**Supplement 2**: Transcriptomic profiling using Affymetrix

Summary report of the pilot study to determine the platform to use for gene expression profiling from FFPE

The HG-U219 Array Plate, which is a perfect-match-only (PM) array, utilizes content with an emphasis on established, well-annotated genes. The sequences used in the design of this array were selected from updated versions of UniGene (March 30, 2009), RefSeq (July 13, 2009) and GenBank (May 12, 2009). The HG-U219 Array Plate covers over 36,000 transcripts or variants but also contain the same 50 normalization control probe sets as the HG-U133 series allowing direct comparison of the two platforms.

Both intra- and inter-platform comparisons have been carried out assuming that gene expression profiles obtained from frozen material using the HG-U133 Plus2.0 Gene Chip is the standard against which any combination of tissue preservation technique or microarray platform can be benchmarked. The experiment was performed on 8 samples for which we had archived frozen and FFPE tissue in our biobank.

To compare gene expression profiles between frozen and FFPE preserved tissues, we computed the Pearson correlation coefficient for each probe set followed by a t-test to determine whether the observed correlation was due to chance. We consider as significant a correlation for which the nominal p-value is lower than 0.05 or the coefficient is higher than 0.51. In the case of inter-platform comparison, we further matched both platforms using the best match between HG-U133 Plus2.0 and HG-U219 provided by Affymetrix. As shown in Figure A, it is clear that the HG-U219 Array Plate gives the most consistent results (green and yellow lines for frozen and FFPE respectively). In
addition, the intensities after normalization of the HG-U219 Array Plate are very similar even when using different preservation techniques.

Furthermore, on computing the average of the intensity values for all frozen or FFPE preserved samples, the HG-U219 Array Plate gave the most consistent results with a Pearson correlation coefficient of 0.851 compared to 0.518 for the HG-U133 Plus2.0 platform. For some clinically useful breast cancer signature scores, the correlation among expression values has been shown to be optimal considering any combination of microarray platforms and preservation techniques. This apply for instance to the expression of ESR1 and HER2 where it was possible to separate samples with high and low gene expression values irrespectively of the platform (U133 Plus 2 or U219) and the starting material used (frozen or FFPE). Of note, these values were concordant with the results obtained by immunohistochemistry.

For example (Figure B), samples with high ESR1 expression (in red) were in fact the ones classified as immunohistochemistry ER-positive; and vice versa, samples marked with low ESR1 expression (in green) were actually ER-negative by immunohistochemistry. The same was true
for HER2 gene and other previously reported gene expression signatures. These results were very important as it enabled us to identify a platform that can provide reliable gene expression data from FFPE. Given the higher correlation between expression values of frozen tissue obtained by the HG-U219 platform and the consistency in the result, we decided to proceed with using this platform to profile the samples of this study.

Transcriptomic analysis using Affymetrix and defining molecular subtypes

All samples were hybridized using Affymetrix® Human Genome U219 array plates in three batches following standard Affymetrix protocols as previously defined. The samples were normalized using expresso method with robust multi-array analysis (RMA) background correction, quantile normalization and median polish summarization. To adjust for potential batch effect, an R implementation of ComBat method available from the bioconductor [www.bioconductor.org] package. Surrogate Variable Analysis (sva) was used. The batch effect corrected expression data was used in all further analysis. To keep only one expression value per gene, we computed the average intensity of each probe set corresponding to the same Ensembl gene identifier.

We used two methods to determine breast cancer molecular subtype; PAM50 and the 3-Gene Model. The PAM50 centroids for the five subtypes, luminal-A, luminal-B, HER2 amplified, basal and normal-like, were retrieved from the R package genefu. After median centering the genes in the current data set, the nearest PAM50 centroid for each sample was determined as the centroid, which gives the maximum Spearman's rank correlation coefficient with the corresponding sample centroid. The subtype of this nearest PAM50 centroid was assigned as the subtype of the sample. The 3-Gene Model was trained using 353 breast cancer microarray data
from EXPO data set [GEO accession: GSE2109]. This trained model was then used to classify each sample from the current data set into four molecular subtypes namely basal, HER2 amplified, luminal-A and luminal-B as previously defined.

References