Using bROC for analysis of differential expression in RNA-seq data

Application Note, June 2013

Introduction

ROC (receiver operating characteristic) is a generally applicable, non-parametric procedure that provides insight into the discriminatory properties of data features for a binary classifier. However, the method is not efficient for gene expression experiments as they generally do not produce a sufficient number of samples. bROC overcomes this limitation by resampling (bootstrapping) the expression data to produce a large number of simulated measurements that preserve the statistical properties of the original data. Thus, bROC can produce detailed curves of sensitivity (probability of true positive detection) vs. 1-specificity (probability of false positive detection) for all features of interest. CONF = 2AUC – 1, where AUC is the area under the ROC curve, is the primary statistics used for detection of regulated features (probes, genes).

The bROC compares expression between two biological states/endpoints (e.g., treatment and control samples, disease and normal, etc.). At least two experimental/biological replicates are required per state. The algorithm is especially useful for analysis of data sets with small number of replicates and large number of features/probes (thousands or, tens of thousands).

The input data must be normalized using algorithms appropriate for the given experimental platform. If needed, the input data are automatically log2 transformed before they are used in ROC analysis. For RNA-seq data, which contain null values, the typically used automatic transforms are unity shift and log2. The current version of bROC (3.0) includes normalization procedure that has been validated for RNA-seq data.

The present approach is in some respects similar to that used in DESeq Bioconductor package [Anders and Huber] with one significant difference that bROC does not assume any functional form of the noise distributions.

bROC algorithm

The approach may be illustrated as resampling from MA plot (Figure 1), where it is assumed that distribution of measurements in an interval around the given measurement point represents the noise distribution for that data point. Here, M is the difference in expression between two samples, and A is the average expression in two samples, for a given gene/feature. In other words, the average expression level serves as a measure of similarity between the features. Thus:

- Noise distributions are not the same for all measured features but depend on the average expression.
- Genes/features with similar (average) expression level (A) are assumed to be measured with similar uncertainty (i.e., similar noise distributions are expected).
- No assumptions are made about the parametric form of noise distributions.
- The method is applicable to any experimental platform which produces a small number of replicate measurements (at least two) for a large number of features (genes, transcripts).

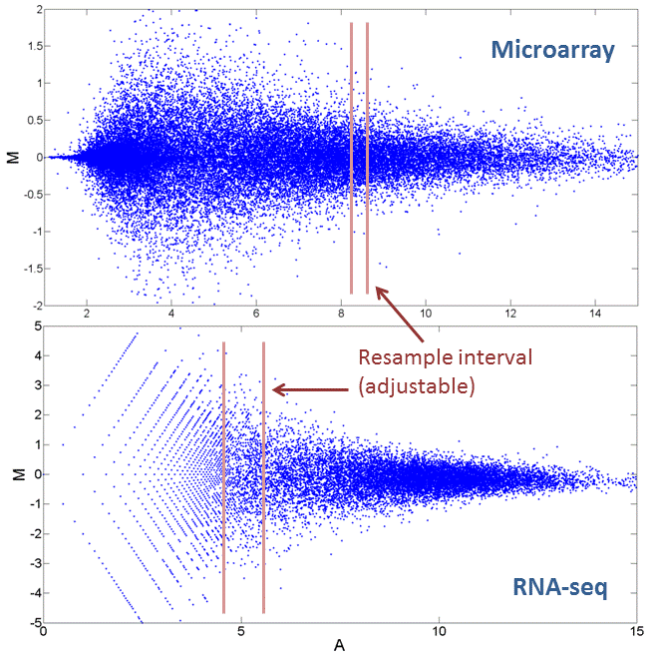

Figure 1. *Examples of MA plots for replicate measurements (same experimental/biological state) for microarray and RNA-seq data. The original data were log2 transformed. For a given measurement point, non-parametric noise distributions are obtained through resampling in the point vicinity, depicted approximately as the resample interval.*

An arbitrary number of simulated measurements may be obtained with resampling and they are used to produce detailed ROC (receiver operating characteristic) curves and associated discrimination measures – Figures 2-3.

For each probe/gene bROC produces the following statistics (discrimination scores):

- CONF = 2 AUC – 1 where AUC is area under the ROC curve. CONF (also known as Gini coefficient) is

equal to twice the area between the ROC curve and the no-discrimination line. CONF = 1 (AUC = 1) indicates perfect separation of the expression measurements between two states and CONF = 0 (AUC = 0.5) indicates no separation (i.e., no differential expression).

- PD (probability of detection) balanced against PFA (probability of false alarm). This value is calculated at the intersection of ROC curve and diagonal of the ROC plot (PFA ≅ 1 – PD).

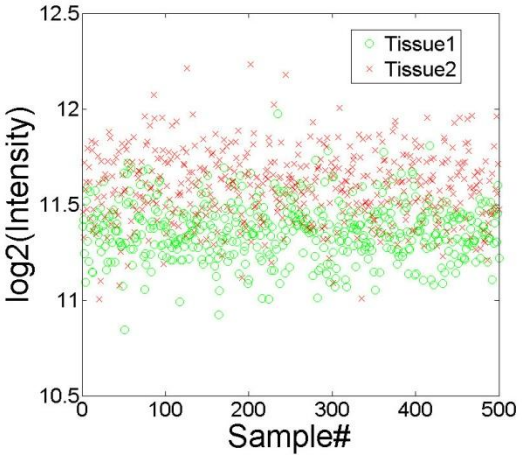

Figure 2. *Large number of simulated measurements may be produced for a given expression feature in two biological states (here, Tissue1 and Tissue2).*

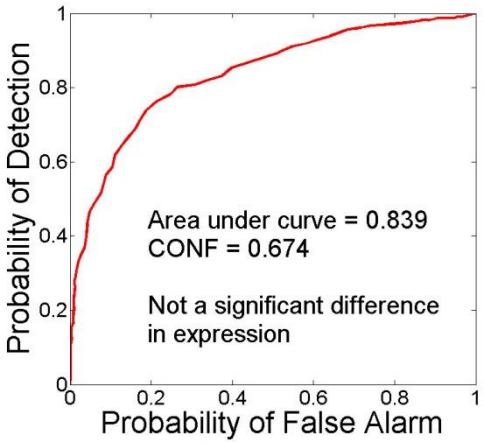

Figure 3. *Detailed ROC curve is constructed using the simulated measurements*

CONF is recommended as the primary statistics in differential expression studies, with detection threshold typically set at 0.95, that is, probes/genes with CONF > 0.95 are considered to be differentially expressed.

Other numerical outputs of the bROC plugin include:

- Estimates of standard deviation for CONF and PD.
- Fold change used in the analysis. In some cases bROC automatically transforms the input data (see *Automatic transformations of input data*) and thus, this quantity may be different from fold change presented by CLC Workbench for the original data.

Data transformations

bROC algorithm works best for logarithmically transformed data. For RNA-seq, many genes are detected with null expression in one of the endpoints being compared and thus, to avoid numerical problems with infinite values, the expression data are transformed as follows:

$$E = \log_2(E_0 + 1),$$

where E_0 are the original expression values.

In CLC plugin implementation of bROC, this transformation is performed automatically. Based on the range of expression data, the algorithm decides if log2 transformation is needed and, if there are null values present, performs unity shift and log2 transformation. (For microarray data, only the log2 transformation, without the shift, is usually performed).

Normalization of RNA-seq data

Normalization of data samples is a crucial step in analysis of expression data, both for microarrays and for the next generation sequencing platforms. The normalization approach used by bROC is described below. Note that normalized data may be optionally saved in the experiment table as *Normalized expression values*.

Median of M-values (MMV) assumes that there is a single scaling factor for each sample of the expression experiment. This assumption is rather common in other approaches to normalization of RNA-seq data [Dillies]. The counts from different samples, which generally are sequenced to different depths, are rendered comparable by application of the normalization (scaling) factors:

$$F_n \mathbf{E}_n \cong F_k \mathbf{E}_k,$$

where F_n is the scaling factor for sample n, and $\mathbf{E}_n$ is the measured expression array for sample n. With log2 transformation this becomes:

$$\mathbf{e}_n + f_n \cong \mathbf{e}_k + f_k,$$

where $\mathbf{e}_n = \log2(\mathbf{E}_n)$, $f_n = \log2(F_n)$, etc. The measured expression array $\mathbf{e}$ can be represented as a sum of 'true' expression $\mathbf{e}^T$ and noise array $\boldsymbol{\delta}$, both of which are not determined. M-value array for samples n and k is then:

$$\mathbf{m}_{nk} = \mathbf{e}^T_n - \mathbf{e}^T_k + \boldsymbol{\delta}_n - \boldsymbol{\delta}_k + f_n - f_k .$$

The noise distribution is arbitrary at this point and does not need to be specified. For samples obtained for the same endpoint, mean over all genes results in:

$$mean(\mathbf{m}_{nk}) = f_n - f_k \cong median(\mathbf{m}_{nk}),$$

because $\mathbf{e}^T_n - \mathbf{e}^T_k = 0$, and $mean(\boldsymbol{\delta}_n) = mean(\boldsymbol{\delta}_k) = 0$, assuming that noise arrays $\boldsymbol{\delta}_n$ and $\boldsymbol{\delta}_k$ are described by a distribution that is symmetrical. If fact, we assume that measurement noise (including biological noise) is described by a set of symmetrical distributions applicable to different levels of expression, which also leads to $mean(\boldsymbol{\delta}) \cong 0$.

For samples obtained for different endpoints we assume that either majority of genes are not differentially expressed between the samples, or there is an approximately equal number of down- and up-regulated genes. Then, *median* is a good approximation of the mean over the genes that are not differentially expressed:

$$median(\mathbf{m}_{nk}) \cong f_n - f_k = N_n.$$

Then, with sample k as a reference sample, the log2-transformed expression array for sample n is normalized as:

$$<\mathbf{e}_n> = \mathbf{e}_n - N_n.$$

For RNA-seq data, where often many features have null values in one of the samples, it is more convenient to use +1 shift preceding the log-transform. However, as we do not make assumption regarding the form of noise distributions, it is not relevant how exactly the $\mathbf{e}_n$ vectors are calculated and the same argument applies.

The normalization algorithm provided by bROC was developed for RNA-seq data. Its utility for microarray data has not been extensively tested or established (i.e., use at your own risk). However, it may be expected that, due to the intrinsic dependence of the bROC analysis on the properties of MA plot, the *median of M-values* may also be useful in the normalization of microarray data, especially when followed by bROC determination of differential expression.

Differential expression analysis workflow for RNA-seq data

The workflow consisting of ***RNA-seq Analysis*** tool available in CLC Genomics Workbench and ***bROC*** provides complete differential expression analysis of RNA-seq count data from raw counts, through alignment and normalization to the list of differentially expressed features. (Here, 'workflow' is to be understood as manual workflow, where the analysis tools are used in sequence but separately. Version 3.0 of bROC is not set up for the CLC workflow framework.) Of course, the *bROC* plugin may be also used with expression data aligned by other methods (outside of the CLC Workbench) and imported into the Workbench. Other normalization methods may be employed, if needed.

With ***RNA-seq Analysis***, two analysis paths, differing in the data normalization approach, are possible:

- Median of M-values (MMV) normalization. This is the preferred and strongly recommended approach. When running *RNA-seq Analysis*, in *Result handling* select *Expression value* Genes: Total gene reads as the output expression value.
- RPKM normalization. In *Result handling* (in *RNA-Seq Analysis*) select *Expression value* Genes: RPKM, which produces expression data normalized using the RPKM method [Mortazavi 2008]. In this case, no normalization should be used when running bROC.

It should be stressed that these two analysis paths produce different (in some cases, significantly different) results for differential expression. This is due to the fact that RPKM does not produce a single scaling factor for the entire sample but a set of scaling factors that depend on the gene length. As a result, the features are rearranged on the MA plot, which alters the noise estimates. Overlap between sets of features declared as differentially expressed when different normalizations are used depends on the experiment and on the sample composition.

To illustrate the use of bROC plugin in the analysis of RNA-seq data we use data published by Dillies *et al.* [Dillies]. The expression data were aligned (quantified) outside of CLC Workbench. Please, refer to the bROC Manual for details on the plugin usage.

Graphical output

Figures 4-6 show results of bROC analysis for *Homo sapiens (Hs)* melanoma data set. Briefly, these data compare melanoma cell line expressing the Microphtalmia Transcription Factor (MiTF) and melanoma cell line in which small interfering RNAs are used to lower the expression of MiTF. The data were produced with Illumina Genome Analyzer IIx. The data set contains expression values for 36,719 genes and three replicates per condition.

bROC produces two graphs that depict genes/features detected as differentially expressed for a given threshold values of CONF: MA plot for values averaged over the experimental groups (Figure 4), and XY plot of mean values for expression in one group versus the other (Figure 5). Additionally, 'volcano' plot is produced, which depicts the dependence of CONF statistic on the fold change.

Figure 6 shows 'volcano' plot for the *Hs* melanoma data set. The fold change (FC) is between averages of two data groups – it is equivalent to the inter-group M-value (for transformed and normalized data). The features (genes) may be identified directly on the graph – mouse pointer placed over a data point or, group of data points, invokes display of feature names, and CONF and fold change values. It is interesting to note that for a given detection threshold there is typically a well-defined minimum FC for genes declared as differentially expressed (in this case about 1, for CONF = 0.95).

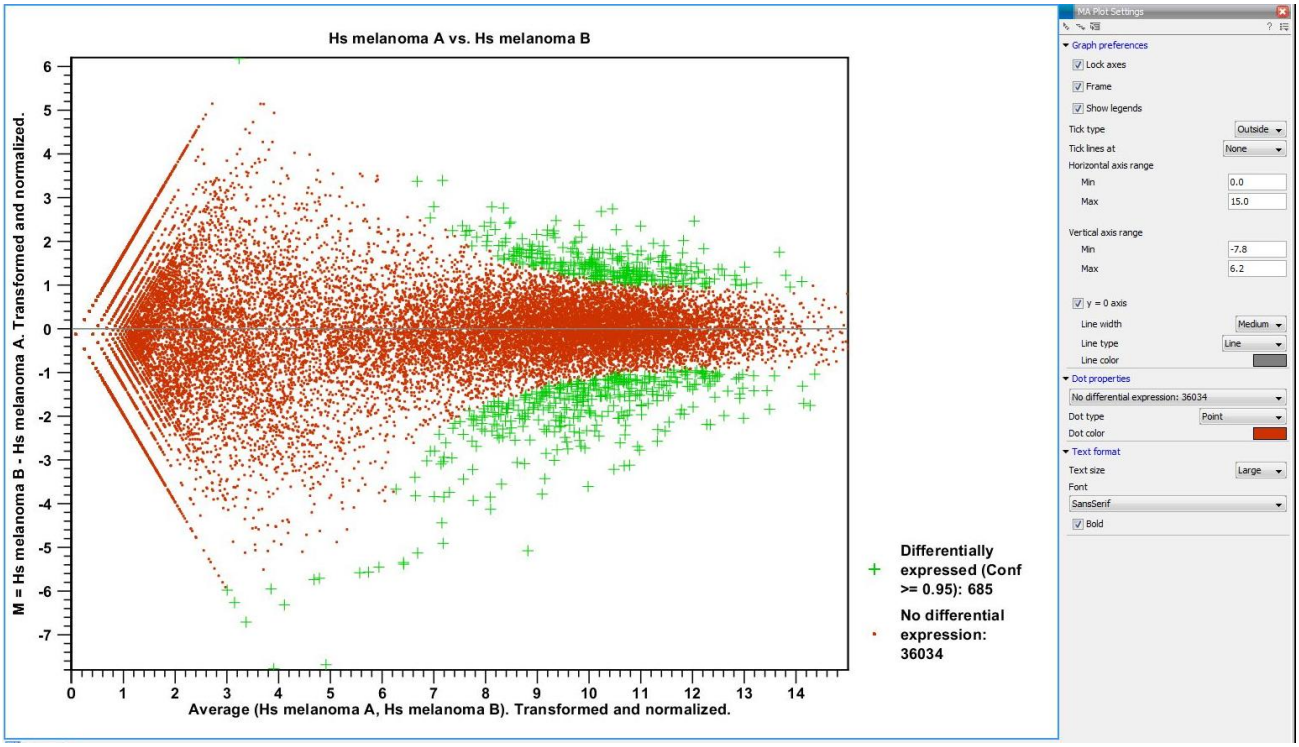

Figure 4. *Example of MA plot generated by bROC, showing features declared as differentially for a given detection threshold (CONF=0.95). The values shown are those used in the bROC analysis – in this example the original data were transformed (+1 and log2) and normalized. The plot appearance may be modified through* Plot Settings *interface on the right – the figures shown here demonstrate a few examples of different plot settings.*

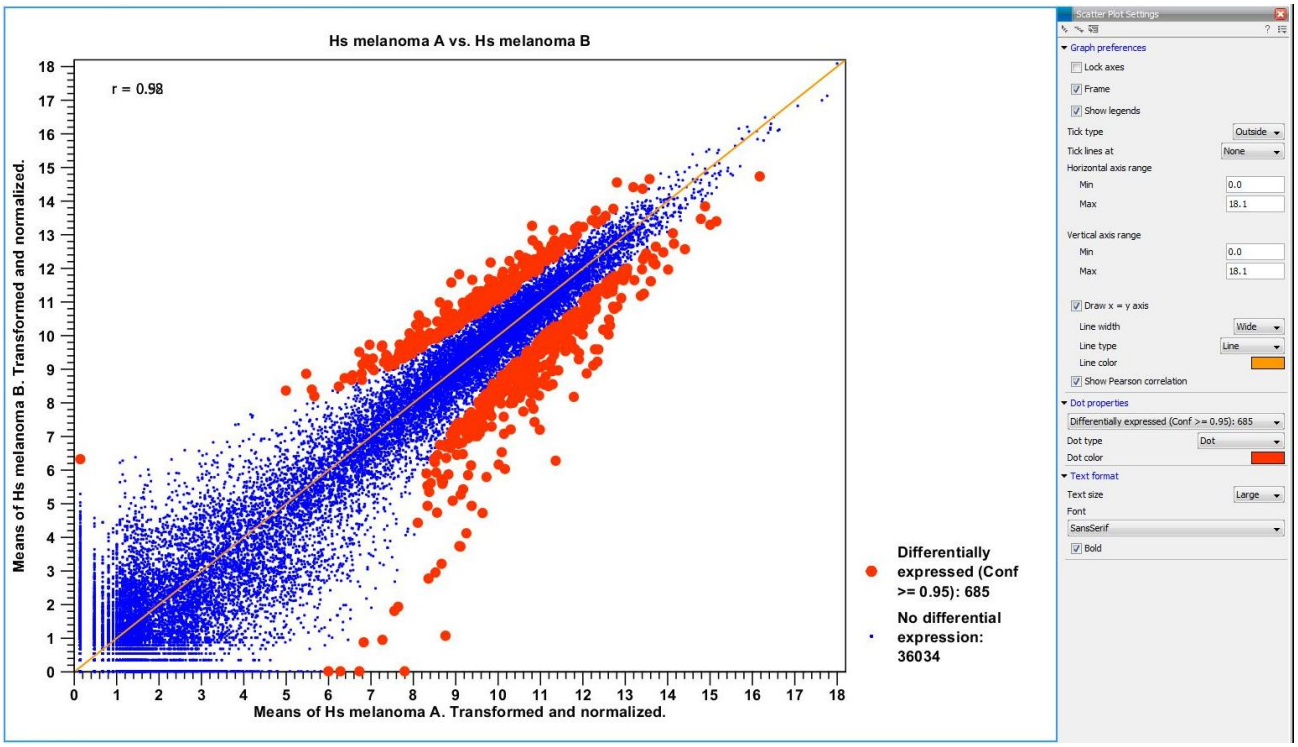

Figure 5. *Example of XY plot generated by bROC, showing features declared as differentially expressed for a given detection threshold (CONF = 0.95).*

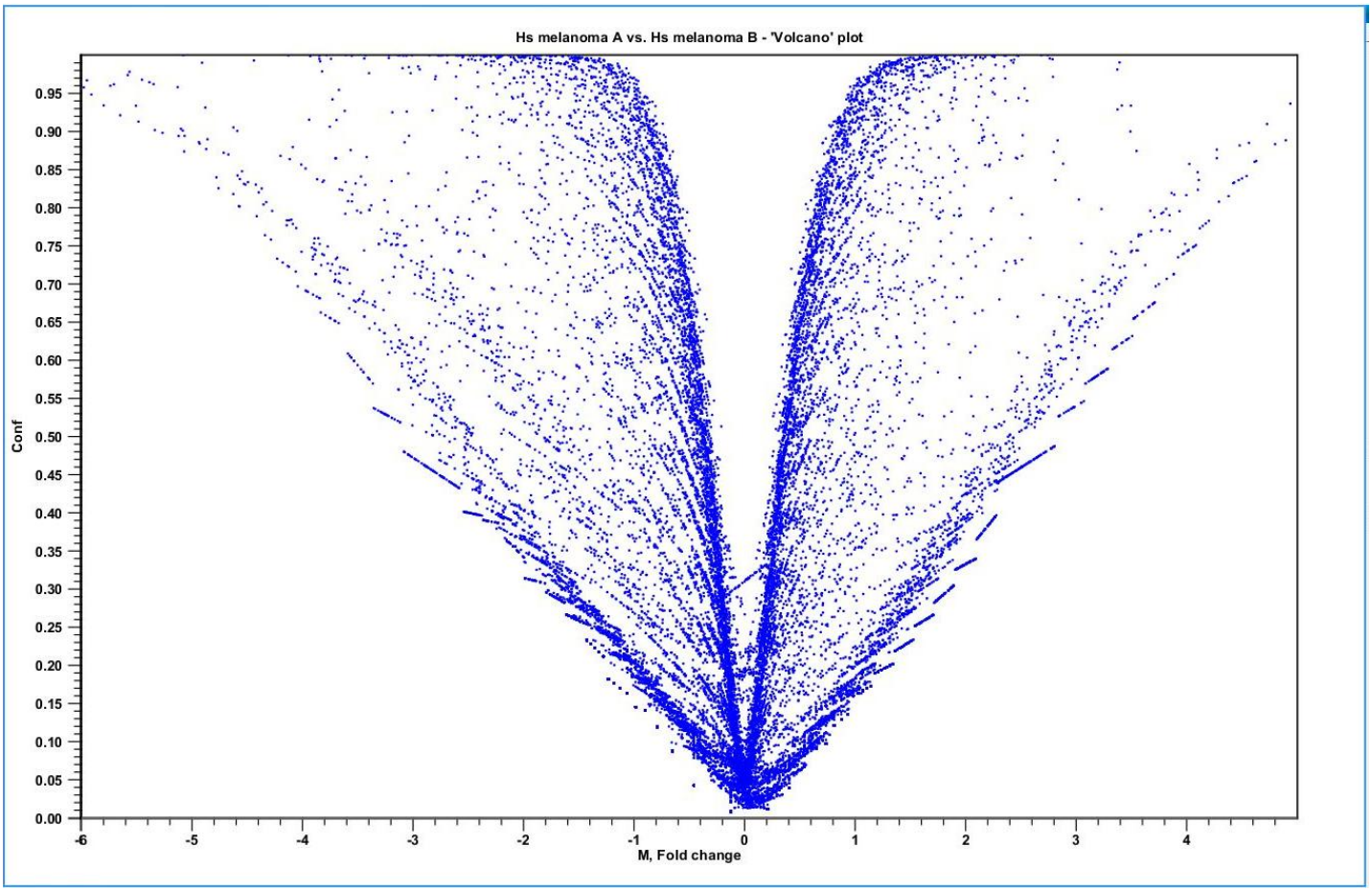

Figure 6. *Example of bROC 'volcano' plot.* Homo sapiens melanoma *data set.*

Effect of normalization: MMV vs. raw counts

This example uses *Aspergillus fumigatus* data from Dillies *et al.* [Dillies]. These sequencing data compare the transcriptome of ubiquitous fungus *A. fumigatus* (strain 1163) in two different growth media ('Af A' and 'Af B'). The data were produced using Illumina HiSeq 2000 machine and contain expression values for 9248 genes and 2 replicates per condition.

The effect of normalization is illustrated in Figure 7, which shows the bROC produced MA plots obtained with MMV normalization (upper panel) and without normalization (lower panel). MMV normalization produces good alignment of (group averaged) M-values to M=0 line and when used with bROC results in detection of 122 differentially expressed genes. With no normalization (raw counts) there is 'downward shift' of the MA plot and no genes are detected as differentially expressed (for the same detection threshold).

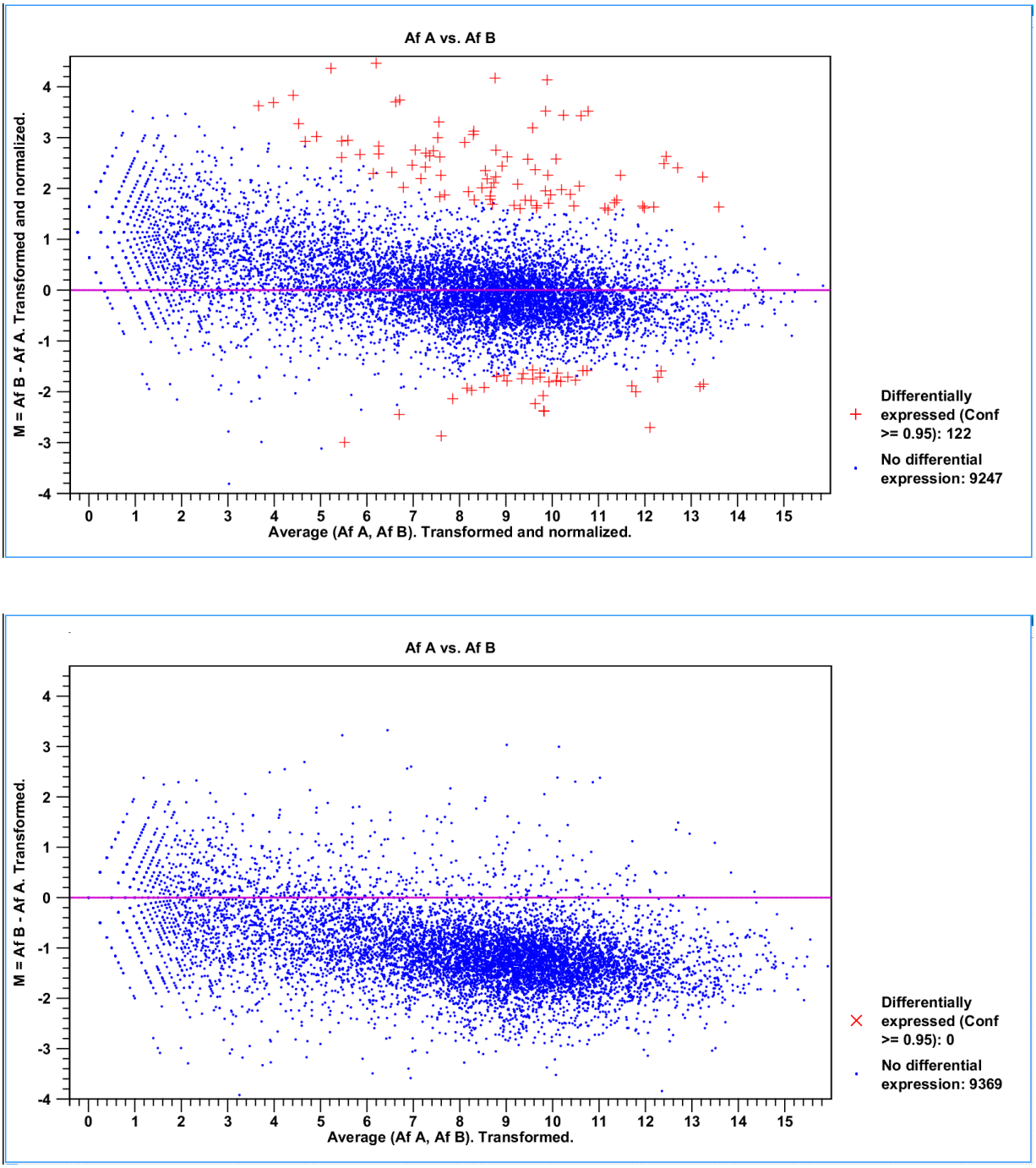

Figure 7. *Effect of normalization on the results of bROC differential expression analysis. Upper panel: MMV normalization, lower panel: no normalization (raw counts). Aspergillus fumigatus data from Dillies et al.*

Effect of normalization: MMV vs. RPKM

For the *Hs Melanoma* data set, RPKM normalization produces similar number of differentially expressed genes (713 as compared to 685 for MMV normalization). However, only 531 genes are detected by both approaches – about 76% overlap.

Figure 8, analogous to Figure 5, shows bROC result for RPKM normalized data. Although, the change in the relative data distribution is noticeable, the overall detection pattern is similar between the two cases. That is, the genes detected as differentially expressed are found in similar positions on the respective scatter plots. However, comparison of Figure 9 with Figure 5 shows substantial differences in the identity of differentially expressed features.

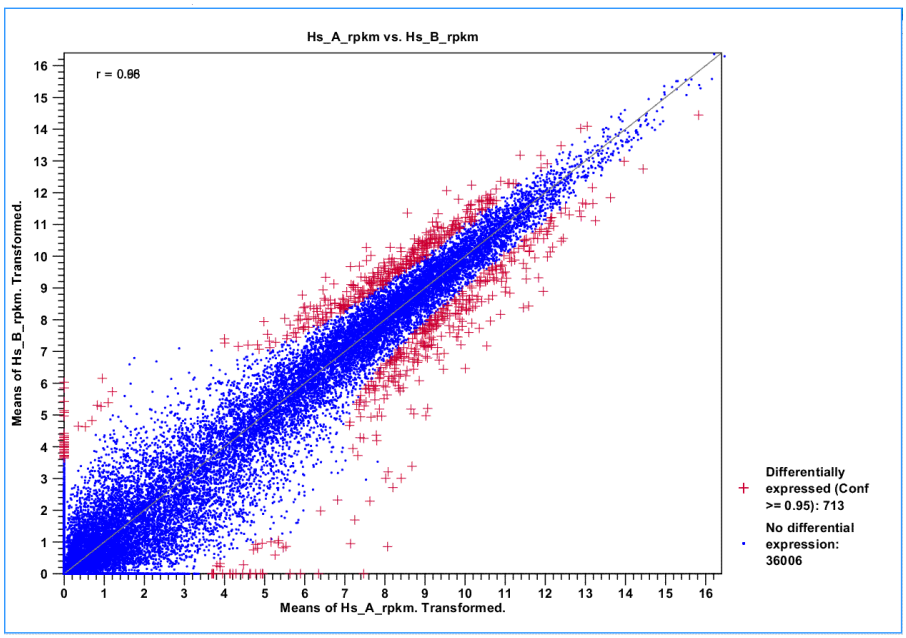

Figure 8. *Result of bROC analysis for Hs Melanoma data set normalized with RPKM.*

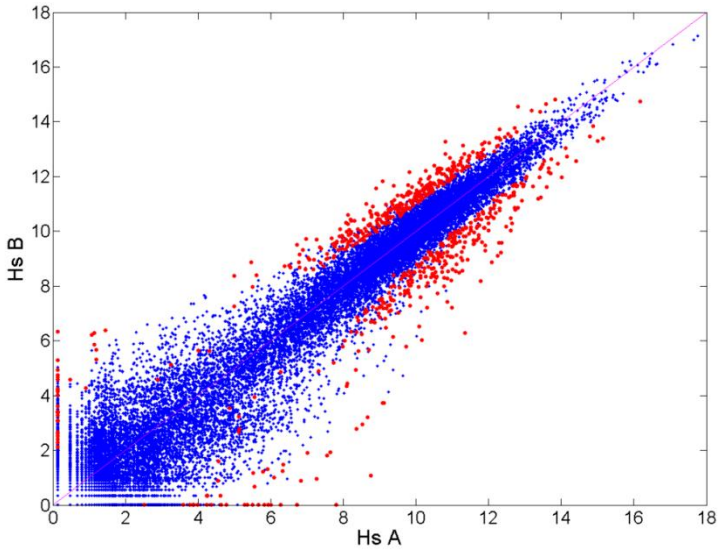

Figure 9. *XY plot for MMV normalized Hs melanoma data showing genes detected with RPKM normalization (red points). To be compared with Figure 9. Plot produced outside of the Genomic Workbench.*

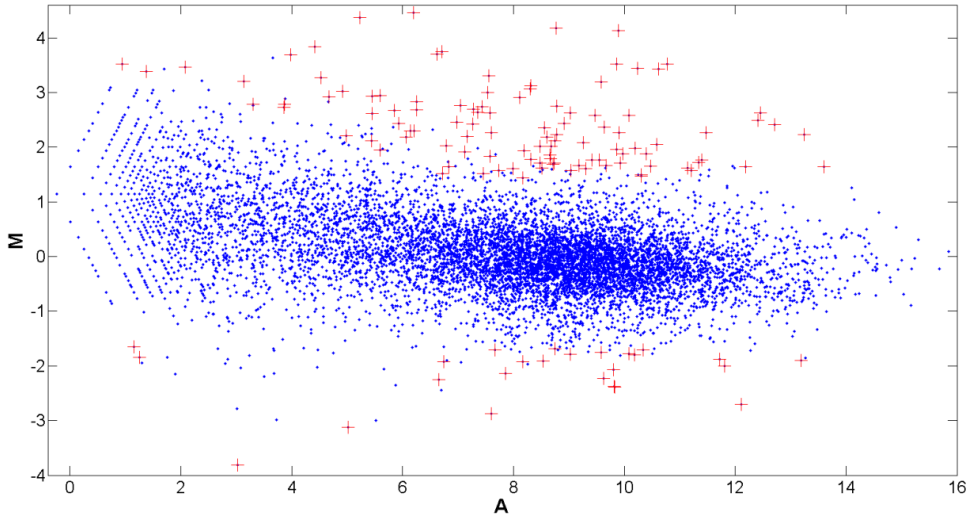

Figure 10. *MA plot for MMV normalized A. fumigatus data showing the features detected by bROC uusing RPKM normalization (red crosses). To be compared with Figure 6 (upper panel).*

Figure 10 shows analogous results for *A. fumigatus* data set but presented on MA plot. Using the standard detection threshold (CONF = 0.95) in the bROC analysis, 132 and 122 genes were detected as differentially expressed with RPKM and MVV normalization, respectively, and 102 genes were detected with both normalizations. Again, comparison of Figure 7 (upper panel) and Figure 10 indicates changes in the identity of genes detected when different normalization methods are used.

Summary and conclusions

The performance of bROC in analysis of RNA-seq data is essentially similar to what was described previously in the context of microarray data analysis (see http://www.clcbio.com/wp-content/uploads/2012/09/bROC_White_Paper_5.pdf). The bROC method offers several attractive features:

(1) Number of detected genes is essentially independent on the number of replicates used in the analysis. For other methods, the number of detected probes increases significantly with the number of replicates when the same detection criteria are used. Our microarray results indicate that not only the number but also the identity of detected genes is only weakly dependent on the number of replicates.

(2) The method is uniquely suitable for analysis of dataset with small number of experimental replicates (however, at least two replicates are required). Although it is true that smaller number of replicates may not produce an adequate picture of experimental and biological noise, the overall pattern of genes discovered as differentially expressed is mostly independent of the number of replicates. The capability to extract informative data from a low number of replicates becomes valuable in situations where specimen availability and amounts are limited.

(3) Differential expression of low-abundance genes can be detected in both microarray and RNA-seq data. It remains to be seen how other statistical methods proposed for this application compare with bROC in this regard.

(4) MMV (Median of M-values) normalization appears to be particularly suitable for bROC, which is inherently based on the analysis (resampling) of MA plots. As an interesting observation, MMV produces normalization factors very similar to the normalization used in DESeq R/Bioconductor package [Anders and Huber], and also to the TMM (trimmed mean of M-values) algorithm [Robinson and Oshlack]. Recent work has indicated that these two normalization approaches are preferable for statistical analysis of differential expression in RNA-seq data [Dillies].

References

Anders S and Huber W. Differential expression analysis for sequence count data. Genome Biology 2010, 11:R106.

Dillies M-A, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, *et al*. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. Briefings in Bioinformatics, Epub Sept 17 2012.

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 2008, 5(7):621-8.

Robinson MD and Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biology 2010, 11:R25.