Chapter-II

Integration of miRNA Expression Datasets: Normalization of Inter-platform Data

Sections of this chapter are part of published work: Bargaje, R., Hariharan, M., Scaria, V., and Pillai, B. (2010) RNA 16(1): 16-25
2.1. INTRODUCTION

miRNAs are small ncRNA molecules, of about 21-23nt in length, that negatively influence post-transcriptional expression of target genes bearing partial complementary regions in their 3'UTR, in eukaryotic cells (Bartel, 2004). Several miRNAs are expressed in specific spatio-temporal patterns during development (Krichevsky et al., 2003) and in adult tissues (Liang et al., 2007). Their role in development and several other biological processes is summarized by Wienholds et. al. (Wienholds and Plasterk, 2005). More recently, reports implying their role in various disease related processes ranging from cancer (Calin and Croce, 2006) to host-virus interaction have emerged from studies by our group (Hariharan et al., 2005;Ahluwalia et al., 2008) and others (Omoto et al., 2004;Lecellier et al., 2005;Jopling et al., 2005). Aberrant expression of miRNAs has been reported in several cancers leading to the suggestion that they could act as bio-markers (Cummins and Velculescu, 2006). The spatio-temporal expression patterns of miRNAs can convey important information regarding their functional role. Expression patterns of miRNAs are important in discovery of biomarkers, functional characterization of miRNAs and in the study of gene regulation in development and disease. Due to the importance of miRNA expression patterns, many platforms have been developed for rapid high-throughput expression profiling of miRNAs (Castoldi et al., 2006;Wang et al., 2007;Chen et al., 2008). However, the expression profiles generated by these methods have not been compared. In the past, comparison of mRNA microarray data has helped in generating valuable reference expression profiles (Shi et al., 2006;Chen et al., 2007). Moreover, miRNA expression profiles are of prognostic value in treatment of cancer and of diagnostic value in molecular classification of cancers and identification of cancer sub-types (Calin and Croce, 2006).

miRNA expression profiling methods measure the expression level of functional and mature miRNAs, ideally distinguishing them from precursor molecules and highly homologous isoforms. At least three high-throughput methods have been applied to identify the expression profiles of miRNAs, in addition to the information on abundance generated indirectly from small RNA cloning. A microarray based detection method which employs complementary oligonucleotide probes specific to
known miRNAs were first reported by Liu et. al. 2004, (Liu et al., 2004a) and is now used widely for high-throughput miRNA detection. miRNA-microarray platforms incorporate different design strategies to improve the specificity of the probes. These include locked nucleic acid based probes (Castoldi et al., 2006) and an extended loop and 5’G to capture 3’C introduced into miRNAs during labeling in miRNA arrays (Wang et al., 2007). RNA-primed, Array-based, Klenow Enzyme (RAKE) assay is a modification of the basic microarray protocol which involves extension of the universal tag containing antisense probe in situ on the microarray by incorporation of fluorescently labeled nucleotides by Klenow enzyme (Lu et al., 2005). The “Bead-based” detection platform uses capture oligonucleotides cross-linked to luminex beads (Nelson et al., 2004). Substantially different from the microarray based methods, this method relies on flow cytometry of beads to uniquely identify and count hybridization events. Besides the high-throughput methods described above, more accurate low throughput methods like Northern hybridization are used in studies on individual miRNAs and to validate findings of high-throughput studies. Real-time PCR based on Taqman assay chemistry has emerged as a gold standard for gene expression assays (Shi et al., 2006). A modified Taqman assay based on stem-loop primer based reverse transcription has been used to detect individual miRNAs as well as to generate tissue specific miRNA expression profiles (Liang et al., 2007). Although there exists a large body of data from miRNA expression profiling studies in public repositories, typically an experimentalist has to use tedious low-throughput methods like northern blotting to establish the expression pattern of a specific miRNA of interest.

In this study miRex: a web based database and analysis tool for miRNA expression data has been developed. The datasets have been individually inspected for data quality, pre-processed and normalized to facilitate comparison of expression profiles between experiments. The expression profiles of the RNAs are displayed graphically and the underlying numerical data is readily retrieved. Expression profiles of entire experiments can be viewed in groups according to publications or the expression of a few miRNAs of interest can be retrieved and visualized separately. Currently, miRex is the only database that contains miRNA expression data spanning 1786 experiments for Human, Mouse, Zebrafish, Arabidopsis and several plants.
Integration of miRNA Expression Datasets: Normalization of Inter-platform Data

Tissue specific expression of a few miRNAs has been revealed by high-throughput experimental analysis using several approaches (Lagos-Quintana et al., 2002; Liu et al., 2004a; Nelson et al., 2004; Lu et al., 2005; Mineno et al., 2006; Nakano et al., 2006; Liang et al., 2007; Landgraf et al., 2007). However, there has not been any systematic comparison of miRNA expression profiles generated by different platforms. Recently, normalization methods originally developed for mRNA expression profiling have been evaluated for miRNA data generated using the Agilent platform (Pradervand et al., 2009). Besides this, there have been no comprehensive studies on inter laboratory and inter-platform comparison of miRNA expression profiles.

Here expression data from high-throughput studies from different laboratories has been normalized and scaled, to create a compendium of miRNA expression profiles for mouse and human tissues, cell lines, cancer samples and developmental stages. Normalization methods were compared to identify methods suitable for normalization of miRNA data. Tissue specific miRNA expression patterns found were in agreement with previously reported profiles and novel expression patterns. The analysis also identified 18 constitutively expressed miRNAs and studied expression of miRNAs of a family in different tissues.

2.2. MATERIALS AND METHODS

2.2.1. Source of data

High throughput experimental data on expression profiling of miRNAs from public repositories, 614 experiments spanning 25 datasets deposited in GEO, the public repository for high-throughput gene expression data hosted by NCBI (Edgar et al., 2002; Barrett et al., 2007), and 1132 experiments from 18 datasets from ArrayExpress (Parkinson et al., 2007) was collected. Besides the microarray based data, there is a dataset consisting of 40 samples analyzed by real time PCR (Liang et al., 2007). The data retrieved from repositories was individually checked and data in the form of fold changes was discarded since such data cannot be compared directly between experiments. The data was further checked for background correction; if the data had been background corrected it was used for further processing. Otherwise a background correction was performed by deducting the background median intensity
Integration of miRNA Expression Datasets: Normalization of Inter-platform Data

from feature intensity. All negative values were converted to null values since negative and null values in expression data correspond to miRNAs expressed at levels lower than the minimum detection sensitivity of the experiment. After this transformation if the percentage of null values exceeded 50%, then the data was discarded. Null values were replaced with imputed values in the remaining datasets since null values would interfere with further numerical operations.

2.2.2. KNNimpute

KNNimpute is a method of imputing missing values by replacing the missing value with a weighted average of “K” other miRNAs of a similar expression profile (nearest neighbors) (Troyanskaya et al., 2001). Missing or null values were imputed using the program for implementation of the KNNimpute algorithm available from Stanford Microarray database. The parameters used include, Euclidean matrix and an optimal K value of 15.

2.2.3. Comparison of expression profiles

miRNA expression data was downloaded and background subtracted signal intensity values were calculated. Mean, median, constitutive miRNA and quantile normalization was carried out using commercial microarray analysis software, Avadis. Boxplots were generated using R. The expression levels were then log transformed and Z scores (Cheadle et al., 2003) were calculated using the formula

\[
Z_{\text{score}g} = \frac{\text{Signal}_{g} - \text{Mean Signal}_{g,\ldots,n}}{\text{Standard Deviation}_{g,\ldots,n}}
\]

Where g is any gene in the microarray.

The Z scores were used for clustering miRNAs using the commercial software for clustering microarray data, Avadis. Z ratios (Cheadle et al., 2003) represent expression level of a miRNA in one tissue relative to its average expression level across all other tissues. For each miRNA, the difference between the Z score in a tissue (\(Z_T\)) and the average Z scores across all other tissues (\(Z_C\)) was divided by the Standard Deviation of the differences.
Integration of miRNA Expression Datasets: Normalization of Inter-platform Data

\[ Z_{\text{ratio}}_{g} = \frac{Z_T - Z_C}{\text{Standard Deviation}} \]

Where, \( g \) is any gene in the microarray, \( Z_T \) is average Z score of gene \( g \) in a tissue (T) and \( Z_C \) is average Z score for gene \( g \) across all other tissues.

2.2.4. Database construction

The database was developed by using in-house Perl scripts for Graphical User Interface (GUI) and using MySQL as back-end data repository. The Z-score values generated for all the experiments is available in form of dynamically generated heatmaps. The user can set cut-offs for visualization. The graphical data is available for download as a text file. A module called the ID converter that allows users to convert miRNA IDs from one nomenclature system to another was created. A master list and linked each miRNA to its aliases arising from different systems of nomenclature like HGNC, RFAM, systematic names from miRBase and miRBase IDs was compiled.

2.2.5. Identification of constitutive miRNAs

Clustering of Z scores was performed using Self Organizing Maps (SOM) (Tamayo et al., 1999) in the microarray analysis software package Avadis. A coefficient of variation (CV) of Z scores was calculated for each miRNA. Since Z scores can be negative, the CV can also be negative. The mean was squared and square root calculated to convert the CV to a positive value. The miRNAs with CV below 0.65 were selected The Z scores of these miRNAs were visualized using a heatmap.

\[ CV_g = \frac{\text{Standard Deviation}_{t1...tn}}{\sqrt{\text{Mean}_{t1...tn}}} \]

Where, \( g \) is any gene in the study. Mean and standard Deviation was calculated across all tissues (t1…tn).
2.3. RESULTS

2.3.1. Data collection

High throughput experimental data on expression profiling of miRNAs from public microarray data repositories like GEO (Edgar et al., 2002; Barrett et al., 2007) and Array Express (Parkinson et al., 2007) was collected. The experiments include several tissue specific expression profiling studies using different platforms on cancer biopsy samples and developmental stages (summarized in Appendix I). The platforms include bead based study and microarray platforms developed by different laboratories. Small RNA cloning and RNA-seq studies were not included since there are very few studies using these techniques. Each dataset was processed through a series of steps summarized in the flow chart provided in figure 2.1.

![Flowchart](flowchart.png)

Figure 2.1: Steps in data generation. A flowchart summarizing the steps carried out to generate the normalized data.
Datasets providing only relative fold change values derived from comparison with reference sample were not used since this data is fundamentally different from expression level measurement and not suitable for comparison with other datasets. The data was subjected to a series of quality control checks to ensure that datasets with a majority of undetected spots were excluded. If the reported data was already background corrected, the data was used directly. In datasets not corrected for background, the median value of background intensities was subtracted from feature intensity. Negative values were substituted with null values and subsequently all null values were replaced with imputed values. KNNimpute, based on imputation of expression level of similar miRNAs was found to be more suitable than other methods which use mean or median (Troyanskaya et al., 2001). The low number of replicates in most datasets renders mean and median measurements unreliable for imputation. Before normalization the data was linearized using log transformation if it had not been previously done.

2.3.2. Data normalization

The numbers of miRNAs in miRNA-microarray experiments are low, typically in hundreds, compared to a few thousands of genes measured in conventional gene expression profiling. Moreover, there is a greater diversity of gene expression profiling platforms already used in miRNA profiling, bead based methods, Massively Parallel Signature Sequencing (MPSS) and oligonucleotide arrays. Both these factors can influence the suitability of conventional normalization methods. The suitability of several normalization methods was investigated, to choose the method most suitable for pre-processing gene expression profiles, giving preference to measures of normalization that are independent of other datasets, since such a strategy would enable one time, internal normalization of each dataset irrespective of subsequent addition of newer datasets.

A subset suitable for testing normalization methods was selected. This subset included miRNA expression profiles of six experiments (two bead based studies and four microarray studies) corresponding to prostate tissue and four (one bead based study and three microarray studies) related to brain tissue. The prostate datasets
Integration of miRNA Expression Datasets: Normalization of Inter-platform Data

included two studies on PC3 cell line that is derived from prostate cancer. The un-normalized expression values showed large differences between different methods and laboratories. In both prostate and brain datasets, normalization to mean (Figure 2.2) or median had comparable effect.

Figure 2.2: Choice of normalization method. The box plots were generated using a subset of normalized data. Four brain samples and six PC3 and prostate samples were selected from bead-based (B1, B2) and oligonucleotide microarray studies (M1–M4). The distribution of log transformed un-normalized values (A), mean normalization (B), quantile normalization (C), normalization with respect to mean expression level of sixteen constitutive miRNAs (D), Z score normalization (E) is shown.
The dataset after normalization had comparable central tendency, but the spread of the data was highly variable. Constitutive genes have earlier been used for normalization of gene expression profiles. Since they are measured during the experiment, they serve as internal controls in quantitating expression. A recent attempt to compare normalization methods for miRNA data generated using Agilent microarrays also reported that normalization based on invariant normalization performed better than other normalization methods. Mean of expression level of a set of sixteen miRNAs that showed minimum variability in these datasets, was used as a normalization factor. This method of normalization was reasonably successful in comparing datasets generated using the same platforms. However, normalization using constitutive miRNAs was ineffective when comparing bead based and microarray based datasets. Quantile normalization and Z scores were both successful in transforming the data sets generating comparable means and scale. The normalization factors in mean, median, invariant and quantile based normalization are derived from multiple datasets. Consequently, addition of new datasets alters the normalized values of the existing datasets making it difficult to create databases of pre-normalized expression profiles. However, Z scores, unlike other normalization methods, are not influenced by the addition of new datasets. Log transformed data from all the datasets was normalized using the Z score method for further studies.

2.3.3. miRex: A web based resource for miRNA expression profiles

A database of Z scores from miRNA expression studies in thirteen organisms including Zebrafish, Arabidopsis, mouse and human, miRex (miRNA expression; http://miracle.igib.res.in/mirex/datasets.html) has been constructed by our group. Z scores of 1786 experiments spanning 44 datasets, used in this study, are available through miRex. miRex is designed to address the needs of both the miRNA research community as well as biologists not directly involved in miRNA detection. The user may arrive upon the data of personal interest from the main page by several alternative queries (Figure 2.3A): (1) selecting from a table of details on the publication or by searching using (2) keywords (3) miRNA name (4) organism (5) platform (description of the experimental condition). The keywords include the terms associated with tissues, cancer conditions, cell lines and treatments used in expression profiling.
Integration of miRNA Expression Datasets: Normalization of Inter-platform Data

Figure 2.3: The user interface of the database. A) Opening page of miRex database, specific options can be chosen to browse through data, with section for recent updates. (B and C) Available datasets and the search interface allows user to select multiple options of the Experiment Type (Microarray, RAKE, Bead-based or Real time), the Organism and then query the server for specific miRNA, or a specific keyword, a particular publication’s author or the tissue of expression. (D and E) Simple search interface and compare datasets makes it feasible to comparison of different datasets. F) User interface for ID convertor module.

The user can browse through available datasets and miRNAs (Figure 2.3B). The datasets are arranged according to dataset IDs and display associated information such as type of experiment, platform ID, contributors and keywords (Figure 2.4A). The dataset IDs are clickable which leads to detailed information of the dataset including, cells or tissues used, technology used, associated publication, sample description and data contributors (Figure 2.4B). The user can explore datasets of interest by using specific keywords, experiment type, organism, specific miRNA, contributor or tissue type (Figure 2.3C). The resulting datasets can further be individually selected and expression profiles can be visualized in the form of heat-map (Figure 2.4C).

In response to the “simple” query (Figure 2.3D) the user is presented with a subset of datasets and experiments that match the search criteria. The user may subsequently select from this list of experiments for further exploration. The results of this query are presented graphically in the form of a heatmap, with each miRNA representing a
Integration of miRNA Expression Datasets: Normalization of Inter-platform Data

row and each experimental condition a column. The expression levels are displayed as a colored matrix. The user can mouse-over the heatmap to see numerical Z-score value as well as the experimental condition associated with it (Figure 2.4C). The column headers are short descriptions of the experiments derived from the original annotation. The user may restrict the results to highly expressed or low expressed miRNAs by defining appropriate Z-score limits.

Figure 2.4: Output from miRex database. A) Explore dataset and dataset available option displays the datasets with brief information about the datasets. The panel above displays the internal dataset ID, the number of experiments performed in the study, the number of miRNAs used in the experiments, the contributors, Z-score range selected. B displays further elaborate information of individual dataset. C) Heatmap representation of the z-score values of various miRNAs in various conditions analyzed in a particular dataset. miRex miRNA ID is displayed in the first column, the z-score values of the experiments in the subsequent columns (mouse-over indicate the z-score value and the particular experiment). The last column indicates the common name of the miRNA. D) User can select different or similar datasets and two different experiments to compare miRNAs expression profile. E) Standard output of ID convertor. The results are available for download as a text file.

In addition to exploring miRNA expression profiles miRex allows comparison and visualization of expression data from independent experiments (Figure 2.3E). The user can define two experiments, either from same or different datasets, to be compared. The results of the comparison are presented as a heatmap with two columns and miRNAs
detected in both experiments as rows (Figure 2.4D). The data represented in the columns can be downloaded as a text file.

The ID converter module (Figure 2.3F) in miRex server gives the user an advantage of (i) Cross-referencing of specific miRNAs across various universal sequence databases, (ii) associating the miRNA sequence families with unique identifiers, (iii) identifying conserved or highly similar sequences of miRNAs across various organisms, (iv) converting the commonly used nomenclature of miRNA sequences to publication format HUGO Gene Nomenclature Committee (HGNC) identifiers and also (v) the ease of querying a large set of identifiers and obtaining the corresponding associated IDs which can be viewed and also downloaded (Figure 2.4E). To the best of our knowledge, there exists no other server which offers the users the ease of cross referring miRNA identifiers across databases.

The following cases briefly illustrate the utility of the ID converter:

(i) Using miRex ID converter, the user can retrieve the various identifiers associated with any particular miRNA in various databases like EMBL, RFAM, HUGO and miRBase.

(ii) The RFAM ID RF00027 links to all the hsa-let-7 family sequences across different organisms hsa-let-7a to –let-7e and also mmu-let-7.

(iii) Common names like hsa-miR-133b can be mapped to HGNC ID 31759 and HGNC Name MIRN133B.

(iv) Pasting a list of miRNA IDs and querying against miRex ID converter to any other database identifier retrieves a list of associated IDs which can be viewed as well as downloaded to the user’s computer.

2.3.4. Tissue specific miRNA expression

The Z score provides a mean centered rank for the expression level in units of standard deviation (Cheadle et al., 2003). The expression levels are represented in terms of ranked normalized data. Z scores have been used previously by our group and others to compare data between experiments. This method is specially suited for database development since the internally normalized values do not change with
Integration of miRNA Expression Datasets: Normalization of Inter-platform Data

subsequent addition of new datasets. Internal normalization of log transformed intensity values was performed by centering them to the respective averages. The Z scores thus provide an index of the expression level of the miRNA with respect to the cellular pool of miRNA and allow classification of miRNA with respect to their basal and tissue specific expression levels. The Z score expression levels, expressed in terms of ranks cannot be used directly to calculate fold change between samples.

Z ratios, the equivalent of fold changes from Z scores were calculated to overcome this limitation. miRNAs up-regulated in a certain tissue were identified using the average of Z scores from replicates of that tissue as the “test” value and the average expression profile of all the other conditions/tissues as the “control” value. A Z ratio of 1.96, corresponding to 95% confidence interval (p value <0.05) (Cheadle et al., 2003), was used as a cutoff to identify tissue specific miRNAs. This approach of comparing normalized miRNA expression profiles was validated by looking for the Z ratios for miRNAs widely acknowledged to be tissue specific. In a majority of cases, our results agree with the reported literature (Table 2.1).

Table 2.1: Previously reported tissue specific miRNAs and respective Z ratios found in our study

<table>
<thead>
<tr>
<th>miRNA</th>
<th>Organism</th>
<th>Tissue</th>
<th>Literature</th>
<th>Z Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>hsa-miR-10b</td>
<td><em>Homo sapiens</em></td>
<td>Kidney</td>
<td>(Beuvink et al., 2007)</td>
<td>2.40</td>
</tr>
<tr>
<td>hsa-miR-146a</td>
<td><em>Homo sapiens</em></td>
<td>Spleen</td>
<td>(Beuvink et al., 2007)</td>
<td>2.15</td>
</tr>
<tr>
<td>hsa-miR-155</td>
<td><em>Homo sapiens</em></td>
<td>B cell lymphoma</td>
<td>(Eis et al., 2005)</td>
<td>3.46</td>
</tr>
<tr>
<td>hsa-miR-205</td>
<td><em>Homo sapiens</em></td>
<td>Breast</td>
<td>(Baskerville and Bartel, 2005)</td>
<td>7.52</td>
</tr>
<tr>
<td>hsa-miR-205</td>
<td><em>Homo sapiens</em></td>
<td>Thymus</td>
<td>(Baskerville and Bartel, 2005)</td>
<td>2.14</td>
</tr>
<tr>
<td>hsa-miR-223</td>
<td><em>Homo sapiens</em></td>
<td>Spleen</td>
<td>(Beuvink et al., 2007)</td>
<td>6.98</td>
</tr>
<tr>
<td>hsa-miR-30c</td>
<td><em>Homo sapiens</em></td>
<td>Kidney</td>
<td>(Beuvink et al., 2007)</td>
<td>3.14</td>
</tr>
<tr>
<td>mmu-miR-1</td>
<td><em>Mus musculus</em></td>
<td>Heart</td>
<td>(Lagos-Quintana et al., 2002;Sempere et al., 2004)</td>
<td>3.15</td>
</tr>
<tr>
<td>mmu-miR-122</td>
<td><em>Mus musculus</em></td>
<td>Liver</td>
<td>(Lagos-Quintana et al., 2002)</td>
<td>4.68</td>
</tr>
<tr>
<td>mmu-miR-124a</td>
<td><em>Mus musculus</em></td>
<td>Brain</td>
<td>(Lagos-Quintana et al., 2002)</td>
<td>3.16</td>
</tr>
<tr>
<td>mmu-miR-133a</td>
<td><em>Mus musculus</em></td>
<td>Heart</td>
<td>(Lagos-Quintana et al., 2002;Sempere et al., 2004)</td>
<td>3.94</td>
</tr>
</tbody>
</table>
The expression of many of these tissue specific miRNAs were also reflected in cell lines derived from these tissues, for instance, hsa-miR-15b and hsa-miR-19b were found in MCF7 cell line and breast carcinoma tissue in high levels. However, the larger number of conditions available in miRex, allowed us to expand the expression profiles of some miRNAs reported to be tissue specific. For instance, hsa-miR-92 identified in a screen for fetal liver (Fu et al., 2005) specific miRNA was found to be constitutively expressed at high levels in all conditions using miRex. On the other hand, hsa-miR-133a was found to be expressed in heart at levels higher than the widely studied heart specific miRNA, hsa-miR-1, thus offering a better tissue specific miRNA signature (Lagos-Quintana et al., 2002). hsa-miR-155, previously reported to be CLL specific (Eis et al., 2005), was found at high levels in normal blood samples studied by different groups.

Figure 2.5: Tissue specific expression profiles of miRNAs. A) Tissue-specific expression profiles of miRNAs based on Z ratios calculated as described in the Materials and Methods section. The red boxes signify a z ratio of >1.96, which corresponds to a p-value of <0.05 for tissue-specific expression. B) A subset of the miRNAs in (A) show a high basal level of expression across different tissues.
Figure 2.5A provides an overview of tissue specific expression profiles of miRNAs, newly identified in miRex. In agreement with independent experimental findings in mouse (Mishima et al., 2008), the microRNA hsa-miR-125b is amongst the most highly expressed miRNAs in testis and ovary. Interestingly, miR-125b, miR-99a and let7c arise from the same transcript but do not show similar expression profiles. Both miR-125b and 99a are expressed highly in the ovary while let7c, in agreement with its role in neurogenesis (Wulczyn et al., 2007) is expressed highly in the fetal brain. Z scores allowed study of expression profiles of miRNAs arising from genomic loci in close proximity to each other. A subset of miRNAs (from figure 2.5A) showing high z scores in many tissues is shown in figure 2.5B. The heterogeneity of cell types in certain tissue samples like liver, lung and blood could result in overlap with cell-type expression profiles. For instance, hsa-miR-223 reported to be expressed highly in macrophages and granulocytes (Chen et al., 2004) was also found in lung, liver and spleen samples.

2.3.5. Identification of constitutive miRNAs

Although there are several reports of tissue specific miRNAs, there have been very few reports of constitutively expressed miRNAs. Averages of Z scores from different laboratories to derive inter-laboratory validated miRNA expression profiles for various tissues were calculated. Subsequently, miRNAs according to their expression profiles in various tissues were clustered using SOM (Tamayo et al., 1999) (Figure 2.6A). Coefficient of variation was used as a measure of fluctuation in expression of miRNAs in different tissues. The miRNAs with lowest coefficient of variation were identified. The expression profiles of these miRNAs show that constitutive miRNAs with invariant low and high levels exist (Figure 2.6B).

Liang et al, have identified universally expressed microRNAs using real time PCR based measurement of 345 miRNAs in 40 tissues (Liang et al., 2007). Constitutively expressed miRNAs identified in our study were compared with those reported in two previous experimental (Liang et al., 2007; Peltier and Latham, 2008) (Appendix II). The miRNAs identified in each study had low coefficient of variation, with four miRNAs hsa-miR-92, hsa-miR-93, hsa-miR-191 and hsa-miR-103 being identified in
Integration of miRNA Expression Datasets: Normalization of Inter-platform Data

multiple studies including our own. It has been suggested that these miRNAs may be involved in regulation of constitutive metabolic functions. Several constitutively expressed miRNAs, for instance, hsa-miR-23b, 26a, 92, 18a have been shown to occur at fragile sites in genomic regions involved in cancer (Calin et al., 2004b). Besides metabolic functions, constitutive miRNAs could perhaps be involved in cell cycle, apoptosis and growth control pathways irrespective of cell type.

Figure 2.6: Expression profile of miRNAs across different tissues. A) Cluster of 130 miRNA expression profiles across 23 different tissues. Z scores for miRNA expression profiles were calculated as described in Materials and Methods. Average Z scores were calculated wherever multiple replicates of the same sample were available. Clustering was performed using SOM. B) Identification of constitutive miRNAs. Coefficient of variation for Z scores of each miRNA was calculated. miRNAs with < 0.65 CV were selected, and their expression profiles visualized by generating a heatmap of Z scores of 23 tissues.

2.3.6. Expression profiles of miRNA family members

Certain miRNA families have undergone significant expansion in the human genome. Comparison with lower organisms has led to the prediction that increase in sequence variants is associated with functional diversification, through increased range of targets against the miRNA (Chen and Rajewsky, 2006). If indeed, increased sequence variants in humans are associated with specialized functions, they may be expected to show diversified expression profiles. The extent of expression diversification amongst
sequence related miRNAs was checked, using the hsa-let-7 family members. The let-7 family of miRNAs consist of intergenic (hsa-let-7a-2, hsa-let-7i, hsa-let-7e, hsa-let-7a-1 and hsa-let-7f-1), intronic (hsa-let-7c, hsa-let-7g, hsa-let-7d and hsa-let-7f-2) and exonic (hsa-let-7a-3 and hsa-let-7b) miRNAs. The eight members of the hsa-let7 family were clustered according to their Z scores from 140 experiments which included samples from cell lines, in vivo tissue samples and clinical cancer samples. Using a cutoff correlation coefficient of 0.6, hsa-let-7b and hsa-let-7e formed a separate group. The hsa-let-7b and hsa-let-7e miRNAs showed significant up-regulation in a 17 cancer samples. The similarity in expression of the members of the hsa-let-7 family (Figure 2.7) is not correlated to their arising from the same genomic locus since hsa-let-7a-3 and hsa-let-7b; hsa-let-7a-1, hsa-let-7f-1 and hsa-let-7d arise from two clusters while hsa-let-7e arises from a different genomic locus.

*Figure 2.7: Expression levels of different members of the let-7 family of miRNAs. Z scores for expression of eight members of the hsa-let-7 family were calculated. The expression profiles for 140 experimental conditions were clustered using SOM.*
2.4. DISCUSSION

Currently, miRBase serves as a miRNA registry and information resource providing sequence and nomenclature information on miRNAs (Griffiths-Jones et al., 2006). However, it is currently difficult to access and visualize information regarding the expression patterns of miRNAs. During the course of our work, efforts to build databases comprising miRNA genomic organization and predicted or validated targets of miRNA includes, Argonaute (Shahi et al., 2006), miRNAMap (Hsu et al., 2006), small RNA expression smiRNAdb (Landgraf et al., 2007), microRNA.org (Betel et al., 2008) miRZ and miRGator (Nam et al., 2008) These databases provide miRNA expression values derived from individual studies. miRNAMap uses the data from Lu et. al. 2005 (Lu et al., 2005) for expression of miRNAs in normal tissues and tumors. miRZ uses the small RNA cloning based data from Landgraf et. al. (Landgraf et al., 2007). The Argonaute database provides literature based information about expression, rather than quantitative measure of miRNA expression. smiRNAdb and microRNA.org use small RNA cloning data for depicting the expression profiles of miRNAs in Human and Mouse from Landgraf et. al. (Landgraf et al., 2007). miRGator, one of latest databases which has a collection for 12 miRNA related datasets from human and mouse is the only tool that currently allows search and retrieval of miRNA expression data from several studies.

Large numbers of datasets have been generated to analyze expression of miRNAs in cells, tissues under different conditions. miRNA expression has been profiled using many high throughput techniques such as microarrays, sequencing, qPCR, etc. However wealth of information generated by these datasets is not completely exploited. These expression profiles can be further analyzed by integrating the expression datasets. This type of cross comparison can be used to find robust miRNA expression signatures. It also makes the data amenable for integration with other genome-wide datasets such as transcript expression profiles, chromatin organization etc. However such integration of data is possible only when the data generated using different platforms is made comparable.

Generation of expression data using high throughput methods is inherently susceptible to problems stemming from a large number of variables including inadequate
Integration of miRNA Expression Datasets: Normalization of Inter-platform Data

replication, technical errors in the implementation of the protocol (e.g.: poor quality of starting sample, probe design), dye bias in the case of two color experiments, tissue heterogeneity and inherent biological variation. It has been suggested that as many as 39 steps of quality control need to be maintained to ensure reproducibility of the data range (Imbeaud and Auffray, 2005). Inter-laboratory and cross-platform validation (Shi et al., 2006) seems to be the only solution until the technology is robust enough to generate highly quantitative and reproducible data.

Here, we have created miRex, a tool for meta-analysis of miRNA expression data. Where, the data was retrieved from public repositories, processed through quality control steps, log transformation and normalization. Different methods for normalization were tried however; Z scores was found to be a better method for building dynamic database such as miRex. The normalized data is made available online through database called miRex, which is likely to become an important and valuable repository for large scale miRNA expression profiling data besides providing ready access to information on expression patterns of individual miRNAs, compiled from various studies.

The data generated in this process was first checked for the signatures of miRNA expression known from literature. Z ratios are used to find out the enrichment of expression in a particular condition, many of predicted tissue specific miRNAs were already proven to be tissue specific miRNAs in literature, which suggests that the data processed for miRex is of good quality. Successively, miRNAs that are constitutively expressed in most of the tissues were studied. sixteen constitutive miRNAs were observed to be expressed constitutively and showed a good overlap with miRNAs that are known to be constitutively expressed in literature (Liang et al., 2007; Peltier and Latham, 2008). These miRNAs can be used as normalization control for qRT-PCR experiments. They are mostly involved in house-keeping functions such as growth, cell division and development. Finally expression profile of miRNAs which are members of miRNA family was studied. One of the well known families of miRNAs, let7 was used for the analysis. The analysis shows that though miRNAs can be divided in different families based on mature miRNA sequences, miRNA family members can be then subdivided based of expression patterns in different tissues.
In summary, the integration of miRNA expression datasets has enabled us to establish pipeline to integrate datasets generated using different platforms. This data is made available to research community online through “miRex”. This data can be then further used for performing different type of analyses, here it has been used for predicting constitutive and tissue specific miRNAs, and was used to subdivide miRNA families on the basis of expression profiles.