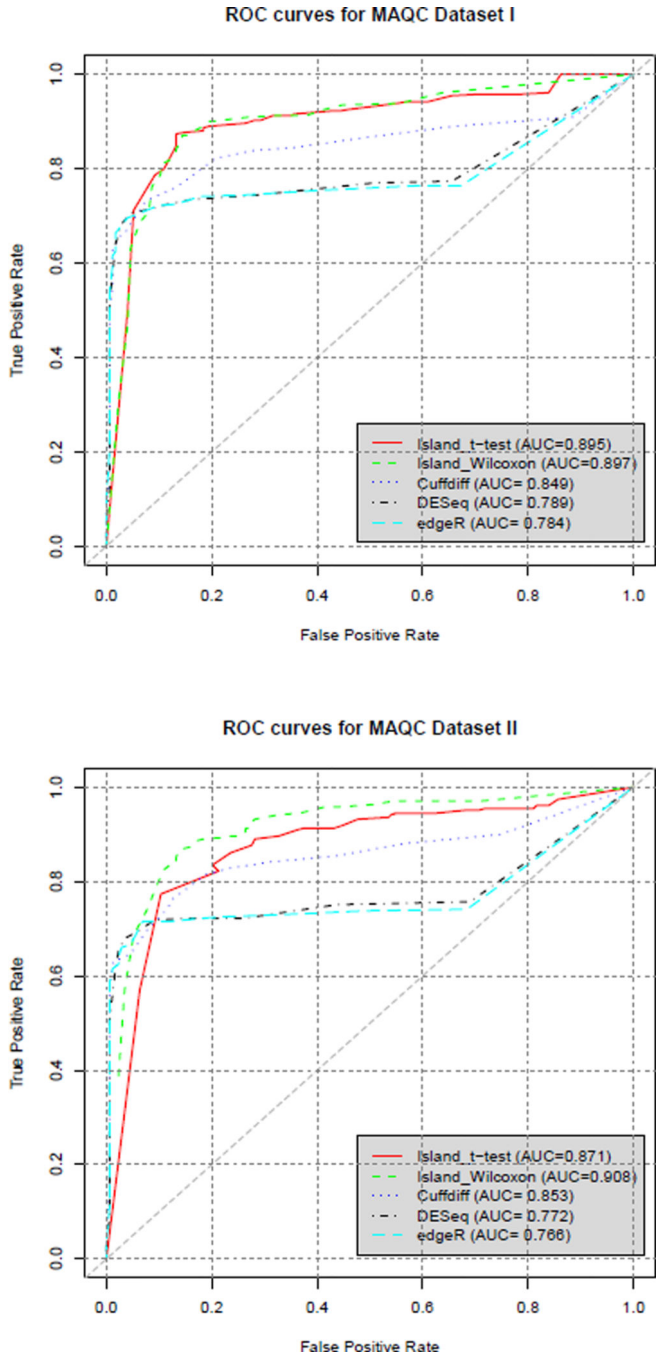


Figure 6.

The ROC curves for the four methods using qRT-PCR validated gene set: (a) The ROC curves for MAQC dataset I; (b) The ROC curves for MAQC dataset II.

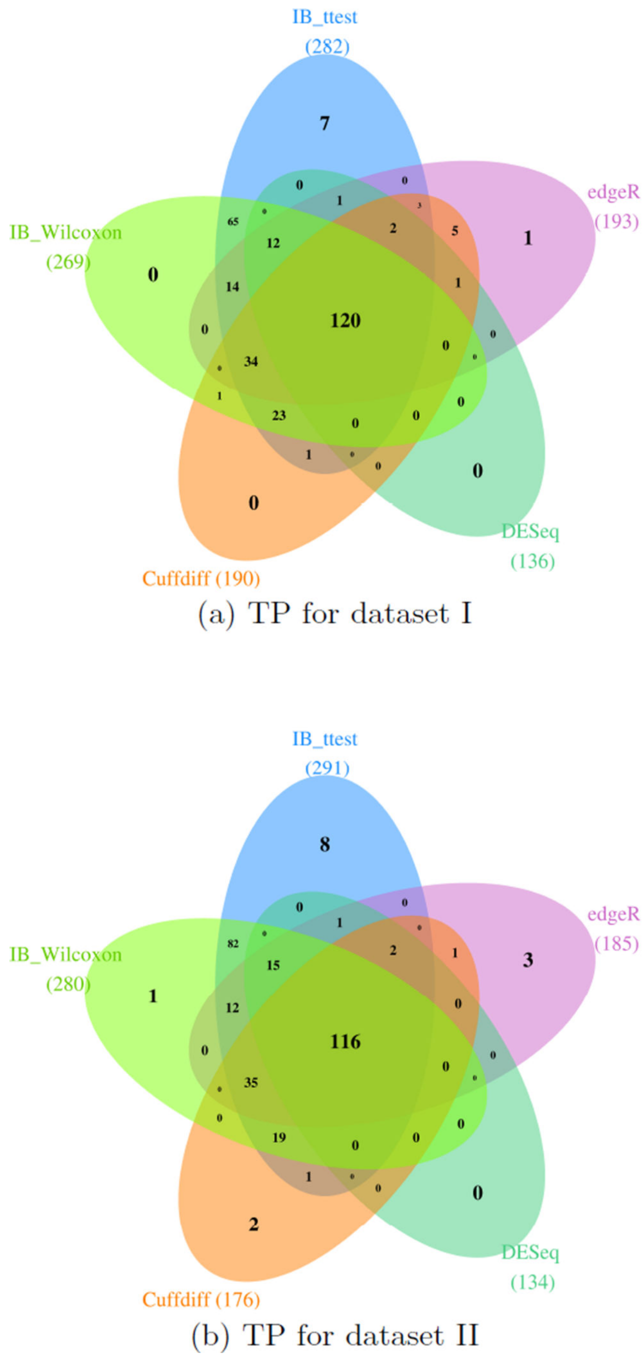
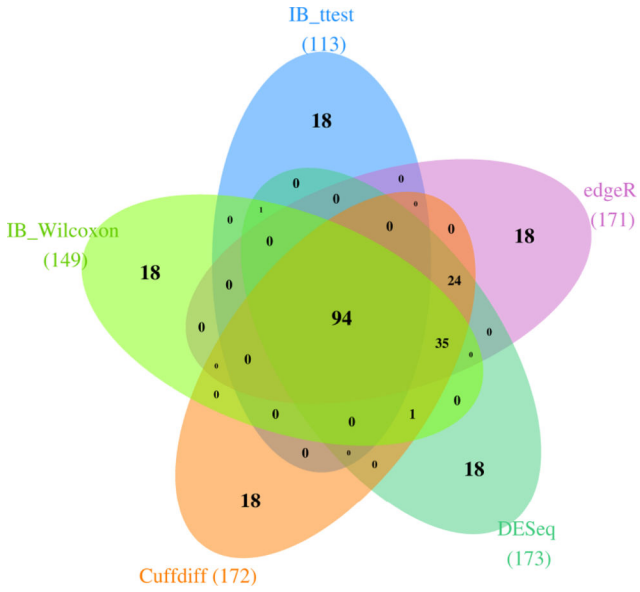


Figure 7.

The overlap between the number of true DE (TP) genes detected by each method for MAQC datasets I and II.



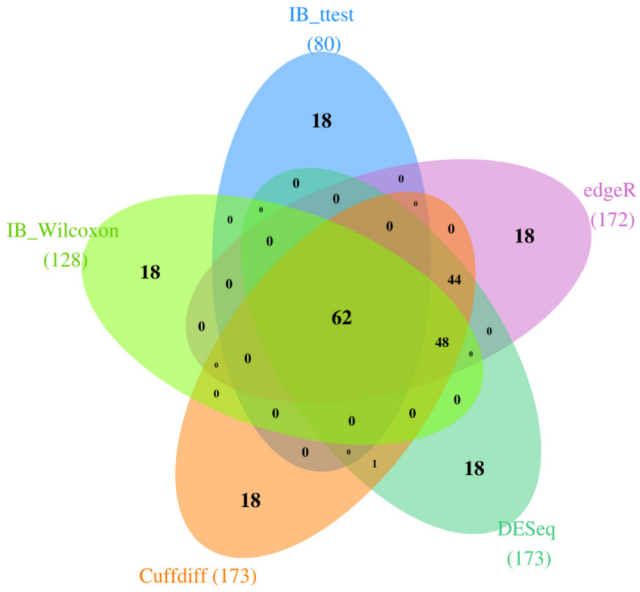


Figure 8.

The overlap between the number of true non-DE (TN) genes detected by each method for MAQC datasets I and II.

Table 1

Detailed information about the constructed islands for each sample using dataset I and II.

Dataset I				
Sample	Number of Islands		Average Island Length	
	High	Low	High	Low
Brain	389,376	389,399	78.4487	3896.63
UHR	391,680	391,703	80.0481	3871.65
Dataset II				
Sample	Number of Islands		Average Island Length	
	High	Low	High	Low
Brain	465,298	465,321	86.0159	3240.47
UHR	446,156	446,179	87.2413	3381.96

Table 2

The number of true DE (TP) and true non-DE genes (TN) found by each method for dataset I and II using *p*-value ≤ 0.05.

Method	TP(D I)	TP(D II)	TN(D I)	TN(D II)
IB_ttest	282	291	113	80
IB_Wilcoxon	269	280	149	128
Cuffdiff	190	176	172	173
DESeq	136	134	173	173
edgeR	193	185	171	172

Table 3

The number of shared true DE genes detected by each method for dataset I and II using p-value ≤ 0.05. The diagonal represents the numbers of true DE genes detected by each method.

Dataset I					
	IB_ttest	**IB_Wilc.**	**Cuffdiff**	**DESeq**	**edgeR**
IB_ttest	**282**	268	183	135	186
IB _Wilc.		**269**	178	132	180
Cuffdiff			**190**	123	165
DESeq				**136**	136
edgeR					**193**
Dataset II					
	IB_ttest	**IB_Wilc.**	**Cuffdiff**	**DESeq**	**edgeR**
IB_ttest	**291**	279	173	134	181
IB _Wilc.		**280**	170	131	178
Cuffdiff			**176**	118	154
DESeq				**134**	134
edgeR					**185**

ACM Conf Bioinform Comput Biol Biomed Inform (2013). Author manuscript; available in PMC 2015 January 26.