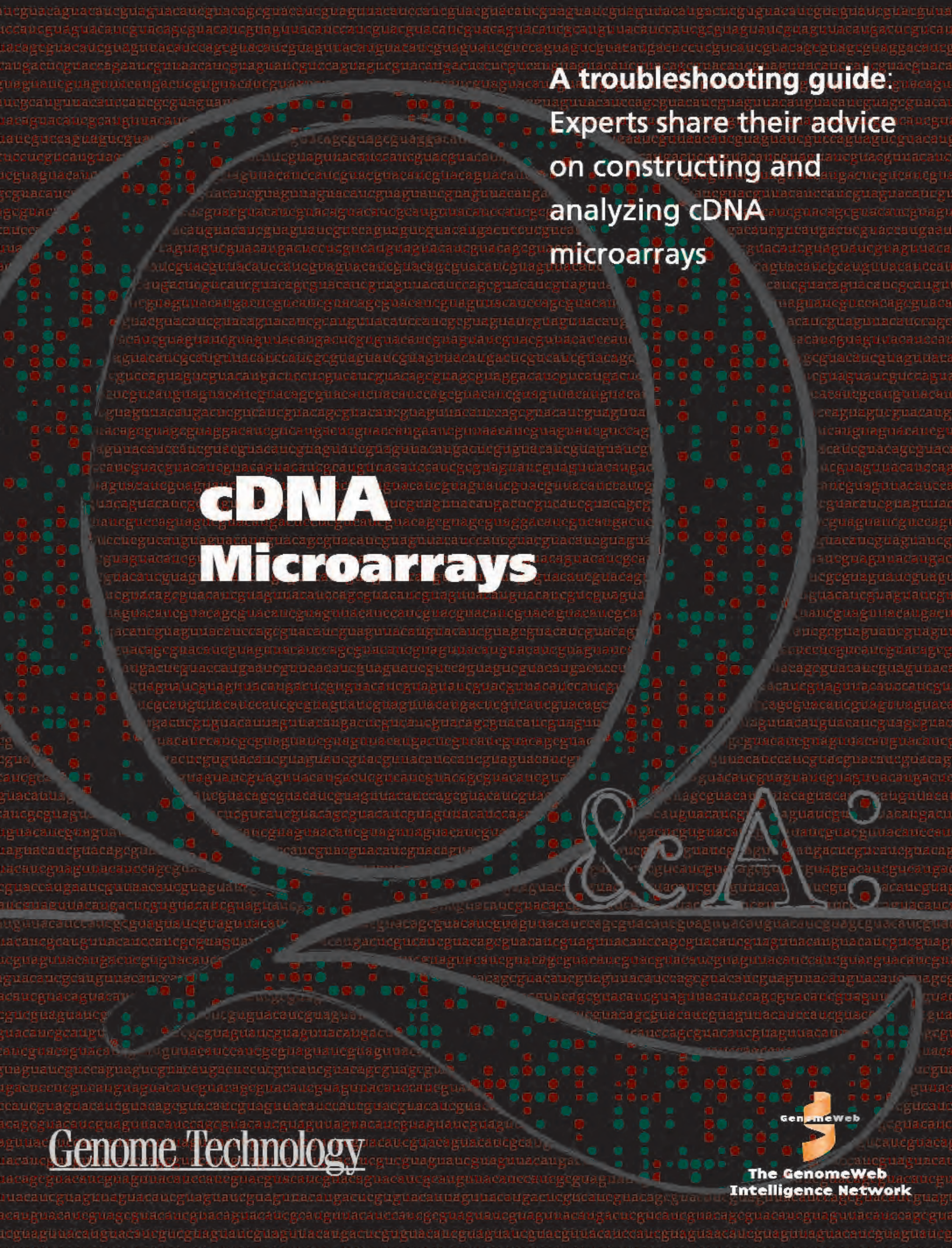


A troubleshooting guide: Experts share their advice on constructing and analyzing cDNA microarrays

cDNA Microarrays

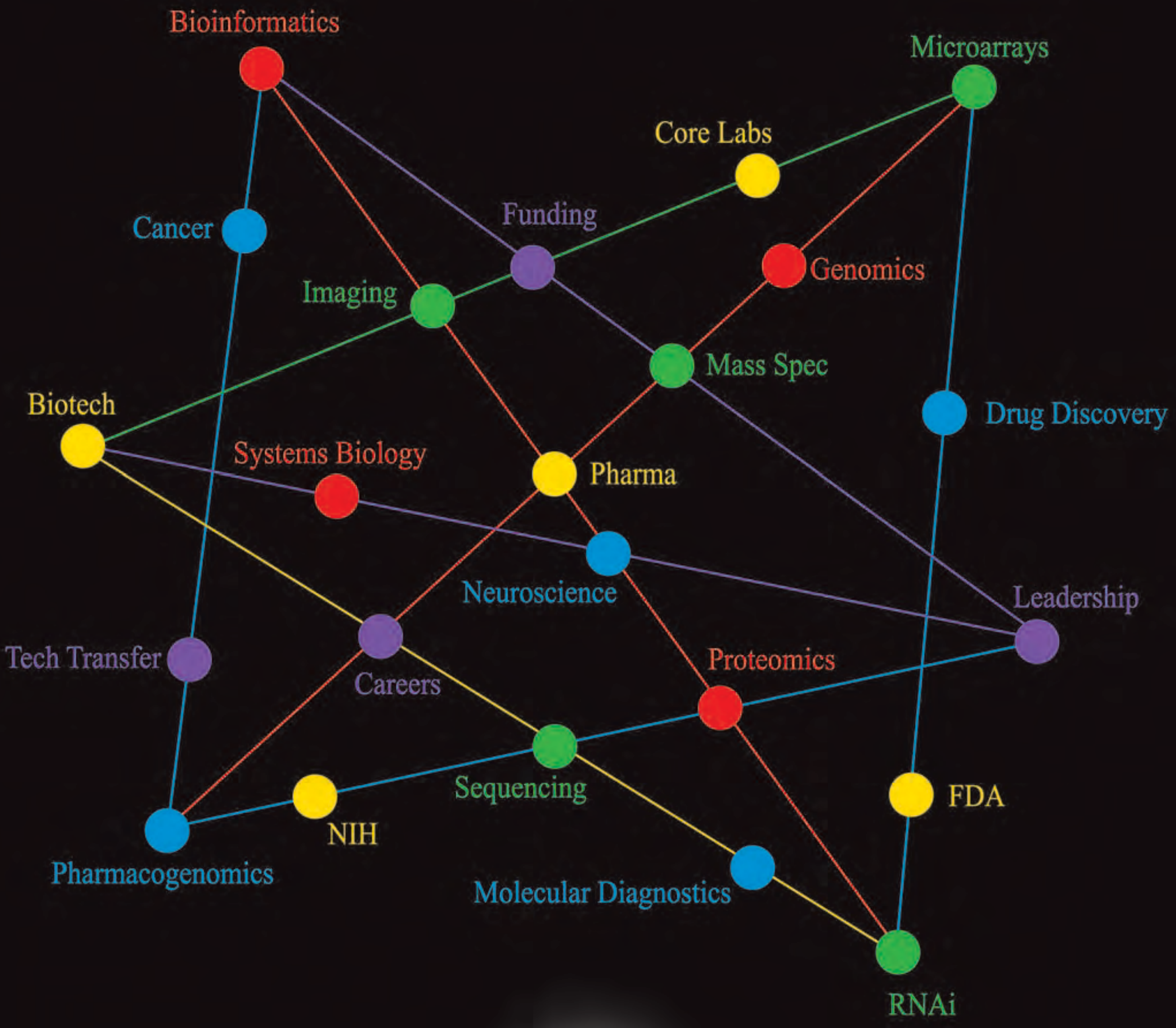
Genome Technology

GenomeWeb

The GenomeWeb Intelligence Network

The GenomeWeb Intelligence Network.

Connecting the dots for researchers worldwide.



GenomeWeb

Table of contents

If you want to increase sensitivity and avoid microarray artefacts...
Talk to Tecan

HS Pro microarray hybridization stations with ABS™

The HS 400™ Pro and HS 4800™ Pro are the perfect solutions for superior microarray processing. Fully automated – from pre-hybridization to slide drying – the instruments maximise the sensitivity and reproducibility you expect from your microarray results.

Only Tecan's HS Pro systems have ABS technology (patent pending) to prevent hybridization artefacts by excluding bubbles from the process. A variety of low volume hybridization chambers will even enable processing 2 independent arrays on 1 slide, on a throughput of up to 48 slides in a run.

The HS Pro systems are part of the Tecan Microarray Suite, a range of instruments and solutions from microarray processing and scanning to data analysis. They are easy to use, require less maintenance, and offer the flexibility to process different microarray applications. HS Pro systems are just one example of how Tecan is in touch with what you need most.

To find out more about HS Pro or any of our products, please talk to Tecan on the number below.

www.tecan.com

TECAN.

Liquid Handling & Robotics | Detection | Sample Management | Components | Services & Consumables

Switzerland +41 44 922 89 22 **USA** +1 919 361 5200 **Other** +43 62 46 89 33

Letter from the editors

Welcome to *Genome Technology*'s latest technical reference guide focusing on cDNA microarrays. Here, we present the thoughts of a variety of experts familiar with maximizing specificity and sensitivity — without sacrificing robust and reproducible data. They give time-tested advice for getting a good look into genome-wide gene expression studies.

This kind of microarray has been kicking around for more than a decade — and remains popular despite rising competition from oligonucleotide and SNP arrays. While technological advances in printing arrays and scanning have made expression analysis more robust, data analysis continues to go head to head with complex issues of data quality, analysis, and storage. Additionally, public repositories of microarray data have made it possible to conduct comparative studies to validate existing data. While there have been major improvements, advances continue in search of more sensitive, specific assays with more reproducible results.

To this end, we've followed our experts in the lab, mining their expertise in the areas of sample preparation, results validation, and data analysis. Keep this guide in close reach for questions pertaining to every problem area, from ensuring you've made reliable levels of cDNA, to making certain you've used the latest analysis software to create a robust and reproducible microarray experiment.

— Ciara Curtin and Jeanene Swanson

Index of experts

Genome Technology would like to thank the following contributors for taking the time to respond to the questions in this tech guide.

Aedín Culhane

Research Associate

Dana-Farber Cancer Institute

Elva Diaz

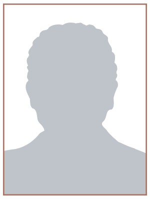
Assistant Professor

University of California, Davis
School of Medicine

Audrey Gasch

Assistant Professor of Genetics
Laboratory of Genetics &
The Genome Center of Wisconsin

University of Wisconsin, Madison



Li Liu

Bioinformatics Director

University of Florida

Coleen Murphy

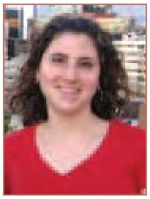
Assistant Professor of Genomics &
Molecular Biology

Lewis-Sigler Institute for Integrative
Genomics, Princeton University

Radhakrishnan Nagarajan

Assistant Professor

University of Arkansas for Medical
Sciences



Renee Rubio

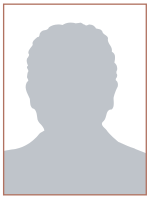
Project Manager

Dana-Farber Cancer Institute

Gordon Smyth

Research Fellow

Walter and Eliza Hall Institute of
Medical Research



Marina Telonis-Scott

Expression Core Director

University of Florida

What sample prep quality control steps do you have in place?

It is very important that RNA and cDNA used in the analysis is high quality to assure that the results are reliable. We test our samples at three stages throughout the process. First is the RNA stage, where we run the samples on an Agilent Bioanalyzer to measure concentration values, A260/280 ratios, and RNA integrity. Second, we test the samples after reverse transcription to be sure cDNA generation from RNA was successful. We use the Nanodrop or a spectrophotometer to determine an OD at A260 and determine the amount of cDNA present. The amount of cDNA present should be extremely close to the amount of RNA starting material. If there's a huge discrepancy, then the reverse transcription reaction was not successful. Third, we assay our samples after Cy3 and Cy5 dye labeling to assure good dye incorporation. Although we could use quantitative measures, a visual test at this stage is satisfactory — if you can see the pink and blue from the Cy3 and Cy5 dyes, respectively, then incorporation was good. Poor quality samples are discarded and not hybridized since they do not provide useful data.

— Aedín Culhane and Renee Rubio

We routinely do three quality control steps: First, all amplified samples are quantified using a NanoDrop spectrophotometer (requires only 1 µl of sample) to determine yield (a minimum of 5 µg of amplified sample is required for each hybridization). Second, all amplified samples are analyzed by gel electrophoresis or Bioanalyzer to assess distribution of amplified products. Third, labeled samples are analyzed using the NanoDrop to measure dye incorporation and yield after purification.

— Elva Diaz

There are two issues with sample prep: one has to do with the biological experiment, the other has to do with the quality of RNA used for labeling. We try to be as consistent in every detail of the experiment (from growing the cells in the same lot of media every time and keeping experimental details exactly the same from day to day, to prepping the RNA consistently every time).

Assessing the quality of the RNA is the easier to address. We routinely check RNA quality by gel and also by absorbance indicators (A260, A280, and A230). As for quality of the biological experiment, this is assessed by 1) reproducibility in replicate experiments and 2) agreement with other data in the lab, including experimental replicates done by others in the lab.

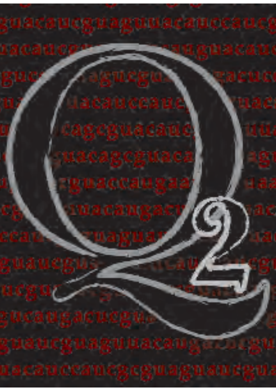
— Audrey Gasch

We are mostly concerned with lifespan and development of *C. elegans*, so we make sure that each sample that we collect would have given us the expected phenotype. Specifically, we take about 100 worms from every sample that we collect (before adding Trizol, of course) and we carry on with the lifespan; this lets us know that the sample was capable of doing what we expected. In the case of embryos, we take a small sample of the collected eggs and check that they hatch normally.

After RNA preps, we check the concentration and we also run the samples on gels to be sure that there is no degradation. For cDNA, we check for yield, and after labeling we check both the concentration and the labeling efficiency (all on the NanoDrop). Finally, we do several replicates and dye-flips to be sure that we have multiple samples to analyze.

— Coleen Murphy

(continued on p. 14)



How do you improve hybridization specificity and sensitivity?

The specificity and sensitivity of hybridization is dependent on the probe sequence and optimization of hybridization conditions. Clones on our cDNA arrays were generated from an Image 40K clone library several years ago and we are no longer printing these arrays. However, when designing arrays, it is important to check the GC content of probes and the cross-hybridization potential of probes to other sequences and to themselves. We have developed protocols that use a variety of blocking agents and which involve a pre-hybridization step that virtually eliminates background from non-specific hybridization to the glass. Our protocols are available online.

It is also important to have several positive and negative control probes on arrays as these can be an important measure of the quality of the assay and the sensitivity and specificity of the conditions used. Similarly, the use of spike-in controls can provide an overall assessment of the entire laboratory protocol (see also question 5).

— Aedín Culhane and Renee Rubio

We usually hybridize our DNA microarrays with 50 percent formamide at 42 degrees to 50 degrees for 16 hours, which we have found to be optimum in terms of specificity and sensitivity for mouse brain samples.

— Elva Diaz

We only do cDNA-DNA hybridizations, which are reported to hybridize with higher specificity than cRNA-DNA hybridizations. We have optimized the protocol and do not deviate from hybridization conditions. Deviations in hybridization conditions from day to day (such as subtle differences in the composition of hybridization solution and variations in hybridization times) can decrease reproducibility.

— Audrey Gasch

We haven't done much of this because our protocols were worked out many years ago, and mostly depend on the sequences that were selected for printing during the array design process. We haven't had any problems that suggested that this aspect needed to be improved. Perhaps with future array designs we will need to revisit this question, however.

— Coleen Murphy

Spotting multiple probes corresponding to the same transcript. Control probes/housekeeping genes. Generating swap-arrays in the case of two-color microarrays can minimize the spurious gene expression that is an outcome of preferential dye binding.

— Radhakrishnan Nagarajan

Unfortunately, as a bioinformatician, I can't comment on hybridization as a chemical process. We do use a range of quality assessment displays, including MA-plots, image plots, intensity plots, and foreground-background plots, to check the data arising from microarray experiments and these are invaluable for highlighting any problems with the hybridizations. Our local arrays routinely include various control spots including spot buffer and poly A to detect non-specific hybridization and often spike-in ratio controls to measure sensitivity. One way to highlight such

(continued on p.14)

Q3 What do you consider when choosing a validation method?

A validation method should be an independent test for the level of gene expression. Either qRT-PCR or northern blots are good methods, although qRT-PCR is our first choice. For validating a collection of genes, the TaqMan 384-well microfluidic RT-PCR cards are useful. More and more, it is also worth examining complementary public datasets. A substantial amount of data is now available in the public microarray repositories, ArrayExpress and GEO, and these datasets are frequently useful in choosing which genes to validate. We have also had some success using tissue microarray arrays (TMA). For clinical applications, TMAs are particularly useful as most labs have large collections of paraffin blocks of tissue. More labs are using TMAs and this is likely to be more widely used as the Human Protein Atlas antibodies are made available.

— Aedín Culhane and Renee Rubio

The main consideration is ease and cost. For simple validation of expression levels, real-time PCR is efficient, easy, and relatively inexpensive (assuming that you have access to the equipment via a core facility). However, most often we validate candidate genes with *in situ* hybridization of tissue sections because we are interested in the cell types that express our transcript of interest.

— Elva Diaz

It depends on what we will do with the data. To validate the expression levels of a specific gene we are interested in, we typically use quantitative RT-PCR — in general we get very good agreement with our spotted array data (within 1.5X agreement). If instead we are using the microarray data to infer that a given pathway of genes is affected or that a certain transcription factor is active, we use statistical analyses on groups of previously defined genes to get statistical confidence in the genomic response.

— Audrey Gasch

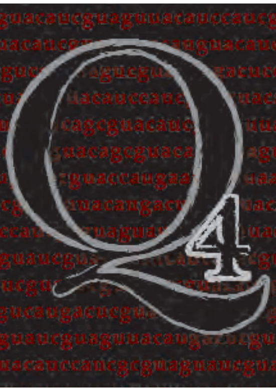
Whether it is actually informative or not! The genes we have checked by RT-PCR have always verified what we have seen by array, which is not surprising — I think that it is safe to say that when one designs a good experiment, the array is likely to give decent results. What I am more interested in is whether the gene expression differences are validated biologically, either through expression analysis and/or through phenotypic analysis. For the former, we use promoter::GFP expression in the relevant mutant backgrounds and check whether we can see differences that are predicted by the array analysis. In the latter case, we are quite fortunate that we can knock down gene function in *C. elegans* by feeding them bacteria containing double-stranded RNA of the gene of interest (RNAi), so we can very quickly check whether a gene plays a role in the phenotype we are interested in. To me, this type of biological validation is much more important than just checking (again) whether arrays really work or not.

— Coleen Murphy

Reproducibility is an important validation criteria. qRT-PCR is used to validate transcripts of interest. The choice of the housekeeping genes in the case of qRT-PCR is again far from trivial.

— Radhakrishnan Nagarajan

(continued on p. 14)



How do you ensure that low-intensity data is not missing?

Data points can be missing due to two reasons. The first is when a gene is not expressed or expressed at a level below the detection limit of the experiment. We can impute a very low level of expression for these genes, perhaps equal to the background level of the array.

The second instance, which is much more of a problem, is when a gene is expressed but not detected. This is a false negative result and is typically due to technical or image analysis problems. Improper aligning of a grid in image analysis can be corrected by repeating the image analysis step. Other times, false negative missing values are due to problems with batch-to-batch variation when printing the arrays. This was common on older cDNA arrays. Array printing processes are much improved now so this is less common now. However, imputing such values is challenging and can really only be done with replicates.

> "Data can be maximized by using the highest-quality RNA samples and arrays, which in turn maximizes the signal-to-noise measured on the arrays."
>
> *— Audrey Gasch*

It is normally impractical to determine whether missing values are due to the first or second case. Therefore it is important to use other approaches to rule out the second case. We would use exploratory data analysis method (see question 7) and would also count the number of missing values on each array/genes. We discard arrays with a high percentage of missing values, as this is normally indicative of an array with QC problems. Such arrays will skew batch normalization. We also discard genes that are missing in a large percentage of the samples as these may indicate a problem with the printing of these spots.

We typically use the TM4 suite of software for analysis of cDNA microarray data. TM4 is easy-to-use, open source software and is freely available. TM4 is a complete suite of packages with modules for normalization (TM4 MIDAS), analysis (TM4 MeV) and storage (TM4 MADAM) of cDNA microarray data. A brief tutorial which describes how to use this software is available in the recent review by Saeed *et al.*, (*Methods Enzymol*. 2006). In TM4 MeV, one can set cut-off levels to either discard or keep low-intensity data or genes with a high percentage of missing values. If some data are missing (rows with less than 15 percent missing) we impute the missing values using an algorithm such as knn impute (Troyanskaya *et al.*, *Bioinformatics* 2001).

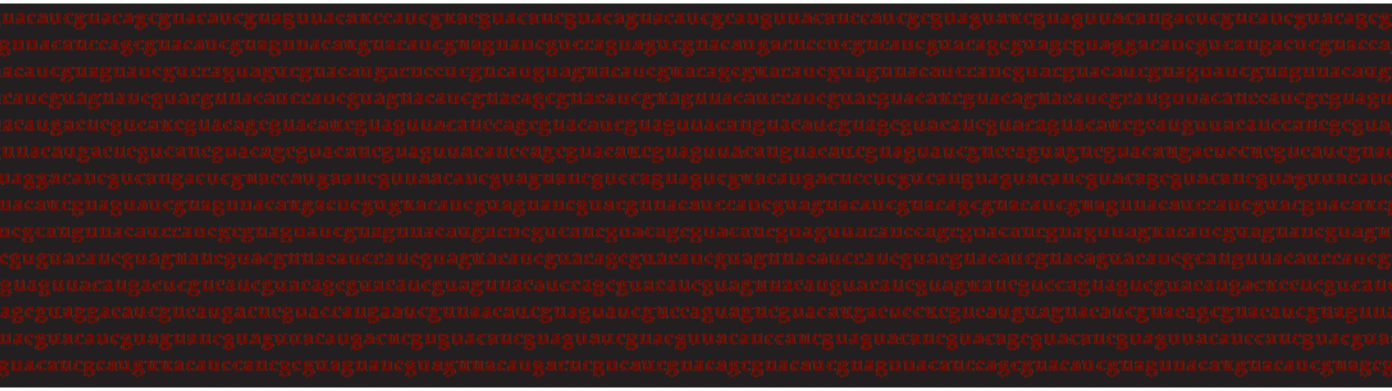
— Aedín Culhane and Renee Rubio

We use "spike-in" controls to determine the level of detection for any given microarray experiment.

— Elva Diaz

Data can be maximized by using the highest-quality RNA samples and arrays, which in turn maximizes the signal-to-noise measured on the arrays and helps to accurately measure low-intensity data.

— Audrey Gasch



Genes will not be thrown away just because they have low signal intensity. Rather, they will be marked as "absent" or "marginal" and enter downstream statistical analysis as categorical data or ranked data. This ensures that genes that are regulated from absent to present or the other way around will still be identified.

— Li Liu

We have not put a lot of emphasis in trying to squeeze out very low-intensity data, since this is somewhat risky. We have done arrays at different concentrations to try to pull out low-intensity data, but if it is too low I'd rather not rely on it. We use different analysis methods, too, and some of them look at significance independent of intensity (as long as the spot passes all of our quality filters), so this is how we are able to pay attention to some of the low-intensity data.

— Coleen Murphy

Discard low-intensity data as they are usually not reproducible. Any statistical conclusions at the noise floor is challenging as these intensities are not reproducible.

— Radhakrishnan Nagarajan

We avoid missing data by using background correction methods which return positive corrected intensities. Where possible we use morphological opening background estimates. These are available from SPOT or GenePix 6.0 image analysis software. If morphological background estimates are not available, then we use the "normexp" adaptive background estimator implemented in the limma package for R to transform whatever background estimate is available, usually the local median background. Both of these background estimators not only avoid negative corrected intensities but also stabilize the variabilities of the log-ratios as a function of intensity, which has many benefits for the downstream analysis. One benefit is that we need to do very little filtering of the data, retaining the ability to detect differential expression at relatively low expression levels.

—Gordon Smyth

The arrays can be scanned on the highest PMT. Also, replication is important to adequately determine low signal intensity from noise.

— Marina Telonis-Scott

Genome Technology Online

Still have questions about your cDNA microarray experiments?

Get answers from your peers at the *Genome Technology Online Forum*

http://www.genome-technology.com/forum/

How do you normalize and confirm data?

Before performing a batch or between-array normalization, we check for consistency between arrays using the approach outlined in question 7 (boxplots, histograms, image, density plots). It is important to exclude any outlier arrays prior to batch normalization. This is because most normalization methods use the assumption that gene expression, on average, is consistent across samples, which follows from assuming that each cell produces a fixed amount of RNA. This assumption has to be tested as an array with an unusual distribution, either due to artifacts in the data or real biological effects, will skew normalization. One example where this assumption would be invalid due to biological reasons would be a study where an important regulator of transcription is knocked down. Frank Holstege and his group have done some nice work on this problem and recommend using spike-in controls on cDNA microarrays (van Bakel *et al.*, *EMBO Reports* 2004).

Having settled on an appropriate dataset, we typically use LOWESS (Locfit) normalization implemented in the MIDAS module of TM4. In most of our experiments, we perform dye reversal replicates to remove any potential dye bias, so we use the "flip dye" option in MIDAS to average over these replicates and to eliminate genes with inconsistent results across replicates. We also typically perform a regularization of the standard deviation across the samples.

Following normalization we again perform the checks outlined in question 7 (boxplots, histogram, hierarchical cluster, correspondence analysis) as a final measure of the overall quality of our data.

Following normalization and exploratory data analysis, we perform a statistical analysis of the data appropriate for our experimental design. Genes that we identify as correlating with our phenotype (*i.e.*, significantly differentially expressed) are then often subjected to a Gene Ontology or pathway analysis to look for biological interpretations that can be used to put the results into context. For example, in MeV we have implemented EASE (Hosack *et al*. *Genome Biol*. 2003), which looks for overrepresented functional classes in the final dataset based on Gene Ontology assignments and maps to cellular pathways.

Finally, selecting significant genes with some biological justification supporting them, we perform qRT-PCR to confirm them. As always, this is best if we can use a completely independent set of samples, but a first-pass confirmation can be performed using the RNA samples we have used for the array experiments themselves.

— Aedín Culhane, Renee Rubio

Data are normalized by print-tip LOWESS normalization method in the marray package. Normalization is confirmed by inspecting the data when graphed as signal versus ratio or when inspecting the spatial distribution of ratios on the microarray itself.

— Elva Diaz

On each spotted microarray, the data in the two fluorescence channels (*e.g.,* the red and green channels) are regionally normalized by setting the center of the red/green distribution to 1.0. Data are confirmed through replicates and other genomic indicators (such as expected enrichment of functionally related genes

(continued on p.14)

What protocols do you use to ensure your data is reproducible?

We have found that it is absolutely essential to have standard lab protocols that are used consistently for all experiments. If this is not done, performing a simple hierarchical clustering on the data often shows that the arrays cluster based on technical covariates (date performed/person, *etc*.). As noted previously, our "Standard Operating Procedures" are available online.

However, one cannot forget experimental design and, briefly, the design must be appropriate for the experimental question and implemented in such a way that covariates are not confounded.

Finally, one has to confirm and validate the results. In some applications, cross-validation approaches can be used, but the application of your results to an independent test set is the only way to assure reproducibility.

— Aedín Culhane and Renee Rubio

> "I'm a big fan of trying to get the same biological data from several different types of experiments — such as using RNAi and many alleles of the same gene — to weed out unimportant changes in the background."
>
> *— Coleen Murphy*

Typically, three biological replicates are required for any microarray experiment to ensure reproducibility of positively differentially expressed genes.

— Elva Diaz

Replicating the experiment (from biological sample collection through arrays) is the best way to verify reproducibility.

— Audrey Gasch

We do many more biological replicates than I think most labs do (and some, but fewer, technical replicates) because I'm more concerned about finding gene expression changes that are robust to slight variations in preparation. Dye-flips didn't give me different data when I used indirect labeling of non-amplified samples, but since started using linear amplification more often, we see that dye-flips are more important. I'm also a big fan of trying to get the same biological data from several different types of experiments — such as using RNAi and many alleles of the same gene — to weed out unimportant changes in the background.

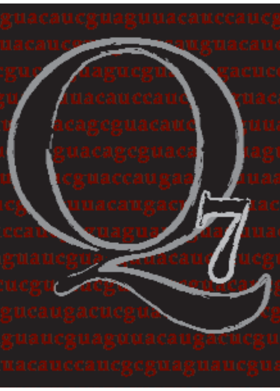
— Coleen Murphy

Generating replicate arrays is the only way to assess reproducibility within a given laboratory and within a given microarray platform.

— Radhakrishnan Nagarajan

It is important to include replicate arrays in the experiment with biologically independent samples. Unnecessary pooling of the biological samples should be avoided so that biological variability can be estimated from the replicate arrays.

— Gordon Smyth



Which statistical techniques ensure that your data is robust?

We use some of the tools in Bioconductor, which is a suite of packages in R for exploratory data analysis, as well as those available in available in TM4 and in many online microarray data analysis suites.

We assess data distribution using a boxplot, density plot, or histogram using the R functions boxplot(), density(), hist(). We also generate a pseudo or false-color images of the data as this is useful for visualization of spatial irregularities on an array. We generally perform both hierarchical cluster analysis (Pearson correlation metric, average linkage) and a dimensional reduction analysis (correspondence analysis) on the data to look for biases and to check for any obvious outliers. We believe it essential to perform these exploratory analysis steps both before and after normalization, and prior to any supervised data analysis (even a t-test). We have described how we use these approaches in a recent review (Brazma and Culhane, in *Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics* 2005). Other useful microarray analysis reviews that are applicable to cDNA microarrays include Quackenbush (*Nat Gen*. 2002), Quackenbush (*N Engl J Med*. 2006), and Steinhoff and Vingron M. (*Brief Bioinform*. 2006).

> "It essential to perform these exploratory analysis steps both before and after normalization, and prior to any supervised data analysis."
>
> — *Aedín Culhane and Renee Rubio*

Finally, most studies wish to extract a list of differentially expressed genes for validation. Different feature selection methods will produce different gene lists. We have compared different approaches in a recent study (Jeffery *et al*. *BMC Bioinformatics* 2006) and find methods which use a modified estimate of variance outperform classical approaches such as Student T-test. Methods which use a modified (or pooled) estimate of variance are thus more likely to yield a robust gene list than classical statistical methods. Although Significance Analysis of Microarray (SAM) analysis (Tusher *et al*., *Proc Natl Acad Sci USA* 2001) does use a modified estimate of variance, methods which estimate a modified variance using a Bayesian or probability approach are better (Jeffery *et al*.). For this we use the limma package in Bioconductor and find it performs well. However, if a study contains too few replicates or has high levels of noise, it is difficult to estimate the mean and variance of each gene in each group. In this case, it may be best to compare the ranks of the genes using a non-parametric rank approach, such as Rank-Products (Breitling *et al*. *FEBS Lett*., 2004; Jeffery *et al*.).

— *Aedín Culhane and Renee Rubio*

We typically use linear models in the limma package in Bioconductor to estimate gene expression differences between samples.

— *Elva Diaz*

First, we make sure the data are reproducible (assessed by linear regression of replicates). Second, we use statistical methods (like t-test and multiple-test correction) to identify genes that are significantly affected in triplicate arrays. Third, we often rely on other statistical methods to look at a genomic response (*e.g.* hypergeometric distribution of Fisher's exact test to assess the enrichment of functionally related genes in a selected set of data).

— Audrey Gasch

We conduct both parametric (*e.g.* t-test, ANOVA) and non-parametric (*e.g.* rank test, permutation) tests on the same data set. Further, false discovery rate will be calculated and used for filtering together with p-values and fold change. For biomarker identification, we go with machine learning algorithms because they deal with co-regulated genes better than traditional statistical analysis.

— Li Liu

"Reproducibility of results across replicate array and across distinct statistical approaches indicate robustness of the findings."

— Radhakrishnan Nagarajan

We use statistical techniques that use multiple hypothesis testing and false-discovery rate (FDR) to be more certain that we are not just chasing down a lot of false positives. When we find genes that are deemed important by several totally different algorithms (they appear in clusters that reflect a biological role and they are also considered significant by SAM, for example) it also gives us confidence that the data are reliable, and should be reproducible by other analyses. In the end, though, we like to verify all of this data biologically to be sure that it is relevant and robust.

— Coleen Murphy

Each of the statistical techniques work under implicit assumptions and lead to spurious conclusions when these assumptions are violated. Reproducibility of results across replicate array and across distinct statistical approaches indicate robustness of the findings.

— Radhakrishnan Nagarajan

The microarray experiments that we analyze are typically differential expression studies with a small number of biological replicates per treatment, usually just two or three replicates. For these small experiments it is crucially important to use measures of statistical significance which borrow information between genes in order to achieve more stable and powerful results for each gene. We use the empirical Bayes t and F-statistics implemented in the limma package for R. The empirical Bayes t-statistics achieve a beneficial compromise between regular t-statistics and just using fold changes.

We also take a linear modeling approach to microarray experiments, which allows us to analyze complete experiments involving many treatment conditions as an integrated whole, rather than making piecemeal comparisons using subsets of the arrays.

— Gordon Smyth

Q1: What sample prep quality control steps do you have in place?

(continued from p. 5)

Agilent 2100 Bioanalyzer is used to determine the quality of RNA prior to hybridization.

— Radhakrishnan Nagarajan

All RNA is inspected for integrity using an Agilent Bioanalyzer prior to sample preparation. Poly A spike-in controls are added to all labeling reactions to determine labeling efficiency and reproducibility. Several hybridization controls are also included and assessed prior to data analysis.

— Marina Telonis-Scott

Q2: How do you improve hybridization specificity and sensitivity?

(continued from p. 6)

control spots is to color-code them in MA-plots.

— Gordon Smyth

I tend to keep up to date with current protocols — for instance, Agilent improved hybridization conditions to increase sensitivity and reduce noise. For glass slides, the MAUI hybridization system is also an option. This system employs air flow to more efficiently mix the hybridization while reducing the amount of labeled cRNA required.

— Marina Telonis-Scott

Q3: What do you consider when choosing a validation method?

(continued from p. 7)

Most of my collaborators use quantitative RT-PCR to validate differential expression discovered by microarrays. We have found that the choice of housekeeping genes for normalizing the PCR results can be crucial. There's no such thing as a universal housekeeping gene, so it may be necessary to screen the housekeeping genes for differential expression in the study under consideration.

— Gordon Smyth

Accuracy, robustness, and repeatability.

— Marina Telonis-Scott

Q5: How do you normalize and confirm data?

(continued from p. 10)

in the group of genes that change in expression in response to the condition tested).

— Audrey Gasch

For Affymetrix data, we use ProbeProfiler or dChip programs. Both of them weigh probes within a probe set and generate more reliable intensity values than MAS5.

For Agilent data (or other microarrays that use one probe to represent one target gene), a standard median-adjusted scaling factor will be applied. If the variance across arrays is big, a scaling factor derived from relative variance will be applied.

— Li Liu

Normalization is generally based on balancing the two channels; in most cases, spatially-based and LOWESS normalization have not significantly altered the data (especially on newer arrays) so we haven't placed a large emphasis on special normalizations.

— Coleen Murphy

Normalization in the case of two-color arrays is accomplished with loess. Those in the case of Affymetrix arrays [are] accomplished with RMA in conjunction with quantile normalization.

— Radhakrishnan Nagarajan

Print-tip loess normalization is our workhorse for genome-scale microarrays. For small boutique arrays we have found it useful to use specially constructed large-library-pool control spots to guide the loess curve (Oshlack *et al.*, *Genome Biology* 2007). For some special designs it is necessary to analyze the individual channel log-intensities instead of the log-ratios. In that case we quantile-normalize the A-values after loess normalizing of the M-values.

— Gordon Smyth

List of resources

There are a number of Web resources and publications germane to microarray confirmation and validation. In addition to our experts' recommendations, we have rounded up a selection of online tools to ensure necessary and sufficient array results.

Articles

Breitling R, Armengaud P, Amtmann A, Herzyk P. (2004)
Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments.
FEBS Lett. 573(1-3):83-92.

Jeffery IB, Higgins DG, Culhane AC. (2006)
Comparison and evaluation of microarray feature selection methods.
BMC Bioinformatics 7:359.

Hosack DA, Dennis G, Jr, Sherman BT, Lane HC, Lempicki RA. (2003)
Identifying biological themes within lists of genes with EASE.
Genome Biol. 4:R70.

Oshlack A, Emslie D, Corcoran LM, Smyth GK. (2007)
Normalization of boutique two-color microarrays with a high proportion of differentially expressed probes.
Genome Biology. 8:R2.1-2.8.

Quackenbush J. (2006)
Microarray analysis and tumor classification.
N Engl J Med. 354:2463-72.

Quackenbush J. (2002)
Microarray data normalization and transformation.
Nat Genet. 32 Suppl:496-501.

Saeed AI, Bhagabati NK, Braisted JC, Liang W, Sharov V, Howe EA, Li J, Thiagarajan M, White JA,

Quackenbush J. (2006)
TM4 microarray software suite.
Methods Enzymol. 411:134-93.

Steinhoff C, Vingron M. (2006)
Normalization and quantification of differential expression in gene expression microarrays.
Brief Bioinform. 7(2):166-77.

Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altman RB. (2001)
Missing value estimation methods for DNA microarrays.
Bioinformatics. 17(6):520-5.

Tusher VG, Tibshirani R, Chu G. (2001)
Significance analysis of microarrays applied to the ionizing radiation response.
Proc Natl Acad Sci USA. 98:5116-21.

van Bakel H, Holstege FCP. (2004)
In control: systematic assessment of microarray performance.
EMBO Reports. 5:964-9.

Web Tools

ArrayExpress
http://www.ebi.ac.uk/arrayexpress

Bioconductor
http://www.bioconductor.org

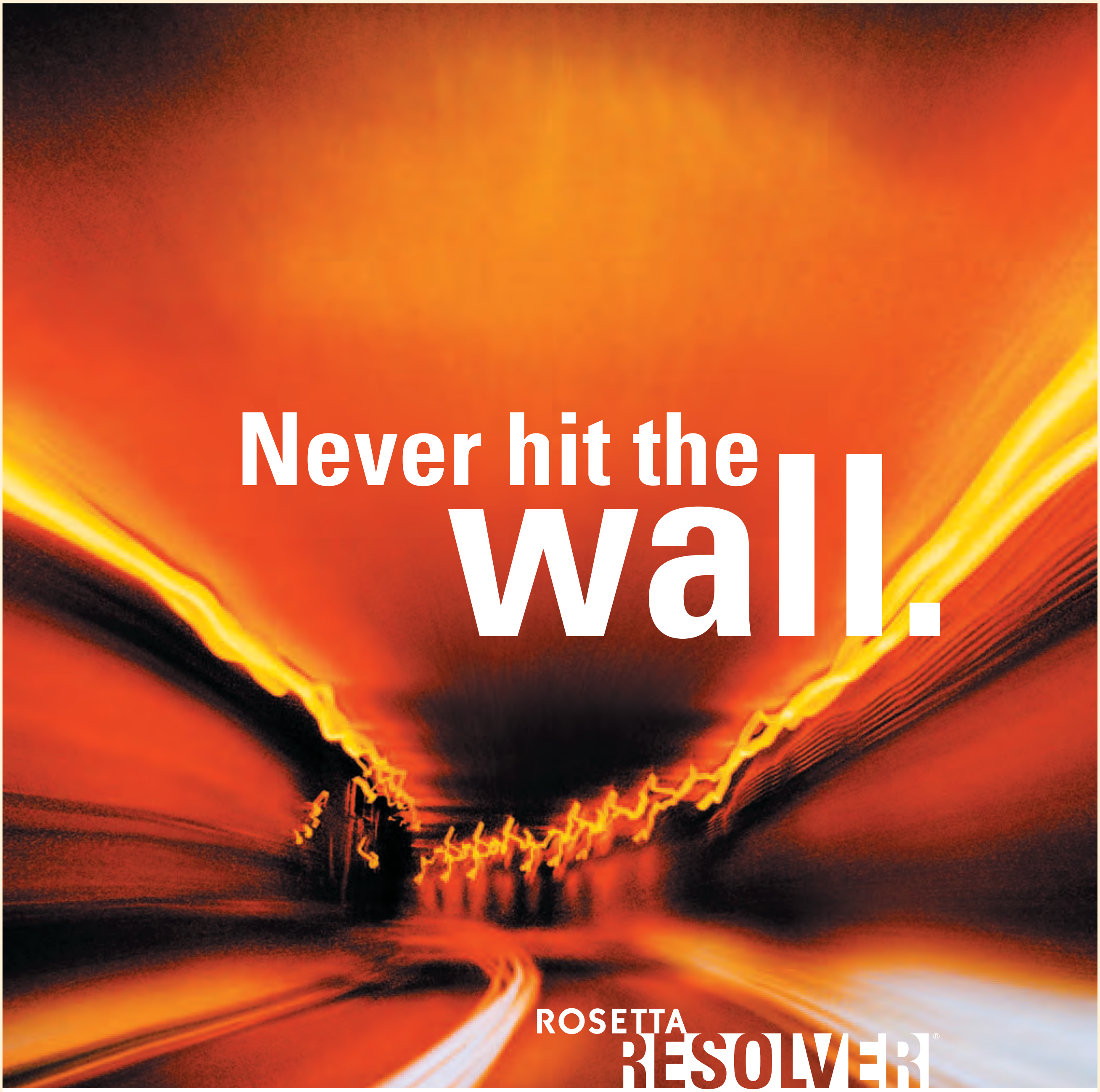
GEO: Gene Expression Omnibus
http://www.ncbi.nlm.nih.gov/geo

Human Protein Atlas
http://www.proteinatlas.org

R Project
http://www.r-project.org

Spot
http://experimental.act.cmis.csiro.au/Spot

TM4 software suite
http://www.tm4.org



Gene Expression Data Analysis System

Finally a gene expression analysis solution that grows with you.
The Rosetta Resolver® system is now available in three configurations: Standard, Extended and Premium. Each edition provides the same comprehensive analysis tools and scalable data management solutions, configured and priced according to your needs. So whether you're a biotech, research or pharmaceutical organization, the Resolver system extends to support the unique research goals of your scientists.

One Resolver system. Three configurations. Zero walls.
Only at www.rosettabio.com.

1.800.RESOLVE (U.S.)
+1.206.926.1220 (Outside the U.S.)
sales@rosettabio.com

ROSETTA™
BIOSOFTWARE

Copyright © 2007 Rosetta Inpharmatics LLC. All rights reserved.
Resolver and the Rosetta Biosoftware convergence icon are registered trademarks of Rosetta Inpharmatics LLC.