AFFYMETRIX
USERFORUM
NEWS • TECHNOLOGY • ANALYSIS

AFFYMETRIX

ARRAYS AND CONTENTS

www.affymetrix.com/userForum/

Article Reprint

# A Core Lab Case Study:

## Exon Array Challenges and Opportunities

By Stuart Pepper[1], Yvonne Hey[1], Gill Newton[1], Michał Okoniewski[2] and Crispin Miller[2]

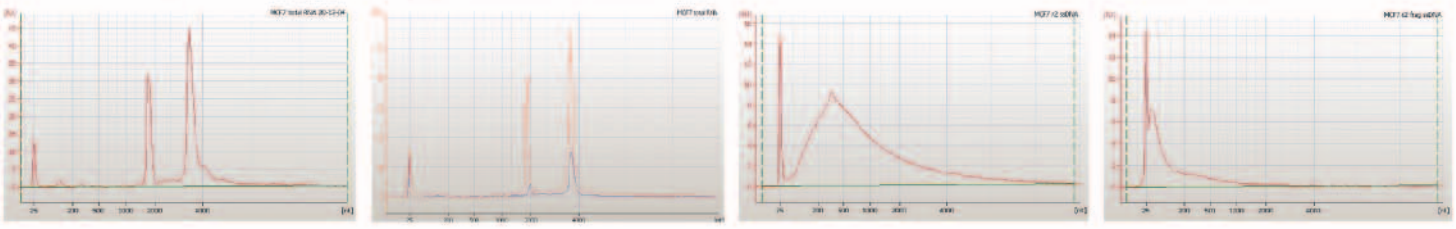
*This article looks at how the Molecular Biology Core Facility at the Paterson Institute for Cancer Research has started integrating the new generation of GeneChip® Exon Arrays for Human, Mouse and Rat into the range of services offered for expression profiling. These arrays offer a significant jump in the content and potential uses of Affymetrix expression arrays, with the number of probe sets increased from approximately 54,000 on the Human Genome U133 Plus 2.0 Array to approximately 1.4 million probe sets on the Human Exon 1.0 ST Array. The positioning of probe sets in each exon of 300,000 transcript clusters allows the mapping of splice variants at a level of detail that no other array platform can equal. For core facilities, these new arrays offer great benefits in cost effectiveness – customers can be offered 30 times as much data using products that are the same price. They also present challenges – the chemistry for these arrays is fundamentally different from the familiar in vitro transcription (IVT) reactions, and quality control (QC) metrics need to be reconsidered. Downstream processing of data is also a challenge, as the data sets for even small projects have now become significantly larger.*

The Cancer Research UK GeneChip® Microarray facility is based at the Paterson Institute for Cancer Research in Manchester. This service is available to all Cancer Research UK grantees and has a throughput of approximately 1,000 expression arrays per year. As with all core facilities, there is a constant drive to maintain cost effectiveness, and to keep the service up to date with developments to the Affymetrix platform. Therefore when the Exon Arrays were released it was important for us to start planning how these arrays could be offered as part of our range of services. For in-house testing we have a stock of RNA from two cell lines, MCF7 and MCF10A, which we use to compare array formats or different labeling protocols. Our first experiment with Exon Arrays was to generate technical replicates for the two cell lines using both the standard 1µg labeling protocol and the low-input 100ng protocol, and then cross map this to existing data generated using GeneChip® Human Genome U133 Plus 2.0 Arrays (HG-U133 Plus 2.0). This approach allowed us to validate a large number of fold change estimates generated using Exon Arrays without performing large numbers of qPCR assays.

> "In developing the Exon Arrays, Affymetrix has had to come up with a new labeling strategy to avoid the potential 3' bias that can results from oligo dT driven cDNA synthesis reactions, as used in the standard IVT labeling assay."

In developing the Exon Arrays, Affymetrix has had to come up with a new labeling strategy to avoid the potential 3' bias that can result from oligo dT driven cDNA synthesis reactions, as used in the standard IVT labeling assay (Full details of this protocol are in the manual, available at:

Figure 1. Bioanalyzer traces of RNA QC Stages



**Panel A:** Total RNA

**Panel B:** After ribosomal reduction step

**Panel C:** Single stranded DNA product

**Panel D:** Post fragmentation

*www.affymetrix.com/support/technical/manuals.affx*). This approach brings with it a problem, as ribosomal RNA will also take part in this reaction. To avoid high levels of contamination it is, therefore, necessary to carry out a ribosomal reduction step prior to cDNA synthesis. As suggested in the protocol, we used a Bioanalyzer to check that the ribosomal reduction step had worked (Figure 1, Panel B).

After the ribosomal reduction and random primed cDNA synthesis steps the protocol is similar to the familiar IVT chemistry, the final major difference being that at the end of the protocol labeled, single-stranded DNA is generated rather than cRNA. The intermediate yield of cRNA provides a useful QC point to confirm that the cDNA and IVT reactions have worked; however, it is worth noting that a yield would still be obtained here if the ribosomal reduction step had failed. The yield of ssDNA is the final QC point, although again, we also ran Bioanalyzer RNA chips to check the ssDNA before and after fragmentation (Figure 1, Panels C and D).

This protocol is more involved than the standard IVT, particularly with the extra ribosomal reduction step at the beginning. As a core facility with fairly high throughput, we were initially concerned about maintaining satisfactory turnaround times for exon projects. However we have managed to label batches of 12 samples over two days by running two rounds of the ribosomal reduction step with six samples in each, and then processing the batch of 12 in parallel. The final potential bottleneck during lab processing is scanning of the arrays, with a scan time of approximately 35 minutes per array. To do this manually would mean that only 12 arrays per day could be processed, so an AutoLoader is crucial to any lab looking to maintain a significant throughput, particularly if both exon and standard arrays are being run in parallel.

> "The results for either 'probes' or 'probe sets' mirror each other closely, with generally higher detection rates for the probe sets; this suggests that increased sensitivity is gained by combining the results for multiple probe sets across a transcript."

For groups who are used to the well established QC parameters of traditional Affymetrix arrays, the new Exon Arrays can be slightly intimidating at first. The QC report format is somewhat more complex, and there are fewer fixed guidelines to help interpret QC data. To fully analyze a set of 12 Exon Arrays takes a significant amount of computer power and time; however as a core facility we were keen to be able to rapidly generate QC reports before handing data back to users. We have used ExACT (available at *www.affymetrix.com/products/software/specific/exact.affx*) to generate QC reports and by using the Sketch normalization method with three probe lists (control.affx, normgene.exon and normgene.intron) it is possible to generate QC reports for 12 Exon Arrays in under five minutes on a standard desktop PC. The first six columns of Table 1 show some of the QC parameters for triplicate labelings of MCF7 and MCF10A. The results show good consistency across the arrays and a significantly higher percentage of probe sets called present in exons versus introns, as would be expected. The results for either 'probe sets' or 'meta probe sets' mirror each other closely, with generally higher detection rates for the meta probe sets; this suggests that increased sensitivity is gained by combining the results for multiple probe sets across a transcript. The bottom line of the table gives the ROC values for each array. These values give a measure of the false positive rate for each array, with 1 being perfect; again the values are very consistent across the arrays.

The first significant data analysis we undertook was to use the different mappings provided by Affymetrix to cross map exon results for MCF7 and MCF10A to existing data derived

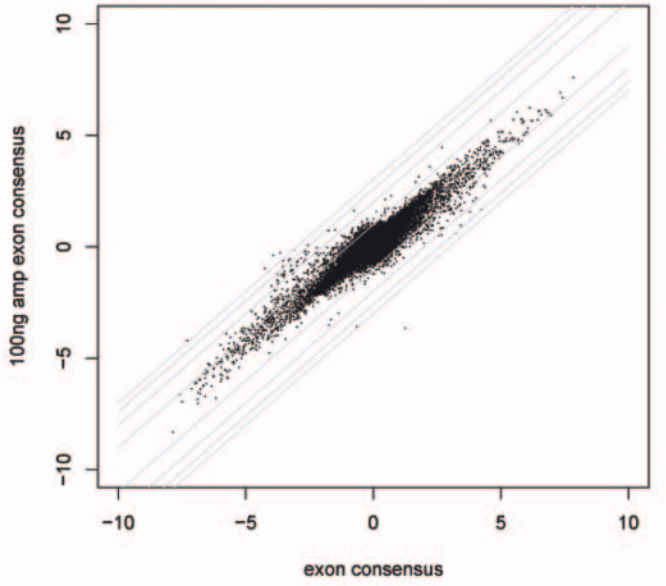
Table 1. QC% Probes Detected Above Background p<= 0.01

| | 1ug total RNA | | | | | | 100ng total RNA | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | MCF7_r1 | MCF7_r2 | MCF7_r3 | MCF10A_r1 | MCF10A_r2 | MCF10A_r3 | MCF7_r1 | MCF7_r2 | MCF7_r3 | MCF10A_r1 | MCF10A_r2 | MCF10A_r3 |
| Total | 45.2 | 41.9 | 42.7 | 44.9 | 41.6 | 41.7 | 38.4 | 34.2 | 34.6 | 39.4 | 39.7 | 34.7 |
| Control set | 71 | 74.5 | 75.7 | 78.8 | 80 | 78.8 | 77.6 | 70.1 | 70.8 | 72.3 | 74.5 | 68.4 |
| Exon set | 78.5 | 76.6 | 77.3 | 81.2 | 78.4 | 78.9 | 73.2 | 69.4 | 70.5 | 72.9 | 73.3 | 68.4l |
| Intron set | 30.6 | 26.4 | 27.3 | 28.7 | 25 | 25 | 22.6 | 18.4 | 18.5 | 24.4 | 24.6 | 19.6 |

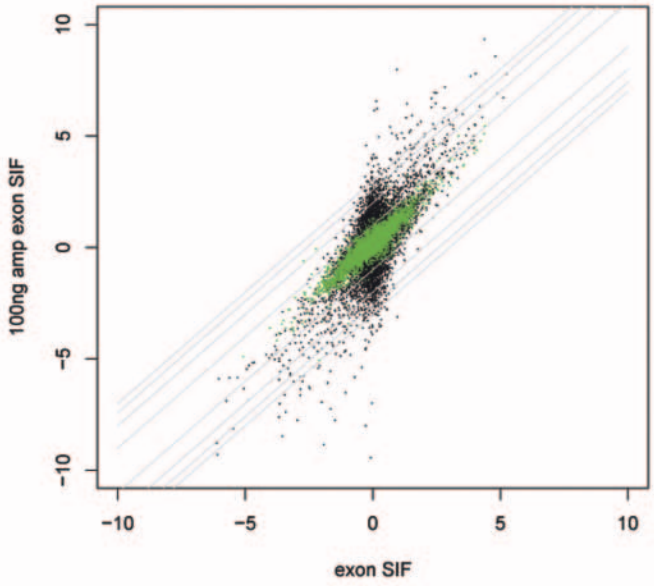
% Probe Sets Detected Table 1.QC Above Background p<= 0.01

| | 1ug total RNA | | | | | | 100ng total RNA | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | MCF7_r1 | MCF7_r2 | MCF7_r3 | MCF10A_r1 | MCF10A_r2 | MCF10A_r3 | MCF7_r1 | MCF7_r2 | MCF7_r3 | MCF10A_r1 | MCF10A_r2 | MCF10A_r3 |
| Total | 74.8 | 68.9 | 69.8 | 71.5 | 66.3 | 67.2 | 65.4 | 58.6 | 59.6 | 67.2 | 68.4 | 58.9 |
| Control set | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |
| Exon set | 93.1 | 92.5 | 92.3 | 93.5 | 92.2 | 92.5 | 90.9 | 87.4 | 88.6 | 90.4 | 90.7 | 88.1 |
| Intron set | 67.1 | 58.9 | 60.2 | 62.2 | 55.3 | 56.5 | 54.5 | 46.3 | 47.2 | 57.3 | 58.9 | 46.5 |
| ROC AUC: pos-neg cont. | 0.880975 | 0.880793 | 0.882748 | 0.8975915 | 0.893318 | 0.8961567 | 0.87245 | 0.864759 | 0.86736 | 0.8649001 | 0.865325997 | 0.8550509 |

Figure 2.




Comparison of input amounts using consensus mapping




Analysis performed using SIF mapping to give exon level data. Green data points are those with DABG < 0.01, black spots are DABG > 0.01.

from HG-U133 Plus 2.0 Arrays. This work is currently under review for publication and cannot be discussed in detail but, the results were very encouraging, with excellent agreement in the fold changes measured by each array type.

The second point of interest for us was how well data from the two different input amounts of RNA would compare. The standard Whole Transcript Assay requires a minimum of 1µg of total RNA (and our experience so far suggests that a minimum of 1.5 µg will give more reliable yields), whereas many microarray projects are based on biopsy material from which it can be difficult to obtain this amount of RNA. When using the low-input protocol the initial ribosomal reduction step is omitted to avoid loss of the small amount of mRNA in the sample. From the point of view of batch processing, this significantly reduces the time required to take samples through the protocol, but does so at the risk of increasing the level of background in the final sample. The right-hand columns of Table 1 show the QC results for the low-input protocol. The arrays show the same level of consistency within the group as the standard protocol, but there does seem to be some drop off in the number of probe sets called present.

Figure 2 shows a comparison of fold changes obtained when analyzing the two cell lines MCF7 and MCF10A when starting with either 1 µg or 100 ng of input RNA, and also points out the importance of analysis on the results. For this figure, only probe sets which can be cross mapped to HG-U133 Plus 2.0 Arrays have been presented, simply as a way of reducing the complexity of the plots. In the left panel the arrays were analyzed using a consensus mapping file (map.hg-u133-2.0-plus-cons.txt) which essentially generates a transcript-level summary of data, yielding approximately 42,000 data points.

For the right panel, a similar SIFsequence-based file (map.hg-u133-2.0-plus-sif.txt) was used to generate exon-level summaries. In this plot, if all the data are included, then there is considerable scatter in the data. However, if the data are selected for probe sets that are present (DABG < 0.01, green spots), then the correspondence between input amounts looks very good. The files for consensus and SIF translation to HGU133 Plus 2.0 Arrays may be found in *www.affymetrix.com/Auth/analysis/downloads/exon/HuEx-1_0-stv2.extras.zip*.

> "One observation suggests that Exon Arrays may prove to be more sensitive than traditional format arrays."

We are still in the early stages of following up exon data; however, one observation suggests that Exon Arrays may prove to be more sensitive than traditional format arrays. As part of a previous study, we had used real-time PCR to show that five transcripts which had all been called as absent in both MCF7 and MCF10A on HG-U133 Plus 2.0 Arrays

were actually detectable. Using the consensus mapping approach we found that four out of five of these transcripts had DABG values below 0.01. Further work will be needed to confirm whether the DABG proves to be robust and reliable but this initial observation looks very positive.

Exon Arrays present some novel challenges to core facilities, with essentially every aspect of the reagents, arrays and software being different to previous Affymetrix reagent. However, our initial results have been very encouraging and we will soon be rolling out Exon Arrays as an integral part of our service.

*The low-input data set has not been submitted for publication but is available on request.*

[1] Cancer Research UK GeneChip Microarray Facility, Molecular Biology Core Facility, Paterson Institute for Cancer Research, TheUniversity of Manchester, Christie Hospital Site, Wilmslow Road, Withington, Manchester, M20 4BX, UK.
[2] Bioinformatics Group, Paterson Institute for Cancer Research, The University of Manchester, Christie Hospital Site, Wilmslow Road, Withington, Manchester, M20 4BX, UK.