University of Louisville

ThinkIR: The University of Louisville's Institutional Repository

Electronic Theses and Dissertations

12-2010

Assessment and performance of VSN-INV normalization on the NCI-60 microRNA expression profiles.

Martin T. Disibio
University of Louisville

Follow this and additional works at: <https://ir.library.louisville.edu/etd>

Recommended Citation

Disibio, Martin T., "Assessment and performance of VSN-INV normalization on the NCI-60 microRNA expression profiles." (2010). *Electronic Theses and Dissertations*. Paper 354.
<https://doi.org/10.18297/etd/354>

This Master's Thesis is brought to you for free and open access by ThinkIR: The University of Louisville's Institutional Repository. It has been accepted for inclusion in Electronic Theses and Dissertations by an authorized administrator of ThinkIR: The University of Louisville's Institutional Repository. This title appears here courtesy of the author, who has retained all other copyrights. For more information, please contact thinkir@louisville.edu.

**ASSESSMENT AND PERFORMANCE OF VSN-INV
NORMALIZATION ON THE NCI-60 MICRORNA EXPRESSION
PROFILES**

By

Martin T Disibio II
B.S, University of Louisville, 2003

A Thesis
Submitted to the Faculty of the
University of Louisville
J.B. Speed School of Engineering
as Partial Fulfillment of the Requirements
for the Professional Degree

MASTER OF SCIENCE

Department of Computer Engineering and Computer Science
J.B. Speed School of Engineering
University of Louisville
Louisville, KY

December, 2010

**ASSESSMENT AND PERFORMANCE OF VSN-INV
NORMALIZATION ON THE NCI-60 MICRORNA EXPRESSION
PROFILES**

Submitted by: _____
Martin T Disibio II

A Thesis Approved on

11-23-10
(Date)

by the Following Reading and Examination Committee:

Dr. Eric C. Rouchka, Thesis Director

Ming Ouyang

Ted Kalbfleisch

ABSTRACT

ASSESSMENT AND PERFORMANCE OF VSN-INV

NORMALIZATION ON THE NCI-60 MICRORNA EXPRESSION

PROFILES

Martin T. Disibio II

November 23, 2010

Multiple normalization methods have been proposed for the analysis of microRNA microarray expression profiles but there is no consensus method. One of the more robust methods, quantile normalization, is commonly used in transcript (mRNA) studies and was therefore used for normalizing the first microRNA expression profiles of the NCI-60 cell panel, published in 2007. In this study the appropriateness of VSN-Inv, a recently proposed alternative normalization method, to the NCI-60 dataset is verified. VSN-Inv normalization results in much increased inter-sample correlations among control groups, and significantly higher intra-chip correlations of duplicate probes, versus quantile and no normalization. Furthermore, VSN-Inv normalization was found to have favorable performance for hierarchical clustering and discovery of miRNA-mRNA interactions, and a lower misclassification rate for predictive analysis based on tissue of origin when using log transformed data (median 0.19, best 0.12).

ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. Eric Rouchka, for his guidance and patience, and especially for the many inconveniently-timed meetings which my schedule required. I thank my wife Kara, for giving me constant support during this thesis and the entire degree, and for forgiving me for being so scarce during those many late nights.

Additionally, I would like to thank Dr. Ming Ouyang and Dr. Ted Kalbfleisch for serving on my review committee, and Dr. Robert Flight for providing valuable feedback and comments regarding the analysis portion.

TABLE OF CONTENTS

I. INTRODUCTION	1
II. BACKGROUND	5
A. Biological Background	5
B. Technology Background	9
III. METHODOLOGY	13
A. Download Data	13
B. Normalization	14
C. Clustering and Predictive Analysis	15
D. Correlations	17
IV. RESULTS	18
A. VSN-Inv Invariants	18
B. Assessing Normalization via Duplicate Probes and Control Samples	21
C. Clustering	26
D. Predictive Analysis	29
E. Correlations	32
V. CONCLUSION	34
VI. FUTURE DIRECTION	35
REFERENCES	37
CURRICULUM VITAE	39

LIST OF FIGURES

FIGURE	PAGE
Figure 1 – Possible mechanisms of microRNA induced silencing complex (miRISC)-mediated repression .	7
Figure 2 – VSN-Inv standard deviation versus mean component plot	19
Figure 3 – Intensities of selected invariant probes	20
Figure 4 – Scatter plot of total sample intensities before and after normalization	21
Figure 5 – Box plot of correlations between the 35 duplicate probes	22
Figure 6 - Box plot of correlations coefficients between all probes	22
Figure 7 – Scatter plot of signal for all 35 duplicate probe pairs in miRNA dataset	24
Figure 8 – Box plot of correlations between control samples	25
Figure 9 –Venn diagram of probes selected for hierarchical clustering	26
Figure 10 – Venn diagram of probes selected for hierarchical clustering (log transformed)	26
Figure 11 – Complete-linkage hierarchical clustering of samples	28
Figure 12 – Complete-linkage hierarchical clustering of samples (log transformed data)	29
Figure 13 – Medoid probes with k=40	30
Figure 14 – Misclassification rates after PAM-PAM with k=40	31
Figure 15 – Medoid probes with optimal k	32
Figure 16 – Misclassifications rates after PAM-PAM with optimal k	32

Figure 17 – Distribution of significant miRNA-mRNA correlations among differentially expressed probes	33
--	----

I. INTRODUCTION

In the past decade a new class of small-interfering RNA, known as microRNA (miRNA), has been discovered and thereafter shown in a multitude of studies to be critical to eukaryotic life. It has been experimentally proven or implicated to play a key role in such diverse systems as embryonic development [1], cardiac tissue[2], nervous system[3], circadian rhythm[4], and immune systems[5]. Understanding this class of molecule and its related processes within the cell could lead to developments such as novel gene therapy techniques and cancer treatments.

Microarray analysis is a widely utilized tool for understanding and quantifying gene expression, and thusly has been adapted to work on miRNA as well. Microarrays allow researchers to quickly and simultaneously measure the presence of hundreds or thousands of unique genes using relatively cheap materials. Before microarray data can be used it must be normalized due to the inherent inaccuracy of the process. There are a number of normalization methods available, such as scaling, loess, variance-stabilization, and quantile [6]. For transcript expression microarrays (genes), a widely used normalization technique is GCRMA, a combination of adjustment for background and non-specific binding, quantile normalization and median polishing, enhanced with pre-computed probe affinities based on sequence. Quantile normalization has been used for miRNA microarray data previously [7], but there is no consensus that it is optimal [6]. Specifically, questions have been raised regarding whether the underlying assumption of

quantile normalization, that the probe intensity distribution is identical between all samples[8], is true for certain miRNA microarray experiments[9]. In experiments involving different tissues of origin, it is expected that a large fraction of the miRNAs are differentially expressed between samples [9].

One such experiment was the release of the first miRNA microarray expression profiles for the NCI-60 cell panel, published in 2007 by Blower, et al. [7]. In that study cluster analysis showed that the different cancer types also corresponded to different miRNA expression profiles, evidenced by the moderate to accurate hierarchical clustering of the samples by cancer type. The expression profiles were processed using quantile normalization. Additionally, by integrating the transcript/protein expression profiles for the NCI-60 published in 2007 by Shankavaram, et al. [10], Blower, et al. attempted to determine whether miRNA-mRNA interactions can be observed in the data by calculating a Wilcoxon rank-sum test against known interactions, as listed in Tarbase V2 (accessed September 2006)[11], and all other probe-probe combinations. A significant difference in the populations was not found ($p=0.28$), indicating that interactions cannot be detected. Blower, et al. suggest that the lack of correlation is due to the frequent targeting of a miRNA sequence to multiple non-coexpressed genes.

Following, Wang and Li performed a more in-depth study addressing the correlations present in the integrated miRNA-mRNA datasets [12]. The same datasets and normalization procedures were used as previously. Guided by target predictions from both TargetScan and miRBase::Targets, they found negative correlations can be found matching predicted interactions. Generally, more negative correlations are found than positive, which are due to premature degradation of transcripts induced by the

targeting miRNA. Intronic miRNAs, meaning miRNAs whose coding sites lies within its host gene's site, were highly positively correlated with their host gene, suggesting co-expression. Furthermore, Wang and Li found, using a newer release of Tarbase, that interactions known to be involved in mRNA degradation showed more negative correlations than interactions involving translational repression. This finding somewhat conflicts the results of Blower, et al. This may be explained by the fact that the newer release of Tarbase contains many more validated targets and therefore a better sampling of data. However, Wang and Li also suggest that due to a large variance between the probe-probe interactions derived from each miRNA-gene interaction, that the data is not highly reliable and a conclusion should not be made.

In working with these two datasets, we also noted some instances where quantile normalization loses significant information in the miRNA dataset. The OSU-CCC V3 microchip used to assay the 60 cell lines contains 35 pairs of probes that have the same oligonucleotide sequence. These probes are intended to measure different stages of miRNA (e.g precursor vs. mature) or related sequences in the same family, as evidenced by their labeling in the array design file, meaning this duplication is probably an unintentional effect of design constraint[13] the small miRNA sequence landscape presents, relative to transcription factors and genes. The probes of the OSU-CCC V3 chip are 40 nucleotides long, while microRNA precursors are generally 80 to 150 nt, and mature miRNAs are around 22nt. As expected, preliminary research shows that in the raw data many of the probe-pairs have strong correlations; however, after quantile normalization the distribution of correlations between probes has been lowered. This finding is discussed more in-depth in the results section.

From these studies, concluding that the underlying assumption of quantile normalization is invalid for this dataset may be justified. An alternative normalization method, VSN-Inv, has recently been published [9], and offers a more limited set of underlying assumptions that appear to be valid. Reassessing some of the earlier studies using a different normalization method could yield different results. In this work we verify the applicability of VSN-Inv normalization to the NCI-60 OSU-CCC expression profiles, and compare the performances of both quantile and VSN-Inv against no normalization (background-subtraction only). Verification of VSN-Inv applicability is performed by checking the invariants selected by VSN-Inv, correlations between duplicate probes, and correlations between control replicates. Performances compared between normalization methods include hierarchical clustering, predictive analysis using shrunken centroid partitioning, and discovery of miRNA-mRNA interactions. An R-language toolset, NCTOOLS, was created to aid with the analysis. This package and all resulting data is available at <http://bioinformatics.louisville.edu/VSN-Inv/>.

II. BACKGROUND

A. Biological Background

Central Dogma of Molecular Biology

The “Central Dogma of Molecular Biology” is the term used to describe the fundamental process shared by all eukaryotic life in which the information encoded in an organism’s DNA is transformed into the proteins and enzymes constituting that organism [14]. Understanding this process is essential to understanding the role that microRNA plays in the cell as well.

The process begins in the cell nucleus where the DNA double-helix resides in the form of chromosomes. The DNA temporarily unwinds at a given point along the helix at a gene region. The sequences of one or both strands are duplicated in order into a new separate nucleotide chain composed of RNA. The resultant RNA molecule is spliced to contain only the active parts of the gene, known as exons, a 5’ cap, and polyadenylated (poly-A) tail, yielding the final messenger RNA (mRNA). Of these steps, polyadenylation is important because it guides the interactions the mRNA will take part in later. This entire step is called transcription.

The mRNA then exits the nucleus through the nuclear envelope, into the main section of the cell where it is guided onto a ribosome. Ribosomes are intra-cellular structures present in every cell, where the final step of protein encoding (translation) occurs. Once an mRNA molecule is incorporated into the ribosome, it is processed three

nucleotides at a time, where each nucleotide triplet, known as a codon, results in a new specific amino-acid being added to the growing protein chain. In each mRNA there are beginning and ending sections that are skipped during this process, known as the untranslated regions (UTRs).

The above steps were reiterated here because it is critical to understand that an mRNA is composed of distinct functional sections that govern the way the mRNA is processed. In order for the mRNA to be successfully translated into protein, the mRNA must be composed of the proper sections (5' cap, UTRs, poly-A tail), and remain intact until it is incorporated into the ribosome. Interference with any one of these can affect the normal expression of the mRNA.

MicroRNA (miRNA)

MicroRNAs (miRNAs) belong to a class of small non-coding RNAs known as small-interfering RNAs (siRNAs). This class of RNA is not translated into a protein via the standard mechanism described previously, but instead serves other purposes within the cell. These purposes discussed here are limited to the manners by which they interfere with the expression mechanism of genes via transcription and translation, which were described previously.

miRNAs were first discovered by Lee, et al., in 1993 [15], as a previously unseen regulatory mechanism in *C. elegans*. They were not recognized as a new class of molecule until much later. Following, miRNAs have been identified in many different species, including *Homo sapiens*, and shown to be fundamental to functioning of key organs and life processes. There are roughly 1000 known miRNAs in humans [16].

miRNAs are distinguished from other siRNAs in that miRNAs have evolved to target specific sets of genes using anti-sense complementarity to 6mer regions in the 3' UTR of the mRNA for that gene [17]. By binding to the 3' UTR, the miRNA regulates the expression of that gene through several mechanisms, thus preventing the successful translation of the mRNA into protein (Figure 1). There are two putative classes of mechanisms by which the mRNA is blocked: premature cleavage (degradation), where the mRNA is degraded before reaching the ribosome; and translational repression, in which the mRNA cannot be processed correctly by the ribosome and is not properly translated [18]. Premature degradation should result in a negative correlation between the expression profiles of the miRNA and its target. In some rare instances a miRNA can up-regulate the expression of a gene [19].

Figure 1 – Possible mechanisms of microRNA induced silencing complex (miRISC)-mediated repression Image reproduced from [18].

miRNAs themselves undergo several steps similar to the transcription of DNA before they reach their mature forms, in that the form originally transcribed from the chromosomal DNA differs from the final form. Only the mature forms interact with mRNA in the manner described above. For this study it is not necessary to understand the steps preceding the mature miRNA.

Many different prediction algorithms and web resources exist for listing putative miRNA-mRNA interactions based on primary sequence and annotation of miRNAs and genes (i.e. non-experimental predictions) [20]. Because the rules of effective targeting of a gene by a miRNA are not fully known, the algorithms differ in their underlying assumptions and factors considered. Some common prediction algorithms are listed in (Table 1). These complexities lead to an overall disagreement between predictions, an inability to predict some uncommon forms of targeting (wobble seeds, bonding outside 3' UTR), and a high false-positive rate.

Table 1 – Common miRNA target prediction algorithms

Prediction Algorithm	Factors
TargetScan	Seed region, site conservation, AU content
PicTar	Seed region, site conservation, free energy
miRanda	Base pairing, free energy, conservation
PITA	Free energy, secondary structure
Rna22	Over-represented hexamers in miRNAs

NCI-60

The NCI-60 is a cell panel consisting of 60 cell lines derived from nine different types of cancer. Originally developed by the National Cancer Institute to screen compounds for anticancer activity, these cell lines provide a superb research platform for two main reasons. (1) The cell panel is very diverse, containing lines derived from breast, colon, central nervous system, kidney, lung, prostate, ovarian, melanoma, and leukemia; (2) the cell lines are homogeneous and can be obtained in unlimited amounts. The wealth of information this cell panel presents is made apparent when datasets from multiple experiments are integrated to allow for more powerful studies and new insights.

Previous studies include pharmacological profiles, transcript profiles, protein expression using gel electrophoresis, reverse-phase lysate arrays, chromosomal aberration characterization, and more. Indeed, Bussey, et al. summarize the notoriety of these cell lines by stating that they have been “more fully characterized at the molecular level than any other set of cells in existence” [21].

The expression profiles of these cell lines used in this study present a unique challenge for analysis due to the diversity of the tissues of origin, and lack of spike-in probes which could establish a known baseline.

B. Technology Background

Microarray

Microarrays are a very common tool developed to simultaneously measure the expression levels of thousands of sequences of the DNA/RNA in a cell in a single step. While various designs of microarrays exist, they all employ the same basic tactic. On the

microarray is a grid containing thousands of small spots (micrometers in diameter), preloaded with synthesized cDNA/RNA of all different, but specific, sequences. Each unique sequence is referred to as a “probe”. When the biological sample being measured washes over the probe, DNA/RNA present in the sample bonds to complementary probes on the microarray. This step is known as hybridization. The array is designed such that each probe uniquely measures a specific gene or other sequence. This is done by giving each probe a unique sequence, where it is most complementary to the desired target, above all other sequences. The process by which the optimum sequences are calculated, maximizing signal and minimizing error, is known as array design.

Measuring the expression level of each probe is done via software, where an optical scan of the chip is read into image processing software. The expression level is determined from the strength of the color in the image, which corresponds to the amount of fluorescent material in each probe, and therefore the amount of hybridized target in each spot.

OSU-CCC V3 Chip

This microRNA microarray chip was designed by Blower, et al. with the help of The Ohio State University Comprehensive Cancer Center. It is a pin-spotted microarray and contains probes for both human and non-human miRNAs. There are 627 human-specific miRNA probes. Each probe is spotted in duplicate on each chip, yielding 1254 data points per sample [7]. Array design was performed according to the previously established method [13].

Each probe is 40 nucleotides long, while miRNA precursors are generally 80 – 150 nt in length. Generally, the probes exist in pairs designed to measure opposing sides of the precursor microRNA stem-loop, which are the also the sites containing the mature microRNAs.

Normalization

Normalization is the process by which error is removed from the raw microarray data, with the result being data that is numerically closer to the real value for each probe in each sample, and therefore more accurate. Error can originate from three sources: biological, systematic, and random noise. Normalization through software attempts to estimate and remove only the systematic error [9]. Several methods of normalization for microarray data exist, e.g. scaling, variance-stabilization, quantile, loess. Each method makes a different set of assumptions about the types of error inherent in the data, and also has a varying level of utility and robustness.

Quantile normalization was originally proposed and advocated by Bolstad, et al. due to its simplicity and efficacy [8]. It is widely used and has been confirmed through several studies to be one of the most robust methods [9]. The underlying assumption of this method is that the distribution of probe intensities (i.e. overall signal intensity) per sample does not change. The name is derived from quantile-quantile (QQ) plots, in which two sets of data can be said to have the same distribution if the plot forms a straight diagonal line. In this case the straight-line is extrapolated to N-space to encompass all samples, and a transformation is calculated.

Variance-Stabilization with Invariants (VSN-Inv) is an extension of the existing variance-stabilization method. This extension, proposed by Pradervand, et al. [9], is a combination of two steps; (1) identification of invariant probes *a priori* from the given data and (2) normalization based on the optimal parameters as calculated from only the invariant set of probes, instead of all data. Pradervand, et al. also note that this method is a way to limit the assumptions that a chosen normalization method makes about the data. In this method, the assumption is that at least a small fraction of probes are invariant over all samples. The ideal invariant probe is one with high mean intensity, and low standard deviation across the dataset. Invariant probes are selected by first using a normal mixture model based on signal to classify each probe as one of two components: high or low intensity. All high-intensity probes selected from this step and then classified again using another mixture model based on standard deviation to classify as either high or low standard deviation, but with an optimal number of components as determined by the clustering algorithm. The first component, the one with lowest standard deviation, contains the final list of invariant probes. Finally, the invariant probes are used to estimate the parameters for standard variance-stabilization, and normalization is done according to those parameters on the full dataset.

III. METHODOLOGY

The analysis in this study comprises four main steps completed in the order listed. The first step consists of the data preparation work, including downloading published data from an online repository and then normalizing it in different ways. The second step is verification of VSN-Inv normalization on the NCI-60 miRNA dataset by analyzing intra-chip and intra-group correlations after normalization. The next step consists of simple analysis and comparison of results in hierarchical clustering, and predictive analysis on both untransformed and log-transformed data, because results in these two areas differ greatly if the data has been log-transformed. The last and final step is an analysis where an additional mRNA expression dataset is integrated, and this step assesses the discovery of significant correlations between the original miRNA data and the integrated mRNA data.

All calculations were performed using the R language and environment [22]. Existing libraries were used where possible, and all custom processing was performed using original source code. All non-standard libraries were installed via the built-in package management of the R environment, or installed as part of the Bioconductor framework [23].

A. Download Data

All datasets were downloaded from ArrayExpress [24]. The mRNA transcript and protein expression profiles were downloaded using the “ArrayExpress” library [25] using accession ID “E-GEOD-5720”. The mRNA dataset contains data from both the

Affymetrix HG-U133A and HG-U133B (human) platforms. In this study only the U133A data were used. The miRNA expression profiles were accessed using accession ID “E-MEXP-1029”. The miRNA dataset cannot be directly imported into R/Bioconductor using the existing library since it lacks proper GPR headers in the data files. It was imported by using the download-only functionality of the ArrayExpress library to retrieve the files, and then manually reconstructing the eSet object using a combination of existing and custom code to parse the expression, sample and array design files. Each dataset is also available from an additional source, GEO and CellMiner, respectively, but ArrayExpress was used due to its automated capabilities within the R environment.

B. Normalization

The mRNA expression profiles were normalized using GCRMA, as performed by the “gcrma” function in the “gcrma” package.

The miRNA data were normalized using three methods: raw background subtraction only, quantile normalization, and VSN-Inv. Normalization was a multi-step process performed according to the steps used with the original publication [7]. (1) background median was subtracted from the median signal; (2) duplicate spots were averaged; (3) normalization; (4) and finally control cell lines were averaged together. For the background subtraction-only data step (3) was excluded. Blower, et al. included an additional step where all values having $\log_2$ expression < 5 were set to the median of such values. In order to best discern the differences between normalization methods on low-intensity probes this step was not performed. Quantile normalization was handled using the “normalize.quantiles” method in the “preProcessCore” package. VSN-Inv was

performed using the source code released with the original publication [9], available at (<http://www.unil.ch/daf1/page58744.html>).

Because VSN-Inv and background subtraction do not involve log transformations while quantile returns log-transformed data, data was stored in both untransformed and log-transformed forms after normalization, to allow comparison in both modes. Log transformation was done in base two, and a small value was added to each dataset to make the minimum value 1 to remove negatives and zeros.

C. Clustering and Predictive Analysis

For clustering and predictive analysis, suitable probes were selected according to the criteria used by Blower, et al. Suitable probes are those probes having expression ≥ 256 ($\log_2$ expression ≥ 8) in at least 10% of the samples. Suitable probes were selected separately for each dataset after normalization.

Hierarchical clustering was performed using the “hclust” method in the default “stats” library. The distance matrix between samples was calculated separately beforehand, using a correlation metric as the distance between samples. The distance is defined as $1 - \text{cor}(\text{sample1}, \text{sample2})$, meaning that two perfectly correlated samples (Pearson’s $r=1.0$) have distance 0.

Predictive analysis and calculation of misclassification rates for predicting tissue of origin were performed using PAM_c-PAM_p (also called simply PAM-PAM), the same technique as Blower, et al. PAM-PAM is a two-step predictive analysis technique, in which first k number of representative probes are found by clustering probes into k clusters using partitioning around medoids (PAM_c), where each cluster is described by its

central-most probe (the medoid), and then inputting the expression levels of the k probes into Predictive Analysis for Microarrays (PAM_p), to perform predictive analysis against a specified classification variable. Partitioning around medoids is a clustering technique similar to k-means, but whereas k-means seeks to minimize the Euclidean distance between samples, PAM_c instead seeks to minimize the dissimilarity. The classification variable used for predictive analysis was the sample diagnosis, i.e. tissue of origin. 17 of the cell lines were excluded following the precedence set by Blower, et al., which is due to their lack of data or lack of differentiation from other cancer lines (breast, lung, prostate, and melanoma LOX IMVI), and therefore low predictive power.

Partitioning around medoids (PAM_c) was performed using the `pam()` method in the “cluster” library. Extending the analysis previously made by Blower, et al., this step of the analysis was performed in two separate ways: (1) specifying the cluster count $k=40$ as done by Blower, et al., which always yields 40 probes as input to the second PAM step, and (2) quantitatively determining the optimal cluster count separately for each dataset, which yields between four and 100 input probes to the second step. Optimal cluster count was found using the `pamk()` method in the “fpc” library. `Pamk()` works by testing each possible cluster size in a given range and selecting the one resulting in the largest average silhouette width. The silhouette width is a metric unique to the “medoid” partitioning technique, which describes the amount of dissimilarity a given variable has from its final assigned cluster, with the range being -1.0 to +1.0. Therefore, the optimal number of clusters also results in the least amount of dissimilarity among all clusters on average, e.g. the highest average silhouette width. In this study, preliminary analysis showed that all datasets resulted in three clusters being optimal. Because this was not

very informative the minimum allowable cluster count was restricted to four, and maximum was set to 100.

Predictive analysis for microarrays was performed using the pamr package. As the predictive analysis step is non-deterministic, predictive analysis was performed with six-fold cross validation and 100 iterations, following the precedent set by Blower, et al., saving the misclassification rate from each iteration. The threshold variable required by pamr was set to zero, meaning that all input probes should be included in the predictive analysis.

D. Correlations

Finally, significant miRNA-mRNA interactions were found by integrating the resultant data from each normalization method with the mRNA expression profiles. Only 59 samples are in common, as the H23 cell line lacks a mRNA expression profile. Candidate probes for each dataset were selected by finding all probes having at least 1 log-fold change between the lowest and highest samples as done previously [12]. Correlations were then calculated on both the log and untransformed data using the built-in “cor” method in R, and p-values for each interaction according to Equation 1. In this equation r is the Pearson’s correlation coefficient, n is the sample size, and pt is the “R” function to perform the Student’s t test. Then p-values were corrected using Benjamini-Hochberg to control the false discovery rate. Significant interactions were determined at three significance values, p-value $\leq$ 0.001, 0.01, and 0.05.

$$p - value = 2 * pt(-|r| * \sqrt{\frac{(n-2)}{1-r^2}}, n-2) \quad (\text{Eq 1})$$

IV. RESULTS

A. VSN-Inv Invariants

First, the appropriateness of VSN-Inv normalization on the NCI-60 dataset was assessed. VSN-Inv assigned 32% of the probes to the high-mean component. Of these probes, 31 were identified as high-mean and low-SD invariants, listed in (Table 2) (Figure 2). The intensities were then plotted to verify (Figure 3). Many of the probes appear to be at or near background level, so the selection may not be optimal. The signal mean cut-off was 409 or greater.

Table 2 – Probes identified as invariants in VSN-Inv normalization

hsa-let-7iNo1	hsa-mir-184-precNo2	hsa-mir-339No2	hsa-mir-424No1
hsa-mir-029a-2No2	hsa-mir-196bNo2	hsa-mir-346No1	hsa-mir-429No1
hsa-mir-124a-1-prec1	hsa-mir-197-prec	hsa-mir-371No1	hsa-mir-431No2
hsa-mir-124a-3-prec	hsa-mir-198-prec	hsa-miR-373*No1	hsa-mir-483No1
hsa-mir-129-2No1	hsa-mir-212-precNo1	hsa-mir-373No2	hsa-mir-498No1
hsa-mir-130bNo2	hsa-mir-302dNo1	hsa-mir-376bNo1	hsa-mir-519dNo1
hsa-mir-135a-1No1	hsa-miR-324-5pNo1	hsa-mir-377No1	hsa_mir_320_Hcd306left
hsa-mir-148bNo2	hsa-mir-328No1	hsa-mir-412No1

Figure 2 – VSN-Inv standard deviation versus mean component plot

This plot is automatically generated by VSN-Inv normalization and shows the classification of probes by mixture models, and the selected invariants (Invariants = SD Component 1). The vertical dashed line shows the mean cut-off from the first clustering step, with the low-mean component being probes to the left of the line, and the high-mean component being the probes to the right. The horizontal line shows the SD cut-off from the second clustering step, with all points below the line belonging to SD component 1, the component with the lowest SD distribution.

Figure 3 – Intensities of selected invariant probes
Lines represent loess fit on sorted data.

Total sample intensities before and after normalization were analyzed (Figure 4). Samples were plotted according to diagnosis, and then sorted within groups in order of increasing intensity. It can be seen that quantile normalization translated all samples to have the same total intensity, as per its design and underlying assumption (visualized as the horizontal line of points in the middle). However, there is more than a three-fold difference in signal intensity between the lowest and highest samples in the ‘raw’ data. Sample intensity does not appear to correlate strongly to diagnosis, and indeed, there are large differences within each group. The assumption of quantile normalization may be incorrect, but it is not evident whether the large differences in signal intensity are due to

technical or biological differences. VSN-Inv normalization sample intensity followed the raw data much more closely.

Figure 4 – Scatter plot of total sample intensities before and after normalization

B. Assessing Normalization via Duplicate Probes and Control Samples

Next all 35 pairs of duplicated probes were selected from each normalized data. They were box plotted to determine the overall distribution (Figure 5). Quantile normalization appears to lower the overall correlation between probes, however some correlations do appear to increase. VSN-Inv normalization greatly increases the correlation between all pairs except for a few outliers.

Figure 5 – Box plot of correlations between the 35 duplicate probes

Figure 6 - Box plot of correlations coefficients between all probes

To determine whether the boost with VSN-Inv normalization was an inherent side effect, all probe-probe correlations were calculated ($627 \times 626 / 2 = 196,251$ total interactions) and plotted (Figure 6). VSN-Inv normalization appears to strengthen correlations both positively and negatively, with a slight imbalance towards positive (notice the widening of top and bottom quartiles beyond the other datasets). Quantile normalization appears to compress correlations towards zero.

Because many of the probes are near background level, and a known side-effect of VSN normalization is that low intensity probes can have a large increase of variability [9], all 35 duplicate probes were scatter plotted (Figure 7). For the high-intensity signals, *has-mir-106aNo1*, *has-mir-106-prec-X*, *has-mir-107No1*, and *has-mir-107-prec-10*, both normalization methods performed virtually identically, and there is nearly perfect agreement between the probes in each pair. For the remaining low-intensity pairs near background level, quantile normalization shows only a very small effect on signal intensity, and very little correlation. VSN-Inv (blue) shows a large influence on the low

intensity signals, yielding strong correlations in nearly every case. It is expected that each probe pair should have a strong correlation, but in this case, due to the extreme disagreement between the two normalization methods, a conclusion cannot be made whether VSN-Inv is making the data closer to the real signal level, or exhibiting strong error.

Figure 7 – Scatter plot of signal for all 35 duplicate probe pairs in miRNA dataset

Lines show loess fit. BG subtraction only (black), quantile (red), VSN-Inv (blue).

Finally, assessment of agreement between control lines in the miRNA dataset was performed. For each group of control lines, all sample-sample correlations were calculated for the five samples in the three control lines, A549, K-562, and PC-3, for each normalization method (Figure 8). In the normalization steps described in the Methods section, step (4), averaging control lines together, was skipped for this analysis. Quantile normalization appears to negatively impact the agreement between control lines. VSN-Inv has no effect on the inter-sample correlations because VSN-Inv linearly transforms each sample, in effect changing the slope of the plot between each pair, but not the strength of the correlation.

Figure 8 – Box plot of correlations between control samples

BG subtraction only (white), quantile (red), VSN-Inv (blue).

C. Clustering

Probes selected for inclusion in hierarchical clustering were distributed as follows: (1) untransformed data: background subtraction only - 280, quantile - 267, and VSN-Inv: 524; (2) log-transformed data: background subtraction only - 301, quantile - 267, VSN-Inv - 501. Overlap between selected probes can be seen in the Venn diagrams (Figures 9-10). There was very high agreement between normalizations, with 266 and 267 probes being selected in all datasets. The high amount of probes selected by VSN-Inv is likely a side-effect, indicating that many probes near background level were elevated above the cut-off. The probes selected for inclusion in quantile normalization, 267, differ from the probes selected in the study by Blower, et al., 279, because this study did not adjust for probe bias or batch effect.

Probes selected for hierarchical clustering

Figure 9 – Venn diagram of probes selected for hierarchical clustering

Probes selected for hierarchical clustering (log)

Figure 10 – Venn diagram of probes selected for hierarchical clustering (log transformed)

The results of hierarchical clustering can be seen in Figure 11-12. Hierarchical clustering on log-transformed data is clearly superior, and clusters samples by disease into much cleaner groups. There are several qualitative ways to compare the results between datasets. (1) Overall discrimination between diseases: melanoma and leukemia cell lines should be tightly coupled, while lung cancer and breast cancer lines are expected to be more dispersed, following the previous hierarchical clustering results based on mRNA expression [7]. Both quantile and VSN-Inv normalizations fit these results, with leukemia cells being linked perfectly together and far from the other lines in the log-transformed data. Additionally the majority of the CNS and CO lines are grouped in all cases, and especially tightly coupled in the log-transformed data. VSN-Inv with log-transformation resulted in the best clustering regarding the CNS and CO lines, outperforming quantile, with only one sample not grouped with the others in both groups. (2) Cell line pairs known to be biologically similar [7] should be clustered together. (a) Lines ME-MDA-MB-435 and ME-MDA-N were closely clustered in the VSN-Inv and background subtracted data regardless of transformation, but only closely by quantile with transformation. (b) The BR-TD47 and BR-MCF7 lines were clustered closely in all cases except quantile normalization without log transformation. (c) The OV-OVCAR-8 and OV-NCI-ADR-RES lines were closely clustered in all data sets. Overall, the clustering performed on the VSN-Inv normalized data with log transformation appears to cluster most cleanly by tissue groups, with quantile normalization with log transformation being a close second.

Figure 11 – Complete-linkage hierarchical clustering of samples

transformation (Figure 14). Out of all results, VSN-Inv normalization had the highest misclassification rate when run on untransformed data (0.35), yet also the lowest misclassification rate when run on log-transformed data (0.12). The high misclassification rate of VSN-Inv on untransformed data is a clear sign of over-fitting, which is likely due to probes near background level. The best classification rate overall is with vsn-inv normalization and log transformation (median rate 0.19). It outperforms quantile normalization with or without log transformation, medians 0.21 (Wilcoxon rank-sum $p < 10^{-7}$) and 0.19 (Wilcoxon rank-sum $p = 0.02$), respectively.

Figure 13 – Medoid probes with k=40

Figure 14 – Misclassification rates after PAM-PAM with $k=40$

Next, partitioning was performed using the optimal cluster count determined by each dataset. For background-subtracted data the optimal k was 11, quantile: 16, and vsn-inv: 6. Only four probes were shared (Figure 15), which shows significant disagreement relative to $k = 40$. The fact that VSN-Inv identified the fewest medoids despite it also having the highest number of candidate probes (524) shows that it is probably over-fitting to low-intensity probes not included in the other datasets. Over-fitting is then evidenced by its high misclassification rate, median 0.51 (Figure 16), well beyond the rates of raw and quantile, medians 0.28 and 0.26 respectively.

Overall, VSN-Inv normalization with log transformation shows the best performance under PAM-PAM, with the lowest misclassification rates of all datasets.

Medoid Probes with Optimal k

Figure 15 – Medoid probes with optimal k

Figure 16 – Misclassifications rates after PAM-PAM with optimal k

E. Correlations

Next the ability to extract meaningful miRNA-mRNA interactions was analyzed. Candidate probes were selected with the following results: mRNA – 15,639 probes, quantile – 627 probes, vsn-inv – 625 probes. Notably, the 627 probes selected here differs from the 555 probes selected in the study by Wang and Li because in this study the step after normalization to set all intensities < 5 to the median of such values was not performed. Correlations were checked using three significance levels, 0.001, 0.01, and 0.05, and counts were compiled over which interactions were found in both or each datasets. Positive and negative correlations were compiled separately.

Similar results were found in both the untransformed and log transformed data. The VSN-Inv data finds significantly fewer interactions at all significance levels, and both signs (Figure 17) (Table 3), despite its effect of significantly raising the variance of low intensity probes, and the wider distribution of inter-probe miRNA correlations as noted earlier.

Figure 17 – Distribution of significant miRNA-mRNA correlations among differentially expressed probes

Table 3 - Distribution of significant miRNA-mRNA correlations among differentially expressed probes

Correlation	-			+		
Significance	0.001	0.01	0.05	0.001	0.01	0.05
Untransformed Data						
Both	11	93	568	7060	13216	23832
Quantile	170	1192	4882	38509	68617	115363
VSN-Inv	25	119	902	15508	26856	46555
Log Data						
Both	527	1817	5459	788	2883	8082
Quantile	3424	9225	23943	3846	10162	25815
VSN-Inv	508	1924	6093	2622	5755	12585

V. CONCLUSION

VSN-Inv normalization shows improvements in inter-group correlations among control groups, and greatly increased intra-chip agreement between duplicate probes on the NCI-60 microRNA dataset. The agreement is significantly different from quantile normalization, adding evidence once again that the choice in normalization can greatly affect the results of a study. Although these results initially favor VSN-Inv for normalization, the results must be weighed in accordance to the details of the remaining analysis, which show a few problem areas. The correlations of intra-chip duplicate probes, which favor VSN-Inv over quantile normalization, comprise many weak signals that are known to be greatly affected by that type of normalization, meaning that they may be false positives and not true correlations. VSN-Inv works well for predictive analysis and classification, but requires a strictly selected probeset and then only on log-transformed data. However, in some specific cases VSN-Inv normalization does yield results that are more accurate. Hierarchical clustering works best on vsn-inv normalized data, and can be done with or without log transformation, and the best classification rate for tissue of origin is achieved using vsn-inv normalization with log transformation. Finally, using vsn-inv normalization for the discovery of miRNA-mRNA interactions results in only a fraction of identified pairs versus using quantile normalization, meaning that it is possibly more discerning in this area.

VI. FUTURE DIRECTION

Future work could enrich the analysis by extending it in several ways. First and foremost, additional normalization methods should be considered. There are other normalization methods which do not have an underlying assumption as strict as the one made by quantile normalization, and therefore may also have better performance than quantile, and comparable performance to VSN-Inv.

Additionally, the relative performance of each normalization method in hierarchical clustering should be quantified so that an objective comparison can be made. Preceding that, a method to measure the correctness of the clustering programmatically would have to be determined. A manual method to measure the correctness is to determine the minimum number of swaps necessary to cluster together all cells with the same tissue of origin, however, this is very similar to the standard predictive analysis, and therefore may not offer much additional insight. Furthermore, the significant miRNA-mRNA correlations found in each dataset should be compared against the list of validated targets in Tarbase, thereby offering a metric by which to compare the accuracy of the normalization methods.

Furthermore, the invariant selection process selected some probes that were weaker than expected. This may be due to the strong restriction that probes be classified in only two clusters, high and low intensity. It is possible that allowing a larger number of clusters to be defined, similarly to the classification of components for standard

deviation, that a smaller set of probes, but of higher mean signal, would be selected as the invariants, and therefore yield a more accurate normalization.

Finally, a recent study shows that the expression levels of some microRNAs correlates to cell doubling time [26]. This may offer light as to the large variance in total sample intensity across the NCI-60 panel, and possibly a more appropriate or novel normalization method.

REFERENCES

1. Wienholds, E., et al., *MicroRNA expression in zebrafish embryonic development*. Science, 2005. **309**(5732): p. 310-1.
2. Couzin, J., *Genetics. Erasing microRNAs reveals their powerful punch*. Science, 2007. **316**(5824): p. 530.
3. Hebert, S.S. and B. De Strooper, *Molecular biology. miRNAs in neurodegeneration*. Science, 2007. **317**(5842): p. 1179-80.
4. Yang, M., et al., *Circadian regulation of a limited set of conserved microRNAs in Drosophila*. BMC Genomics, 2008. **9**: p. 83.
5. Rodriguez, A., et al., *Requirement of bic/microRNA-155 for normal immune function*. Science, 2007. **316**(5824): p. 608-11.
6. Meyer, S.U., M.W. Pfaffl, and S.E. Ulbrich, *Normalization strategies for microRNA profiling experiments: a 'normal' way to a hidden layer of complexity?* Biotechnol Lett, 2010.
7. Blower, P.E., et al., *MicroRNA expression profiles for the NCI-60 cancer cell panel*. Mol Cancer Ther, 2007. **6**(5): p. 1483-91.
8. Bolstad, B.M., et al., *A comparison of normalization methods for high density oligonucleotide array data based on variance and bias*. Bioinformatics, 2003. **19**(2): p. 185-93.
9. Pradervand, S., et al., *Impact of normalization on miRNA microarray expression profiling*. RNA, 2009. **15**(3): p. 493-501.
10. Shankavaram, U.T., et al., *Transcript and protein expression profiles of the NCI-60 cancer cell panel: an integromic microarray study*. Mol Cancer Ther, 2007. **6**(3): p. 820-32.
11. Papadopoulos, G.L., et al., *The database of experimentally supported targets: a functional update of TarBase*. Nucleic Acids Res, 2009. **37**(Database issue): p. D155-8.
12. Wang, Y.P. and K.B. Li, *Correlation of expression profiles between microRNAs and mRNA targets using NCI-60 data*. BMC Genomics, 2009. **10**: p. 218.
13. Liu, C.G., et al., *An oligonucleotide microchip for genome-wide microRNA profiling in human and mouse tissues*. Proc Natl Acad Sci U S A, 2004. **101**(26): p. 9740-4.
14. Crick, F., *Central dogma of molecular biology*. Nature, 1970. **227**(5258): p. 561-3.
15. Lee, R.C., R.L. Feinbaum, and V. Ambros, *The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14*. Cell, 1993. **75**(5): p. 843-54.
16. Griffiths-Jones, S., et al., *miRBase: tools for microRNA genomics*. Nucleic Acids Res, 2008. **36**(Database issue): p. D154-8.
17. Brennecke, J., et al., *Principles of microRNA-target recognition*. PLoS Biol, 2005. **3**(3): p. e85.

18. Carthew, R.W. and E.J. Sontheimer, *Origins and Mechanisms of miRNAs and siRNAs*. Cell, 2009. **136**(4): p. 642-55.
19. Vasudevan, S., Y. Tong, and J.A. Steitz, *Switching from repression to activation: microRNAs can up-regulate translation*. Science, 2007. **318**(5858): p. 1931-4.
20. Thomas, M., J. Lieberman, and A. Lal, *Desperately seeking microRNA targets*. Nat Struct Mol Biol, 2010. **17**(10): p. 1169-74.
21. Bussey, K.J., et al., *Integrating data on DNA copy number with gene expression levels and drug sensitivities in the NCI-60 cell line panel*. Mol Cancer Ther, 2006. **5**(4): p. 853-67.
22. Team, R.D.C., *R: A language and environment for statistical computing*. 2010, R Foundation for Statistical Computing: Vienna, Austria.
23. Gentleman, R.C., et al., *Bioconductor: open software development for computational biology and bioinformatics*. Genome Biol, 2004. **5**(10): p. R80.
24. Parkinson, H., et al., *ArrayExpress update--from an archive of functional genomics experiments to the atlas of gene expression*. Nucleic Acids Res, 2009. **37**(Database issue): p. D868-72.
25. Kauffmann, A., et al., *Importing ArrayExpress datasets into R/Bioconductor*. Bioinformatics, 2009. **25**(16): p. 2092-4.
26. Gaur, A., et al., *Characterization of microRNA expression levels and their biological correlates in human cancer cell lines*. Cancer Res, 2007. **67**(6): p. 2456-68.

CURRICULUM VITAE

Martin Thomas Disibio, II

EDUCATION: M.S., Computer Science, Complete December 2010
University of Louisville, Louisville, KY, GPA 4.0

B.S., Computer Engineering and Computer Science 2003
University of Louisville, Louisville, KY, GPA 3.6

**PROFESSIONAL
EXPERIENCE:**

Senior Software Developer
Yum! Brands, Inc. KFC IT
2003 – Present

Design and development of vertically integrated suite of software,
operating system and infrastructure, supporting financial, inventory
and other business-related processing.

PRESENTATIONS: Poster presentation at SAS data mining conference M2010.