

Data Analysis

The Glycoarray designed by Core E for gene expression analysis is a custom Affymetrix GeneChip constructed with the same probe-set-design algorithms that were used for the construction of the recently released human genome U133 GeneChip set (Affymetrix.com). Our Glycoarray has oligonucleotide probe sets consisting of 11 probe pairs to interrogate each targeted mRNA sequence. Each probe pair consists of one perfect match (PM) and one mismatched (MM) 25-base oligonucleotide. The PM oligonucleotide is complementary to a given portion of the targeted gene, typically within 600 bases of the 3'-polyadenylation signal of the transcript. The MM oligonucleotide is identical in sequence to the PM probe, except for a single mismatched base at the 13th position (center) of the probe. The difference between the PM and MM probe signals among all probe pairs for a given gene is used to calculate the hybridization signal. This signal is a quantitative metric (a weighted average) calculated for each probe set that represents the relative abundance of a transcript. In addition, a "detection p value" is calculated based on an algorithm using the probe-pair intensity distribution within a given probe set relative to a user-definable threshold. A one-sided Wilcoxon Signed Rank Test is applied to this probe-pair intensity distribution to generate the p value. Thus, a given transcript can be categorized as present ($p < 0.04$), absent ($p > 0.06$), or marginal ($0.04 < p < 0.06$) on the basis of predefined p -value thresholds. More information on Affymetrix algorithms is available at the following Web site:

http://www.affymetrix.com/support/technical/technotes/statistical_reference_guide.pdf.

The PM/MM strategy in combination with interrogating each transcript with multiple probe pairs increases specificity. This technology is very sensitive for detecting low-level transcripts. Affymetrix GeneChip arrays have been reported to detect as little 0.075 pM target mRNA (1 transcript per 7 cells) [1]. A more recent report suggests that 0.5 pM target mRNA (1 transcript per cell) may be a more realistic assessment of this technology's practical detection limit [2]. The precise sensitivity and specificity of individual probe sets are highly variable and data should be interpreted cautiously. Because the manufacturing and sample processing steps of the Glycoarray are highly standardized, experience using this GeneChip will allow researchers to gain confidence in interpreting results and characterizing the performance of individual probe sets based on data reproducibility, independent corroborative data, and review of the annotation describing the probe and target sequences.

Microarray Data Analysis

Microarray data analysis is a rapidly evolving area of research, which involves both image analysis and the application of statistical techniques to evaluate, classify, and cluster large amounts of microarray information to extract meaningful biological interpretations. Numerous software tools have been developed by software and data analysis companies and academic laboratories. In fact, Y. F. Leung's Functional Genomics Web site (<http://ihome.cuhk.edu.hk/~b400559/>) lists at least 63 different software tools that are currently available for microarray data mining. Core E routinely uses the following 4 software tools (each discussed below) to analyze Glycoarray data:

- Microarray Suite 5.0 (Affymetrix, Santa Clara, CA)
- GeneSpring (Silicon Genetics, Redwood City, CA)
- BRB ArrayTools (Biometric Research Branch, NCI/NIH, <http://linus.nci.nih.gov/BRB-ArrayTools.html>)
- DNA-Chip Analyzer [3, 4]

Consortium members may schedule a consultation with a member of Core E for assistance and advice in using these software tools.

Microarray Suite 5.0

This software provides instrument control for the GeneArray Scanner and Fluidics Station for processing Glycoarrays and performing array image acquisition and analysis. This software performs statistical analyses of probe set data (described above) and generates both signal intensity values (a weighted average difference between the signals of the PM and MM probes that comprise each probe set) representing the relative expression level of each gene on the array and associated “detection p values” representing the confidence that the targeted transcript is indeed detected. This detection p value is particularly useful when evaluating signal intensities at the lower end of the spectrum where the correlation between the detection p value and signal intensity is weaker. For example, signals for 2 different genes may be equal to 100 (a relatively weak signal for arrays scaled to a median signal target intensity of 250). Both genes may be present based on detection p -value thresholds, but the actual detection p value can raise or lower confidence in the signal measurement for gene transcripts detected with such low signals.

In addition to hardware control and transcript signal and detection, Microarray Suite 5.0 also performs GeneChip “Baseline Comparison Analysis.” In this type of analysis, each probe pair from 2 GeneChips is directly compared to each other for each probe set. Probe pairs with saturated fluorescence signals are eliminated from the analysis and a Wilcoxon Signed Rank Test is applied to the differences between PM and MM intensities and between PM and background intensities to compute the change in expression p values (http://www.affymetrix.com/support/technical/technotes/statistical_reference_guide.pdf). This rigorous comparison analysis has been shown to significantly reduce array variability relative to comparisons of expression signals alone [5]. Methods using this pair-wise comparison analysis for calculating expression values for all genes on all arrays from data sets with more than one chip have been developed [5] and may be applied to data sets generated with the Glycoarray. Expression values generated by this method may be analyzed with statistical software such as BRB ArrayTools (see below).

GeneSpring

This commercially available software allows rapid, platform-independent analysis of microarray data. Multipoint and time-course experiments are easily visualized, queried, and analyzed for genes showing similar expression profiles. This software is ideal for quickly and easily handling large microarray data sets, performing statistical tests, and creating Venn diagrams to compare the relationships of gene lists generated by clustering.

BRB ArrayTools

We highly recommend this free, platform-independent software program developed by the Biometric Research Branch, a statistical and biomathematical component of the Division of Cancer Treatment and Diagnosis at the National Cancer Institute. This powerful data analysis software is simple to download and easy to learn with the help of the well-written user’s manual. BRB ArrayTools is an integrated package for the visualization and statistical analysis of DNA microarray gene expression data. BRB ArrayTools uses an Excel front end that is integrated into Excel as an “add-in”. Data input consists of spreadsheets that provide expression values, gene annotation, and user-specified descriptions for each sample. The analytic and visualization tools themselves are developed in the R statistical system with Visual Basic for Applications that integrate C and FORTRAN programming and Java applications to run the analytic methods. This software can perform paired and unpaired t-tests, f-tests, and permutation tests on microarray data sets. In addition, BRB ArrayTools can perform hierarchical clustering, classification, class prediction, and multidimensional scaling of data.

DNA-Chip Analyzer

The DNA-Chip Analyzer is a software package that implements model-based expression analysis of oligonucleotide arrays and several high-level analysis procedures [3, 4]. The principal advantage of this software is that the probe-set response is modeled with data from all arrays within an experimental set (as opposed to Microarray Suite 5.0, which independently calculates signals for each gene on each array). It is possible to assess standard errors for the expression signals by pooling information that is collected from multiple arrays. This approach also allows automatic probe selection in the analysis stage to reduce errors from cross-hybridizing probes and image contamination. Preliminary analysis of DNA-Chip Analyzer by Core E indicates that in most cases, this software correlates well with the Microarray Suite 5.0 signal (DNA-Chip Analyzer uses transcript presence and absence calls generated by Microarray Suite 5.0) and provides the additional advantage of assessing standard errors for the expression signals and straightforward corrections for cross-hybridizing probes and image contamination. Outlier arrays and probes are flagged for further analysis.

Generally, either DNA-Chip Analyzer or Microarray Suite 5.0 may be used as the principal tool for calculating expression signals for GeneChip arrays. DNA-Chip Analyzer is free to academic users, although learning to use it is not trivial and the user's manual is not nearly as straightforward as the BRB ArrayTools' manual. It may be useful to derive expression data and perform statistical analyses using both of these software packages to compare results. This can provide increased confidence in results that are consistent across these 2 analysis tools. The expression of genes that appear to change in only one of the analyses can be followed up in another study.

References

1. Lipshutz, R.J., et al., *High density synthetic oligonucleotide arrays*. Nat Genet, 1999. **21**(1 Suppl): p. 20-4.
2. Chudin, E., et al., *Assessment of the relationship between signal intensities and transcript concentration for Affymetrix GeneChip arrays*. Genome Biol, 2002. **3**(1): p. RESEARCH0005.
3. Li, C. and W.H. Wong, *Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection*. PNAS, 2001. **98**(1): p. 31-36.
4. Li, C. and W. Hung Wong, *Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application*. Genome Biology, 2001. **2**(8): p. research0032.1-research0032.11.
5. Welle, S., A.I. Brooks, and C.A. Thornton, *Computational method for reducing variance with Affymetrix microarrays*. BMC Bioinformatics, 2002. **3**(1): p. 23.