# Human Gene 1.0 ST Array Performance

## I. Introduction

The Human Gene 1.0 ST Array, Human Exon 1.0 ST Array, and Human Genome U133 Plus 2.0 Array were observed to have comparable gene-level performance. This is based on an evaluation of true positive rates (finding a detected change givenreal changes ) and false positive rates (finding false changes given no real change). Real and apparent performance was evaluated with a Latin Square spike-in experiment and with a tissue mixture experiment. In the spike-in experiment, Receiver Operator Characteristic (ROC) plots, which reflect both true and false positive rates, reveal comparable performance between the three array types with respect to a limited number (28) of spiked transcripts (Figure 2). The tissue mixture experiment revealed similar overall detection rates (Figure 8), fold changes (Figure 6), and comparable ROC plots for discriminating known changes based on the full data set (Figure 9 and Figure 10).

## II. Array Platforms

The Human Genome U133 Plus 2.0 Array (3' biased design), the Human Exon 1.0 ST Array (WT based design), and the Human Gene 1.0 ST Array (WT based design) were used in the performance evaluation presented below. Additional information about each of these array designs can be obtained from the respective Technote for each design available from http://www.affymetrix.com/support/.

To account for genome version differences, probes from the Human Exon 1.0 ST Array and the Human Gene 1.0 ST Array were mapped onto Build 35 of the human genome and grouped into gene bounds defined by RefSeq, Ensembl, and complete coding sequence (CDS) transcripts from Genbank. (See the Exon Probeset Annotations and Transcript Cluster Groupings white paper available from http://www.affymetrix.com/support/ for more information.)

For the spikes used in the Latin Square experiment, 11 probe pair probesets on HG-U133 Plus 2.0 were identified for each of the spikes. For the Human Exon 1.0 ST Array and the Human Gene 1.0 ST Array, probesets were created for each spike based on probes on the array which match the spike. The reason for this was that about half the spikes are very short, only covering the very 3' end of the gene.

For the tissue mixture experiment, only those probesets for genes in common across all three array types were used in the analysis. This produced a set of 15,865 probesets; these probesets tend to correspond to the better known genes and in general exhibit higher signal than other probesets on each of the arrays. It should be noted that probesets on each array type are not equivalent with regard to the biological targets they interrogate, particularly between the WT-based

arrays and HG-U133 Plus 2.0. The HG-U133 Plus 2.0 array can have multiple probesets mapping to the same gene (e.g. for interrogating alternative 3' UTRs) whereas the WT based arrays have a single probeset. Where such multiple mappings exist we chose one probeset at random to represent the gene.

# III. Data Sets

## *III.A. Spike in Latin Square Study*

The spike-in study used 28 RNA spikes that were observed to be absent in HeLa RNA using rtPCR (Ct ≥ 35).[1] The spikes were added to HeLa RNA at 7 different concentrations in a Latin Square design.

For this study, 7 different spike pools were created to contain the different spike RNA concentrations. Each pool was added to HeLa RNA, split into triplicates, and processed with the standard assay appropriate to the array type. The resulting samples were hybridized to the appropriate arrays. Hence 21 separate arrays were run for each array type.

Spike concentrations range from 0 to 12 pM. The non-zero concentrations were in factors of 2, with the 6 non-zero concentrations ranging between 0.375 pM to 12 pM (Table 1).

| Pool | Dilution | Copies Per Cell | Conc. (pM) |
|---|---|---|---|
| 1 | 0 | 0 | 0 |
| 2 | 1:400,000 | 0.88 | 0.38 |
| 3 | 1:200,000 | 1.75 | 0.75 |
| 4 | 1:100,000 | 3.50 | 1.50 |
| 5 | 1:50,000 | 7.00 | 3.00 |
| 6 | 1:25,000 | 14.00 | 6.00 |
| 7 | 1:12,500 | 28.00 | 12.00 |

Table 1: Range of RNA spike concentration used in the Latin Square spike-in experiment

1 pM concentration corresponds to about 2.33 copies of RNA transcript per cell, so this range corresponds to less than 1 copy per cell ranging up to 28 copies per cell.

The concentrations used in this study are intentionally low so that we can use them to assess signal response to differing RNA transcript concentrations that approach the limit of detection of these array platforms.

[1] rtPCR primers were based on the HG-U133 Plus 2.0 target region (region of the transcript sequence spanned by the probes). Thus it is possible that the gene is expressed, just not an isoform detected by the primers used. Based on Human Exon 1.0 ST Array expression data, we suspect that there is endogenous expression in the HeLa total RNA for three of the spikes (A7, B12, and C9).

### *III.B. Tissue Mixture Study*

The tissue mixture study used commercially available total RNA for human Brain and human Heart. The RNAs were mixed at 9 different mixture levels: 100% Heart, 95% Heart/5% Brain, 90% Heart/10% Brain, 75% Heart/25% Brain, 50% Heart/50% Brain (repeated 3 times), 25% Heart/75% Brain, 10% Heart/90% Brain, 5% Heart/95% Brain and 100% Brain.

Each mixture was split into triplicate and separately amplified and labeled using the standard array type protocols, and then hybridized on the arrays, for a total of 33 chips per array type.

## IV. Spike-in Latin Square Data Set Results

### *IV.A. Comparable Repeatability*

Signal from all array types in the spike-in study was highly repeatable. Most of the probesets measure the expression of endogenous transcripts from HeLa (not the spikes) and the reproducibility is dominated by this set. Only the intersection of genes represented on all three arrays were considered and the repeatability presented takes into account both array and target preparation variability. Pearson correlations between arrays of approx. 0.995 or better were observed (Figure 1). The Human Exon 1.0 ST Arrays have similar pairwise correlations to those observed on the Human Gene 1.0 ST and HG –U133 Plus 2.0 Arrays except that there are 3 chips in this particular data set that have slightly lower performance. These 3 chips show up as red crosses in the boxplots for the other arrays.

Another measure of repeatability is the Coefficient of Variation[2], which also showed that these arrays have very repeatable signal. The Human Gene 1.0 ST Array showed a slightly lower CV at 5.5% while the Human Exon 1.0 ST and HG-U133 Plus 2.0 were at 7.2% and 7% respectively for this specific data set.

[2] The Coefficient of Variation (CV) shown in Figure 1 is calculated only for expressing probesets. Expressing probesets were defined to be those with a mean signal above the 40$^{th}$ %-tile. The rationale is that conventional wisdom only expects about 60% of the genes to express, so we ignore the lowest 40%. For each probeset, its mean and standard deviation are calculated across the 21 chips from the Latin Square spike-in experiment. The ratio is taken and an estimate of CV is found by taking the median over the expressing probesets.

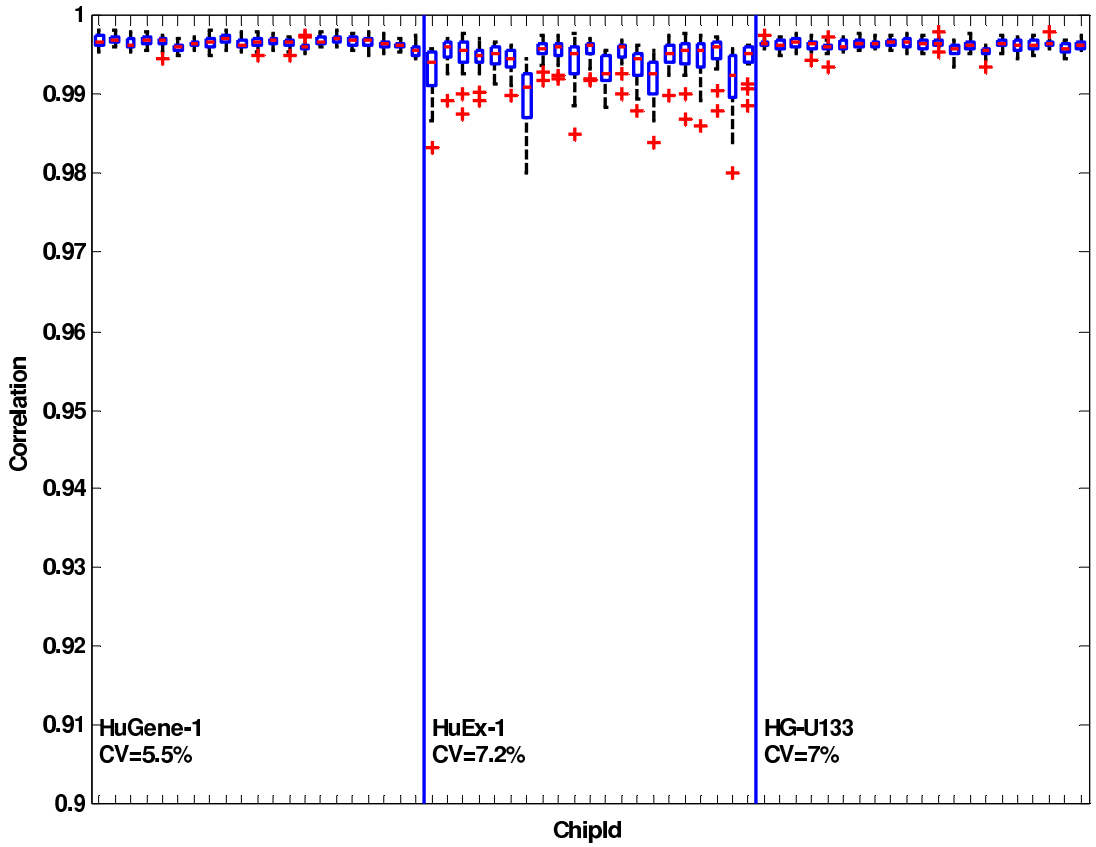


Figure 1: RMA Signal correlations between arrays arranged by array type and excluding spike-in probesets. Only probesets for the intersection of genes on each array were used. Each box plot summarizes the 20 correlations with other arrays of the same type.

## *IV.B. Comparable Detection Thresholds*

### IV.B.1 ROC curves for spike-ins

One method of assessing different platforms is to use ROC (Receiver Operating Characteristic) curves. Given a statistic for finding a difference between two conditions, plot the true positive rate (i.e., the proportion of detected differences among things that truly changed, i.e., the spike transcripts), versus the false positive rate (the proportion of detected differences among things that truly did not change, i.e., transcripts entirely absent or present in the unchanging HeLa background[3]). The statistic we used is the classic t-statistic

[3] The spike concentrations are in the lower range of detectable transcript levels, so we selected non-spike transcripts represented by probesets on the array by taking the mean signal across all arrays for each probeset and choosing those probesets with mean signal in the lower 60%. This also avoids probesets subject to saturation effects that vary across individual chips. Including these highly expressed probesets will tend to skew the ROC curves downwards as they will drive up the false positive rate.

(i.e.: $t = (\overline{x_1} - \overline{x_2}) / \sigma_{pooled}$ with $df = n_1 + n_2 - 2$). Here the means are calculated as the mean probeset signal for a spike at concentration level 1 and the mean probeset signal for the same spike at concentration level 2.

So for any particular value of $t_0$ we find the true positive rate using all spikes that have a t value exceeding $t_0$. The false positive rate was determined using probesets in the HeLa background that have t values exceeding $t_0$.[4]

ROC plots by convention graph the true positive rate on the Y axis versus the false positive rate on the X axis, parameterized by all possible values of $t_0$. Larger areas under the curve (AUC) correspond to better results.

Finally, we chose to process all arrays using full quantile normalization and generated signal using RMA (i.e., median polish using array and probeset factors with $\log_2$ probe intensities adjusted using the standard RMA background correction). Probeset signal was not exponentiated so its units are with respect to the log2 intensities.

[4] While we can derive a separate empirical false positive rates for the t-statistics for each of the 21 different possible comparisons between the 7 pools in this analysis we aggregated all the 21 comparisons into a single empirical t-stat distribution. While a much simpler method, this one size fits all approach ignores the fact that individual comparisons of pairs of pools can have substantial differences in the distribution of their t-statistics.

## IV.B.1.1 Comparable 2-fold Detection

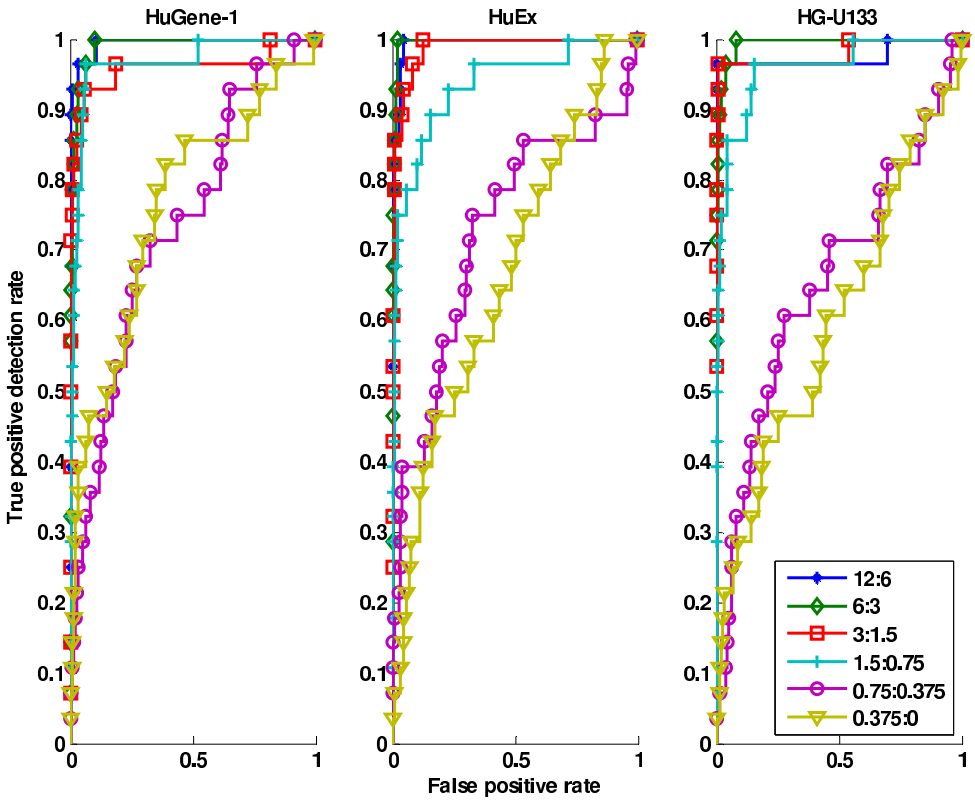


Figure 2: True Positive rates (Y axis) plotted versus False positive (X axis) for gene-level probesets using RMA signal

The ROC plots in Figure 2 represent each array type's ability to discriminate between presence of the spike and absence of the spike. Each curve plots false positive rates versus true positive rates; this shows how well a given concentration level can be detected versus its next lower concentration level.

For example, consider the left most panel in Figure 2. Even at a false positive rate of 0.1, 40% of spikes are detected when looking at 2-fold changes between 0.375 pM to 0.75 pM (from a little less than 1 copy of RNA transcript per cell to less than 2 copies per cell).

The three array types all showed similar characteristics. Detection between 2-fold changes at higher concentrations was quite good while detection of very low fold changes (0.375 pM versus 0 pM and 0.375 pM and 0.75 pM) was slightly better than random given 3 replicates. Note that this performance assessment is based on the technical variability of both the target preparation and the arrays and does not account for real biological variability.

In practice, only the very left hand portion of the ROC plot is of interest to most scientists; that is, users will usually select a $t_0$ such that the false positive rate is

low. Therefore, to compare the proportion of true positives for each array, a vertical line was drawn at the 0.1 false positive rate relative to Figure 2. The values for each array were then plotted versus the concentration ratio. As seen in Figure 3, the three array types exhibited very similar detection rates across the range of concentration differences.

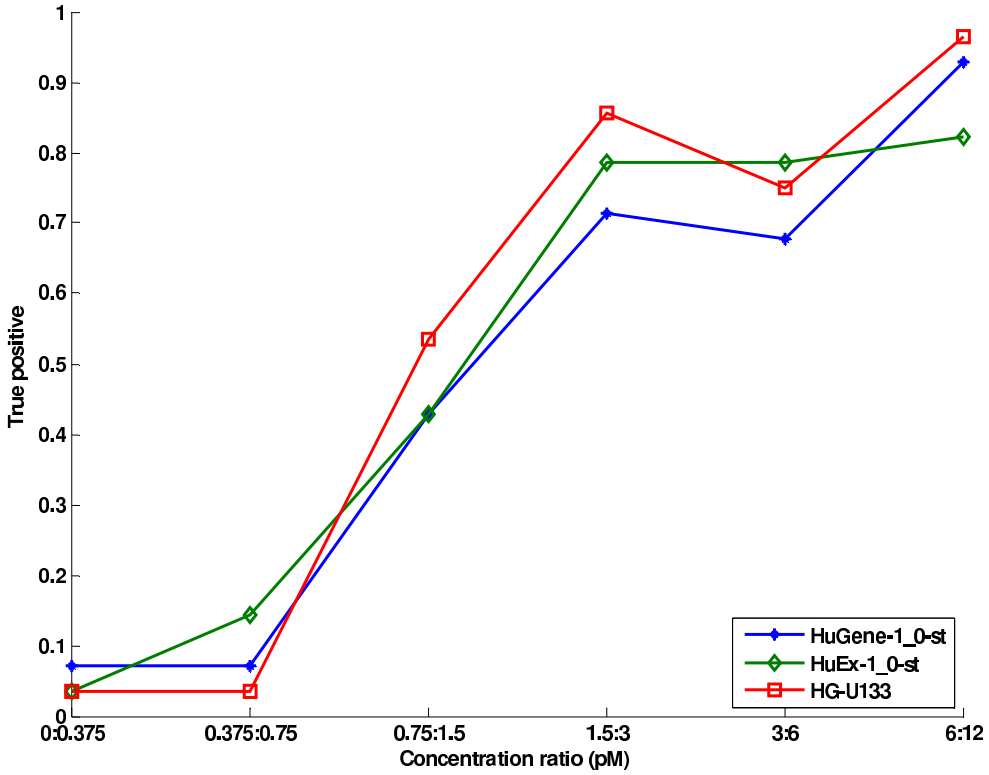


Figure 3 : Detection rate between concentration levels for gene-level probeset signal by array type (.01 level)

Figure 3 plots the proportion of true positives successfully detected at a false positive rate of 0.01 across 2-fold differences in concentration (except lowest concentration shows difference between .375 pM and 0), The three array types exhibit very similar detection rates across the range of concentration differences.

### IV.B.1.2 Comparable Absent/Present Detection

A very similar analysis can be made for detection of presence by looking at various spike levels versus zero rather than 2 fold changes. Figure 4 shows such a plot of detection rate of the spikes given a false positive rate of 0.01. All three

arrays exhibit similar detection rates.[5] At a concentration of 3.5 pM and greater all array types successfully detect 80% or more of the spikes.

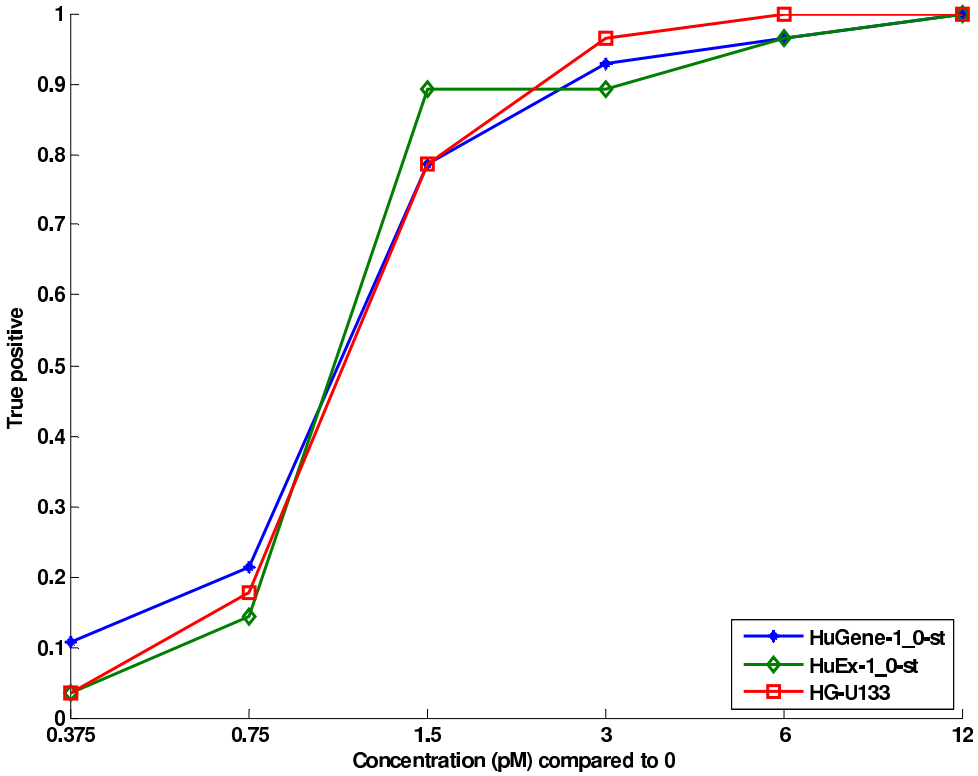


Figure 4 : Detection rate from zero for different concentrations of gene-level probesets by array type (.01 level)

## IV.B.1.3 An Observation Regarding the Use of t

The use of the t-statistic as a method to discriminate between different transcript concentrations is motivated by its widespread use and acceptance. In the ROC plots there are no underlying distributional assumptions as we can directly count false positive rates since we know that the background is unchanging.

However in real experiments it is not possible to find such an unchanging background. It has been our experience that the theoretical false positive rate as predicted by the t-statistic agrees reasonably closely with the empirical false

[5] The true positive rates at the .01 significance level at the lower level comparison, 0 to 0.375 pM appear different in Figure 3 and Figure 4, even though these are actually the same comparisons. This is due to slightly different data being used to calculate the empirical false positive rate. Empirical false positive rates were calculated using unspiked probesets and using the same pairwise comparisons as used to find the ROC curves, and the pairwise comparisons are different when comparing zero to other levels instead of between levels. Hence the apparent improvement of HuGene-1 may be a statistical artifact.

positive rate. For example, on the Human Exon 1.0 ST Array a theoretical rate of 0.05 actually corresponded to a true false positive rate of a little over 0.04. Similarly, the theoretical rate of 0.01 corresponded to a true rate of around 0.008.

Note that this closeness of the empirical distribution of the t-statistic to the theoretical t distribution occurs when we pool all 3 vs 3 comparisons; there may be considerably poorer agreement for any particular 3 versus 3 comparison. This variation was ignored in our analysis.

# V. Tissue Mixture Data Set Results

Signals for 15,865 genes represented on all three arrays were estimated by fitting RMA using the Affymetrix Power Tools (APT) apt-probeset-summarize command line program (version 1.6.0) using default settings over all the mixture levels and triplicates for each mixture level (described in Section III.B), but repeating the non-50/50 mixture level CEL files 3 times to prevent the signal estimation being dominated by the three pools of the 50/50 mixture level triplicates.

## *V.A. Real vs. Apparent Performance – Caveats of Tissue Mixture Study*

We refer to the performance observations from the tissue mixture study as "apparent" performance. The reason for this is that unlike the Latin Square experiment where we know what truth is, for a mixture study we are evaluating performance without knowing what is actually changing between our two tissues. For example, we compute the ability of each platform to detect changes between the two tissues; however we are unable to take into account the actual transcript diversity accounting for those differences. This is illustrated in Figure 5 where HG-U133 Plus 2.0 is only measuring the expression of a single transcript isoform whereas the Human Gene 1.0 ST Array is measuring the expression of both isoforms. As a result, the HG-U133 Plus 2.0 Array has an inflated apparent performance due to the 10 fold change observed for Variant 1 versus the 2 fold change observed for the gene as a whole.

| | Heart | Brain |
|---|---|---|
| Variant 1 | 10 copies per cell | 1 Copies per cell |
| Variant 2 | 1 copy per cell | 20 copies per cell |
| Total | 11 copies per cell | 21 copies per cell |
| HuGene | 5.3 | 6.3 |
| HG U133 | 6.1 | 1.2 |

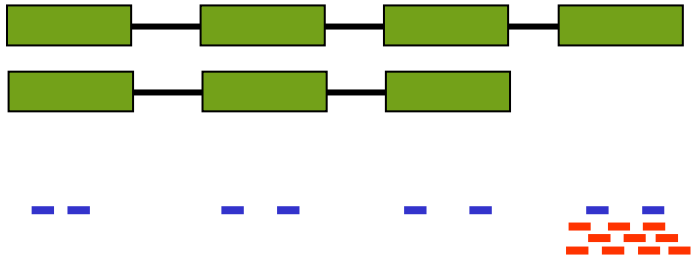

Figure 5: Evaluation of performance using only a tissue mixture study can be misleading due to the fact that the different array platforms are measuring different sets of transcripts. This figure shows the gene structure of two different transcript variants in green. Human Gene 1.0 ST Array probes are shown in blue while HG-U133 Plus 2.0 Array probes are shown in red. In this

example, the observed performance for the Human Gene 1.0 ST Array would be based on overall two fold change between heart and brain. In contrast, the HG-U133 Plus 2.0 Array would be based on an overall 10 fold change.

Another caveat to be aware of is that one may observe differences between the two tissues which are due to differences in background as opposed to real differences in the intended target. Thus a platform with less specificity may have good apparent performance driven in part by changes in background as opposed to real improvements in specificity and sensitivity. These effects are minimized by the various array design strategies used for these three different arrays.

The main benefit to a tissue mixture study is that the evaluations are done over a much larger set of array content (versus a limited number of spikes) and the samples reflect real biological changes such as changes in splice variants.

## *V.B. Analysis Motivation*

The tissue mixture study lends itself to an analysis that exploits the known mixing ratios. Given a mixture of $x$ Brain and $(1-x)$ Heart starting RNA, if the concentration of RNA from gene $g$ in Heart is $H_g$ and the concentration in Brain is $B_g$, then concentration in the mixture is $C_{p_H} = xB_g + (1-x)H_g$. If signal *y* is proportional to the true concentration for gene *g* with proportionality constant $\alpha_g > 0$ (which may be different for different array types, as well as being gene-specific) with a background $\beta_g > 0$ that is independent of the tissue type, then signal for mixture level x is:

$$y = (\alpha_g B_g + \beta_g)x + (\alpha_g H_g + \beta_g)(1-x) = \alpha_g (B_g - H_g)x + (\alpha_g H_g + \beta_g) \quad (1)$$

## *V.C. Good Fold Change Concordance Observed Between All Three Platforms*

If we ignore the background term in Equation (1) (i.e. assume background is negligible with respect to expression level) then forming the ratio of the estimated signals when x=0 and when x=1 is given by:

$$(\alpha_g H_g + \beta_g)/(\alpha_g B_g + \beta_g) \approx \alpha_g H_g / \alpha_g B_g = H_g / B_g \quad (2)$$

This is an estimate of Fold Change between Heart and Brain. Note that this estimate is independent of $\alpha_g$ and hence in theory should be similar across different array types. Figure 6 shows the pairwise scatter plots for the log Fold Change estimates[6] between Heart and Brain across the different array types. Agreement between Human Exon 1.0 ST and Human Gene 1.0 ST Arrays is

[6] Log fold change is problematic if signal is close to zero. We set signal to be 1 if estimated signal was less than 1. Such cases also give rise to large apparent log(FC) , if both Heart and Brain signal is in the noise. To account for these we set log(FC)=0 for probesets that showed no significant difference between Heart and Brain (at the 5% level).

excellent. Agreement between the WT based array designs and HG-U133 Plus 2.0 is comparable to previous expression platform changes (e.g. HG-U95 versus HG-U133). In this particular data set, a small number of genes showed an unusual patter of discordance between the WT based arrays and HG-U133 Plus 2.0. These genes were observed to have high expression in brain and low/no expression in heart for both platforms; however HG-U133 Plus 2.0 showed extremely high fold changes due to very low estimated expression levels in heart. The asymmetry of this observation suggests something particular to the biology (heart versus brain transcripts) and how the biology is interrogated on the arrays (whole gene interrogation versus interrogation at a specific 3' end). It should be noted that there is no variance stabilization for these fold changes values (i.e. we have FC values where the denominator is approaching 1).

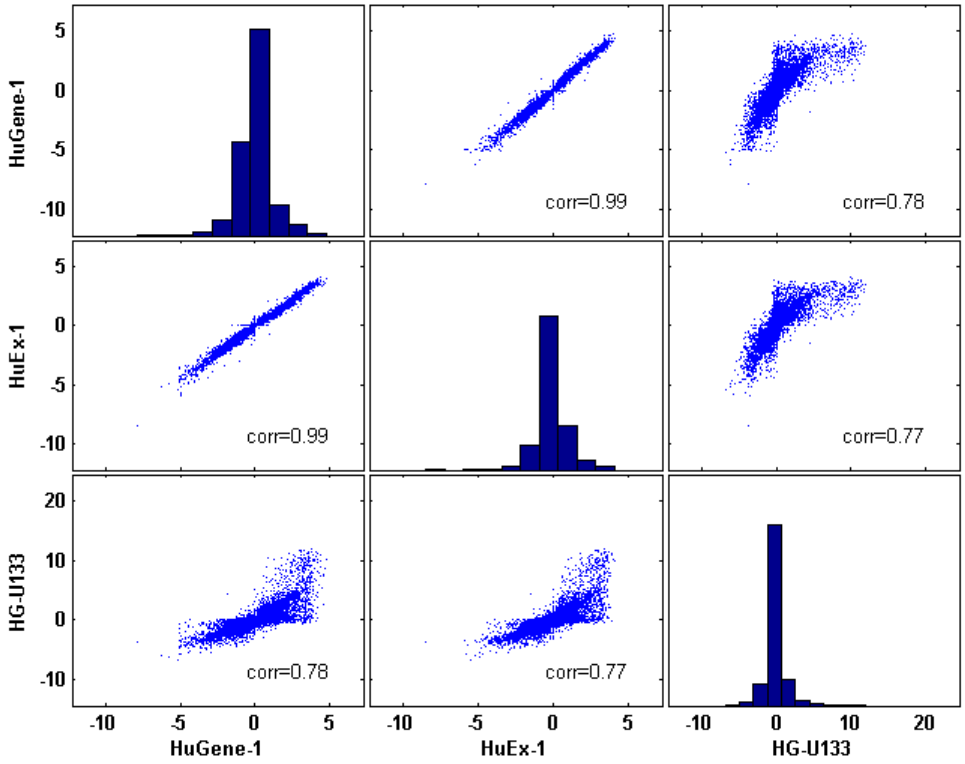


Figure 6 Log FC estimates between Heart and Brain. Histograms on the diagonal show the distribution of observed FC values within a single platform.

## *V.D. Comparable Absolute Apparent Sensitivity Observed Between All Three Platforms*

Equation (1) is in the form of a regression with known x values (since we know the mixture levels). Note that when $x = 0$ the *y* value (the intercept) is the signal from Heart and when $x = 1$ the *y* value is the signal from Brain. Figure 7 shows

an example for a specific gene on Human Gene 1.0 ST Array. Finding different expression levels between Heart and Brain is equivalent to fitting a regression and checking if the slope is significantly different from zero.

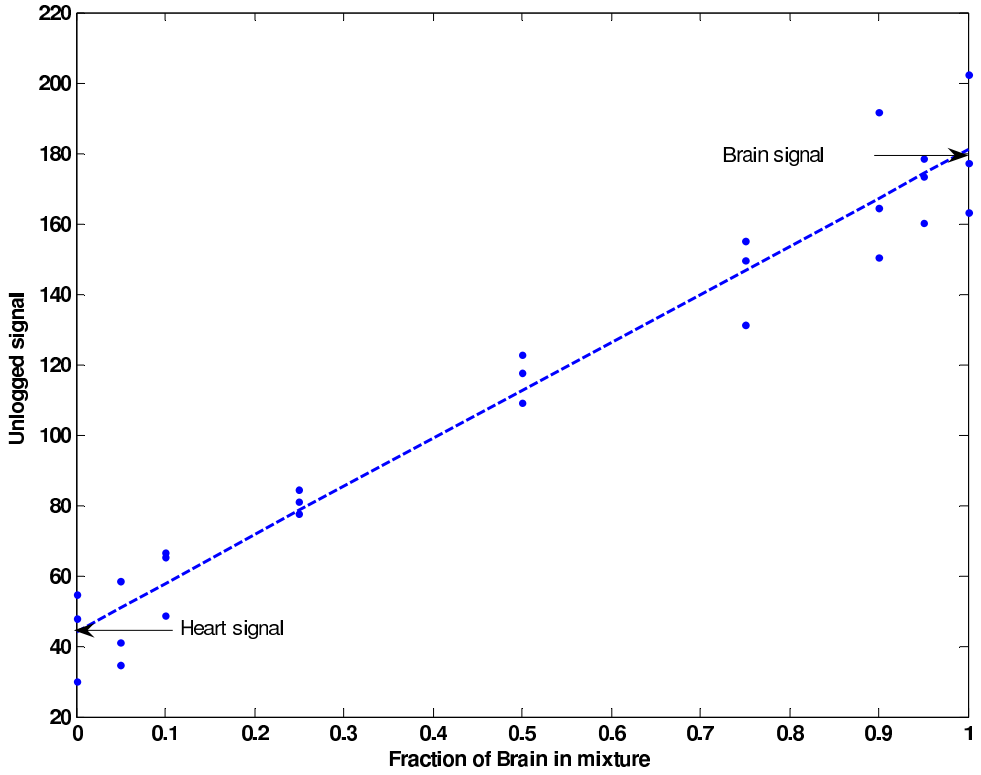


Figure 7: Signal regression for a gene level probeset on the Human Gene 1.0 ST Array across the different mixture levels.

We performed this regression to find probesets with likely different signal between Heart and Brain using all mixture levels except the 50/50 which has three pools and where two of the pools were withheld from the signal estimation for an analysis of variability (used to correct t-statistics in Section V.D).

A simple measure of sensitivity of arrays is to simply count the number of probesets with significant differences in signal between Heart and Brain. This measure works because this specific data set is based on a single biological sample each of heart and brain. Thus there is no biological variability within each tissue. The result is shown in the Figure 8, where similar levels of detection are observed in all three platforms. The comments above regarding caveats should be noted; specifically we are able to evaluate the behavior of each platform here, but not the absolute performance. For the 15,865 genes common to all platforms the detection rate at the 0.01 level was 69%-70% for the HG-U133 Plus 2.0 Array, 67%-69%, for the Human Gene 1.0 ST Array, and 63%-64% for the Human Exon 1.0 ST Array.

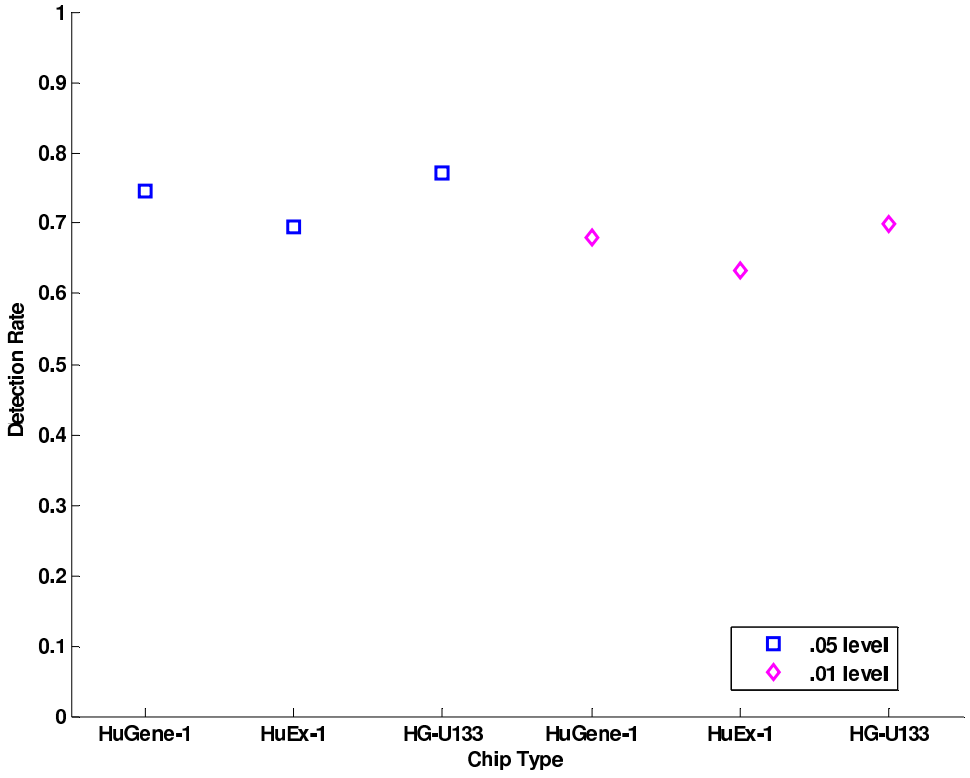


Figure 8 Detection Rate by Array type at .05 and .01 significance

While the analysis in Figure 8 ranks the array types by sensitivity to systematic changes in RNA pools, it may overstate biological sensitivity. Probesets that respond only to certain background RNA transcripts (i.e., their true target RNA

transcript is absent) may show differential expression because these elements are present in different concentrations in Heart and Brain.

## *V.E. Detection Rates for Small Changes are Similar*

The mixture experiment allows us to simulate an experiment with much smaller changes between experimental conditions. We don't know all the genes that change between Heart and Brain, but we can estimate those that are likely to have changed by using the regression method described above. At minimum, this group will be greatly enriched for genes that exhibit real change in expression level between Heart and Brain.

Our aim in this section to compare the apparent sensitivity of the different arrays, so we further restrict this group to genes that not only exhibit significant change in expression level between Heart and Brain for all array types, but also where the direction of change is the same for all array types. This procedure hopefully minimizes any systematic effects that might boost sensitivity of one array type over another. Setting the significance level for the regression at 5% gives a subset of 8,141 genes or about 51% of all genes in common on the three array types. This subset of 8,141 genes is assumed to be truly different between Heart and Brain based on the complete mixture data set.

### *V.E.1.1.1 5% comparison (simulating changes in expression level)*

As each mixture level was run in triplicate, we then calculated t-statistics (as defined in IV.B.1) between each mixture level and a baseline: the mixture with 95% of a tissue type. In this case all 8,141 genes will have expression at some level in both mixture levels and the detection rates will only be between different expression levels. Sensitivity for each array type can then be assessed using the proportion of genes in this subset actually found to be changing (as assessed by the proportion of t-statistics exceeding a given significance level, which we empirically determined by using a bootstrap of 3 vs 3 replicates pulled from the 9 replicates of the 50% mixture). Higher proportions, especially for comparisons with the mixture levels closest to the baseline 95% tissue type, correspond to a more sensitive array type.

The result is shown in Figure 9. The left panel is the result of comparing different mixtures with 95% Heart, 5% Brain, and the right panel is the result of comparing different mixtures with 95% Brain, 5% Heart.

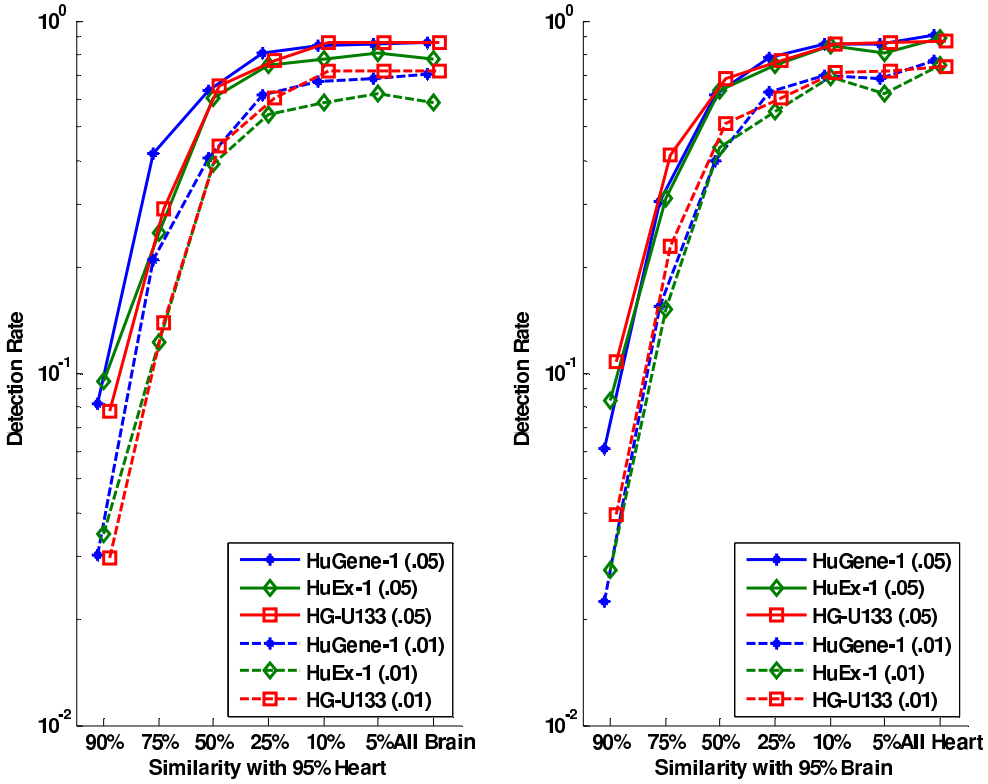


Figure 9: Detection Rates for change using t-statistics at 5% and 1% level between each mixture level and a 95% RNA mix for a tissue

This result shows detection rates rising as the difference between the 95% mixture and the other mixture level gets more extreme. The three different array types showed similar apparent sensitivity for the 15,865 genes in common to all three platforms.

### *V.E.1.1.2 Pure tissue comparison*

If we instead compare a baseline of a pure tissue type to mixtures, a large (though unknown) number of genes might be turned on in addition to expression changes. In each true mixture both Heart and Brain genes will have expression, but in the Heart only RNA only Heart genes express (and in the Brain only RNA only Brain genes express). Apparent sensitivity for each array type is then assessed using the proportion of genes in this subset actually found to be changing. Higher proportions, especially for comparisons with the mixture levels closest to the pure tissue type, correspond to greater apparent sensitivity.

The result of this assessment in shown in Figure 10, where detection rates of genes that changed are plotted by array type for two different significance levels (.05 and .01), using Heart as one baseline (left panel), and Brain as the other (right panel).

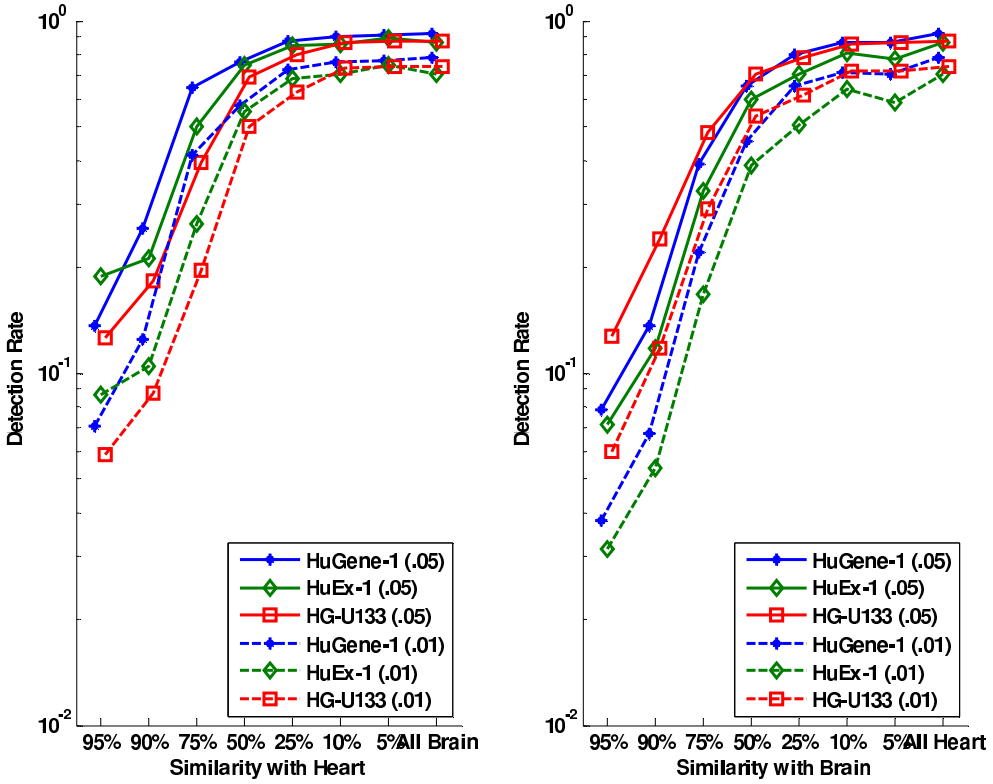


Figure 10: Detection Rates for change using T-statistics at 5% and 1% level between each mixture level and a tissue only sample

As with the previous analysis, all three platforms showed comparable apparent sensitivity. This result shows detection rates rising as the difference between the pure tissue and the mixture level gets more extreme. The pure tissue comparison shows markedly elevated detection rates between a pure tissue and a 5% admixture of the other tissue type (Figure 9) when compared with the detection rates between the 5% admixture and 10% admixture mixtures (Figure 10). This is true for both the .01 and .05 significance levels and for both tissue types. One possible explanation is that the detecting presence of transcripts even at a low level versus its absence is qualitatively a different task for these assays than detecting small changes in expression level. Hence the elevated detection rate in the Pure Tissue comparison can be ascribed to detection of presence versus absence of transcripts.

A similar effect can sometimes be seen in spike-in studies, where presence at the lowest spike level versus absence has higher detection rates for spike-ins than detection rates for a 2-fold difference in spike-in concentration at the lowest two concentrations. This effect can be seen in Figure 3 and Figure 4 of the spike-in study for both the Human Gene 1.0 ST and Human Exon 1.0 ST Arrays.

# VI. Summary

In summary, comparable gene level performance was observed for all three array platforms examined: Human Gene 1.0 ST, Human Exon 1.0 ST, and the Human Genome U133 Plus 2.0 Arrays. This evaluation accounted for both the differences in target preparation and in array format and content. The Gene and Exon arrays have the added benefit of interrogating the whole locus whereas the Human Genome U133 Plus 2.0 array is only interrogating discrete 3' ends. The Exon array has the further advantage of enabling exon level analysis and includes substantially more discovery content.