Minor change in library files for Ecoli and Ecoli_ASv2 has negligible effect on data

A minor difference between the historic and modern library files for the Ecoli and Ecoli_ASv2 commercial arrays has been identified. (For information about the four affected Custom Arrays please contact Affymetrix Support). The modern library files use a pre-computed feature pitch (*i.e.*, grid spacing) from the CIF file, that differs slightly from the feature pitch computed by either GCOS or AGCC when it is not provided in the CIF file. This is due to small rounding differences between the two computation methods. Most summarization routines (*e.g.*, MAS 5, RMA, and PLIER) are robust to individual feature changes, and therefore, even less change is observed at the signal level. As a result this small change in feature pitch has a negligible effect on the biological results.

To confirm this, an analysis was performed using the Latin Square data on U133A expression array where the feature pitch was intentionally removed from the library files before analysis to recreate the issue. It is important to note that the U133A array used **is not** affected by the feature-pitch issue (see list of affected arrays), and is used merely as a representative array type to illustrate the scale of the problem. Using the different feature pitches resulted in some intensity differences in all 42 CEL files examined. About 60% of the features on the array showed small intensity differences, but only 5% of the features had a relatively large proportional change in intensity (10% or greater) (Figure 1). Furthermore, those with the largest intensity differences were mostly confined to the lower intensity features. Again, we note that most signal summarization techniques average out changes to any individual feature. The typical correlation of Signal (summarized over probe sets using RMA) is 0.998 between the original feature pitch setting and the deliberately altered setting. This compares very favorably to the correlation between technical replicates of experiments which is normally between 0.97-0.99 (data not shown). While the % present call rate produced by the MAS5 algorithm is not always an easily-interpretable metric, people generally monitor its behavior. Figure 2 displays the % present for the feature pitch defined in the CIF file in blue, and for calculated value in red. The largest % present difference observed using the two difference values for the feature pitch was 0.7% (average 0.21%) compared to 4.6% (average 1.27%) for the largest % present difference between the technical replicates. After comparing the signal values and %P calls using the pre-computed feature pitch and the feature pitch calculated by either GCOS or AGCC we conclude that there is negligible impact on expression performance.

In summary, a small rounding difference has been identified between the modern and historic library files for the Ecoli and Ecoli_ASv2 arrays for the feature pitch value. This results in small changes in the calculated feature intensities in the CEL files. Since summarization routines (*e.g.*, MAS 5, RMA, and PLIER) robustly combine multiple features into a single probeset value, minimizing the observed effect at the signal level. Examination of the signal values and the %P calls revealed very minimal differences resulting from using the two different values for feature pitch. In fact, the minimal technical variation seen in the experiments was greater than the variation observed using the different values for feature pitch. Therefore we conclude that the impact of using the modern library files is negligible on expression performance.

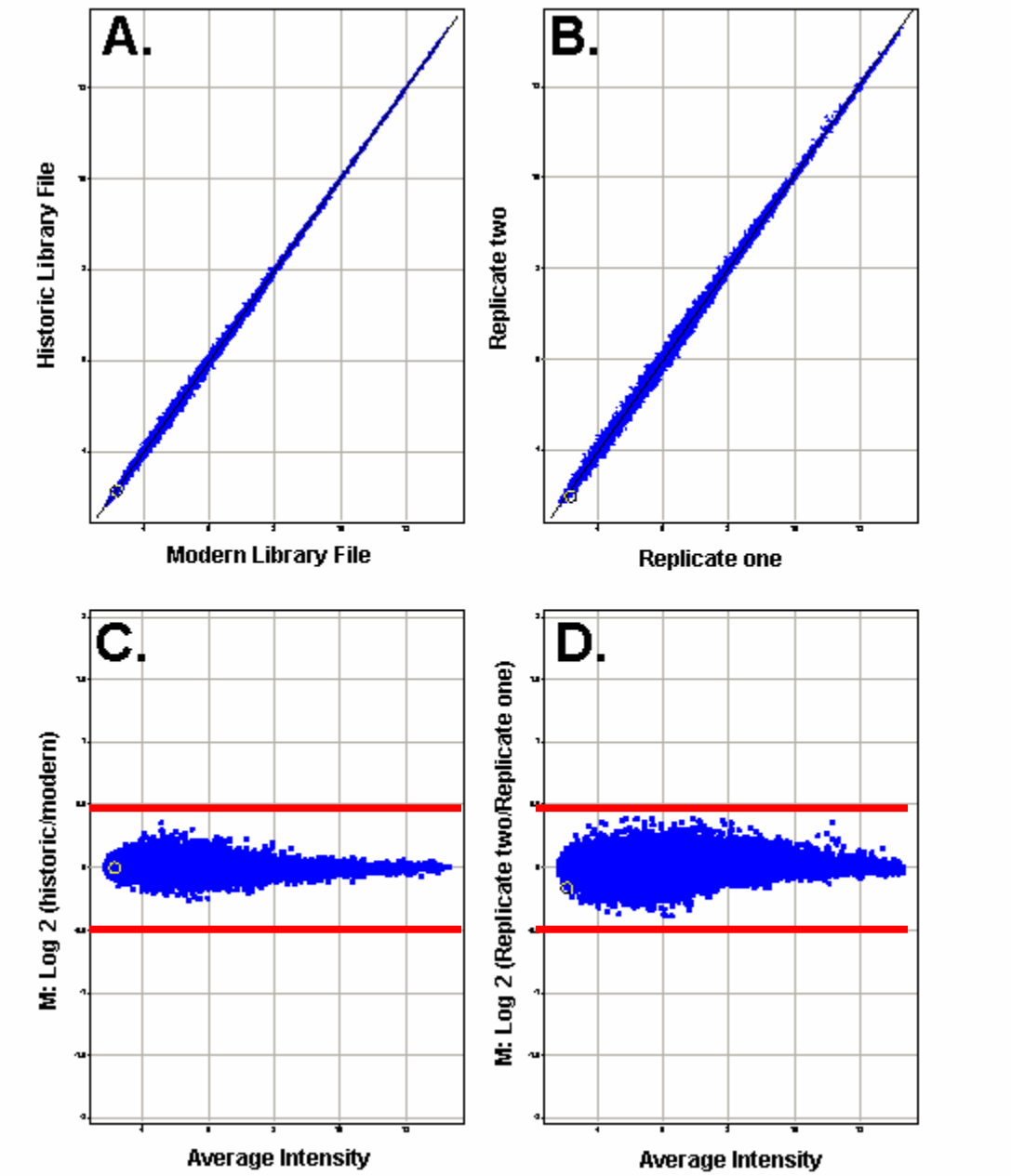

Figure 1: Comparison of technical variation and the intensity difference resulting from using the pre-computed feature pitch from the CIF file and the dynamically generated feature pitch. RMA signals were generated from a CEL file using a CIF file with the undefined feature pitch are plotted against RMA signals from a CEL from the same DAT file generated using a CIF file with the feature pitch defined in the file **A**. The R^2 value is 1.00 and the linear fit is y=-0.008+0.999*x indicating that the two arrays are very highly correlated and there is no overall bias (*i.e.*, no significant impact on derived fold change). In **B** a similar plot for two technical replicates (including the one used in A) with the defined feature pitch. The R^2 value is 0.999 and the linear fit is y=-0.055 + 1.008*x. **C** and **D** show MvA plots for arrays in **A** and **B** respectively. The absolute value of fold change in both cases is less than 1.4 (red line). The variation introduced by the small difference in the value used for feature pitch is less than that observed for the technical replicates, and the ratios.

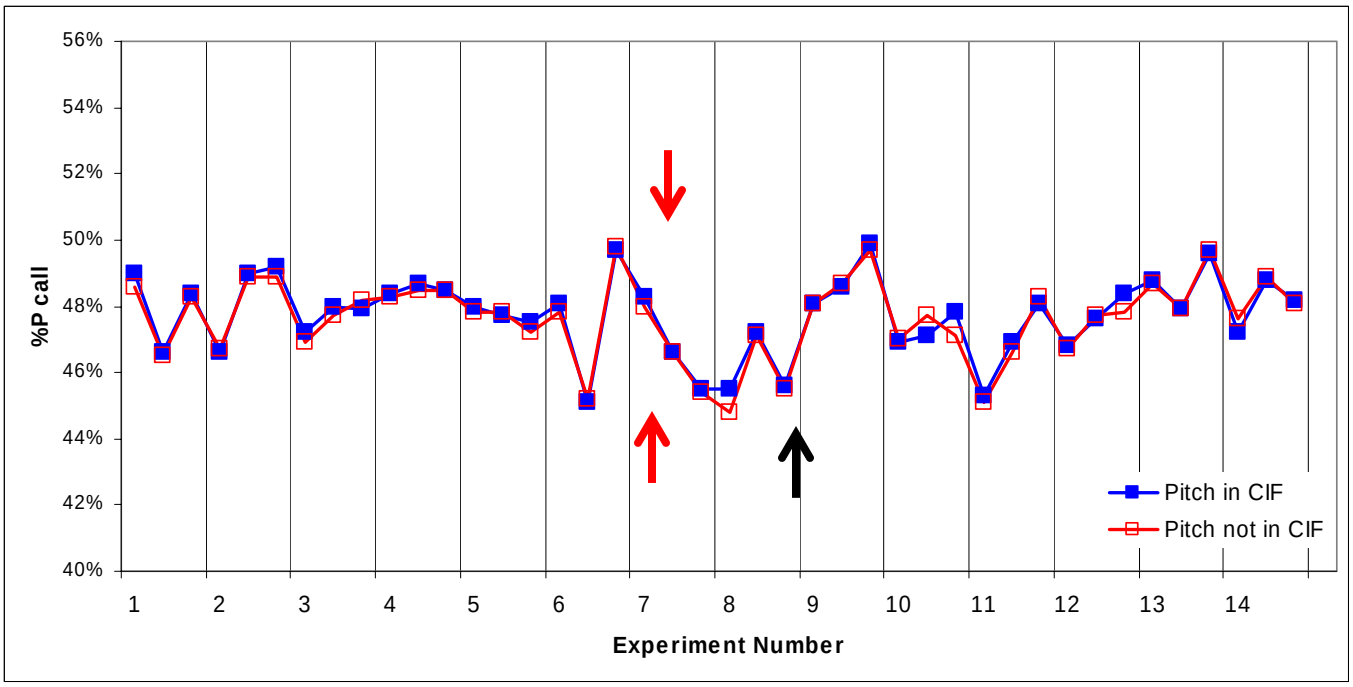

Figure 2: Comparison of the MAS5.0 %P calls using the pre-computed feature pitch from the CIF file and the dynamically generated feature pitch. The MAS5.0 algorithm was used to generate the %P calls for the 42 arrays. The y-axis contains the %P call and the x-axis is the individual experiments, with the three technical replicates grouped together by the vertical gridlines. The %P for arrays analyzed with the pre-computed feature pitch in the CIF file are shown in Blue, and those with the dynamically calculated feature pitch are shown in Red. The largest difference observed difference using the two difference feature pitches was 0.7% (black arrow). In general the differences in %P between technical replicates for the same experiment were larger than the difference between the gridding algorithms (red arrows).