GS01 0163
Analysis of Microarray Data

Keith Baggerly and Kevin Coombes
Section of Bioinformatics
Department of Biostatistics and Applied Mathematics
UT M. D. Anderson Cancer Center
kabagg@mdanderson.org
kcoombes@mdanderson.org

27 September 2005

Lecture 7: Normalization and Affymetrix

- What is Normalization?
- What Methods have People Suggested?
- How Can These Methods be Tested?
- What do I Recommend?

What is Normalization?

Broad question: How do we compare results across chips?

Focused goal: Getting numbers (quantifications) from one chip to mean the same as numbers from another chip.

Why is Normalization an Issue?

amount of RNA

efficiencies of RNA extraction, reverse transcription, labeling, photodection

PCR yield

DNA quality

variation that is obscuring as opposed to interesting.

What Has Been Tried?

Housekeeping genes – start with a set of genes whose expression you believe shouldn't change, and scale the other expression values accordingly.

What Has Been Tried?

Housekeeping genes – start with a set of genes whose expression you believe shouldn't change, and scale the other expression values accordingly.

Spike-ins – introduce a set of markers whose relative intensities you can control, and use these to calibrate the remaining intensities.

What Has Been Tried?

Housekeeping genes – start with a set of genes whose expression you believe shouldn't change, and scale the other expression values accordingly.

Spike-ins – introduce a set of markers whose relative intensities you can control, and use these to calibrate the remaining intensities.

Simple scaling – multiply all of the intensities from one chip so that a summary of the result matches the summary value from another chip or some standard.

Some Difficulties

Housekeeping genes –

Some Difficulties

Housekeeping genes – what are they? Are there any genes whose expression does not change across a wide variety of conditions? Do those conditions include cancer (broad distortion)?

Spike-ins –

Some Difficulties

Housekeeping genes – what are they? Are there any genes whose expression does not change across a wide variety of conditions? Do those conditions include cancer (broad distortion)?

Spike-ins – How do we regulate the amount of the spike-in relative to the amount of the material of interest?

Simple scaling –

Some Difficulties

Housekeeping genes – what are they? Are there any genes whose expression does not change across a wide variety of conditions? Do those conditions include cancer (broad distortion)?

Spike-ins – How do we regulate the amount of the spike-in relative to the amount of the material of interest?

Simple scaling – Does a log scale MA plot look flat? What scaling factor do we use? Median? Total amount of RNA present? Some other quantile?

An Alternative?

Combining Spiking and scaling?

Use spike-ins covering a very broad dynamic range, and use this to try to define “linearity” of expression, after which you scale.

Housekeeping: Can this be Inferred?

What is a housekeeping gene?

Typical assumption – something that is vital to the stable functioning of the cell, present at constant levels. In most suggested cases, this level has been high.

Housekeeping: Can this be Inferred?

What is a housekeeping gene?

Typical assumption – something that is vital to the stable functioning of the cell, present at constant levels. In most suggested cases, this level has been high.

Slightly different definition – a gene that retains roughly the same rank in a sorted list of expression values.

Choose a subset with “invariant ranks” and use these to determine our mapping (Schadt et al.)

What is Our Baseline?

When we are assessing invariance, we have to define “with respect to what?”

For the most part, this has meant choosing a single chip as a canonical “reference” or “baseline”.

Typically, this chip is taken to be a “middle” chip by some metric, such as the chip having the median “median intensity”.

This is the default in dChip.

Invariance and Nonlinearity

Using an invariant set effectively applies a whole bunch of scaling factors, with the factor changing depending on the intensity (scaling at high levels is different than scaling at low levels).

For the most part, this serves to map the quantiles of one set of intensities over to another.

Similar to producing an MA-plot, fitting a smooth curve (e.g. loess) to the center of the data, and subtracting the curve off.

An Implicit Assumption

Is matching the 90th percentile reasonable if one sample causes no gene expression whatsoever, and a second produces expression in half of the genes?

An Implicit Assumption

Is matching the 90th percentile reasonable if one sample causes no gene expression whatsoever, and a second produces expression in half of the genes?

The assumption that mapping using quantiles or scaling is reasonable is based on the assumption that “most genes don’t change”, and quantiles use this more extensively than scaling.

If this underlying assumption is doubtful, then using the above methods is shaky.

Cases Where Most Genes Do Change

Fed/Starved experiments

Cases Where Most Genes Do Change

Fed/Starved experiments

Heat shock experiments

Cases Where Most Genes Do Change

Fed/Starved experiments

Heat shock experiments

Experiments comparing different organs

How Can We Test Normalization Methods?

Most of the following is based on a paper by Bolstad, Irizarry, Astrand and Speed (Bioinformatics, 2003, p.185-193)

When can we know what the true answer should be?

How Can We Test Normalization Methods?

Most of the following is based on a paper by Bolstad, Irizarry, Astrand and Speed (Bioinformatics, 2003, p.185-193)

When can we know what the true answer should be?

If we're looking at the same stuff, or stuff that differs only in a way that we view as irrelevant to the biology, then things should look the same after normalization.

Find some good datasets that work like this.

Affymetrix Data

The Affymetrix Latin Square Experiment (Hu95Av2, repeated on Hu133A)

roughly 42 chips which have the same stuff printed modulo changes in spike-ins.

Gene Logic Dilution Data

75 Hu95Av2 arrays, two tissue sources, liver and central nervous system. 30 chips each for each tissue alone, in 5 blocks of size 6, with each block at a different dilution level. Remaining 15 chips are mixtures of liver and CNS, in 3 blocks of 5, in ratios of 25:75, 50:50, and 75:25.

4 of the liver arrays are included with the BioConductor affydata package.

A Gene Logic Spike In Dataset

98 Hu95Av1 arrays, with 11 cRNA spike-ins. For 26 of these chips, with a common AML source, the level was nominally the same for all of the spike-ins, but this level was altered from chip to chip (i.e., chip 1 had all 11 at 0, chip 2 had all 11 at 0.5, chip 3 at 0.75, etc.) Remaining chips were set up in two Latin Square experiments with 3 replicates at each level.

With the first 26, nothing should change except the spike-ins.

What was Used?

5 of the liver arrays from the Dilution experiment (all at the same dilution level)

Should normalization correct for dilution?

all 26 arrays from the spike-in experiment where the spike-in levels were nominally the same.

Aside: Why is Gene Logic Doing This?

Gene Logic is the biggest single consumer of Affy chips, and they have assembled very large databases of expression profiles, which they market.

Aside: Why is Gene Logic Doing This?

Gene Logic is the biggest single consumer of Affy chips, and they have assembled very large databases of expression profiles, which they market.

They are trying to improve the analysis (which helps them) and to advertise their own approaches (again).

The datasets can be requested (they'll send you a cd).

Given Data, What Tests Should we Use?

Scaling a la Affy, equating 96% trimmed means and choosing a baseline chip (median of median). (Affy's algorithm performs this scaling on the summary measures, but probe levels were used here).

Given Data, What Tests Should we Use?

Scaling a la Affy, equating 96% trimmed means and choosing a baseline chip (median of median). (Affy's algorithm performs this scaling on the summary measures, but probe levels were used here).

Nonlinear invariant rank approach, a la Schadt et al., again using a baseline chip (the same one).

Given Data, What Tests Should we Use?

Scaling a la Affy, equating 96% trimmed means and choosing a baseline chip (median of median). (Affy's algorithm performs this scaling on the summary measures, but probe levels were used here).

Nonlinear invariant rank approach, a la Schadt et al., again using a baseline chip (the same one).

Some "Complete Data Methods" that they introduce

Complete Data: Eliminating Baseline

Cyclic Loess

Contrasts

Quantiles

All of these work by treating the chips in a symmetric fashion, and all of them are implemented as normalization methods in the BioConductor affy package.

Cyclic Loess

Start with MA plots, fit a loess smooth for each pair of chips.

Let $M_k = log_2(x_{ki}/x_{kj})$ for arrays i, j, and let $\hat{M}_k$ denote the fitted loess curve for this pair of chips. Then the adjusted value should be $M'_k = M_k - \hat{M}_k$.

How much should we adjust each of the chips? Use a symmetric approach

$x'_{ki} = 2^{A_k + 0.5 * M'_k}$, $x'_{kj} = 2^{A_k - 0.5 * M'_k}$

Repeat for all pairs, then refit and repeat. (2-3 iterations) This is Slow.

Contrasts

Start with vectors of log intensities from each chip. Use a rotation matrix (a matrix of orthogonal *contrasts*) to change the basis. The first column of the converted matrix is the average of all of the log intensities. For the remaining $n - 1$ columns (n arrays), fit loess curves, flatten out, and reverse the transformation.

Central focus is the average on the log scale (the geometric mean).

Still slow.

Quantiles

Assume that the distributions of probe intensities should be completely the same across chips.

Quantiles

Assume that the distributions of probe intensities should be completely the same across chips.

Start with n arrays, and p probes, and form a $p \times n$ matrix X.

Quantiles

Assume that the distributions of probe intensities should be completely the same across chips.

Start with n arrays, and p probes, and form a $p \times n$ matrix X.

Sort the columns of X, so that the entries in a given row correspond to a fixed quantile.

Quantiles

Assume that the distributions of probe intensities should be completely the same across chips.

Start with n arrays, and p probes, and form a $p \times n$ matrix X.

Sort the columns of X, so that the entries in a given row correspond to a fixed quantile.

Replace all entries in that row with their mean value.

Quantiles

Assume that the distributions of probe intensities should be completely the same across chips.

Start with n arrays, and p probes, and form a $p \times n$ matrix X.

Sort the columns of X, so that the entries in a given row correspond to a fixed quantile.

Replace all entries in that row with their mean value.

Undo the sort.

This procedure, sorting and averaging, is comparatively fast.

Quantile Argument - Why the Mean?

Why did they choose the mean? If we work in 2d, then perfect quantile agreement produces a straight line.

If we plot n-vectors of quantiles, then we want to see a straight line in n-space along the main diagonal.

Projecting the observed n-vector onto this central axis suggests using the mean value.

My Quibble: Scale

The argument suggesting the mean does not address the issue of what scale the observations are on, and works equally well for both the raw and log scales, but the mean values differ.

I prefer the median, which is much more scale-invariant, but for the most part there is little practical difference.

Does Baseline Matter?

Actually, yes.

from Astrand (J. Comp. Biol., 2003, p.95-102)

Starting with a pair of arrays (Mu11K), they normalized in dChip using array A as baseline, and again using array B as baseline.

Of the predicted log ratios, roughly 18% of the ratios (1603/8799) differed by more than 10%.

A Related View...

Of the lists produced using dChip's filters, there were 469 genes one way, and 298 the other.

The overlap was 265.

A Related View...

Of the lists produced using dChip's filters, there were 469 genes one way, and 298 the other.

The overlap was 265.

We have seen similar results with more chips.

How Do We Define Success?

We want the results for a given *probeset* to be the same from one chip to the next.

For each method, compute expression values, (put them on the log scale ?), and compute the variance and the average (this is akin to what is required for a multivariate MA plot).

These are our method summaries.

How Do We Define Success?

For each pair of methods, produce a plot with the average of the log means on the x-axis and the log ratio of the variances on the y-axis.

Fit a smooth curve, and look for cases where one method's variance is much smaller than the other's.

What Preprocessing Do We Employ?

To make sure that differences in the summary method employed do not drive the results, pick one method and use it throughout.

They use RMA with the default BG correction throughout.

Some Results

In general, pairwise MA plots did show some curvature before normalization, suggesting that a nonlinear fit might help.

Some Results

In general, pairwise MA plots did show some curvature before normalization, suggesting that a nonlinear fit might help.

All of the methods that they used reduced the variance of the numbers substantially from using just the raw numbers without normalization.

Some Results

In general, pairwise MA plots did show some curvature before normalization, suggesting that a nonlinear fit might help.

All of the methods that they used reduced the variance of the numbers substantially from using just the raw numbers without normalization.

The complete data methods also improve on the baseline chip methods, and they improve more on simple constant scaling than they do on nonlinear fits such as the invariant rank set.

More Results

With respect to the three “complete data” methods, the differences are fairly small, with the quantile method looking very slightly better.

As the quantile method is definitely the fastest, this is the one that they recommend.

Do I Agree?

Of course!

Do I Agree?

Of course!

In this case, the assumptions underlying this type of approach are met, and most genes should not change.

Do I Agree?

Of course!

In this case, the assumptions underlying this type of approach are met, and most genes should not change.

But this assumption is not always met, and in cases where I am suspicious of this I try to use something that is less dependent on a large number of assumptions.

This can be simple scaling, or scaling for both mean and spread.

What Do I Do?

In general, the first thing I do is check the context of the experiment, to see if I am persuaded that most genes might be the same.

What Do I Do?

In general, the first thing I do is check the context of the experiment, to see if I am persuaded that most genes might be the same.

Next, I check the pairwise MA plots. If I don't see substantial nonlinearity, then I don't have to use quantiles so I don't.

What Do I Do?

In general, the first thing I do is check the context of the experiment, to see if I am persuaded that most genes might be the same.

Next, I check the pairwise MA plots. If I don't see substantial nonlinearity, then I don't have to use quantiles so I don't.

If I do see nonlinearity, I check images of the CEL files to look for artifacts, and if I see nothing I'll go with quantiles.

What's Next?

This has been a "concept" lecture.

On Thursday, we're going to try a comparison of these measures (repeating the above study) using the 4 chips in the Dilution dataset.