TCGA, miRNA, and mRNA

Keith A. Baggerly
Department of Bioinformatics and Computational Biology
UT M. D. Anderson Cancer Center
kabagg@mdanderson.org

MDACC TCGA Analysis Group, 7 July 2010
Why Are We Here?

We want to talk about *microRNAs*.

More specifically, we want to talk about
(1) what microRNAs are,
(2) what types of associations we want to measure,
(3) what microRNA (and mRNA) data we have in TCGA
(4) the distinctions between different levels of data
(5) whether we can simply correlate level 3 data across assays
(6) what other tests we might try
(7) whether the TCGA data suggests associations are present
(8) other questions...
In the Beginning was the Worm...

In the mid-90’s, people were exploring what mutations did to *C. elegans*.

This worm has 4 larval stages (L1-L4), and folks noticed that mutations in a gene called *lin-4* screwed up the timing: L1 characteristics show up in later stages. Similar problems were seen when a gene called *lin-14* was knocked down.

The fun thing was that the product of *lin-4* was not a protein, but rather a 22bp chunk of noncoding RNA.

This product partially complements the 3-prime UTR of *lin-14*.
The Family Expands

In 2000, *let-7*, which encodes a 21bp RNA, was also discovered to regulate developmental transitions in *C. elegans*.

For *let-7*, however, cross-species conservation was observed, suggesting this might not be worm-specific.

Also around this time, other short RNAs were shown to induce transcription through RNA interference (RNAi), and it turns out these small interfering RNAs (siRNAs) make use of much of the same machinery.

miRNA and siRNA

What We Want to Know

Can we relate miR expression to mRNA regulation?

Is this regulation altered in cancer?

Can TCGA data let us address these questions?
The Data At Hand (miRNA)

8 data folders (4 per tissue) acquired as of May 7 2010:

Level 1 Data (Agilent output)
Level 2 Data (Probe-level data, 2421/1510 obs)
Level 3 Data (miR-level data, 799/534 obs)
MageTab (SDRF) File

The first three all contain files for 529 (ovary) or 386 (GBM) samples at various stages of processing.

The last contains mappings of names used at various levels.
The Data At Hand (Affy mRNA)

71 data folders (3 each for 23 batches, and 1 mapping per tumor type) acquired as of Jul 3 2010:

- **Level 1 Data** (HT_HGU133A CEL files)
- **Level 2 Data** (Probe-level data, 22277 obs)
- **Level 3 Data** (Gene-level data, 12042 obs)
- MageTab (SDRF) File

Arrays were reportedly processed with **RMA**.

Can we simply look for high correlations between Level 3 measurements across assay types?
Some GBM MiR Data Looks Odd...
Let’s Add the Batch Labels

L2 to L3: log2-transform, average, center. *Forgot one.*
Attempts at Understanding

Posted documentation (sites)
http://cancergenome.nih.gov/data/types/
genomic:description/
http://bioinfostore.unc.edu/tcga

Contacting the originators

“The Level 3 Data were obtained after processed the Level 1 data with the following procedures: normalize with quantiles, collapse by probes, collapse by genes, log2 transformation, and median center.”
Level 1 Data: What Column Was Used?

- gProcessedSignal*
- gMeanSignal
- gMedianSignal
- gBGSubSignal (same as Processed above)
- gNetSignal*
- gTotalProbeSignal**
- gTotalGeneSignal**

Guesses?
Level 1 Data: What Column Was Used?

gProcessedSignal*
gMeanSignal
gMedianSignal
gBGSubSignal (same as Processed above)
gNetSignal*
gTotalProbeSignal**
gTotalGeneSignal**

Guesses?

gMeanSignal
L1 to L2: What Normalization Was Employed?

Quantile Normalization is the only one explicitly mentioned.

If this was applied to the Level 1 data, then the vectors of spot intensities will have the same set of values for all samples, just in different orders.

We can test this by summing the Level 2 data, and examining the results across files, since sums don’t care about order.
Level 2 Sums: Do They Look Constant?
Level 2 Sums: A Pattern?
Level 2 Fits: Are We OK?
Checking Some Misfits...

Correlations > 0.999 For 6 Odd Samples

Names Reported

636 629 633 631 628 630

Names Derived

636 629 633 631 628 630
L3 From L2: How?

Cropped Zoom for hsa–miR–329 (y = 7.4 + x shown)

Average of log2(level 2 values).
What About GBM Level 3 mRNA Data?

Some shifts here
What About Ovary Level 3 mRNA Data?

and here
What About Combined Level 3 mRNA Data?

and here
Summary and Options

*Batch effects* are clearly visible in the L3 data for both types of assays.

It’s not clear the data is always on comparable scales.

**Suggestion:**

For each miR (mRNA), take the L3 data and rank and center the values, *batch by batch*.

Then check all the resultant miR/mRNA Pearson correlations.

Of course, before we do that, we’ve got to combine the data...
Mapping the Data

Using the SDRF files, we attempted to replace all L2 and L3 “array names” with their TCGA sample ids for consistent mapping.

**Problem 1**: some GBM samples were run multiple times (almost all in batches 3 and 4). In some cases Affy replicates occur in more than one batch.

**Solution 1**: keep the first rep of each sample, such that the batch information agrees with that posted on the TCGA data matrix.
Mapping the Data (2)

**Problem 2:** The ordering of samples is different for the different assays, and different batch boundaries are used.

**Solution 2:** We use the ordering from the MiRNA SDRF files, which more closely tracks with the ordering on the TCGA data matrix. Batch boundaries were determined from the matrix.

**Problem 3:** Some samples were run on only one assay.

**Solution 3:** Drop these. Restricting attention to the common samples reduces the GBM miR data from 371 samples to 366, and Ovary from 529 samples to 526.
Mapping the Data (3)

Problem 4: The ordering of probesets and gene names is not necessarily the same across tissues.

Solution 4: Explicitly enforce this matching in R.

Problem 5: Different versions of the Agilent miRNA array were used for GBM and Ovary.

Solution 5: Keep only the common probes (Level 2) and common miRs (Level 3).

After all this, are the Level 3 correlations big? We checked (1) the combined data, (2) just the GBM data, and (3) just the Ovary data.
Level 3 Corrs, Combined Data

50 Null Simulations, Real Superimposed

Base takes 10 sec to compute. Set cutoffs at 0.3.
Level 3 High Corrs, Combined Data

Note: more high corrs than low!
Not All MiRs are Equal!

|Cor| > 0.3; Blue < -0.3, White > 0.3. 2348 Genes, 173 miRs
Some Lists

combinedOrderedMirs[166:173]
[1] "hsa-miR-22"  "hsa-miR-142-3p"  "hsa-miR-142-5p"
[5] "hsa-miR-185"  "hsa-miR-155"  "hsa-miR-146a"
combinedOrderedMirs[1:20]
[1] "hsa-miR-9"  "hsa-miR-20b"  "hsa-miR-20a"
[5] "hsa-miR-19b"  "hsa-miR-93"  "hsa-miR-106b"
[9] "hsa-miR-18a"  "hsa-miR-130b"  "hsa-miR-15b"
[13] "hsa-miR-19a"  "hsa-miR-152"  "hsa-miR-18b"
[17] "hsa-miR-29b"  "hsa-miR-29c"  "hsa-miR-29a"
Level 3 Corrs, GBM Data

50 Null Simulations, Real Superimposed

Fewer samples, broader nulls. Set cutoffs at 0.4.
Again, positive bias.
Some Lists

```r
> gbmOrderedMirs[1:20]
[1] "hsa-miR-222"   "hsa-miR-23a"   "hsa-miR-221"   "hsa-miR-155"
[5] "hsa-miR-204"   "hsa-miR-221"   "hsa-miR-34a"   "hsa-miR-214"
[9] "hsa-miR-210"   "hsa-miR-34b"   "hsa-miR-21"    "hsa-miR-223"
[13] "hsa-miR-22"   "hsa-miR-142-5p" "hsa-miR-142-3p" "hsa-miR-9*
[17] "hsa-miR-9"    "hsa-miR-95"    "hsa-miR-181"   "hsa-miR-185"
> gbmOrderedMirs[166:185]
[1] "hsa-miR-29c"   "hsa-miR-7"    "hsa-miR-240"   "hsa-miR-250"
[5] "hsa-miR-504"   "hsa-miR-138"  "hsa-miR-410"   "hsa-miR-485-5p"
[9] "hsa-miR-329"   "hsa-miR-433"  "hsa-miR-130b"  "hsa-miR-106b"
[13] "hsa-miR-93"   "hsa-miR-15b"  "hsa-miR-25"    "hsa-miR-18a"
[17] "hsa-miR-20a"  "hsa-miR-106a" "hsa-miR-19a"   "hsa-miR-19b"  
```
Level 3 Corrs, Ovarian Data

Intermediate n’s, intermediate nulls. Set cutoffs at 0.35.
and again, positive bias!
Some Lists

> ovOrderedMirs[1:10]
[1] "hsa-miR-506"  "hsa-miR-507"  "hsa-miR-514"
[6] "hsa-miR-18a"  "hsa-miR-106b"  "hsa-miR-25"

> ovOrderedMirs[143:162]
[1] "hsa-miR-22"  "hsa-miR-152"  "hsa-miR-214"
[5] "hsa-miR-409-3p"  "hsa-miR-431"  "hsa-miR-382"
[9] "hsa-miR-145"  "hsa-miR-409-5p"  "hsa-miR-411"
[13] "hsa-miR-381"  "hsa-miR-432"  "hsa-miR-299-3p"
[17] "hsa-miR-136"  "hsa-miR-376a"  "hsa-miR-154"
There are outliers all around the circumference!
Does Intensity Matter?

For this, we used corrs with the L2 miR data
Do Batches Really Matter?

again, using L2 miR data (log2!)
Summary

There are some real associations here, and they show up even using a coarse batch adjustment.

Batch adjustment of some type is required.

There are differences in association by tissue.

The most extreme correlations are positive, not negative.

Tracking the precise processing used (and employing it automatically) is likely important.
Open Questions

Using frozen RMA? (mapping the HT-HGU133A)

Checking for known differences (e.g., gender)?

Checking for pathways?

Comparing with known miRs?

Comparing fits with predictions from different models?

Using identified associations to cluster samples