High–throughput, quantitative multiplex gene expression assay accelerates the validation of breast cancer biomarkers

Kathryn Sciabica[1], Yong Wu[1], Mervin R. Gutierrez[1], Anna Ivshina[2], Cynthia L. Boardman[1],
Scott Mack[1] and Lance D. Miller[2]
[1]Beckman Coulter, Inc., Fullerton, CA USA; [2]Genome Institute of Singapore, Singapore

INTRODUCTION

Gene expression signatures of tumor biomarkers, generally involving 10 or more genes, are often used to elucidate the underlying molecular mechanisms of tumor progression. Many microarray studies have demonstrated that changes in gene expression associated with tumor progression are predictive of tumor progression and can therefore be used to direct treatment strategy and improve patient outcome. Although high-density microarray analysis is a powerful tool for discovering biomarker candidates associated with a particular disease, it is not practical for detailed biomarker validation with increased sample throughput. In this study, we generated a multiplex assay that targets 20 genes selected from a prior breast cancer microarray study. This multiplex contains primers for 11 genes associated with tumorigenesis, apoptosis, cell-cycle regulation and cell proliferation along with 9 reference genes. Gene expression profiles of 50 breast tumors of varying histological grades (G1, G2 and G3) were characterized with this 20-gene multiplex. Distinctive patterns of gene expression were identified and these patterns correlated with tumor histological grade. The multiplex assay results were highly concordant to a larger gene probe set previously established to differentiate G1 and G3 tumors by microarray analysis. This new approach not only offers savings in sample, time and expense with increased throughput; but also provides superior quantitation linearity with R^2 values greater than 0.99. This multiplex, quantitative assay effectively transfers biomarker discovery from a large-scale microarray platform into a more sensitive, quantitative and high-throughput validation method; which better fits a routine testing environment.

The eleven genes of interest in this assay include five genes, which Ivshina, *et al.* (2006) previously demonstrated to have the capability to classify between low (G1) and high-grade (G3) breast tumors using gene expression profiles. These five genes were selected from 264 genes identified as potential tumor biomarkers. Histological grade G2 tumors were shown to be hybrids that demonstrate either G1-like or G3-like properties of molecular signature and survival.[2] Using gene expression data from microarray technology, this small subset of five genes was able to separate histological grade G2 tumors into two subtypes of G2a and G2b (Figure 1A) when analyzed by pattern recognition algorithms.[1] Additional data analysis illustrated that this classification system based on gene expression profiling is an excellent prognosticator of disease recurrence (Figure 1B). Our assay demonstrates the ability to further subtype histological grade G2 tumors based on the gene expression signature from a small subset of genes. Furthermore, the study identified distinct patterns of gene expression that correlate with tumor progression using a novel quantitative multiplexing methodology.

§ The PCR process is covered by patents owned by Roche Molecular Systems, Inc. and F. Hoffman La Roche, Ltd.

METHODOLOGY

Tumor Samples. Samples were obtained from the tissue repositories of the Genome Institute of Singapore. A total of 50 breast carcinomas samples of varying histological grades 1, 2 and 3 were analyzed in this study. For comparison, three normal breast tissues were also examined.

RNA Amplification. Following RNA extraction of the tumor biopsies, each sample was subjected to anti-sense RNA (aRNA) synthesis through *in vitro* transcription. The aRNA was then quantified for concentration.

Multiplex Primer Design. Primers for the 20-gene multiplex assay were designed by importing the target gene ID or sequence into the designer module of the GenomeLab™ eXpress Profiler software.

Multiplex Universal Priming Strategy and RT-PCR. cDNA was synthesized from aRNA by using the GenomeLab™ GeXP Start Kit from Beckman Coulter. A total 5 ng of aRNA was used per RT reaction and performed in four technical replicates. Each target anti-sense mRNA is detected by a gene-specific sequence in a chimeric forward primer in the RT reaction. PCR amplification is predominantly carried out by universal forward and reverse primers. All gene targets in the multiplex panel were amplified by universal primers. The forward universal primer is fluorescent dye labeled, enabling subsequent fluorescence detection of amplicons.

Separation by Capillary Electrophoresis (CE). PCR product separation, detection and quantitation was performed on the GenomeLab™ GeXP Genetic Analysis System by capillary electrophoresis.

Fragment Analysis and Gene Expression Signature Analysis. After amplified fragments were separated, the data were initially analyzed using the Fragment Analysis module of the GeXP system software (Figure 2). The data were imported into the analysis module of eXpress Profiler software and normalized against a reference gene. Prior to normalization, peak areas are calculated for each particular fragment. Data for each fragment and technical replicate were averaged and %CV calculated. The results were subjected to further analyses in Microsoft* Excel.

Class Prediction/Pattern Recognition Algorithm Analysis. Normalized Gene Expression Ratios were exported into the data-mining Microsoft Excel macro program, XLMiner* for Windows (Resampling Stats, Inc.), for analysis. Each tumor sample was binary coded according to the existing four class scheme of tumor grades described by Ivshina, *et al.* of G1, G2a, G2b, and G3 – a four class prediction scheme. Normalized gene expression ratios for all 11 genes of interest were subjected to pattern recognition by the Fisher discriminate analysis and neural network algorithms. The data from this 20-gene multiplex assay were compared to previous data derived from microarray studies.

*All trademarks are property of their respective owners.

BACKGROUND

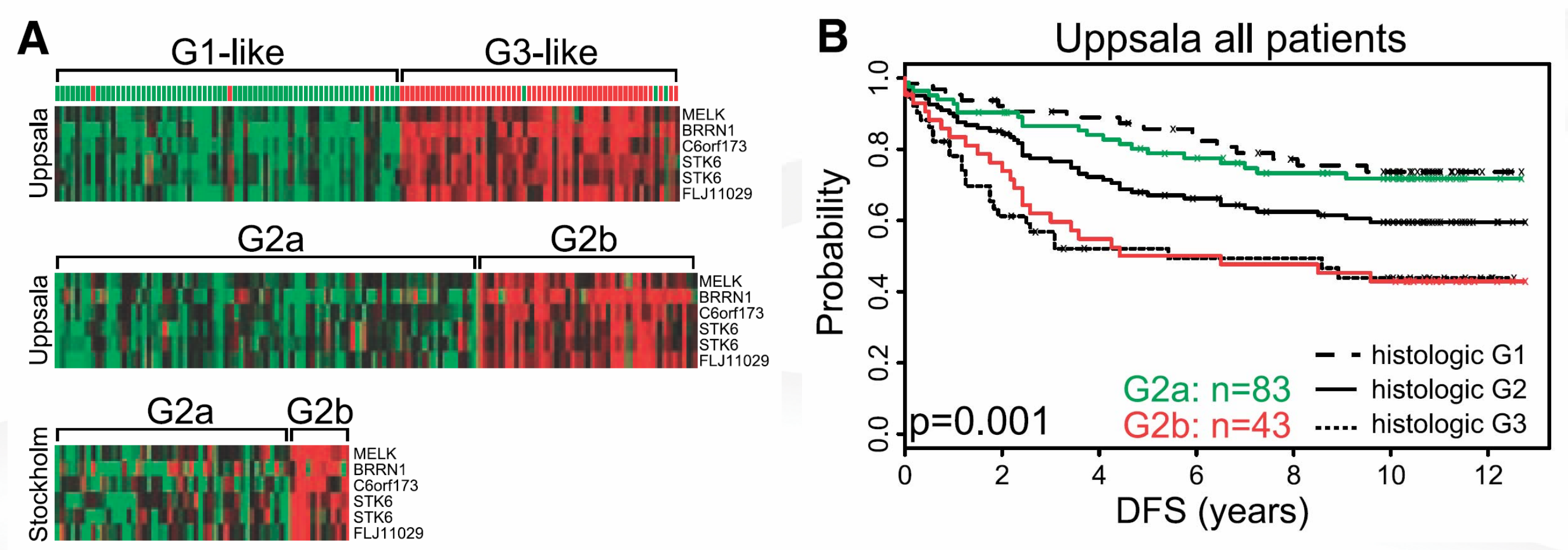

Figure 1. . Previous microarray studies narrowed down a 264-gene classifier into a small subset of five genes within six probesets that could be used to stratify tumor samples by molecular subtype referred to as "genetic grade".[1] Class prediction algorithms, prediction analysis of microarray (PAM) and statistically weighted syndromes (SWS), were utilized to classify low and high grade tumors (Figure 1A). Histological grade G2 tumors were further separated into two subtypes of G2a (G1-like) or G2b (G3-like) by gene expression profiling. This molecular reclassification into high- and low-grade tumor is supported by associated survival data (Figure 1B).

MULTIPLEX GENE LIST

Gene #	Gene Name	Function	PCR Product Size
G1	STK_6	Protein Serine/Threonine Kinase (Signal Transduction)	131
G2	orf173	orf associated with breast cancer	143
G3	PRR11	Hypothetical protein associated with breast cancer	157
G4	BRRN1	Cell Cycle Regulation	170
G5	C10orf56	orf associated with breast cancer	176
G6	CIRBP	Apoptosis	194
G7	FOS	Cell Proliferation	212
G8	PRC1	Cytokinesis	229
G9	FOSB	Cell Proliferation	244
G10	MELK	Protein Serine/Threonine Kinase (Signal Transduction)	255
G11	STK6	Protein Serine/Threonine Kinase (Signal Transduction)	295
R1	SF3A	Reference Gene	162
R2	PPIB	Reference Gene	187
R3	GUSB	Reference Gene	220
R4	PSMC	Reference Gene	260
R5	PUM1	Reference Gene	267
R6	RPLPO	Reference Gene	278
R7	RPS3	Reference Gene	286
R8	phMGFP	Reference Gene	301
R9	RPL19	Reference Gene	340

Table 1. The 20-gene multiplex assay consists of eleven genes of prognostic interest and nine reference "housekeeping" genes (shaded gray) for increased flexibility during normalization. A previously established small gene subset (type in **blue**) by Ivshina, *et al.* (2006) consisting of five genes and six primer pairs is contained within the 20-gene assay.

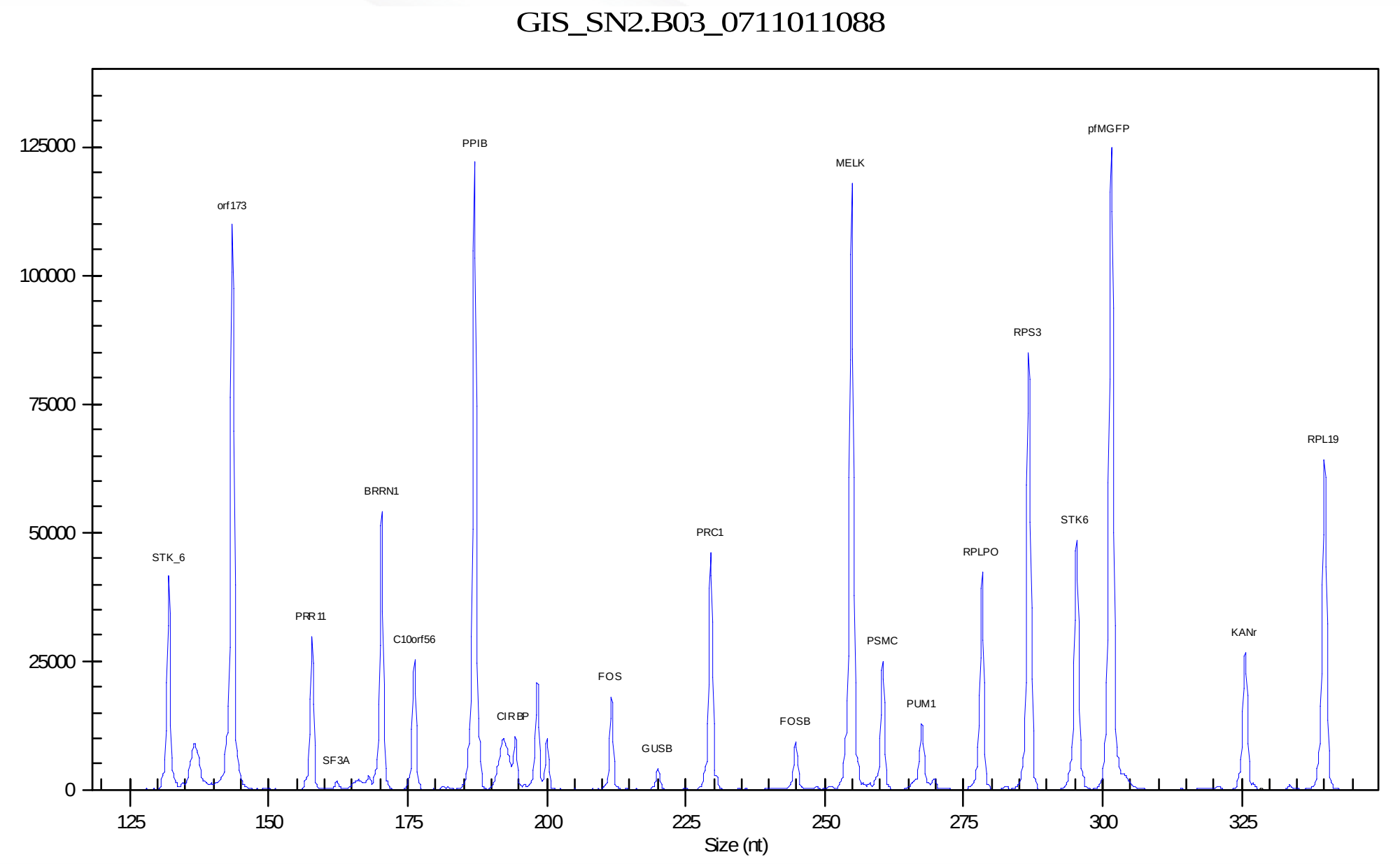

Figure 2. A representative electropherogram from the 20-gene multiplex assay shows detection of all 20 genes in a histological grade 3 breast tumor sample. The KAN^r peak at 325 nucleotides size serves as an internal control for the multiplex RT-PCR reaction.

DISCUSSION

Many genome-wide microarray studies aim at identifying smaller subsets of genes that can accurately classify tumors as efficiently as their larger gene set counterparts. We demonstrate the ability to classify tumor samples using a multiplex assay, based on a small number of genes, by subjecting data from these gene expression profiles to pattern recognition algorithms. The potential strength of this assay lies in its ability to use small amounts of RNA from tumor biopsies acquired during surgery and then quickly analyze the gene expression profiles of a small subset of highly informative genes to classify the patient sample. The time, cost and amount of RNA needed to employ current microarray-based expression profiling may be prohibitive for widespread adaptation in routine testing environments. The high reproducibility (Table 2) obtained with a smaller, relevant gene set demonstrate the potential of developing innovative assays using this platform that can be used to better understand how underlying molecular mechanisms of gene expression relate to tumor progression. Further studies are needed to clarify the value of these gene expression signatures combined with conventional histological analysis to potentially enhance patient prognosis. Larger population studies that correlate gene expression signatures to corresponding survival data will be the key to understanding the value of using these molecular tests as routine assays in the clinic.

SUMMARY OF KEY FINDINGS

- A 20-gene multiplex gene expression assay has the potential to be as informative as higher-density microarray assays in classifying breast cancer progression.
- The assay creates distinct patterns of gene expression that can be used to classify tumor samples with the use of class prediction and pattern recognition algorithms, such as discriminate analysis and neural networks.
- A multiplexed assay (up to 30 genes) allows for an increased sample throughput.
- Multiple internal reference genes available in this single assay offer great flexibility for selecting the best combination of reference genes for a specific study.
- Averaged total %CV for the entire multiplex assay is < 10%, which demonstrates superb reproducibility for multiplex gene expression analysis.

RESULTS

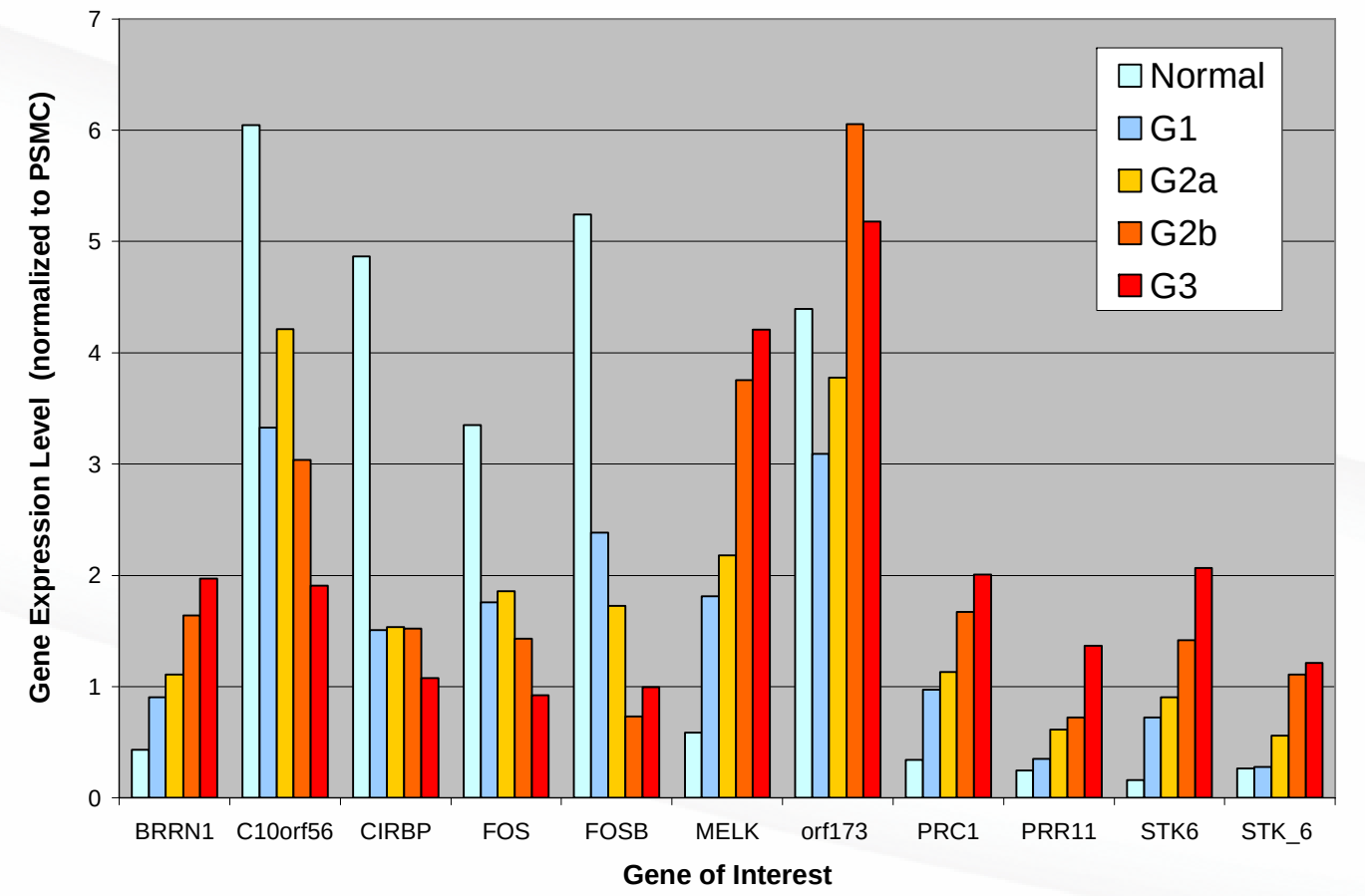

Figure 3. The average normalized gene expression value for 50 breast tumor and three normal breast samples. Genes involved in DNA replication and signal transduction such as BRRN1, MELK, PRC1, and STK6 genes are upregulated during the progression of breast cancer. In contrast, downregulation occurs in genes that control cell cycle and apoptotic function, such as FOS, FOSB, and CIRBP, in higher grade tumors.

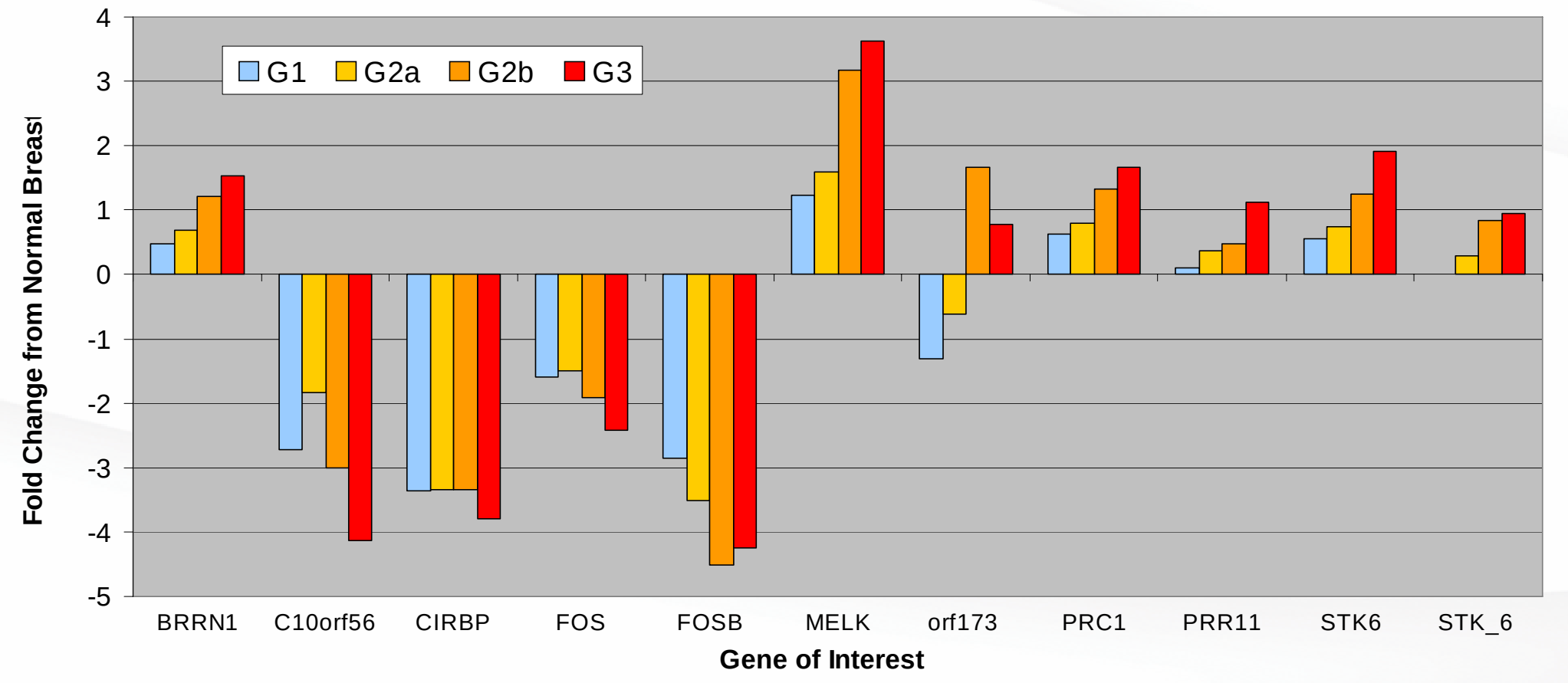

Figure 4. Fold change of gene expression between normal tissue and molecularly subtyped tumors of G1, G2a, G2b, and G3 stages in the 11 genes of interest. The graph illustrates increasing changes in gene expression from normal breast tissue during tumor progression. The average gene expression ratio for each tumor subtype was compared to the average gene expression ratio of the normal breast samples to calculate the fold change.

Comparison between Genetic Stages	p-value
G1-G2a	0.788819
G2a-G2b	0.000725
G2b-G3	0.07561

Table 3. Significance of fold change difference among molecular subtypes. A significant difference was observed in the fold change calculations between molecularly subtyped G2a and G2b samples. The two-tail p-values were calculated from the fold change (Figure 4) differences from all the genes of interest between molecular classifications. These molecularly subtyped G2a tumors have been previously described as G1-like and G2b as G3-like in gene expression profile and survival analysis.[1] No significance in the average gene expression difference between G1-G2a or G2b-G3 subtypes reflects the similarities of their respective gene expression profiles.

Gene Name	Sample Name NB2 aRNA	BT18 aRNA	BT19 aRNA	BT52 aRNA	BT120 aRNA	Average %CV per gene
BRRN1	11.31	13.59	4.63	2.46	2.08	6.81
C10orf56	12.23	8.86	11.20	11.95	1.08	9.06
CIRBP	12.85	8.34	4.63	6.74	4.85	7.48
FOS	8.28	8.03	13.53	13.69	12.88	11.28
FOSB	8.36	1.02	5.89	8.06	4.69	5.60
GUSB	4.46	9.16	8.89	3.45	3.40	5.87
MELK	5.66	3.37	4.99	4.72	6.24	4.99
orf173	5.61	9.07	19.45	3.90	10.28	9.66
phMGFP	7.60	12.75	4.60	6.35	5.84	7.43
PPIB	6.35	9.54	5.30	4.47	3.95	5.92
PRC1	9.49	4.11	7.81	6.65	6.48	6.91
PRR11	16.02	24.32	5.39	3.22	3.81	10.55
PSMC	NA	NA	NA	NA	NA	NA
PUM1	18.16	10.22	13.76	18.10	14.40	14.93
RPL19	10.74	13.72	1.46	5.97	6.47	7.67
RPLPO	2.90	14.01	2.89	4.54	2.76	5.42
RPS3	6.39	17.44	7.06	4.31	7.27	8.49
SF3A	21.11	28.16	21.88	15.18	54.68	28.20
STK_6	34.68	16.72	26.57	3.14	18.67	19.96
STK6	14.32	16.14	5.73	9.08	5.87	10.23
Average %CV per sample	11.40	12.03	9.25	7.16	9.25	9.81

Table 2. % CV is an Indicator of Overall Reproducibility. The % CV for genes in each sample (values in **black**) are calculated from four technical replicates initiated at cDNA synthesis and carried out through normalized gene expression ratio analysis. Peak areas representing expression of each gene were normalized to the reference gene, PSMC, to attain a normalized gene expression ratio. In this particular experiment, one normal breast tissue and four tumor tissue samples were used to calculate average % CV for the assay. An overall average % CV of less than 10% (yellow) demonstrates the superb reproducibility of this assay.

Discriminate Analysis—Classification Confusion Matrix				
	Predicted Class			
Actual Class	G1	G2a	G2b	G3
G1	6	3	0	0
G2a	1	9	0	2
G2b	0	1	6	2
G3	0	0	0	20

Error Report			
Class	# Cases	# Errors	% Error
G1	9	3	33.33
G2a	12	3	25.00
G2b	9	3	33.33
G3	20	0	0.00
Overall	50	9	18.00

Table 4. Discriminate analysis (DA) among four molecular classifications of tumor. Inherent differences of gene expression signatures in the 20-gene multiplex assay can classify tumors into the four subtypes. Actual class notations were previously assigned using microarray analysis and statistically weighted syndromes (SWS) class prediction algorithm.[1] A class error rate of 18% represents the discrepancy between the 20-gene multiplex assay and the previous study.

Neural Networks—Classification Confusion Matrix				
	Predicted Class			
Actual Class	G1	G2a	G2b	G3
G1	4	5	0	0
G2a	0	12	0	0
G2b	0	0	9	0
G3	0	0	0	20

Error Report			
Class	# Cases	# Errors	% Error
G1	9	5	55.56
G2a	12	0	0.00
G2b	9	0	0.00
G3	20	0	0.00
Overall	50	5	10.00

Table 5. Neural networks (NN) analysis among four molecular classifications of tumor. There is a 10% error rate when compared to previous microarray studies in a four-class prediction schema using neural network analysis, a learning pattern recognition algorithm. All mismatches resulted from G1 tumors being classified as the closely-related classified G2a tumors.

REFERENCES

1. Ivshina, *et al.*. Genetic Reclassification of Histologic Grade Delineates New Clinical Subtypes of Breast Cancer. Cancer Res 2006; 66:10292-301.
2. Kuznetsov, *et al.*. Statistically Weighted Voting Analysis of Microarrays for Molecular Pattern Selection and Discovery Cancer Genotypes. IJCSNS 2006; 6:73-83.
3. Ma, *et al.*. Gene expression profiles of human breast cancer progression. PNAS 2003; 100:5974-9.
4. Perou, *et al.*. Molecular portraits of human breast tumours. Nature 2000; 406:747-52.

GeXP QUANTITATIVE ANALYSIS METHOD

The GeXP Genetic Analysis System from Beckman Coulter offers multiplexed, quantitative gene expression analysis capable of examining up to 30 genes in a single reaction from as little as 5 ng total RNA. Here, we demonstrate that the GeXP system produces gene expression data with superb linearity and is sensitive enough to precisely detect even small changes in gene expression. A superior linear correlation between the amount of RNA and gene expression quantitation value was generated for each gene in a multiplex with an average correlation coefficient (R^2) well above 0.99. In addition, we verified that ten consecutive 0.5-fold increases in RNA concentration were accurately quantified by the GeXP system for all 24 genes in a multiplexed assay. The capacity of the GeXP system to deliver multiplexed, sensitive and precise gene expression analysis opens a new door for scientists to explore subtle, yet biologically meaningful changes in an extremely effective and efficient manner.

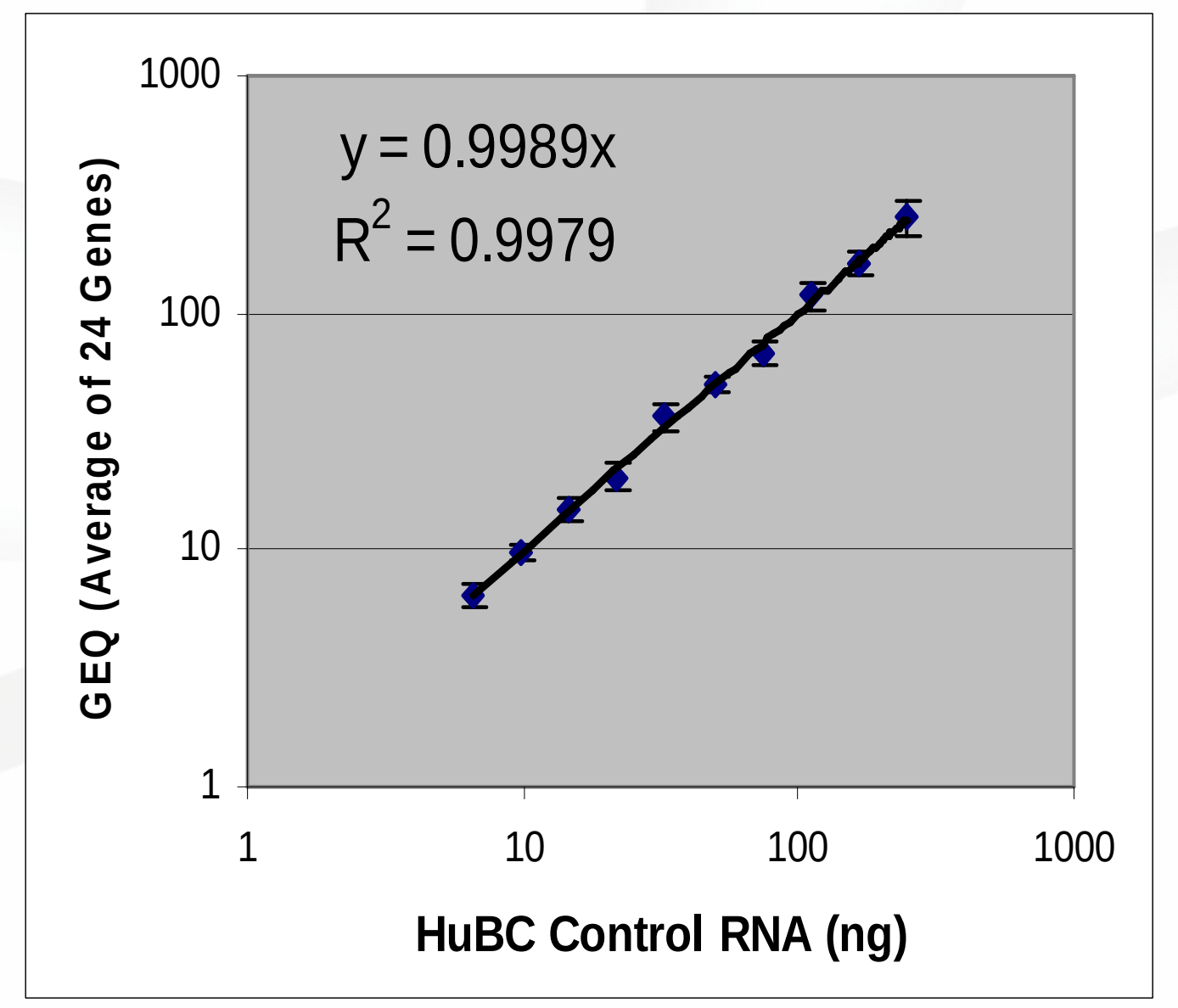

Figure 4. The linear correlation between the amount of HuBC control RNA and the gene expression quantitation (GEQ) value on detecting 0.5-fold change of target mRNAs. Error bars represent one standard deviation from the mean for four technical replicates. The correlation coefficient (R^2) is shown on the chart.

HuBC control RNA (ng)	6.5	9.8	14.6	21.9	32.9	49.4	74	111	167	250
GEQ value	6.5	9.8	15.0	20.4	36.2	49.8	67.8	118.3	162.7	250.6
RA (%)	99.8	99.9	97.2	93.3	89.9	99.3	91.6	93.4	97.4	99.8
Average RA (%)	96.2									

Table 3. Relative Accuracy (RA) of the GeXP system in detecting 0.5-fold change in amount of HuBC control RNA. The relative accuracy (RA) of a GEQ value is defined as: RA = [(RNA amount - Absolute(RNA amount – GEQ value)) / RNA amount] x 100%.

Gene Symbol	Accession#	R2
PPIA	BC000689	0.9984
GAPDH	NM_002046	0.9992
IGFBP5	L27560	0.9992
CP	NM_000096	0.9994
ESR1	NM_000125	0.9996
EXT1	NM_000127	0.9942
KRT18	NM_000224	0.9965
OXCT1	NM_000436	0.9985
DCK	NM_000788	0.9927
ACTB	NM_001101	0.9964
AP2B1	NM_001282	0.9945
GNAZ	NM_002073	0.9869
MYBL2	NM_002466	0.999
RFC4	NM_002916	0.9989
TGFB3	NM_003239	0.9993
CDC42BPA	NM_003607	0.9862
ALDH4A1	NM_003748	0.9977
WISP1	NM_003882	0.9811
PRC1	NM_003981	0.9963
KNTC2	NM_006101	0.9926
RAB6B	NM_016577	0.9913
EGLN1	NM_022051	0.9988
bbc3	U82987	0.9976
HSC4A2 IV2	X05610	0.9976
Mean		0.9955
Median		0.9976

Table 4. The correlation coefficient (R^2) for each gene in Human Breast Cancer*Plex* for all 24 genes, the average and median values for the R^2 were 0.9955 and 0.9976, respectively. These results demonstrate that the GeXP system delivers highly sensitive and precise quantitative gene expression analysis.

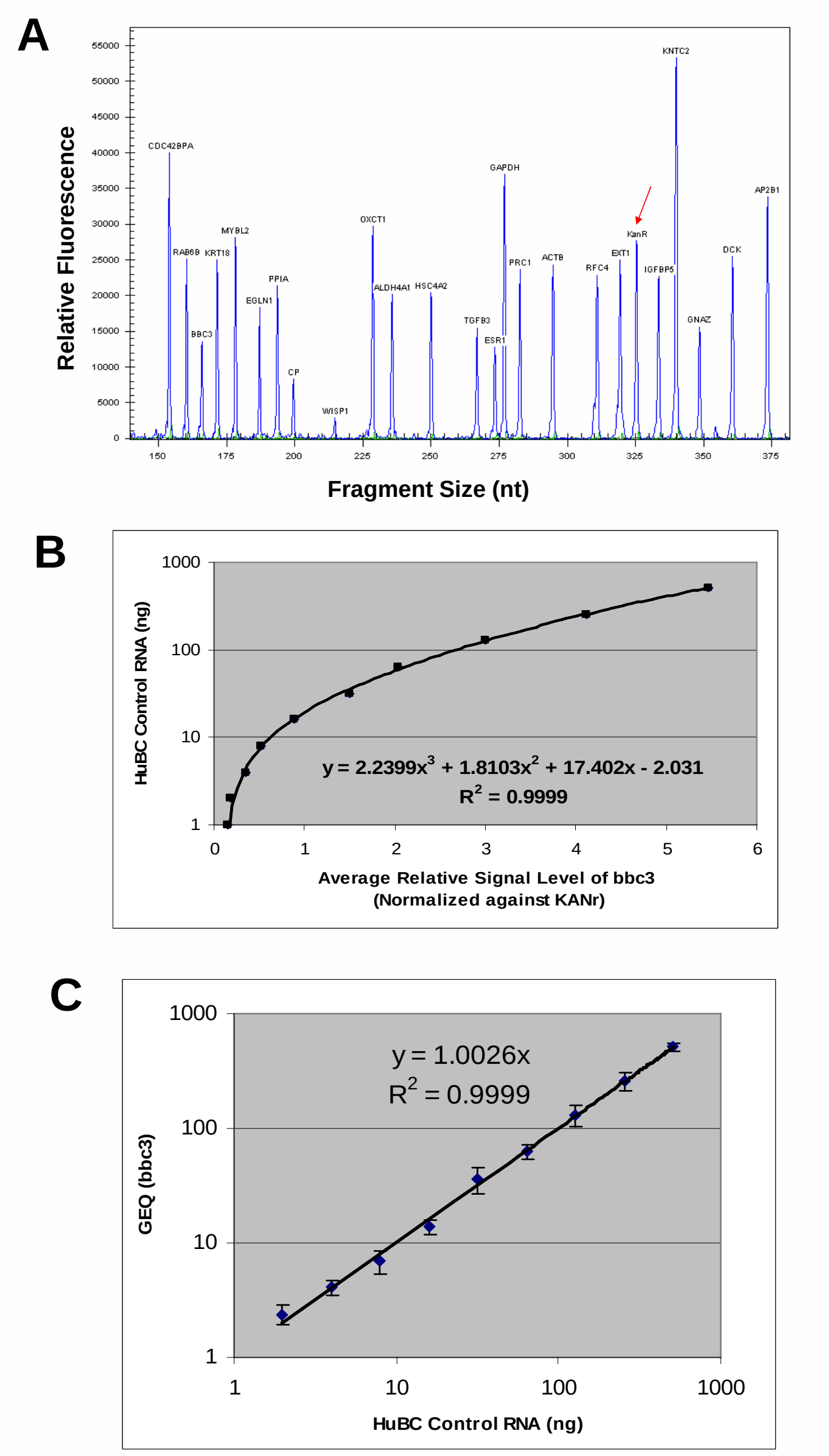

Figure 5. The quantitative data analysis procedure for the GeXP system. A) A reaction internal control, KAN^r (red arrow), is used to determine the relative signal level of each gene in the multiplex reaction. B) The correlation between the average of relative signal level for six technical replicates of *bbc3* in the Human Breast Cancer*Plex* and the amount of HuBC control RNA was fit to a third-order polynomial model. The fitting equation and R^2 are shown on the chart. The GEQ value (y) is calculated using the equation from this standard curve with relative signal level (x) for a particular RNA concentration. C) The relationship between the amount of HuBC control RNA and the GEQ value of *bbc3* was plotted. A linear regression model was applied to data points from 2 ng to 512 ng total input RNA. Error bars represent standard deviation of six technical replicates. The correlation coefficient (R^2) is displayed on the chart.

Note: The GenomeLab GeXP Genetic Analysis System from Beckman Coulter is intended For Research Use Only; not for use in diagnostic procedures.