

GENE EXPRESSION PREDICTORS OF BREAST CANCER OUTCOMES

Erich Huang², Skye H Cheng¹, Holly Dressman², Jennifer Pittman⁵, Mei-Hua Tsou¹,
Cheng-Fang Horng¹, Andrea Bild², Edwin S Iversen^{4,5}, Ming Liao⁵, Chii-Ming Chen¹,
Mike West⁵, Joseph R Nevins^{2,6} and Andrew T Huang^{1,3}

¹Koo Foundation Sun Yat-Sen Cancer Center, Taipei, Taiwan

²Department of Molecular Genetics and Microbiology, Duke University Medical Center

³Department of Medicine, Duke University Medical Center

⁴Department of Biostatistics and Bioinformatics, Duke University Medical Center

⁵Institute of Statistics and Decision Sciences, Duke University

⁶Howard Hughes Medical Institute

SUMMARY

Background The integration of currently accepted risk factors with genomic data carries the promise of focusing the practice of medicine on the individual patient. Such integration requires interpreting the complex, multivariate patterns in gene expression data, and evaluating their capacity to improve clinical predictions. We do this here, in a study of predicting nodal metastatic states and relapse for breast cancer patients.

Methods DNA microarray data from samples of primary breast tumors were analyzed using non-linear statistical analyses to evaluate multiple patterns of interactions of groups of genes that have predictive value, at the individual patient level, with respect to lymph node metastasis and cancer recurrence.

Findings We identify aggregate patterns of gene expression (metagenes) that associate with lymph node status and recurrence, and that are capable of honestly predicting outcomes in individual patients with about 90% accuracy. The identified metagenes define distinct groups of genes, suggesting different biological processes underlying these two characteristics of breast cancer. Initial external validation comes from similarly accurate predictions of nodal status of a small sample in a quite distinct population group.

Interpretation Multiple aggregate measures of gene expression profiles define valuable predictive associations with lymph node metastasis and disease recurrence for the individual patient. These results indicate the potential for gene expression data to aid in achieving more accurate individualized prognosis. Importantly, this is evaluated in terms of precise numerical predictions, via ranges of probabilities of outcome, for the individual patient. Such precise and statistically valid assessments of patient-specific risk will ultimately be of most value to clinical practitioners faced with treatment decisions.

INTRODUCTION

Calibrating therapeutic intervention to an individual's prognosis is central to effective oncologic treatment. In breast cancer, invasion into axillary lymph nodes is the most significant prognostic factor in breast cancer (1;2). Dissection of axillary nodes is consequently a crucial component of the therapeutic decision-making process. Newer, less invasive modalities for assessing lymph node status, such as sentinel node biopsy, are gaining acceptance (1), but it remains clear that clinico-pathologic parameters such as the presence or absence of positive axillary nodes represent the best means available to classify patients into broad subgroups by recurrence and survival (3-5). Even so it remains an imperfect tool. Among patients with no detectable lymph node involvement, a population thought to be in a low-risk category, between 22 and 33% develop recurrent disease after a 10-year follow-up (6). Properly identifying individuals out of this group who are at risk for recurrence is beyond current capabilities.

The question of lymph node diagnosis is part of the broader issue of more accurately predicting breast cancer disease course and recurrence. Though current clinical predictors are useful, they are just not accurate enough for prediction at the individual patient level. Genomic measures of gene expression, using microarrays and other technologies, provide new information that is now understood to identify patterns of gene activity that sub-classify tumors (7-10). Such patterns may correlate with the biological and clinical properties of the tumors, so it is of interest to investigate whether, and how, such data might add predictive value to current clinical predictors. Credible predictive evaluation is critical in establishing reproducible results and a key step towards integrating complex genomic data into prognoses for the individual patient (11-14).

The studies we report here move towards this goal in studying gene expression patterns predictively related to lymph node involvement and breast cancer recurrence in defined patient subgroups. We focus on predictions for the individual patient and aim to provide quantitative measures – in terms of probabilities of clinical phenotype and disease outcome -- that summarize the genomic information relevant to predicting at the individual level.

METHODS

MIAME (minimal information about a microarray experiment)-compliant information regarding the analyses performed here, as defined in the guidelines established by MGED (www.mged.org), is detailed in the following sections.

Experimental design. The analysis involved the use of a total of 89 tumor samples (details described below) for comparative gene expression measurements. The goal of the analysis was to identify those gene expression patterns characteristic of particular sets of tumor samples within the group based on the statistical analysis methods described below. These samples represent a heterogeneous population, and were selected based on clinical parameters and outcomes with the view to generating cases suitable for two focused studies, as reported here. Details of clinical characteristics of the 89 patients are provided in Table 1. For the lymph node study, external validation involved predicting outcomes on a subset of tumors from our previous Duke breast cancer study; full clinical and protocol details of this study are as previously reported (11). Each sample was hybridized once.

Samples used, extract preparation, and labeling. The 89 samples are from primary tumor biopsies at the Koo Foundation Sun Yat-Sen Cancer Center (KF-SYSCC) in Taipei, collected and banked between 1991-2001. Samples were collected under Duke (IRB# 3157-01) and KF-SYSCC (9/21/01, see Supplementary Material) Institutional Review Board guidelines. Total RNA was extracted from tumor tissue with Qiagen RNEasy kits, and assessed for quality with an Agilent Lab-on-a-Chip 2100 Bioanalyzer. Hybridization targets (probes for hybridization) were prepared from total RNA according to standard Affymetrix protocols.

Hybridization procedures and parameters. The amount of starting total RNA for each reaction was 20 µmcg. Briefly, first strand cDNA synthesis was generated using a T7-linked oligo-dT primer, followed by second strand synthesis. An in vitro transcription reaction was performed to generate the cRNA containing biotinylated UTP and CTP, which was subsequently chemically fragmented at 95°C for 35 min. The fragmented, biotinylated cRNA was hybridized in MES buffer (2-[N-morpholino]ethansulfonic acid) containing 0.5 mg/ml acetylated bovine serum albumin to Affymetrix GeneChip Human U95Av2 arrays at 45°C for 16hr, according to the Affymetrix protocol (www.affymetrix.com and www.affymetrix.com/products/arrays/specific/hgu95.affx). The arrays contain over 12,000 genes and ESTs. Arrays were washed and stained with streptavidin-phycoerythrin (SAPE, Molecular Probes). Signal amplification was performed using a biotinylated anti-streptavidin antibody (Vector Laboratories, Burlingame, CA) at 3 µmcg/ml. This was followed by a second staining with SAPE. Normal goat IgG (2 mg/ml) was used as a blocking agent.

Measurement data and specifications. Scans were performed with an Affymetrix GeneChip scanner and the expression value for each gene was calculated using the Affymetrix Microarray Analysis Suite (v5.0), computing the expression intensities in ‘signal’ units defined by software. Scaling factors were determined for each hybridization based on an arbitrary target intensity of 500. Scans were rejected if the scaling factor exceeded a factor of 25, resulting in only one reject. Files containing the computed single intensity value for each probe cell on the arrays (CEL files), files containing experimental and sample information (control info files), and files providing the signal intensity values for each probe set, as derived from the Affymetrix Microarray Analysis Suite (v5.0) software (pivot files), can be found in the Supplementary Material on the project web site.

Array design. All assays employed the Affymetrix Human U95Av2 GeneChip. The characteristics of the array are detailed on the Affymetrix web site (www.affymetrix.com/products/arrays/specific/hgu95.affx).

Statistical analysis. Analysis uses predictive statistical tree models (15). This begins by applying k-means correlation-based clustering following an initial screen to remove genes varying at low levels, targeting a large number of clusters that are then used to generate a corresponding number of *metagene* patterns. Each metagene is the dominant singular factor (principal component) within a cluster, evaluated using the singular value decomposition (SVD). We identify 496 such factors this way, each representing the key common pattern of expression of the genes in the corresponding cluster. This strategy extracts multiple such patterns while reducing dimension and smoothing out gene-specific noise through the aggregation within clusters. Formal predictive analysis then

uses these metagenes in a Bayesian classification tree analysis. This generates multiple recursive partitions of the sample into subgroups (the “leaves” of the classification tree), and associates Bayesian predictive probabilities of outcomes with each subgroup. The analysis is applicable to even very small samples, and is developed to generate parsimonious models that are automatically resistant to over-fitting, as detailed previously (15). Overall predictions for an individual sample are then generated by averaging predictions, with appropriate weights, across many such tree models. We perform iterative out-of-sample, cross-validation predictions: leaving each tumor out of the data set one at a time, refitting the model (both the metagene factors and the partitions used) from the remaining tumors, and then predicting the hold-out case. This rigorously tests the predictive value of a model and mirrors the real-world prognostic context where prediction of new cases as they arise is the major goal.

Supplementary information

Additional information, including full details of all metagenes and complete details of the statistical tree methodology, is available at the project web site: <http://cgt.duke.edu/>.

Role of funding source

Partial support came from Koo Foundation Sun Yat-Sen Cancer Center (KFSYSCC) Research Fund. Several coauthors are personnel at KFSYSCC, and were involved in clinical data collection and design, and the writing and submission of this report.

RESULTS

Gene expression patterns in primary breast tumors that predict lymph node metastasis

The first study compares traditional “low-risk” versus “high-risk” patients, primarily based on lymph node status in order to evaluate the predictive associations of gene expression patterns with aggressive versus more benign tumors. Among ER positive individuals, the “high-risk” clinical profile is represented by advanced lymph node metastases (10 or more positive nodes); the “low-risk” profile identifies node-negative women of age greater than 40 years with tumor size below 2cm, precisely as currently used in clinical prognostic practice (15). Our data provides expression profiles on 18 high-risk and 19 low-risk cases (37 of the 89 total in Table 1) to which we applied the Bayesian statistical tree analysis. Figure 1 displays summary predictions from the resulting total of 37 cross-validation analyses. For each individual tumor, this graph illustrates the predicted probability for “high-risk” versus “low-risk” (red versus blue) together with an approximate 90% confidence interval, based on analysis of the 36 remaining tumors performed successively 37 times as each tumor prediction is made. It is important to recognize that each sample in the data set, when assayed in this manner, constitutes a validation set that accurately assesses the robustness of the predictive model. The metagene model accurately predicts nodal metastatic potential; about 90% (with 95% CI 79-99%) of cases are accurately predicted based on a simple threshold at 0.5 on the estimated probability in each case. Case number 7 is in the intermediate zone, exhibiting patterns of expression of the selected metagenes that relate equally well to those of “high” and “low-risk” cases, while case 22 is a clinical “high-risk” case with genomic expression patterns that relate more closely to “low-risk” cases. In contrast, node negative patients 5 and 11 have gene expression patterns more strongly indicative of “high-risk”,

and are key cases for follow-up investigations. The details of clinical information in these apparently discordant cases are shown in Table 2.

Clinical features of these few cases are illuminating, and suggestive of how a broader investigation of clinical data combined with molecular model-based predictions may aid in the eventual decision-making process. Case 22 did in fact recur, 6 years post-surgery; this patient's classification as high-risk for recurrence based on purely clinical parameters was moderated by a lower risk based on metagenes, as demonstrated by this patient having survived recurrence-free for a longer time. Thus the lower probability prediction assigned to patient 22 based on the gene expression profiles is reflected in the clinical behavior of her disease. The clinically "low-risk" patient 7 recurred at 31 months, and patient 11 at 38 months, whereas case 5 is currently disease-free after only 12 months of follow-up. Cases 7 and 11 thus partly corroborate the predictions based on genomic criteria. With such predictions as part of a prognostic model, more intensive or innovative post-surgical therapy would have been indicated for these two cases.

A critical aspect of the analyses described here is allowing the complexity of distinct gene expression patterns to enter the predictive model. Tumors are graphed against metagene levels for three of the highest scoring metagene factors (Figure 2). This analysis highlights the need to analyze multiple aspects of gene expression patterns. For example, if the low-risk cases 1, 3 and 11 are assessed against metagene 146 alone, their levels are more consistent with high-risk cases. However, when additional dimensions are considered, the picture changes. The second frame (upper right) shows that low-risk is consistent with low levels of metagene 130 *or* high levels of metagene 146; hence, cases 1 and 3 are not inconsistent in the overall pattern, though case 11 is consistent. An analysis that selects one set of genes, summarized here as one metagene, as a "predictor"

would be potentially misleading, as it ignores the broader picture of multiple interlocked genomic patterns that together characterize a state. In the predictions, these two metagenes play key roles: low levels of metagene 146 coupled with higher levels of metagene 130 are strongly predictive of high-risk cases. Metagene 330 also plays a role and it is the combined use of multiple metagenes, in the context of the tree selection model building process, that ultimately yields a pattern that has the capacity to accurately predict the clinical outcome.

External validation of lymph node metastasis predictors

To extend this analysis to an independent data set, we used a small but relevant subset of the patient samples studied in a previous Duke breast cancer analysis (11). This is a limited initial study, but most supportive of the basic conclusion of predictive value of multiple metagene patterns. Relative to the Asian cohort, the Duke study patients had rather different characteristics: the racial difference, and the facts that the US women were generally much older and had much larger tumors at surgery. Further, the numbers of extreme (>9) lymph nodes are very small, so we relaxed the criteria for the two risk groups (ignoring age, reducing the number of positive nodes for the high-risk group, and substantially increasing the maximum tumor size for the low-risk group) in order to generate meaningful numbers of cases for study. This led to 6 low-risk cases (lymph node negative, ER+, tumor sizes less than 3.5cm which is the median size of the whole group) and 7 high-risk cases (at least 4 positive nodes, rather than 10). Additional complications are due to the fact that the expression data for this older study were obtained on an earlier Affymetrix microarray, so represent different though overlapping genes; full details of the process of mapping to the metagenes defined by the current study are provided as

Supplementary Material. In spite of these complications, and the resulting expectation that predictive accuracy would be reduced, the predictions based on precisely the model fitted to the Asian data are very accurate: one of the low-risks cases appears more consistent, in terms of metagene expression, with the high-risk cases, whereas the remaining 12 cases are very accurately predicted to lie within their defined risk groups. Interestingly, the apparently discrepant low-risk case (#42) has the largest tumor (3.5cm) of the group. Figure 3 exhibits the three key metagenes, in a format similar to Figure 2 but now including also these external validation cases, where concordance with the Asian samples is clear.

Gene expression patterns that predict recurrence of disease in breast cancer

The second analysis concerns 3 year recurrence following primary surgery among the challenging and varied subset of patients with 1-3 positive lymph nodes. Such patients typically receive adjuvant chemotherapy alone, and uniformly across this risk group, so that it is of interest to explain variations in outcome within this subgroup based on predictors other than treatment regimen. This is a critical subgroup as more than 20% suffer relapse within five years (5). Hence, improved prognosis for this heterogeneous group is of critical importance; patients identified with a high probability of relapse could be targeted for more intensive treatment. Our dataset provides expression profiles on 52 cases in this lymph node category (34 non-recurrent, 18 recurrent). The aggregate predictions from the sets of generated statistical tree models defines a rather accurate picture; once again, there is an approximate 90% (with 95% CI 82-99%) overall predictive accuracy in the 52 separate one-at-a-time, cross-validation prediction assessments (Figure 4).

Based on the gene expression analysis, the 3 year non-recurrent cases 6 and 23, having profiles more akin to recurrent cases, would be candidates for intensive treatment. These patients did receive adjuvant chemotherapy based on additional clinical risk factors (especially tumor size). Thus traditional clinical risk factors other than lymph node status also indicate higher risk of recurrence for these two cases, consistent with the molecular predictions. Each actually survived recurrence-free for over three years; case 6 recurred at 42 months and case 23 remains disease-free after over 6 years. Cases with low genomic criteria for recurrence would be 36, 38 and 42. They, however, each recurred within three years. These are cases that, under prognosis informed by only the genomic model, would have been indicated as more benign and not candidates for intensive treatment, whereas such a treatment might have proven to be more beneficial. Evidently, there is much yet to learn about the combinations of integrated genomic and clinical characteristics that will improve our capacity to identify such critical cases.

Genes implicated in lymph node and recurrence studies

Subsets of genes related to the metagene predictors of lymph node involvement are replete with those involved in cellular immunity including a high proportion of genes that function in the interferon pathway. They include genes that are induced by interferon such as various chemokines and chemokine receptors (Rantes, CXCL10, CCR2), other interferon-induced genes (IFI30, IFI35, IFI27, IFI44, IFIT1, IFIT4, IFITM3), as well as interferon effectors (2'-5' oligoA synthetase), and genes encoding proteins mediating the induction of these genes in response to interferon (STAT1 and IRF1). This connection is intriguing given the role of interferon as a mediator of the anti-tumor response and, together with the fact that many genes involved in T cell function (TCRA, CD3D, IL2R,

MHC) are also included within the group that predict lymph node metastasis. Possibly, this may reflect the distinct nature of these tumors that have acquired a metastatic potential that elicits an anti-tumor response that is ultimately unsuccessful or an aberration of the normal anti-tumor response. Both of the key metagenes, 146 and 330, contain a number of these interferon related genes.

There is little intersection between the lists of genes defined by key metagenes here and those from the Duke lymph node study (11), which is perhaps not surprising given the relative heterogeneity of the patients in the Duke study. However, when the method of analysis used previously (11) is reapplied to the restricted subset of 6 low versus 7 high risk cases identified in the external validation study reported above, the 100 genes that most strongly relate to the categorization of lymph node status do indeed overlap with the top few metagenes of the current study. In particular, these include several genes already noted that are involved in an interferon response (STAT1, MX1, IFIT1, ISG115, IFI27, and IFI44).

Genes implicated in recurrence prediction do not exhibit such a striking functional clustering but do include many examples previously associated with breast cancer. Moreover, this group of genes is clearly distinct set from those that predict lymph node involvement. They include genes associated with cell proliferation control, both cell cycle specific activities (CDKN2D, Cyclin F, E2F4, DNA primase, DNA ligase), more general cell growth and signaling activities (MK2, JAK3, MAPK8IP, and EF1?), and a number of growth factor receptors and G-protein coupled receptors, some of which have been shown to facilitate breast tumor growth (EpoR). Possibly, the poor prognosis with respect to survival reflects a more vigorous proliferative capacity of the tumor.

We conclude that genes implicated in the prediction of lymph node metastasis and overall recurrence of disease, although clearly representing interrelated phenomena, nevertheless reflect the participation of distinct biological processes. The modeling approach we take here is flexible in this regard. The tree models select only those metagenes that are most relevant to the prediction in hand.

DISCUSSION

Personalized medicine aims to characterize those variables unique to the individual that determine disease susceptibility, response to therapy, and eventual disease outcome. We address this in assessing complex, multivariate patterns in gene expression data from primary tumor biopsies, and in exploring the value of such patterns in predicting lymph node metastasis and relapse. The resulting predictive accuracy of about 90%, and additional understanding of individual outcomes generated by the analysis, confirm the utility of gene expression patterns as prognostic factors in breast cancer. We stress the focus on predictions made in terms of numerical probabilities of outcomes for individual patients, with associated measures of uncertainties.

The lymph node risk group analysis defines metagene patterns capable of predicting high versus low risk cases with good accuracy, in both internal and external validation studies. In reanalysis of the small subset of samples from our early study (11) that relate most closely to the risk categories defined in this current study, we find improved predictions relative to our earlier methods and also a number of genes, including interferon-induced genes and others, in common. This provides additional support for the biological relevance of the metagene predictors identified, and suggests

areas for further pathway studies. The concordance between genomic predictors found between the Asian and US samples, though preliminary, is also a positive finding.

A related recurrence study (13) defines a single summary of gene expression related to breast cancer recurrence (though not nodal metastasis), generating a 70 gene predictor. We have been unable to identify more than 17 of these 70 genes on the Affymetrix array used here, and none of these appears in the key metagenes in our recurrence study. It will be of some interest to develop serious comparative studies that deal with cross-technology issues, and to develop future studies that combine and compare alternative summary predictors of outcome. The analysis approach used in (13) follows our own earlier work (11) in developing a single predictor based on an initial screen for genes most correlated with outcome. One distinction of our current work relative to these prior studies is the view that multiple measures of gene expression – multiple metagenes – may be involved in explaining differences and defining predictions. Investigation several metagenes, defining distinct patterns in the data relevant to the outcome, show how the combined effect of several views of clinico-biological data can highlight the similarities between patients while also identifying their differences. The non-linear statistical analysis aids in the elucidation of such patterns as they shed light on individual cases, as well as providing for informed predictions based on multiple patterns.

This latter point relates to the broader question of utilizing gene expression profiles into prognostic settings. We believe that it is the integration of genomic data with clinical risk factors that will determine the strategy for treating patients as individuals with distinct genomic disease features. Genomic data will not replace traditional clinical risk factors but will add significant detail to this clinical information, especially in a context such as breast cancer where multiple, interacting biological and environmental processes

define physiological states, and individual dimensions provide only partial information. As one initial example, our recurrence study here focuses on the 1-3 positive lymph node group where the analysis defines metagenes optimized for prediction within that group; predicting other subgroups, such as higher-risk cases in terms of lymph node count or subgroups stratified by additional clinical factors, will involve exploration of metagenes that optimally relate to outcomes within those subgroups.

Reliably improved predictions of disease course, including lymph node metastasis or recurrence, will profoundly affect the clinical decision process. Several studies indicate that 22-33% of node negative tumors behave in a manner similar to node positive tumors (6). Whether an issue of timing or of the inability to recognize histopathologic involvement of tumor material in the lymph nodes, a capacity to identify these cases as requiring more intensive clinical intervention could lead to an improvement in cancer survival. Previous attempts to correlate characteristics of primary tumors such as S-phase fraction, tumor grade, ploidy, *c-erbB-2* overexpression, and hormone receptor status with lymph node metastasis have proven unsuccessful (16-18). The ability to appropriately utilize gene expression profiles provides opportunity to add enormous additional detail to the few, currently used biological attributes in tumor characterization. Finally, genes implicated in these analyses generate information of value for future pathway studies, with the potential to identify new targets that may feed into improved therapeutic strategies as well as improved understanding of genes related to the biology of metastasis and tumor evolution.

Acknowledgements

Research was supported by Synpac (North Carolina) and the Koo Foundation Sun Yat-Sen Cancer Center Research Fund, and by NSF under grants NSF DMS-0102227 and NSF DMS-0112340. We appreciate the constructive comments of three anonymous reviewers of the original submission.

Conflict of interest statements

There are no conflicts of interest.

Reference List

1. Krag D, Weaver D, Ashikaga T, Moffat F, Klimberg VS, Shriver C et al. The sentinel node in breast cancer - a multicenter validation study. *N.Engl.J.Med.* 1998;339:941-6.
2. Singletary SE, Allred C, Ashley P, Bassett LW, Berry D, Bland KI et al. Revision of the American Joint Committee on Cancer Staging System for Breast Cancer. *J Clin Oncol* 2002;20:3628-36.
3. Overgaard M, Hansen PS, Overgaard J, Rose C, Andersson M, Bach F et al. The Danish Breast Cancer Cooperative Group 82b Trial: Postoperative Radiotherapy in High-Risk Premenopausal Women with Breast Cancer Who Receive Adjuvant Chemotherapy. *N.Engl.J.Med.* 1997;337:949-55.

4. Jatoi I, Hilsenbeck SG, Clark GM, Osborne CK. Significance of axillary lymph node metastasis in primary breast cancer. *J Clin Oncol* 1999;17:2334-40.
5. Cheng SH, Tsou MH, Liu MC, Jian JJ, Cheng JC, Leu SY et al. Unique features of breast cancer in Taiwan. *Breast Cancer Res Treat.* 2000;63:213-23.
6. Polychemotherapy for early breast cancer: an overview of the randomised trials. Early Breast Cancer Trialists' Collaborative Group. *Lancet* 2001;352:930-42.
7. Bhattacharjee A, Richards WG, Staunton J, Li C, Monti S, Vasa P et al. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc.Natl.Acad.Sci.USA* 2001;98:13790-5.
8. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* 2000;403:503-11.
9. Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA et al. Molecular portraits of human breast tumors. *Nature* 2000;406:747-52.
10. Yeoh E-J, Ross ME, Shurtleff SA, Williams WK, Patel D, Mahfouz R et al. Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell* 2002;1:133-43.
11. West M, Blanchette C, Dressman H, Huang E, Ishida S, Spang R et al. Predicting the clinical status of human breast cancer by using gene expression profiles. *Proc.Natl.Acad.Sci.,USA* 2001;98:11462-7.

12. Spang R, Zuzan H, West M, Nevins JR, Blanchette C, Marks J. Prediction and uncertainty in the analysis of gene expression profiles. *In Silico Biol.* 2002;2:0033.
13. van'T Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 2002;415:530-6.
14. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 1999;286:531-7.
15. Pittman J, Liao M, Huang E, Nevins JR, West M. Binary prediction tree modeling with many predictors. ISDS Discussion paper 2002;submitted for publication.
16. Mitra I, MacRae KD. A Meta-analysis of reported correlations between prognostic factors in breast cancer: does axillary lymph node metastasis represent biology or chronology? *Eur.J.Cancer* 1991;27:1574-83.
17. McGuire WL. Prognostic factors for recurrence and survival in human breast cancer. *Breast Cancer Res Treat.* 1987;10:5-9.
18. Tandon AK, Clark GM, Chamness GC, Ullrich A, McGuire WL. HER-2/neu oncogene protein and prognosis in breast cancer. *J.Clin.Oncol.* 1989;7:1120-8.

FIGURE LEGENDS

Figure 1. Cross-validation probability predictions of lymph node status. Samples (tumors) are plotted by index number, and the plotted numbers are marked on the vertical scale at the estimated predictive probabilities of high-risk (red) versus low-risk (blue). Approximate 90% uncertainty intervals about these estimated probabilities are indicated by vertical dashed lines.

Figure 2. Gene expression patterns from the major metagenes that predict lymph node status. Levels of metagenes for samples are plotted by sample index number and by color (color coding as in Figure 1).

Figure 3. Gene expression patterns from the major metagenes that predict lymph node status from current and earlier Duke breast cancer study. Levels of metagenes as in Figure 2, with current study samples now colored cyan (low-risk) and magenta (high-risk). External validation samples from the 2001 Duke breast cancer study appear as red (high-risk) and blue (low-risk).

Figure 4. Cross-validation probability predictions of 3-year recurrence. Samples (tumors) are plotted by index number, and the plotted numbers are marked on the vertical scale at the estimated predictive probabilities of 3 year recurrence (red) versus 3 year recurrence free survival (blue). Approximate 90% uncertainty intervals about these estimated probabilities are indicated by vertical dashed lines.

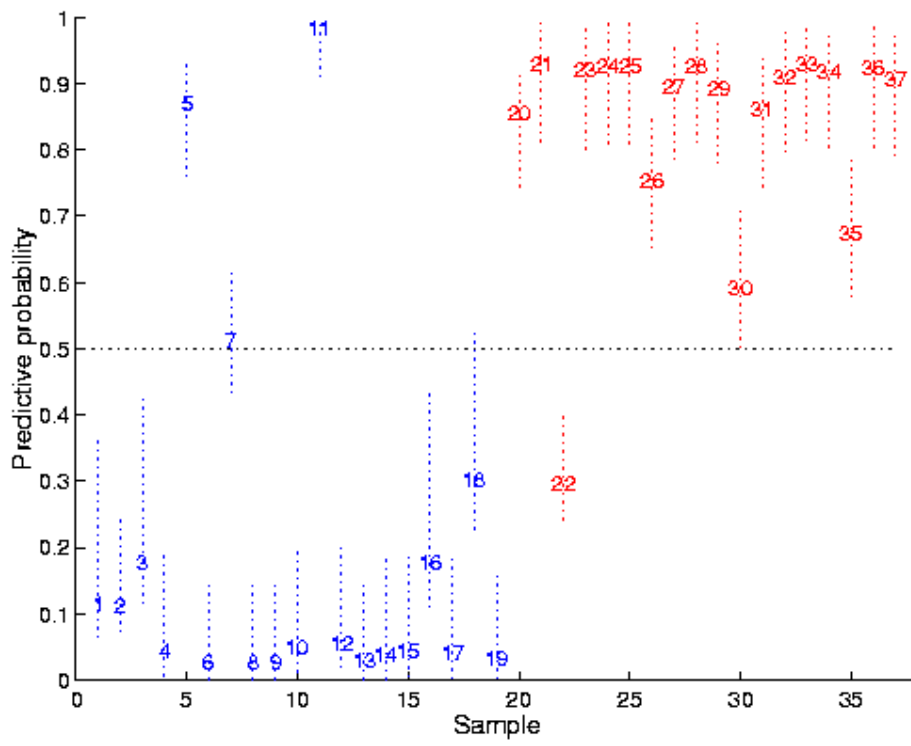


Figure 1

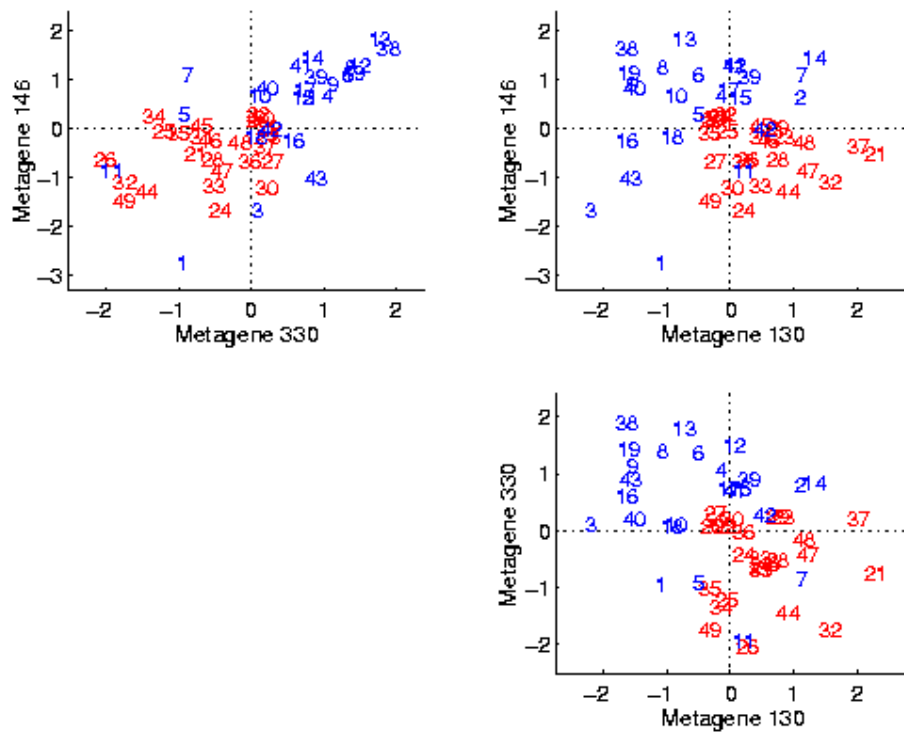


Figure 2

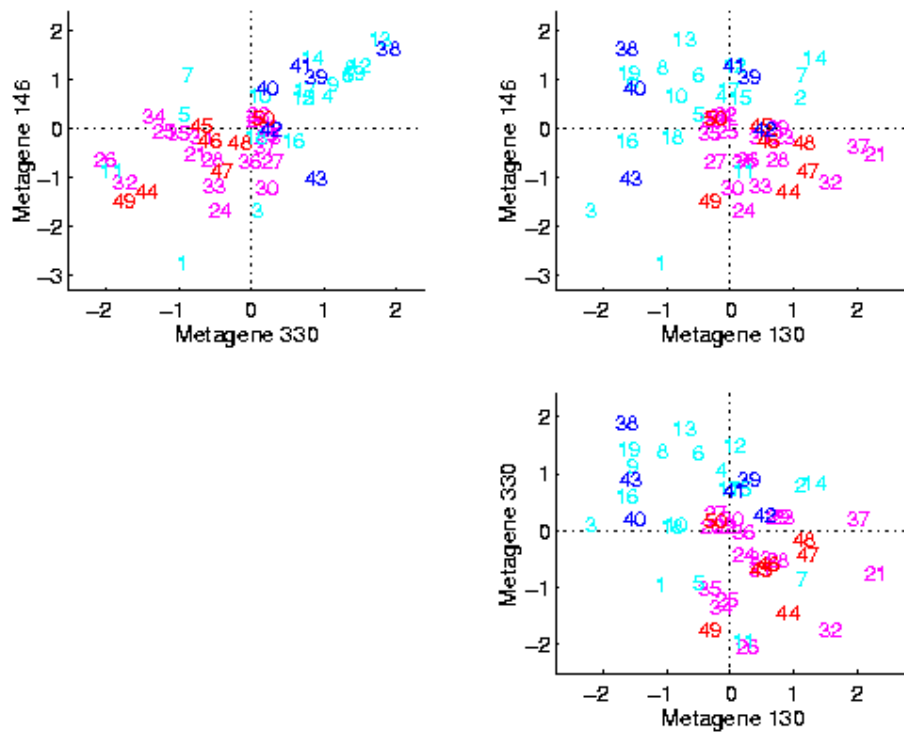


Figure 3

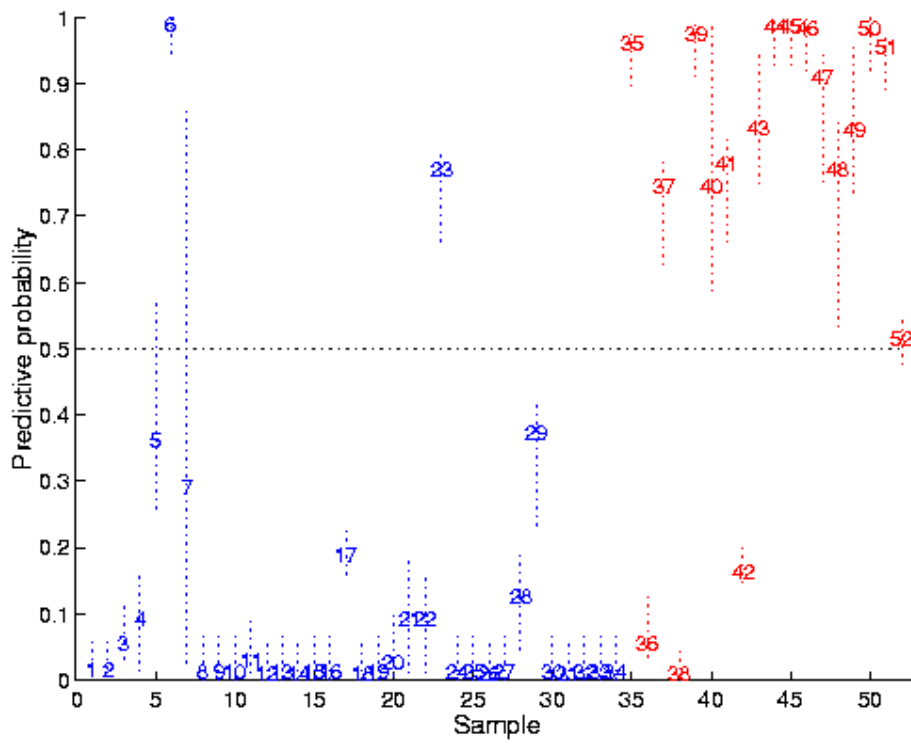


Figure 4

Table 1. Clinical characteristics of patients in the study

	Number	Percentage
Age		
< 40	27	30.3
41-50	26	29.2
51-60	19	21.4
> 60	17	19.1
Histology type		
Infiltrating Ductal Carcinoma	78	87.6
Infiltrating Lobular Carcinoma	2	2.3
Papillary Carcinoma	2	2.3
Tubular Carcinoma	1	1.1
Cribriform Carcinoma	1	1.1
Apocrine Carcinoma	1	1.1
Others (mixed of histologies)	4	4.5
Pathological tumor size	Number	Percentage
< 1 cm	6	6.8
1 – 2 cm	31	34.8
2 – 5 cm	47	52.8
> 5 cm	5	5.6
Lymph node positive		
0	19	21.4
1 – 3	52	58.4
4 – 9	0	0
> 10	18	20.2
Nuclear grade		
Grade I	15	16.8
Grade II	24	27.0
Grade III	50	56.2
LVI (peritumoral and intratumoral)		
Absent	35	39.3
Focal	16	18.0
Prominent	38	42.7
ER status		
Positive	74	83.1
Negative	15	16.9

Table 2. Clinical information on discordant cases

Case#	Surgery	RT	CT	Histology	Tumor size	Nodes	ER	PR	Relapse
LN-5	MRM	N	CMF	IDC	2	0	+++	++	NED, 12 months
LN-7	MRM	N	No	IDC	1.7	0	+++	+++	Yes, 32 months
LN-11	BCS	Y	No	IDC	0.5	0	+	+++	Yes, 38 months
LN-22	MRM	Y	CEF	IDC	3	10	+	+	Yes, 75 months

Case#	Surgery	RT	CT	Histology	Tumor size	Nodes	ER	PR	Relapse
Rec-38	MRM	N	No	TC	1.8	2	+	++	Yes, 11 months
Rec-23	MRM	N	CAF	IDC	3	1	-	-	NED, 74 months
Rec-6	MRM	N	CMF	ILC	3.1	2	+	+	Yes, 44 months
Rec-36	MRM	N	No	IDC	3.5	1	+	-	Yes, 6 months
Rec-42	MRM	N	CEF	IDC	3	2	+	+	Yes, 16 months

Abbreviations: MRM, modified radical mastectomy; RT, adjuvant Radiotherapy; CT, adjuvant chemotherapy; BCS, breast conserving surgery; NED, no evidence of disease; IDC, infiltrating ductal carcinoma; ILC, infiltrating lobular carcinoma; TC, tubular carcinoma

SUPPLEMENTARY INFORMATION

Supplementary material on genes and metagenes

Table 1. Genes associated with metagene predictors of lymph node metastasis.

Table 2. Genes associated with metagene predictors of breast cancer recurrence.

Table 3. Full list of genes defining all metagenes.

Supplementary material on statistical methods and data processing

Details on the specifics of data processing to evaluate metagene summaries for utilization in statistical analysis are provide here. Additional supplementary material includes the full technical report (ref. 15) that describes the statistical tree model methodology in complete detail.

Metagene summaries of gene expression profiles are obtained, for this breast cancer analysis, by combining standard clustering with also standard singular value decomposition (principal components) analysis. The precise steps taken in the study reported here are as follows:

?? Raw data are the 12,625 signal intensity measures of expression of genes on the Affymetrix HU95aV2 DNA microarray, with signal intensities based on the Affymetrix V5 software then transformed to the log-base 2 scale. An initial screen reduces this to a total of 7,030 genes to remove sequences that vary at low levelsor minimally. Specifically, this screens out genes whose expression levels

across all samples varies by less than two-fold, and whose maximum signal intensity value is lower than nine on a log-base 2 scale.

?? The set of samples on these 7,030 genes are clustered using k-means correlated-based clustering. Any standard statistical package may be used for this; our analysis uses the xcluster software created by Gavin Sherlock at Stanford University (<http://genome-www.stanford.edu/~sherlock/cluster.html>). We defined a target of 500 clusters and the xcluster routine delivered 496 in this analysis.

?? We extract the dominant singular factor (principal component) from each of the 496 clusters. Again, any standard statistical or numerical software package may be used for this; our analysis uses the reduced singular value decomposition function (svd) in Matlab (<http://www.mathworks.com/products/matlab>).

?? These 496 metagene predictors are input to the tree model analysis as described in Pittman et al. 2002 (Ref 15) and available as a technical report in Supplementary Material. A key ingredient is the generalized likelihood ratio, or Bayes' factor, measure of association between metagenes and binary outcomes (Section 2.1 of the statistics paper). An initial ordering of metagenes is provided by the Bayes' factor values on all the data (at the root node of the tree). "Top" metagenes are those with highest Bayes' factor in this sense, and several "top" metagenes were selected to define the lists of genes (accompanying material) as described further below. Specific parameters defined to create the precise tree models in the two breast examples are as follows (again with reference to Section 2 of the statistics paper). The tree model analysis as reported utilised a Bayes' factor threshold of 3 on the log scale, allowed up to 10 splits of the root node and then up to 4 at each

of nodes 1 and 2. Trees were allowed to grow to at most 2 levels consistent with the relatively small sample size of the data sets.

?? Predictions for individual patients were performed as described in the paper: the analysis was repeated for each patient, holding out from the model fitting the expression and outcome data for that patient, and then developing the statistical tree model analysis based on only the remaining data. Then, the hold-out patient was predicted (using the statistical analysis as described in Section 2.4 and 2.5 of the statistics paper). We note that the model fitting, including the statistical evaluation of which metagenes are most predictive and the roles they play in the analysis (i.e., the “feature selection process”) is repeated anew for each of these analyses. Were this not done, and metagene selection based on all the data, then the predictions would appear much more accurate, but incorrectly and misleadingly so. This critical perspective, which we have termed “honest prediction” in the cross-validation context, is one we have taken pains to stress in our work (e.g., reference 11) and one that defines our approach to critical model evaluation when prediction is a primary focus.

?? The lists of genes were generated precisely as follows, for each of the recurrence and metastasis analyses separately. From the statistical tree model fit to all the data, the "top" 4 metagenes were selected, based on the marginal Bayes' factor association measure as described. This defines 4 clusters of genes that are the initial basis of the list. The list was extended by adding in additional genes that are most highly correlated (standard linear correlation) with each of these 4 metagenes; the set of unique genes in the resulting lists are reported and form part

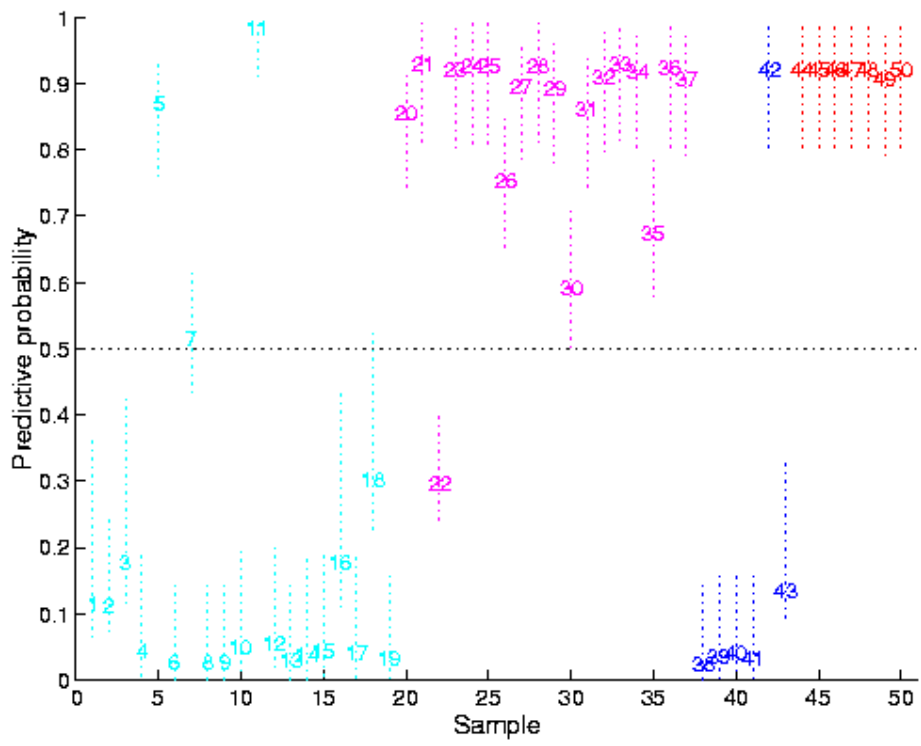
of this supplementary material, as are full details of all genes defining each of the 496 metagenes.

?? In the lymph node metastasis external validation test, the predictions of the sample of cancers from the Duke 2001 PNAS study were performed directly using the tree model fitted only to the data from the current study (as described). That is, predictions were performed entirely out-of-sample with no modification at all to the definition of metagenes, the model or the details of analysis, so paralleling the "real life" circumstances of predicting new patients and providing a completely honest out-of-sample assessment of generalization and predictive validity.

?? The metagene data for the Duke breast cancer samples used for external validation via out-of-sample prediction were evaluated as follows. The samples are from a 2000 study and gene expression profiles are on the early Affymetrix HU6800 array. The first step was then to identify all genes on that array (7,129 genes) that are also represented among the 12,625 genes on the U95av2 array. This was done using the chip-to-chip key available at the Affymetrix web site. This allows for the identification of genes on the HU6800 array that map to genes within each of the 496 metagene clusters from the current study. For example, the key metagenes 330, 146 and 130 have precisely 30, 37 and 8 genes, respectively; mapping these genes to the earlier HU6800 array identifies sets of 26, 42 and 4 genes, respectively (note that there are duplicates in some cases, as for metagene 146 here). These sets of genes on the HU6800 array define the metagene clusters and the corresponding value of the metagenes are evaluated precisely as described,

using the dominant singular factor (principal component) from each of the 496 clusters.

Supplementary Figure. Cross-validation and external validation probability predictions of lymph node status. Samples (tumors) are plotted by index number, and the plotted numbers are marked on the vertical scale at the estimated predictive probabilities of high-risk versus low risk. Color coding is as in Figure 3: predictions for the cases in the current study are the same in Figure 1, but now color coded as magenta (high-risk) and cyan (low risk), the cases from the Duke (PNAS 2001) study are correspondingly color coded red (high-risk) and blue (low-risk). Approximate 90% uncertainty intervals about these estimated probabilities are indicated by vertical dashed lines.



Supplementary Figure

Table 1. Genes associated with metagene predictors of lymph node metastasis

Acc. No.	Symbol	Gene name	GO Function
M12959	TCRA	T cell receptor alpha locus	
M13755	ISG15	interferon-stimulated protein, 15 kDa	
D43767	CCL17	small inducible cytokine subfamily A (Cys-Cys), member 17	G-protein linked receptor protein signalling pathway, developmental processes, cell-cell signaling, chemotaxis
D45248	PSME2	proteasome (prosome, macropain) activator subunit 2 (PA28 beta)	
L03840	FGFR4	fibroblast growth factor receptor 4	FGF receptor signalling pathway
U22970		interferon, alpha-inducible protein (clone IFI-6-16)	
M21121	CCL5	small inducible cytokine A5 (RANTES)	exocytosis, oxidative stress response, cell motility, chemotaxis, inflammatory response, cellular defense response, cell-cell signalling, immune response, response to viruses, signal transduction, calcium ion homeostasis, cell adhesion
L05148	ZAP70	zeta-chain (TCR) associated protein kinase (70 kD)	
D00596	TYMS	thymidylate synthetase	deoxyribonucleoside monophosphate biosynthesis, nucleobase, nucleoside, nucleotide and nucleic acid metabolism
D11086	IL2RG	interleukin 2 receptor, gamma (severe combined immunodeficiency)	protein complex assembly, immune response, cell proliferation, signal transduction
J04088	TOP2A	topoisomerase (DNA) II alpha (170kD)	
U73379	UBE2C	ubiquitin-conjugating enzyme E2C	degradation of cyclin, ubiquitin-dependent protein degradation, protein modification, positive control of cell proliferation
U37352	PPP2R5C	protein phosphatase 2, regulatory subunit B (B56), gamma isoform	
M31303	STMN1	stathmin 1/oncoprotein 18	
X13293	MYBL2	v-myb myeloblastosis viral oncogene homolog (avian)-like 2	anti-apoptosis, cell cycle control, developmental processes, transcription from Pol II promoter
M13194	ERCC1	excision repair cross-complementing rodent repair deficiency, complementation group 1 (includes overlapping antisense sequence)	DNA repair, nucleotide-excision repair, embryogenesis and morphogenesis
U09937	PLAUR	plasminogen activator, urokinase receptor	
U28014	CASP4	caspase 4, apoptosis-related cysteine protease	apoptosis, induction of apoptosis, proteolysis and peptidolysis
X73066	NME1	non-metastatic cells 1, protein (NM23A) expressed in	
L40387	OASL	2'-5'-oligoadenylate synthetase-like	
J04152	TACSTD2	tumor-associated calcium signal transducer 2	
U58515	CHI3L2	chitinase 3-like 2	
AI701049	FARP1	FERM, RhoGEF (ARHGEF) and pleckstrin domain protein 1 (chondrocyte-derived)	
AB018280	KIAA0737	KIAA0737 gene product	
X53280	BTF3	basic transcription factor 3	transcription from Pol II promotor
X63527	RPL19	ribosomal protein L19	protein biosynthesis

AF026947	AKR7A2	aldo-keto reductase family 7, member A2 (aflatoxin aldehyde reductase)	aldehyde metabolism, carbohydrate metabolism, oncogenesis
D37931	RNASE4	ribonuclease, RNase A family, 4	
AL080076	SSBP2	single-stranded DNA binding protein 2	
M55543	GBP2	guanylate binding protein 2, interferon-inducible	immune response
D26070	ITPR1	inositol 1,4,5-triphosphate receptor, type 1	small molecule transport, signal transduction
M24594	IFIT1	interferon-induced protein with tetratricopeptide repeats 1	
M97935	STAT1	signal transducer and activator of transcription 1, 91kD	signal transduction, caspase activation, JAK-STAT cascade, NIK-1-kappaB/NF-kappaB cascade, STAT protein dimerization, STAT protein nuclear translocation, tyrosine phosphorylation of STAT protein, cell cycle control, response to pest/pathogen/parasite, transcription from Pol II promotor
M97935	STAT1	signal transducer and activator of transcription 1, 91kD	
L13435		glycosyltransferase AD-017	
AF060228	RARRES3	retinoic acid receptor responder (tazarotene induced) 3	negative control of cell proliferation
U88964	ISG20	interferon stimulated gene (20kD)	cell proliferation
M97936	STAT1	signal transducer and activator of transcription 1, 91kD	
M97936	STAT1	signal transducer and activator of transcription 1, 91kD	
AL049977	CLDN8	claudin 8	
AB002390	LYSAL1	lysosomal apyrase-like 1	nucleobase, nucleoside, nucleotide and nucleic acid metabolism
AI761567	KIAA1254	KIAA1254 protein	
D12485		ectonucleotide pyrophosphatase/phosphodiesterase 1	
AJ225089	OASL	2'-5'-oligoadenylate synthetase-like	
L07919	DLX2	distal-less homeo box 2	brain development
AI670788	MAP-1	modulator of apoptosis 1	
D14678	KNSL2	kinesin-like 2	
AL039458	LRIG1	ortholog of mouse integral membrane glycoprotein LIG-1	
AL050197	DKFZP586D0623	DKFZP586D0623 protein 623	
AF011468	STK6	serine/threonine kinase 15	protein phosphorylation, oncogenesis, mitosis
AF016266	TNFRSF10B	tumor necrosis factor receptor superfamily, member 10b	induction of apoptosis via death domain receptors, cell surface receptor linked signal transduction
Y13323	ADAMDEC1	ADAM-like, decysin 1	
AB002345	PER2	period homolog 2 (Drosophila)	circadian rhythm
X53281	BTF3	basic transcription factor 3	
AF030514	CXCL11	small inducible cytokine subfamily B (Cys-X-Cys), member 11	response to pathogenic fungi, cell-cell signaling, chemotaxis, inflammatory response, signal transduction
AF019225	APOL1	apolipoprotein L	lipid metabolism
AB011143	GAB2	GRB2-associated binding protein 2	
Z15008	LAMC2	laminin, gamma 2 (nicein (100kD), kalinin (105kD), BM600 (100kD), Herlitz junctional epidermolysis bullosa))	epidermal differentiation
U54558	EIF3S7	eukaryotic translation initiation factor 3, subunit 7 (zeta, 66/67kD)	translational regulation, initiation
W28256	DKFZP586M1120		

X99699	HSXIAPAF1	XIAP associated factor-1	
AB001451	SLI	neuronal Shc adaptor homolog	central nervous system development, peripheral nervous system development, signal transduction
X87342	LLGL2	lethal giant larvae homolog 2 (Drosophila)	
M55542	GBP1	guanylate binding protein 1, interferon-inducible, 67kD	
AI525393	ARPC3	actin related protein 2/3 complex, subunit 3 (21 kD)	cell motility
D78130	SQLE	squalene epoxidase	
AF004230	LILRB1	leukocyte immunoglobulin-like receptor, subfamily B (with TM and ITIM domains), member 1	response to viruses
AF087036	MSC	musculin (activated B-cell factor-1)	transcription from Pol II promotor
AF006621	C4orf1	chromosome 4 open reading frame 1	
AB011084	ALEX2	armadillo repeat protein ALEX2	
AL080213			
AJ000882	NCOA1	nuclear receptor coactivator 1	transcription
U26174	GZMK	granzyme K (serine protease, granzyme 3; tryptase II)	
U53831	IRF7	interferon regulatory factor 7	
AJ131693	AKAP9	A kinase (PRKA) anchor protein (yotiao) 9	synaptic transmission, signal transduction, small molecule transport
AL040446	OSBPL1A	oxysterol-binding protein-related protein 1	
X51985	LAG3	lymphocyte-activation gene 3	
M34455	INDO	indoleamine-pyrrole 2,3 dioxygenase	tryptophan catabolism, pregnancy, defense response
M63193	ECGF1	endothelial cell growth factor 1 (platelet-derived)	DNA replication, mitochondrial genome maintenance, pyrimidine nucleotide metabolism, cell-cell signaling, cell surface receptor linked signal transduction
AF001691	PPL	periplakin	cell shape and cell size control
AL022237	BIK	BCL2-interacting killer (apoptosis-inducing)	
AB000115	C1orf29	hypothetical protein, expressed in osteoblast	
U70063	ASAH1	N-acylsphingosine amidohydrolase (acid ceramidase)	ceramide metabolism, fatty acid metabolism
M33882	MX1	myxovirus (influenza) resistance 1, homolog of murine (interferon-inducible protein p78)	defense response, signal transduction, induction of apoptosis, pathogenesis
J02923	LCP1	lymphocyte cytosolic protein 1 (L-plastin)	
M62800	SSA1	Sjogren syndrome antigen A1 (52kD, ribonucleoprotein autoantigen SS-A/Ro)	pathogenesis
M17016	GZMB	granzyme B (granzyme 2, cytotoxic T-lymphocyte-associated serine esterase 1)	
M85276	GNLY	granulysin	cellular defense response
U95626		chemokine (C-C motif) receptor 2	
AB013924	LAMP3	lysosomal-associated membrane protein 3	oncogenesis, cell proliferation
X72755	CXCL9	monokine induced by gamma interferon	defense response, immune response, inflammatory response, chemotaxis, cell-cell signaling, signal transduction, cellular defense response, G-protein linked receptor protein signalling pathway
D87071	KIAA0233	KIAA0233 gene product	
X58536	HLA-C	major histocompatibility complex, class I, C	
AL022723		major histocompatibility complex, class I, F	

AJ001634	CCL13	small inducible cytokine subfamily A (Cys-Cys), member 13	signal transduction, calcium ion homeostasis, cell-cell signaling, chemotaxis, immune response, inflammatory response
X98834	SALL2	sal-like 2 (Drosophila)	histogenesis and organogenesis
D28915	IFI44	interferon-induced, hepatitis C-associated microtubular aggregate protein (44kD)	
L13210	LGALS3BP	lectin, galactoside-binding, soluble, 3 binding protein	cellular defense response, signal transduction
AF072468	JRK	jerky homolog (mouse)	
Z47553	FMO5	flavin containing monooxygenase 5	
AL080078			
U19523	GCH1	GTP cyclohydrolase 1 (dopa-responsive dystonia)	neurotransmitter synthesis and storage, nitric oxide biosynthesis
AI436567	ATP5D	ATP synthase, H ⁺ transporting, mitochondrial F1 complex, delta subunit	
AL096842	ATIP1	AT2 receptor-interacting protein 1	
AL049367	LOC55970		
X59892	WARS	tryptophanyl-tRNA synthetase	tryptophanyl-tRNA biosynthesis, protein biosynthesis, negative control of cell proliferation
U90548	BTN3A3	butyrophilin, subfamily 3, member A3	
AA808961	PSMB9	proteasome (prosome, macropain) subunit, beta type, 9 (large multifunctional protease 2)	proteolysis and peptidolysis
AA919102	CD3D	CD3D antigen, delta polypeptide (TiT3 complex)	cell surface receptor linked signal transduction, cellular defense response
M11810		2',5'-oligoadenylate synthetase 1 (40-46 kD)	
X04371	OAS1	2',5'-oligoadenylate synthetase 1 (40-46 kD)	
AA203213	ISG15	interferon-stimulated protein, 15 kDa	
M87503	ISGF3G	interferon-stimulated transcription factor 3, gamma (48kD)	cell surface receptor linked signal transduction, transcription from Pol II promoter
AF026941	cig5		
AF026939	IFIT4	interferon-induced protein with tetratricopeptide repeats 4	
AL047596		capicua homolog (Drosophila)	
U97502			
U90546	BTN3A2	butyrophilin, subfamily 3, member A2	
AD001528	SMS	spermine synthase	methionine metabolism, polyamine metabolism
W26226	DJ971N18.2	hypothetical protein	
M11119			
AF097738	TNK1	tyrosine kinase, non-receptor, 1	
D44497	CORO1A	coronin, actin binding protein, 1A	phagosome formation, transport, mitosis, cell motility, cell shape and cell size control
AL031178			
D28137	BST2	bone marrow stromal cell antigen 2	humoral defense mechanism, cell proliferation, cell-cell signaling, developmental processes
M87434	OAS2	2'-5'-oligoadenylate synthetase 2 (69-71 kD)	nucleobase, nucleoside, nucleotide and nucleic acid metabolism
M87284	OAS2	2'-5'-oligoadenylate synthetase 2 (69-71 kD)	
AF070632			
AJ001902	MGC10558	thyroid hormone receptor interactor 6	
U09825	TRIM26	tripartite motif-containing 26	

U33267	GLRB	glycine receptor, beta	small molecule transport, cell surface receptor linked signal transduction
AI597616	MRPL33	mitochondrial ribosomal protein L33	
AA402538	MGC2749	hypothetical protein MGC2749	
D78134	CIRBP	cold inducible RNA binding protein	cold response
D21337	COL4A6	collagen, type IV, alpha 6	oncogenesis
AL031983		gamma-aminobutyric acid (GABA) B receptor, 1	
M74447	TAP2	transporter 2, ATP-binding cassette, sub-family B (MDR/TAP)	cellular defense response, peptide transport, defense response
AB012917	KLK11	kallikrein 11	
AF020202	UNC13	unc-13-like (C. elegans)	excretion, induction of apoptosis, signal transduction, apoptosis
X57522	TAP1	transporter 1, ATP-binding cassette, sub-family B (MDR/TAP)	defense response, cellular defense response, peptide transport
D88153	HYA22	HYA22 protein	
U64197	CCL20	small inducible cytokine subfamily A (Cys-Cys), member 20	antimicrobial humoral response, immune response, inflammatory response, chemotaxis, signal transduction, cell-cell signaling
Y09048		peroxisomal farnesylated protein	protein-peroxisome targeting, peroxisome organization and biogenesis
AA883502	UBE2L6	ubiquitin-conjugating enzyme E2L 6	protein modification
Y00062	PTPRC	protein tyrosine phosphatase, receptor type, C	cell surface receptor linked signal transduction
M91670	E2-EPF	ubiquitin carrier protein	protein modification
AL021683		SCO cytochrome oxidase deficient homolog 2 (yeast)	
M16336	CD2	CD2 antigen (p50), sheep red blood cell receptor	cell adhesion, signal transduction, antimicrobial humoral response
AB018288	RANBP16	RAN binding protein 16	
AL035494		hypothetical protein FLJ10097	
AI651806	CRIM1	cysteine-rich motor neuron 1	neurogenesis
U05875	IFNGR2	interferon gamma receptor 2 (interferon gamma transducer 1)	resistance to pathogenic bacteria, response to viruses, cell surface receptor linked signal transduction
D45248	PSME2	proteasome (prosome, macropain) activator subunit 2 (PA28 beta)	
X87344		major histocompatibility complex, class II, DM alpha	
AL049417	DUSP3	dual specificity phosphatase 3 (vaccinia virus phosphatase VH1-related)	protein dephosphorylation
D32129	HLA-A	major histocompatibility complex, class I, A	
AF035282	C1orf21	chromosome 1 open reading frame 21	
X57352	IFITM3	interferon induced transmembrane protein 3 (1-8U)	immune response
AB023194	KIAA0977	KIAA0977 protein	
AA149431	DKFZp761F2014	hypothetical protein DKFZp761F2014	
X67325	IFI27	interferon, alpha-inducible protein 27	
X02530	CXCL10	small inducible cytokine subfamily B (Cys-X-Cys), member 10	signal transduction, chemotaxis, cell motility, circulation, muscle development, positive control of cell proliferation, cell-cell signaling, inflammatory response, signal transduction, cell surface receptor linked signal transduction
U72882	IFI35	interferon-induced protein 35	
L78833		breast cancer 1, early onset	

L39874		dCMP deaminase	pyrimidine nucleotide metabolism
L05072		interferon regulatory factor 1	oncogenesis, transcription from Pol II promotor
J04164	IFITM1	interferon induced transmembrane protein 1 (9-27)	negative control of cell proliferation, cell surface receptor linked signal transduction, cell cycle control
AB006782	LGALS9	lectin, galactoside-binding, soluble, 9 (galectin 9)	
D13435	PIGF	phosphatidylinositol glycan, class F	GPI anchor formation
M30818	MX2	myxovirus (influenza) resistance 2, homolog of murine	defense response
M91670	E2-EPF	ubiquitin carrier protein	
M91670	E2-EPF	ubiquitin carrier protein	
M24594	IFIT1	interferon-induced protein with tetratricopeptide repeats 1	
J03909	IFI30	interferon, gamma-inducible protein 30	
Y10032	SGK	serum/glucocorticoid regulated kinase	sodium transport, stress response, protein phosphorylation

Table 2. Genes associated with metagene predictors of breast cancer recurrence

Acc. No.	Symbol	Gene name	GO Function
U50648	PRKR	protein kinase, interferon-inducible double stranded RNA dependent	
U37055	MST1	macrophage stimulating 1 (hepatocyte growth factor-like)	
K03183	CGB7	chorionic gonadotropin, beta polypeptide 7	
J03069	MYCL2	v-myc myelocytomatosis viral oncogene homolog 2 (avian)	
M36711	TFAP2A	transcription factor AP-2 alpha (activating enhancer binding protein 2 alpha)	Signal transduction, developmental processes, transcription regulation from Pol II promoter, oncogenesis, ectoderm development
X69699	PAX8	paired box gene 8	histogenesis and organogenesis, embryogenesis and morphogenesis
L38929	PTPRD	protein tyrosine phosphatase, receptor type, D	protein dephosphorylation, transmembrane receptor protein tyrosine phosphatase signalling, phosphate metabolism
L76568	ERCC4	excision repair cross-complementing rodent repair deficiency, complementation group 4	
U11870	IL8RA	interleukin 8 receptor, alpha	
M36067	LIG1	ligase I, DNA, ATP-dependent	DNA repair, embryogenesis and morphogenesis, DNA metabolism
D16105	LTK	leukocyte tyrosine kinase	Signal transduction, protein phosphorylation
U12779	MAPKAPK2	mitogen-activated protein kinase-activated protein kinase 2	protein phosphorylation, MAPKKK cascade
L34059	CDH4	cadherin 4, type 1, R-cadherin (retinal)	cell adhesion
U22028	CYP2A13	cytochrome P450, subfamily IIA (phenobarbital-inducible), polypeptide 13	
U27193	DUSP8	dual specificity phosphatase 8	protein dephosphorylation, inactivation of MAPK
L24559	POLA2	polymerase (DNA-directed), alpha (70kD)	
U40343	CDKN2D	cyclin-dependent kinase inhibitor 2D (p19, inhibits CDK4)	regulation of CDK activity, negative control of cell proliferation, cell cycle arrest
Z36714	CCNF	cyclin F	cell cycle control
U18334	NOS2C	nitric oxide synthase 2C	
U31317	JAK3	Janus kinase 3 (a protein tyrosine kinase, leukocyte)	protein phosphorylation, mesoderm development, cell growth and maintenance
U07375	ITGAV	integrin, alpha V (vitronectin receptor, alpha polypeptide, antigen CD51)	
M64231	SRM	spermidine synthase	
AF023614	TNFRSF13B	tumor necrosis factor receptor superfamily, member 13B	cell surface receptor linked signal transduction
U34806	GPR15	G protein-coupled receptor 15	G-protein linked receptor protein signalling pathway
U26209	SLC13A2	solute carrier family 13 (sodium-dependent dicarboxylate transporter), member 2	small molecule transport
AL031983	GABBR1	gamma-aminobutyric acid (GABA) B receptor, 1	
AL031983	OR2H2	olfactory receptor, family 2, subfamily H, member 2	
AL031983	OR2H5P	olfactory receptor, family 2, subfamily H, member 5 pseudogene	
AL031983	OR211P	olfactory receptor, family 2, subfamily I, member 1 pseudogene	

Y16788	KRTHA3A	keratin, hair, acidic, 3A	cell shape and cell size control
AF047485	LOC90586	amine oxidase pseudogene	cell adhesion, inflammatory response, amine metabolism
X57282	EPOR	erythropoietin receptor	Signal transduction
AF027957	GPR35	G protein-coupled receptor 35	G-protein linked receptor protein signalling pathway
AF007194	MUC3A	mucin 3A, intestinal	
D86979	KIAA0226	KIAA0226 gene product	
U32645	ELF4	E74-like factor 4 (ets domain transcription factor)	transcription from Pol II promotor
L14754	IGHMBP2	immunoglobulin mu binding protein 2	DNA replication, DNA repair, single stranded DNA binding, DNA recombination
X63096	RHCE	Rhesus blood group, CcEe antigens	
X74143	FOXG1A	forkhead box G1A	brain development
Z82180	EAN57	hypothetical protein EAN57	
AF017995	PDPK1	3-phosphoinositide dependent protein kinase-1	protein phosphorylation, insulin receptor signalling pathway, actin cytoskeleton reorganization
U06088	GALNS	galactosamine (N-acetyl)-6-sulfate sulfatase (Morquio syndrome, mucopolysaccharidosis type IVA)	
AB028950	TLN1	talin 1	
AB009398	PSMD13	proteasome (prosome, macropain) 26S subunit, non-ATPase, 13	
AB020706	AP2A2	adaptor-related protein complex 2, alpha 2 subunit	
L35318	GRM2	glutamate receptor, metabotropic 2	synaptic transmission, adenylyate cyclase inhibition
D10704	CHK	choline kinase	lipid metabolism, lipid transport
X64116	PVR	poliovirus receptor	
U04810	TROAP	trophinin associated protein (tastin)	cell adhesion
X83127	KCNAB1	potassium voltage-gated channel, shaker-related subfamily, beta member 1	potassium transport
L76703	PPP2R5E	protein phosphatase 2, regulatory subunit B (B56), epsilon isoform	
AJ237672	MTHFR	5,10-methylenetetrahydrofolate reductase (NADPH)	circulation, amino acid metabolism
M58378	SYN1	synapsin I	
AB007943	RAP1GA1	RAP1, GTPase activating protein 1	
AF003837	JAG1	jagged 1 (Alagille syndrome)	
AB023167	LFG	lifeguard	
AF054185	PSMA7	proteasome (prosome, macropain) subunit, alpha type, 7	
AF052177	KIAA1719	KIAA1719 protein	
M63962	ATP4A	ATPase, H+/K+ exchanging, alpha polypeptide	
U48861	CHRNB4	cholinergic receptor, nicotinic, beta polypeptide 4	Signal transduction, small molecule transport, synaptic transmission, cholinergic
D38081	TBXA2R	thromboxane A2 receptor	G-protein linked receptor protein signalling pathway, respiration, muscle contraction
AC004755	LOC148229	similar to Glt-P1	
AC004755	ONECUT3	one cut domain, family member 3	
AL080150	GEMIN4	gem (nuclear organelle) associated protein 4	ribosome biogenesis, rRNA processing
M31525	HLA-DOA	major histocompatibility complex, class II, DO alpha	
X73079	PIGR	polymeric immunoglobulin receptor	protein secretion
AJ012590	H6PD	hexose-6-phosphate dehydrogenase (glucose 1-dehydrogenase)	
AB028953	KIAA1030	KIAA1030 protein	

S80071	SLC6A7	solute carrier family 6 (neurotransmitter transporter, L-proline), member 7	small molecule transport, synaptic transmission, proline transport
D84307	PCYT2	phosphate cytidyltransferase 2, ethanolamine	
D49738	CKAP1	cytoskeleton-associated protein 1	
U33849	PCSK7	proprotein convertase subtilisin/kexin type 7	
U47927	USP5	ubiquitin specific protease 5 (isopeptidase T)	deubiquitylation
X84746	ABO	ABO blood group (transferase A, alpha 1-3-N-acetylgalactosaminyltransferase; transferase B, alpha 1-3-galactosyltransferase)	
U30185	OPRL1	opiate receptor-like 1	sensory perception, G-protein signalling, adenylate cyclase inhibiting pathway
X82634	KRTHA3B	keratin, hair, acidic, 3B	cell shape and cell size control
U14187	EFNA3	ephrin-A3	
D88587	FCN3	ficolin (collagen/fibrinogen domain containing) 3 (Hakata antigen)	
Z97029	RNASEH2A	ribonuclease H2, large subunit	DNA replication, RNA catabolism
M91592	ZNF76	zinc finger protein 76 (expressed in testis)	transcription regulation from Pol II and Pol III promotor
X17094	PACE	paired basic amino acid cleaving enzyme (furin, membrane associated receptor protein)	proteolysis and peptidolysis, cell-cell signaling
AC004523	CYP4F12	cytochrome P450, subfamily IVF, polypeptide 12	
X81832	GIPR	gastric inhibitory polypeptide receptor	
U30894	SGSH	N-sulfoglucosamine sulfohydrolase (sulfamidase)	proteoglycan metabolism
Z14000	RING1	ring finger protein 1	
Y00970	ACR	acrosin	acrosome reaction
AJ003147	MEFV	Mediterranean fever	
AJ003147	OR1F2	olfactory receptor, family 1, subfamily F, member 2	
AJ003147	MMPL1	olfactory receptor, family 1, subfamily F, member 2	
AJ003147	ZNF200	zinc finger protein 200	
AJ003147	OR1F1	olfactory receptor, family 1, subfamily F, member 1	
AJ010901	MUC4	mucin 4, tracheobronchial	
U90841	SSX4	synovial sarcoma, X breakpoint 4	
AF035531	STX10	syntaxin 10	
J05500	SPTB	spectrin, beta, erythrocytic (includes spherocytosis, clinical type I)	cell shape and cell size control
AB020649	KIAA0842	KIAA0842 protein	
AB009698	SLC22A6	solute carrier family 22 (organic anion transporter), member 6	organic anion transport, alpha-ketoglutarate transport
X67734	CNTN2	contactin 2 (axonal)	cell adhesion
U40391	AANAT	arylalkylamine N-acetyltransferase	
X74614	ODF1	outer dense fibre of sperm tails 1	
AL050220	KLK13	kallikrein 13	
U29943	ELAVL2	ELAV (embryonic lethal, abnormal vision, Drosophila)-like 2 (Hu antigen B)	transcription regulation
U85647	SOLH	small optic lobes homolog (Drosophila)	
AB014590	KIAA0690	KIAA0690 protein	
U78521	AIP	aryl hydrocarbon receptor interacting protein	
M83751	ARMET	arginine-rich, mutated in early stage tumors	oncogenesis
X14640	KRT13	keratin 13	epidermal differentiation
M88468	MVK	mevalonate kinase (mevalonic aciduria)	protein phosphorylation, isoprenoid biosynthesis
D14720	MPZ	myelin protein zero (Charcot-Marie-Tooth neuropathy 1B)	
AB018352	KIAA0809	KIAA0809 protein	
D87463	PHYHIP	phytanoyl-CoA hydroxylase interacting protein	

X60364	ALAS2	aminolevulinate, delta-, synthase 2 (sideroblastic/hypochromic anemia)	heme biosynthesis
AB018258	ATP10B	ATPase, Class V, type 10B	
U15131	ST5	suppression of tumorigenicity 5	
X53742	FBLN1	fibulin 1	
X55448	G6PD	glucose-6-phosphate dehydrogenase	glucose 6-phosphate utilization
X55448	FAM3A	family with sequence similarity 3, member A	
D84110	RBPMS	RNA-binding protein gene with multiple splicing	RNA processing
Z22865	DPT	dermatopontin	
AF059252	DOM3Z	dom-3 homolog Z (C. elegans)	
AF112219	ESD	esterase D/formylglutathione hydrolase	
X71874	PSMB10	proteasome (prosome, macropain) subunit, beta type, 10	proteolysis and peptidolysis, humoral defense mechanism
AB021638	APBA3	amyloid beta (A4) precursor protein-binding, family A, member 3 (X11-like 2)	
M29273	MAG	myelin associated glycoprotein	
AD001530	DXS9928E	DNA segment on chromosome X (unique) 9928 expressed sequence	
AF027204	TM4SF5	transmembrane 4 superfamily member 5	N-linked glycosylation
S75174	E2F4	E2F transcription factor 4, p107/p130-binding	cell cycle control
AJ000730	SLC7A4	solute carrier family 7 (cationic amino acid transporter, y+ system), member 4	small molecule transport, amino acid metabolism
AB023205	TBCD	tubulin-specific chaperone d	protein folding, beta tubulin folding
AF050145	IDS	iduronate 2-sulfatase (Hunter syndrome)	
AF045800	CKTSF1B1	cysteine knot superfamily 1, BMP antagonist 1	developmental processes, neurogenesis
X97671	EPOR	erythropoietin receptor	
AF033105	ARR3	arrestin 3, retinal (X-arrestin)	Signal transduction, vision
X07695	KRT4	keratin 4	cell shape and cell size control, epidermal differentiation
AB014522	CLASP1	cytoplasmic linker associated protein 1	
AF071748	CTSF	cathepsin F	proteolysis and peptidolysis
X69550	ARHGDI2A	Rho GDP dissociation inhibitor (GDI) alpha	cell adhesion inhibition, RHO protein signal transduction
AB009288	CPNE6	copine VI (neuronal)	lipid metabolism, synaptic transmission, neurogenesis, vesicle transport
AF091890	RE2	G-protein coupled receptor	
AF091890	RE2	G-protein coupled receptor	
U57352	ACCN1	amiloride-sensitive cation channel 1, neuronal (degenerin)	synaptic transmission, peripheral nervous system development, monovalent inorganic cation transport, central nervous system development,
AL096740	UBE3B	ubiquitin protein ligase	
U10868	ALDH3B1	aldehyde dehydrogenase 3 family, member B1	lipid metabolism, alcohol metabolism
L41498	EEF1A1L14	eukaryotic translation elongation factor 1 alpha 1-like 14	
AL050369	PRPF31	PRP31 pre-mRNA processing factor 31 homolog (yeast)	
AB023202	RPH3A	likely ortholog of mouse rabphilin 3A	
AF007134	MAPK8IP1	mitogen-activated protein kinase 8 interacting protein 1	
AJ000342	DMBT1	deleted in malignant brain tumors 1	
AF062529	NUDT3	nudix (nucleoside diphosphate linked moiety X)-type motif 3	cell-cell signalling, diadenosine polyphosphate catabolism
AB018274	KIAA0731	KIAA0731 protein	

X65633	MC2R	melanocortin 2 receptor (adrenocorticotrophic hormone)	G-protein linked receptor protein signalling pathway, G-protein signalling, linked to cyclic nucleotide second messenger
U20391	FOLR1	folate receptor 1 (adult)	
J04948	ALPPL2	alkaline phosphatase, placental-like 2	
J03071	GH1	growth hormone 1	
J03071	GH2	growth hormone 2	
J03071	CSH1	chorionic somatomammotropin hormone 1 (placental lactogen)	
J03071	CSH2	chorionic somatomammotropin hormone 2	
J03071	CSHL1	chorionic somatomammotropin hormone-like 1	
M55405	MUC3A	mucin 3A, intestinal	
M55405	MUC3A	mucin 3A, intestinal	
M55405	MUC3A	mucin 3A, intestinal	
M37435	CSF1	colony stimulating factor 1 (macrophage)	developmental processes, positive control of cell proliferation, cell proliferation, antimicrobial humoral response
M62302	LASS1	LAG1 longevity assurance homolog 1 (S. cerevisiae)	
M62302	GDF1	growth differentiation factor 1	