A validated gene expression profile for detecting clinical outcome in breast cancer using artificial neural networks

L. J. Lancashire, D. G. Powe, J. S. Reis-Filho, E. Rakha, C. Lemetre, B. Weigelt, T. M. Abdel-Fatah, A. R. Green, R. Mukta, R. Blamey, et al.

Breast Cancer Res Treat (2010) 120:83–93
DOI 10.1007/s10549-009-0378-1

PRECLINICAL STUDY

A validated gene expression profile for detecting clinical outcome in breast cancer using artificial neural networks

L. J. Lancashire · D. G. Powe · J. S. Reis-Filho · E. Rakha · C. Lemetre · B. Weigelt · T. M. Abdel-Fatah · A. R. Green · R. Mukta · R. Blamey · E. C. Paish · R. C. Rees · I. O. Ellis · G. R. Ball

Received: 11 February 2009 / Accepted: 13 March 2009 / Published online: 4 April 2009

Abstract Gene expression microarrays allow for the high throughput analysis of huge numbers of gene transcripts and this technology has been widely applied to the molecular and biological classification of cancer patients and in predicting clinical outcome. A potential handicap of such data intensive molecular technologies is the translation to clinical application in routine practice. In using an artificial neural network bioinformatic approach, we have reduced a 70 gene signature to just 9 genes capable of accurately predicting distant metastases in the original dataset. Upon validation in a follow-up cohort, this signature was an independent predictor of metastases free and overall survival in the presence of the 70 gene signature and other factors. Interestingly, the ANN signature and CA9 expression also split the groups defined by the 70 gene signature into prognostically distinct groups. Subsequently, the presence of protein for the principal prognosticator gene was categorically assessed in breast cancer tissue of an experimental and independent validation patient cohort, using immunohistochemistry. Importantly our principal prognosticator, CA9, showed that it is capable of selecting an aggressive subgroup of patients who are known to have poor prognosis.

Keywords Breast cancer · Prognosis · Bioinformatics · Survival · Hypoxia · Biomarkers

Abbreviations

ANN	Artificial neural networks
BCSS	Breast cancer specific survival
CA9	Carbonic anhydrase IX
EGF	Epidermal growth factor
DFI	Disease-free interval
EST	Expressed sequence tag

L. J. Lancashire and D. G. Powe contributed equally.

Electronic supplementary material The online version of this article (doi:10.1007/s10549-009-0378-1) contains supplementary material, which is available to authorized users.

L. J. Lancashire
Clinical and Experimental Pharmacology, Paterson Institute for Cancer Research, University of Manchester, Manchester M20 4BX, UK
e-mail: llancashire@picr.man.ac.uk

L. J. Lancashire · C. Lemetre · R. C. Rees · G. R. Ball (✉)
John Van Geest Cancer Research Centre, Nottingham Trent University, Clifton Campus, Clifton Lane, Nottingham NG11 8NS, UK
e-mail: graham.balls@ntu.ac.uk

D. G. Powe · E. Rakha · T. M. Abdel-Fatah · A. R. Green · R. Mukta · E. C. Paish · I. O. Ellis (✉)
Department of Histopathology, Nottingham University Hospitals Trust and University of Nottingham, Nottingham NG7 2UH, UK
e-mail: ian.ellis@nottingham.ac.uk

J. S. Reis-Filho
The Breakthrough Breast Cancer Research Centre, Institute of Cancer Research, Chester Beatty Laboratories, 237 Fulham Road, London SW3 6JB, UK

R. Blamey
Department of Surgery, Breast Institute, City Hospital Nottingham, Nottingham NG5 1PB, UK

B. Weigelt
Signal Transduction Laboratory, London Research Institute, Lincoln's Inn Fields Laboratories, 44 Lincoln's Inn Fields, London WC2A 3PX, UK

HR	Hormonal receptors
HIF-1α	Hypoxia induced factor 1 alpha
ROC	Receiver operating characteristic
RMH	Royal marsden hospital
TMA	Tissue microarray
TNP	Triple negative phenotype
AUC	Area under the curve

Introduction

Breast cancer is a heterogeneous disease where the outcome and response to therapy is often uncertain due to the complex network of overlapping and interacting molecular pathways. New strategies are needed to maximise therapeutic outcomes while limiting unnecessary over-treatment, achievable through customised treatment regimens. Previous studies have shown the ability of microarrays [1] to successfully predict clinical outcome in a variety of malignancies [2–6]. In particular, the molecular classification of malignant breast tumours using high throughput technologies including expression arrays and immunohistochemistry screening on tissue microarrays (TMAs), has successfully identified a number of biologically relevant subgroups [5, 7–10], showing good association between group membership and prediction of clinical outcome, targeted treatment and sensitivity to therapeutics [11–13]. Although there has been little overlap between different studies, more recent meta-analyses have demonstrated that different signatures identify similar groups of patients who have tumours with high proliferation rates [14, 15]. However, these meta-analyses have also demonstrated that most signatures reported to date have a relatively poor discriminatory power in oestrogen receptor negative disease [15]. Determining an optimal subset of predictive markers from microarray data is daunting due to the number of potential biomarker combinations present in these complex datasets. As an example, the seminal gene expression array data of van't Veer et al. [13] comprised in excess of 24,000 variables (gene transcripts) per sample. More recent generations of gene chip now contain in excess of one million variables, further highlighting the requirements for robust computational analysis methods and emphasising the difficulties in translating these results to routine clinical practice.

Given the obvious advantages of analysing high density microarrays offering large (or even complete) genome coverage, powerful approaches are required for determining prognostic gene subsets in breast cancer. One such approach utilises Artificial Neural Networks (ANNs) to assess the prognostic potential of each gene transcript individually in a univariate procedure, and then adding further genes in a sequential, multivariate manner to improve upon the classification accuracy [16]. ANNs are a form of artificial intelligence inspired by learning in human neuronal systems and have been shown to be capable of modelling complex systems with high predictive accuracies on blind data [3, 17–19]. ANN models are developed by iteratively changing a network of weights, in response to predictive error. Predictions are made by mathematically modifying weights generated from input values (e.g. gene transcript intensity), in turn producing a predicted output value (for example, predicted survival). Moreover, the importance of the individual inputs in generating these predictions may be determined to define optimal subsets of biomarkers within the system being analysed. In a previous study [16], we developed a novel iterative stepwise approach to ANN modelling. In this study, we have applied ANN to van't Veer's dataset [13] to determine a minimal set of biomarkers required for the prediction of metastasis in patients with breast cancer. We identified a panel comprising just nine genes predicting tumour metastasis with 98% accuracy. The principal prognostic indicator had a prediction accuracy of 70% when used independently in the model and was found to be the hypoxia-associated enzyme carbonic anhydrase IX (CA9). The prognostic gene panel was validated on a second gene expression dataset consisting of 295 cases [20], with CA9 expression displaying an accuracy of 63% in predicting the development of metastasis in a categorical yes/no fashion. This increased to 66% when the remaining genes in the signature were included and was shown to be an independent predictor of both overall survival and metastasis free survival in this second cohort. Consequently, we investigated the immunohistochemical protein expression of CA9 as a prognostic and predictive indicator in an independent patient TMA containing 552 unselected breast cancers, and in 390 full-face breast excision tumour blocks comprising an experimental and validation cohort of 160 and 230 patients, respectively.

Materials and methods

ANN model development to identify a prognostic gene signature for metastasis

The ANN modelling used a supervised learning approach, multi-layer perceptron architecture with a sigmoidal transfer function, where weights were updated by a back propagation algorithm [21]. Learning rate and momentum were set at 0.1 and 0.5 respectively. The ANN architecture utilised five hidden nodes in the hidden layer and randomised initial weights. The output node was coded as 0 if

the patient showed no evidence of metastasis within 5 years, and 1 if metastasis was evident. Data were downloaded in Microsoft Excel format from http://www.rii.com/publications/2002/vantveer.html. This initial set consisted of 78 samples each with 24,481 corresponding variables specifying the Log_{10} expression ratio of each gene. Prior to ANN training, the data was randomly divided into three subsets; 60% for training, 20% for validation (to assess model performance during the training process) and 20% for testing (to independently test the model on data completely blind to the model). This Monte–Carlo cross validation procedure [22] avoids over-fitting of the data, and has been shown to outperform and to be more consistent than the commonly used leave-one-out cross validation [23, 24], which may be a poor candidate for estimating the prediction error [25].

The forward stepwise approach to biomarker identification using ANNs has been previously described in detail (for specific details the reader is referred to [16]). This method develops a predictive model containing a parsimonious gene expression signature accurately classifying the cases according to the development of metastasis. Receiver Operating Characteristic (ROC) curves were generated to provide statistics regarding the sensitivity, specificity and area under the curve (AUC) of the model.

Patient selection and TMA preparation

Six paraffin processed TMA blocks containing 555 consecutive primary operable invasive breast carcinomas from patients involved in the Nottingham Tenovus Primary Breast Carcinoma Series between 1986 and 1993, were used as detailed previously [10]. The TMA construction involved sampling donor tissue cores from the tumour periphery and avoiding regions of obvious necrosis. In addition, 160 full face paraffin blocks of breast cancer were selected for comparison because of observed heterogeneity of CA9 distribution using immunohistochemistry. All cases used in this study are well characterised and have data on tissue protein expression for tumour-relevant biomarkers, comprehensive pathology and long term clinical follow-up data [10] including information on local, regional and distant tumour recurrence, and survival outcome. Patients with ER positive tumours were treated with adjuvant endocrine therapy whereas patients with a moderate and poor Nottingham Prognostic Index received chemotherapy.

CA9 protein expression was further validated on a cohort of 245 patients diagnosed and managed at the Royal Marsden Hospital (RMH) between 1994 and 2000. Patients were selected on the basis of being eligible for therapeutic surgery, being followed up at the RMH, having representative histological blocks in the RMH pathology files, and receiving standard anthracycline-based adjuvant chemotherapy. All patients were primarily treated with therapeutic surgery followed by anthracycline-based chemotherapy. Adjuvant endocrine therapy was prescribed for patients with ER positive tumours (tamoxifen alone in 96.4% of the patients for the available follow-up period). Complete follow-up was available for 244 patients, ranging from 0.5 to 125 months (median = 67 months, mean = 67 months). Tumours were graded according to a modified Bloom–Richardson scoring system [26] and size was categorised according to the TNM staging criteria. The project was approved by the Ethics and R&D committees at NUH and RMH.

CA9 immunohistochemistry and morphometry

Four micron thick paraffin-processed TMA and full face sections were subjected to microwave antigen retrieval in citrate buffer (pH 6.0), and then immunohistochemically stained with an antibody against CA9 on a TechMate immunostainer (DakoCytomation, Cambridge, UK). The CA9 rabbit polyclonal antibody (Abcam 15086, Cambridge, UK) was used at an optimised working dilution of 1:2,500 with a labelled streptavidin biotin (LSAB) technique. Sections were counterstained in haematoxylin and mounted using DPX mounting medium. Negative control sections had non-immune serum substituted for the primary antibody and positive control sections comprising high-grade ovarian cancer with necrotic foci were included in each immunohistochemistry run.

The immunohistochemically stained TMA and full face sections were scored with observers blinded to the clinicopathological features of tumours and patients' outcome. Staining was assessed in the cell membrane of morphologically unequivocal neoplastic cells of tumours and in stromal fibroblasts. The presence of CA9 staining in stromal fibroblasts was recorded because it has previously been suggested to be of prognostic significance [27]. Presence of tumour membrane and fibroblast CA9 staining was recorded '1' for affirmative and '0' for negative. Damaged tissue cores and those that did not contain invasive carcinoma were excluded from scoring.

Univariate and multivariate statistics

The Chi square test was used for testing the association between CA9 protein expression and other biomarkers scored as categorical variables, to produce contingency tables (Version 15, SPSS Inc., IL, USA). Similarly, the presence or absence of tumour-associated membranous and normal stromal cell cytoplasmic CA9 staining was categorically scored as positive or negative, regardless of its extent or staining intensity. Kaplan–Meier survival plots were produced to estimate disease-free interval (DFI), breast cancer specific survival (BCSS) and the

development time for metastasis formation. DFI was expressed as the number of months from diagnosis to the occurrence of invasive local recurrence, local LN relapse or distant relapse. Survival rates were compared using the log rank (Mantel–Cox) test. A *P*-value of less than 0.05 was deemed significant with 95% confidence intervals.

Results

Development of a signature to predict development of distant metastasis using ANNs

ANN analysis identified a gene expression signature consisting of nine genes which predicted patient prognosis with 98% sensitivity and 94% specificity, with an AUC of 0.971 when assessed by ROC curve analysis. The overall screening process assessed over eleven million individual models. A summary of performance for the models at each step is shown in Table 1 and Supplementary Fig. 1. To further validate the model, an additional set of 19 samples were downloaded from the same location as the first series and used as a second order validation set, as in the original manuscript [13]. This set consisted of 7 patients who remained metastasis free, and 12 who developed metastases within 5 years. The novel nine gene expression signature correctly diagnosed all 19 samples, further emphasising the models predictive power. The response curves for these genes were also analysed, with seven of the nine having strong discriminatory responses (Supplementary Fig. 2 shows the response curve for CA9. The association between increased expression and development of metastases is clearly seen).

As seen in Table 1, four of the nine genes showed a positive association between increased expression and the probability of developing distant metastases, as output by the model. Of those four genes, CA9 gave the highest accuracy (70%) for predicting metastases. On the contrary, three genes showed an inverse association between increased expression and the predicted likelihood of metastases. In addition, two genes showed a weak response in the predicted probability of developing metastases, possibly modulating the responses of other genes in an additive fashion.

Validation of ANN findings

Since the ANN gene signature was capable of predicting the development of metastases to a high degree, the expression of these genes were further explored and validated using the NKI295 dataset [20] which includes gene expression data for a 295 patient cohort. Using the ANN 9 gene signature to classify this series of cases into two groups showed a significantly reduced overall survival ($P < 0.001$) and metastasis free survival ($P < 0.001$) between groups in univariate Kaplan–Meier analysis (Supplementary Fig. 3). Interestingly, the ANN signature was also able to split the groups defined by the original 70 gene signature into prognostically distinct groups ($P < 0.001$). In a multivariate Cox regression model adjusted for age, nodal status, tumour size, ER status, therapy type (chemotherapy or hormonal) and van't Veer's 70 gene signature, the ANN signature was shown to be an independent predictor of metastasis free survival ($P = 0.003$, Hazard ratio = 1.92) and overall survival ($P = 0.012$, Hazard ratio = 1.89) in this larger cohort (Supplementary Table 1a, b). Furthermore, analysis of CA9 gene expression in the NKI295 dataset showed a significant positive association with tumours of a basal-like phenotype ($P < 0.001$) and an inverse association with luminal type cancers ($P < 0.001$). These findings led us to investigate if

Table 1 Summary of the nine genes used in the gene expression signature at each step of model development

Step	Input added	Gene name	Description	Cumulative accuracy (%)	Error	Response
1	NM_001216	CA9	Carbonic anhydrase IX	70	0.44	Positive
2	Contig52778_RC		EST	80	0.38	Weak
3	Contig35076_RC		EST	83	0.38	Negative
4	Contig40557_RC	FLJ13409	EST	87	0.35	Positive
5	AB032973	LCHN	LCHN protein	80	0.40	Positive
6	AB004064	TMEFF2	Transmembrane protein with EGF-like and two follistatin-like domains 2	95	0.23	Positive
7	NM_006101	HEC/KNTC2	Kinetochore associated 2	95	0.22	Weak
8	AF161451	HSPC333	HSPC337	96	0.17	Negative
9	Contig33475		EST's	98	0.15	Weak

Table details the identity of the input added at each step, the gene name (where known) and description. The model accuracy and error when applied to the independent validation data splits are also shown, together with the direction of response of the gene as it correlated with metastases

our gene expression findings could be translated into a routine immunohistochemistry practice for the principal prognosticator CA9.

CA9 protein distribution in breast tumours within the Nottingham patient cohort

CA9 staining was heterogeneously distributed in the cell membrane of tumour tissue and in the cytoplasm of stromal fibroblast. CA9 staining of tumours was predominantly associated with necrotic glandular foci (Fig. 1a, b) but in contrast, positively stained fibroblasts did not always show close association with necrotic malignant tissue. In TMAs, 552 cores were readable but comparison with the full face sections showed lack of concordance (Supplementary Table 1). Membranous CA9 expression was under-represented in TMAs due to heterogeneity in CA9 localisation and because of avoidance of necrotic regions during TMA construction. Membranous CA9 expression was identified in 26/552 (4.7%) TMA cores compared with 29/160 (18.1%) full face cases. For these reasons, only data from the full face sections was used.

Associations between CA9 expression and other clinicopathological variables

Membranous CA9 staining was significantly increased in younger patients with high histological grade cancers ($P < 0.001$; Table 2). However, membranous CA9 expression showed no significant association with menopausal

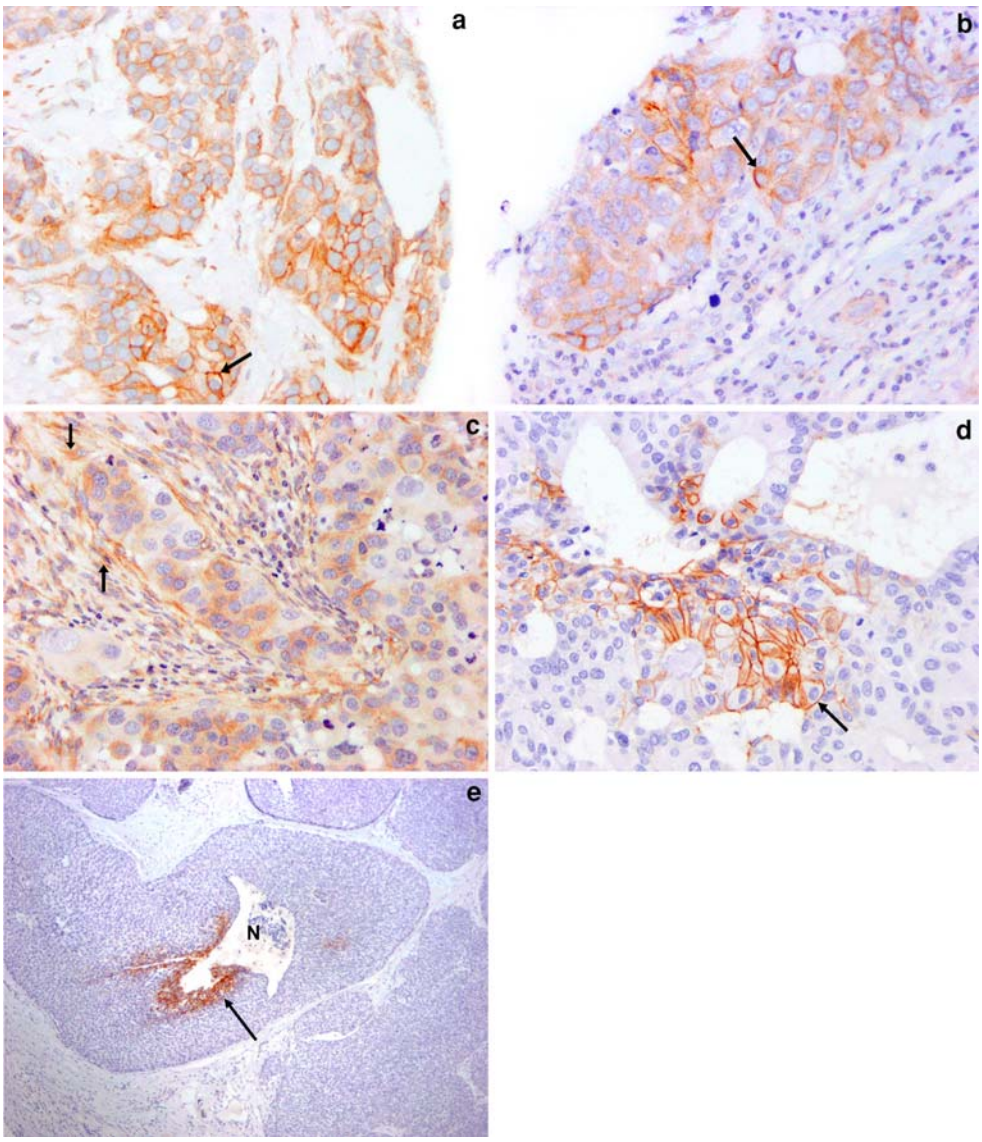

Fig. 1 CA IX immunostaining (*arrow*) was detected in breast tumour cell membrane (**a**), cytoplasm (**b**) and stroma (**c**) in TMA sections but its frequency was reduced due to its heterogeneous localisation. The latter is demonstrated in full face sections of ductal cancer (**d**), especially in cases showing glandular necrosis (*N*) (**e**) associated with hypoxia. Original magnification **a**–**d** 20×; **e** 4×

Table 2 Distribution of patients according to localisation of CA9 immunostaining by age, tumour grade and lymph node involvement

	Cytoplasmic staining	Membrane staining	Fibroblast staining
Age distribution			
Grade 1	51 (39–59)	44 (44)	55.6 (54–57)
Grade 2	54.5 (28–69)	53.5 (51–56)	49 (41–69)
Grade 3	47 (25–66)	48.8 (25–66)	51.1 (28–67)
CA9 distribution (%)	CAIX (+) CAIX (−)	CAIX (+) CAIX (−)	CAIX (+) CAIX (−)
Grade 1	8/89 (95.5) 20/71 (71.8)	1/29 (3.4) 26/131 (19.8)	3/26 (11.5) 24/134 (17.9)
Grade 2	26/89 (30.8) 31/71 (50)	3/29 (10.3) 54/131 (41.2)	7/26 (26.9) 49/134 (36.5)
Grade 3	55/89 (57.5) 20/71 (15)	25/29 (86.2) 49/131 (37.4)	16/26 (61.5) 55/134 (41)
Node involvement (%)			
Lymph node negative	85/89 (95.5) 51/71 (71.8)	28/29 (96.5) 106/131 (80.9)	25/26 (82) 105/134 (78.3)
Lymph node positive	4/89 (4.5) 20/71 (28.2)	1/29 (3.4) 23/131 (17.5)	1/26 (3.8) 23/134 (17.1)

status, tumour size, lymphovascular invasion ($P = 0.056$) or lymph node metastases ($P = 0.051$; Table 3a, b).

Membrane expressing tumours showed a strong negative association with the steroid hormonal receptors (HR) [ER, PgR and androgen receptor ($P < 0.001$ each)], and the luminal cytokeratin CK19 ($P = 0.015$). Importantly, tumours expressing membranous CA9 showed a triple negative phenotype (ER−, PgR−, HER2−) [28] and expressed basal-like markers [CK5/6 ($P = 0.001$), CK14 ($P = 0.02$), BRCA1 nuclear positivity ($P = 0.002$), p53 ($P = 0.001$) and P-cadherin ($P = 0.01$). No association with E-cadherin expression was seen (Table 3a, b).

Fibroblast expression

Stromal fibroblast CA9 staining was seen in 26 (16%) cases; 5 of them showed coexisting membranous and stromal cell CA9 expression (Table 2). Stromal expression showed no significant association with tumour size or menopausal status (Table 3c, d). CA9 expression showed a trend towards association with p53 ($P = 0.06$) and lymph node involvement ($P = 0.051$), but showed no significant associations with the other clinicopathological variables including HR, E-cadherin, HER2, CK56 or CK14 (Table 3c, d).

Survival analysis

No significant association between membranous CA9 immunohistochemical expression in cancer cells or stromal cells was observed with BCSS, DFI or local/regional recurrence.

CA9 protein expression in the validation patient cohort

CA9 protein expression was validated in a cohort of 245 patients, of which 230 could be evaluated for CA9 immunohistochemical expression. Membranous CA9 protein expression was present in 29 cases. Similar to the experimental patient group, the validation cohort showed a significant negative association with ER and PgR expression (Table 3a, b), and was significantly associated with triple negative basal-like tumours ($P < 0.001$). Similar to the Nottingham patient group, the validation group showed no significant association between CA9 expression in cancer cells or fibroblasts and other clinicopathological variables including tumour size, vascular invasion, or patients' outcome in terms of BCSS and DFI. However, CA9 staining in the validation group differed in showing a negative significant association with ER ($P = 0.033$), CK5/6 ($P = 0.01$), and CK14 ($P = 0.001$), and an absence of borderline association with lymph node involvement ($P = 0.268$; Table 3c, d).

Discussion

The aim of our study was to derive a minimal gene expression signature predictive of the outcome of breast cancer patients by applying an ANN approach to analyse a previously published dataset of breast cancer [13]. We hypothesised that this signature would be capable of predicting survival to at least the degree of accuracy obtained in the original study. Using an ANN approach developed specifically for the identification of optimal biomarker subsets in complex data, we found just nine genes were necessary to predict metastatic spread with sensitivity of 98%. This compares favourably with the computational approach used in the original manuscript [13] that resulted in the identification of a prognostic panel comprising 70 genes with a prediction accuracy of 83%. The principal prognostic indicator in our signature was identified as CA9, and this gene correctly predicted metastasis in 70% in the original cohort (van't Veer's) and in 63% of the validation

Table 3 Association between CA9 IHC protein expression with biological markers and clinical parameters assessed in full face sections of breast tumours, according to the cytoplasmic, membranous, or stromal staining pattern of localisation

Parameter	Experimental cohort				Validation cohort			
	Number of samples (%)		χ^2	*P* value	Number of samples (%)		χ^2	*P* value
	CA9(−)	CA9(+)			CA9(−)	CA9(+)		
a								
Tumour size								
Small	59 (88.1)	8 (11.9)	3.194	0.074	180 (86.5)	28 (13.5)	3.117	0.210
Large	70 (76.9)	21 (23.1)			12 (85.7)	2 (14.3)		
Menopausal status								
Premenopausal	41 (74.5)	14 (25.5)	4.567	0.102	–	–	–	–
Postmenopausal	72 (85.7)	12 (14.3)			–	–	–	–
ER								
Negative	24 (53.3)	21 (46.7)	27.196	<0.001	28 (66.6)	14 (33.3)	16.946	<0.001
Positive	87 (91.6)	8 (8.4)			161 (90.9)	16 (9.1)		
PgR								
Negative	38 (64.4)	21 (35.6)	13.746	<0.001	41 (74.5)	14 (25.5)	8.586	.003
Positive	73 (90.1)	8 (9.9)			148 (85)	16 (15)		
AR								
Negative	23 (56.1)	18 (43.9)	16.401	<0.001	–	–	–	–
Positive	79 (87.8)	11 (12.2)			–	–	–	–
P-cadherin								
Negative	53 (86.9)	8 (13.1)	4.110	0.043	–	–	–	–
Positive	56 (72.7)	21 (27.3)			–	–	–	–
E-cadherin								
Negative	47 (87.0)	7 (13.0)	2.984	0.084	60 (85.7)	10 (14.3)	0.055	0.973
Positive	66 (75.0)	22 (25.0)			111 (86.7)	17 (13.3)		
b								
c-erbb2								
Negative	96 (79.3)	25 (20.7)	0.283	0.413	161 (87.5)	23 (12.5)	2.003	0.157
Positive	21 (84.0)	4 (16.0)			25 (78.1)	7 (21.9)		
CK5/6								
Negative	111 (87.4)	16 (12.6)	27.806	<0.001	172 (89.5)	20 (10.5)	11.685	0.001
Positive	8 (38.1)	13 (61.9)			14 (63.6)	8 (36.4)		
CK14								
Negative	103 (83.1)	21 (16.9)	6.456	0.011	179 (89)	22 (11)	13.089	<0.001
Positive	11 (57.9)	8 (42.1)			12 (60)	8 (40)		
P53								
Negative	91 (87.5)	13 (12.5)	14.276	<0.001	130 (86.6)	20 (13.4)	0.207	0.649
Positive	23 (59.0)	16 (41.0)			48 (84.2)	9 (15.8)		
Vascular invasion								
Absent	87 (79.8)	22 (20.2)	1.353	0.508	61 (83.5)	12 (16.4)	0.762	0.408
Present	42 (87.5)	6 (12.5)			130 (87.8)	18 (12.2)		
Lymph node involvement								
Absent	106 (79.1)	28 (20.9)	3.801	0.051	68 (85)	12 (15)	0.481	0.531
Present	23 (95.8)	1 (4.2)			121 (88.3)	16 (11.7)		
Tumour recurrence								
Absent	112 (83.6)	22 (16.4)	2.208	0.137	–	–	–	–
Present	17 (70.8)	7 (29.2)			–	–	–	–
Overall survival	–	–	2.976	0.085	–	–	1.310	0.253
DFI	–	–	2.756	0.097	–	–	2.870	0.093

Table 3 continued

Parameter	Experimental cohort				Validation cohort			
	Number of samples (%)		χ^2	*P* value	Number of samples (%)		χ^2	*P* value
	CA9(−)	CA9(+)			CA9(−)	CA9(+)		
c								
Tumour size								
Small	58 (86.6)	9 (13.4)	0.827	0.363	170 (81.7)	38 (18.3)	3.703	0.157
Large	73 (81.1)	17 (18.9)			14 (100)	0 (0)		
Menopausal status								
Premenopausal	42 (76.4)	13 (23.6)	4.567	0.102	–	–	–	–
Postmenopausal	71 (85.5)	12 (14.5)			–	–	–	–
ER								
Negative	36 (80.0)	9 (20.0)	0.183	0.669	30 (71.4)	12 (28.6)	4.562	0.033
Positive	78 (114)	16 (17.0)			151 (85.3)	26 (14.7)		
PgR								
Negative	51 (86.4)	8 (13.6)	0.986	0.321	42 (23.2)	139 (76.8)	2.023	0.155
Positive	64 (80.0)	16 (20.0)			13 (34.2)	25 (65.8)		
AR								
Negative	34 (85.0)	6 (15.0)	0.288	0.592	–	–	–	–
Positive	73 (81.1)	17 (18.9)			–	–	–	–
P-cadherin								
Negative	54 (88.5)	7 (11.5)	1.258	0.262	–	–	–	–
Positive	62 (81.6)	14 (18.4)			–	–	–	–
E-cadherin								
Negative	23 (82.1)	5 (17.9)	0.029	0.865	163 (92.6)	13 (7.4)	1.692	0.429
Positive	96 (83.5)	19 (16.5)			106 (82.8)	22 (17.2)		
d								
c-erbb2								
Negative	9,100 (83.3)	20 (16.7)	0.161	0.688	151 (82)	33 (18)	0.100	0.751
Positive	20 (80.0)	5 (20.0)			27 (84.4)	5 (15.6)		
CK5/6								
Negative	104 (82.5)	22 (17.5)	0.129	0.72	164 (85.4)	28 (14.6)	6.692	0.010
Positive	18 (85.7)	3 (14.3)			14 (63.6)	8 (36.4)		
CK14								
Negative	101 (82.1)	22 (17.9)	1.932	0.165	173 (86)	28 (14)	16.621	<0.001
Positive	18 (94.7)	1 (5.3)			10 (50)	10 (50)		
P53								
Negative	90 (87.4)	13 (12.6)	3.533	0.060	129 (86)	21 (14)	5.570	0.025
Positive	29 (74.4)	10 (25.6)			41 (71.9)	16 (28.1)		
Vascular invasion								
Present	91 (83.5)	18 (16.5)	1.482	0.477	123 (83.1)	25 (16.9)	0.029	0.865
Absent	39 (83)	8 (17)			60 (82.2)	13 (17.8)		
Lymph node involvement								
Absent	106 (79.1)	28 (20.9)	3.801	0.051	63 (78.8)	17 (21.2)	1.226	0.268
Present	23 (95.8)	1 (4.2)			116 (88.5)	21 (11.5)		
Tumour recurrence								
Absent	113 (85.0)	20 (15.0)	1.460	0.227	–	–	–	–
Present	18 (75.0)	6 (25.0)			–	–	–	–
Overall survival	–	–	1.989	0.158	–	–	0.120	0.7280
DFI	–	–	1.431	0.232	–	–	0.700	0.4034

P values refer to χ^2 or log rank test for overall survival. Significance level = <0.05

cohort [20]. In this validation cohort, the ANN 9 gene signature was showed to be an independent predictor of both metastasis free and overall survival, and interestingly, was able to split the groups defined by the original 70 gene signature into prognostically distinct groups.

A further aim of our study was to investigate if our ANN-derived minimal gene panel for predicting poor prognosis in breast cancer could be successfully translated into routine practice. To test this, we studied the immunohistochemical localisation of the principle prognosticator CA9 in unselected breast cancer. Carbonic anhydrases are induced by hypoxia induced factor 1 alpha (HIF-1α) and assist cancer cells in avoiding death by neutralising acid pH conditions associated with hypoxia-induced glycolysis. Furthermore, it has been proposed that CA9 promotes tumour migration and invasion via its role in extracellular matrix degradation and through the induction of growth factors [29]. These important roles suggest that not only is CA9 a key candidate prognostic biomarker for determining clinical outcome, but because of its resistance to degradation, it could be a more robust marker of hypoxia than HIF-1α protein [30]. Previously, a number of studies have shown that over-expression of CA9 is functionally important in several tumour types including colorectal [31], cervical [32] and uterine [33] cancers, and sarcomas [34]. Although the contribution of CA9 as a prognostic marker in breast cancer has been obscured by conflicting reports, some authors [35] demonstrated that its expression is associated with tumours characterised by a basal-like phenotype and showing reduced patients' survival, emphasising the relationship between CA9 expression and poor prognosis.

In this study, we found membranous expression of CA9 is associated with tumours showing aggressive features including younger patients' age, high grade ductal cancers, basal-like phenotype (CK5/6+, CK14+; ER−, PgR−, HER2−) and BRCA1 positivity. Such patients showed a tendency towards reduced breast cancer specific survival and disease free interval even in the absence of lymph node involvement. It should be noted, however, that immunohistochemical expression of CA9 was not significantly associated with outcome of breast cancer patients.

Immunohistochemical assessment of CA9 was shown to be heterogeneously distributed and was frequently associated with regions showing necrotic foci. Donor tissue used in TMA construction specifically avoided necrotic regions resulting in under-representation of CA9 expression. For this reason, results of full face sections were considered in our study. Supporting our concern about the unsuitability of TMAs for studying CA9 expression, Brennan et al. [35] also identified a reduced frequency (11%) of membranous expression in TMAs when compared with larger samples of tumours.

In agreement with others [27, 36, 37], CA9 expression was identified in the cell membrane of tumour cells and in the cytoplasm of stromal fibroblast cells. The experimental and validation patient cohorts were concordant for membrane staining. In agreement with other studies [35, 38] our data provide further evidence that CA9 occurs in tumours with features of aggressive clinical behaviour, including loss of hormonal receptors, showing poor response to adjuvant endocrine therapy [38].Previously, it was reported that hypoxia can down-regulate ER expression via transcriptional nuclear factors and this might explain the observation seen in the current study [39]. In addition, hypoxia is reported to promote basal tumour-like features (ER−/HER2-negative, CK5-positive) due to up-regulation of SLUG gene expression [40]. Here, our data showed that 62% membrane CA9-expressing tumours significantly associate with the basal markers CK5/6 [41], and have a triple negative phenotype (TNP) [28], supporting the recent findings of Van den Eynden et al. [42]. More recently, it has been proposed that the use of five immunohistochemical markers (ER−, PgR−, HER2−, CK5/6+, EGFR+) can identify a basal subgroup with a worse prognosis (10 year BCSS, 62%) than that seen in TNP (10 year BCSS, 67%) [43]. We showed that 12/29 (41.3%) cases of membranous CA9 fall in the five marker subgroup and, similar to Nielsen et al. [43], we found no lymph node involvement despite their poor prognosis. In addition, 16/26 (61.5%) of the membrane CA9 group were positive for BRCA1 nuclear IHC positivity [44].

The biological significance of CA9 localization in fibroblasts is not readily understood but it has been proposed that it might be caused by the effect of HIF-1α induction factors in these cells due to reasons other than hypoxia [45]. Further work is required to explore the significance of fibroblast CA9 staining.

Other genes identified in our expression signature were more compatible with a tumour suppressor function, including *TMEFF2* and *HEC*. *TMEFF2* encodes for a transmembrane protein containing an epidermal growth factor (EGF)-like motif and two follistatin domains. Our data showed a negative correlation between *TMEFF2* expression and the development of distant metastases, supporting the study of Gery et al. [46] who showed that *TMEFF2* could suppress the growth of prostate cancer cells. More recently [47], it was proposed that *TMEFF2* suppression may contribute to the oncogenic properties of c-Myc, thereby promoting cell proliferation, differentiation, and apoptosis. *HEC* (also known as kinetochore-associated 2), was shown here to be associated with metastases with increased expression. Similar findings have been reported [48] where *HEC* was identified as part of an 11 gene signature predictive of disease recurrence and distant metastasis in prostate and breast cancer.

Furthermore, elevated *HEC* expression has been shown to be associated with poorer prognosis in non-small cell lung carcinomas [49], and therefore a potential target for treatment of cancers, highlighted further still by Gurzov and Izquierdo [50]. Four of the nine genes identified in our panel represent expressed sequence tags (EST's) and the associated gene is therefore of unknown function. However, given their predictive capability with regard to survival, further analysis is justified.

To conclude, using powerful ANN methodologies, we have identified a minimal gene signature that is predictive of outcome at least with a similar degree of accuracy to that obtained in van't Veer's study [13]. Interestingly, this gene signature was shown to have a similar accuracy in predicting the development of metastasis and to be an independent predictor of outcome (metastasis free and overall survival) in a larger validation series from the same group [21]. Moreover, using immunohistochemistry we confirmed its practical and translational application. In agreement with van't Veer et al. [13] we have shown that whilst single genes are capable of discriminating between different disease states, multiple genes in combination enhance the predictive power of these models. Our signature predicted the hypoxic marker CA9 as the principal indicator of poor clinical outcome and although assessment of CA9 protein expression showed no significant association with patients' outcome when compared with our prediction gene panelta, CA9 expression showed association with variables of poor prognosis and aggressive behaviour. In particular, CA9 is associated with basal-like and triple negative cancers. Further studies of all nine genes in combination using immunohistochemistry are warranted to assess the prognostic value of this signature in routine practice.

Acknowledgments This study was supported by a grant from the UK HEFCE and ENACT (The European Network for the identification and validation of antigens and biomarkers in cancer and their application in clinical tumour immunology) and the Breast Cancer Campaign (2005). The authors thank Dr. Kay Savage for production of the Royal Marsden tumour tissue sections. Thanks also to the John and Lucille Van Geest Foundation.

References

1. Brown PO, Botstein D (1999) Exploring the new world of the genome with DNA microarrays. Nat Genet 21((1)(Suppl)):33–37. doi:10.1038/4462
2. Bhattacharjee A, Richards WG, Staunton J et al (2001) Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. Proc Natl Acad Sci USA 98(24):13790–13795. doi:10.1073/pnas.191502998
3. Khan J, Wei JS, Ringner M et al (2001) Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat Med 7(6):673–679. doi:10.1038/89044
4. Rosenwald A, Wright G, Chan WC et al (2002) The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. N Engl J Med 346(25):1937–1947. doi:10.1056/NEJMoa012914
5. Sorlie T, Perou CM, Tibshirani R et al (2001) Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc Natl Acad Sci USA 98(19):10869–10874. doi:10.1073/pnas.191367098
6. West M, Blanchette C, Dressman H et al (2001) Predicting the clinical status of human breast cancer by using gene expression profiles. Proc Natl Acad Sci USA 98(20):11462–11467. doi:10.1073/pnas.201162998
7. Callagy G, Cattaneo E, Daigo Y et al (2003) Molecular classification of breast carcinomas using tissue microarrays. Diagn Mol Pathol 12(1):27–34. doi:10.1097/00019606-200303000-00004
8. Paik S, Shak S, Tang G et al (2004) A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. N Engl J Med 351(27):2817–2826. doi:10.1056/NEJMoa041588
9. Perou CM, Sorlie T, Eisen MB et al (2000) Molecular portraits of human breast tumours. Nature 406(6797):747–752. doi:10.1038/35021093
10. Abd El-Rehim DM, Ball G, Pinder SE et al (2005) High-throughput protein expression analysis using tissue microarray technology of a large well-characterised series identifies biologically distinct classes of breast cancer confirming recent cDNA expression analyses. Int J Cancer 116(3):340–350. doi:10.1002/ijc.21004
11. Masters JR, Lakhani SR (2000) How diagnosis with microarrays can help cancer patients. Nature 404(6781):921. doi:10.1038/35010139
12. Naderi A, Teschendorff AE, Barbosa-Morais NL et al (2007) A gene-expression signature to predict survival in breast cancer across independent data sets. Oncogene 26(10):1507–1516. doi:10.1038/sj.onc.1209920
13. van 't Veer LJ, Dai H, van de Vijver MJ et al (2002) Gene expression profiling predicts clinical outcome of breast cancer. Nature 415(6871):530–536. doi:10.1038/415530a
14. Fan C, Oh DS, Wessels L et al (2006) Concordance among gene-expression-based predictors for breast cancer. N Engl J Med 355(6):560–569. doi:10.1056/NEJMoa052933
15. Wirapati P, Sotiriou C, Kunkel S et al (2008) Meta-analysis of gene expression profiles in breast cancer: toward a unified understanding of breast cancer subtyping and prognosis signatures. Breast Cancer Res 10(4):R65. doi:10.1186/bcr2124
16. Lancashire LJ, Rees RC, Ball GR (2008) Identification of gene transcript signatures predictive for estrogen receptor and lymph node status using a stepwise forward selection artificial neural network modelling approach. Artif Intell Med 43(2):99–111. doi:10.1016/j.artmed.2008.03.001
17. Ball G, Mian S, Holding F et al (2002) An integrated approach utilizing artificial neural networks and SELDI mass spectrometry for the classification of human tumours and rapid identification of potential biomarkers. Bioinformatics 18(3):395–404. doi:10.1093/bioinformatics/18.3.395
18. Lancashire L, Schmid O, Shah H et al (2005) Classification of bacterial species from proteomic data using combinatorial approaches incorporating artificial neural networks, cluster analysis and principal components analysis. Bioinformatics 21(10):2191–2199. doi:10.1093/bioinformatics/bti368
19. Matharoo-Ball B, Ball G, Rees R (2007) Clinical proteomics: discovery of cancer biomarkers using mass spectrometry and bioinformatics approaches—a prostate cancer perspective. Vaccine 25(Suppl 2):B110–B121

20. van de Vijver MJ, He YD, van 't Veer LJ et al (2002) A gene-expression signature as a predictor of survival in breast cancer. N Engl J Med 347(25):1999–2009. doi:10.1056/NEJMoa021967
21. Rumelhart DE, Hinton GE, Williams RJ (1986) Learning representations by back-propagating errors. Nature 323:533–536
22. Picard RR, Cook RD (1984) Cross-validation of regression models. J Am Stat Assoc 79(387):575–583
23. Xu QS, Liang YZ, Du YP (2004) Monte Carlo cross-validation for selecting a model and estimating the prediction error in multivariate calibration. J Chemometr 18(2):112–120. doi:10.1002/cem.858
24. Shao J (1993) Linear model selection by cross-validation. J Am Stat Assoc 88(422):486–494. doi:10.2307/2290328
25. Efron B (1986) How biased is the apparent error rate of a prediction rule? J Am Stat Assoc 81(394):461–470. doi:10.2307/2289236
26. Bloom HJ, Richardson WW (1957) Histological grading and prognosis in breast cancer; a study of 1409 cases of which 359 have been followed for 15 years. Br J Cancer 11(3):359–377
27. Colpaert CG, Vermeulen PB, Fox SB et al (2004) The presence of a fibrotic focus in invasive breast carcinoma correlates with the expression of carbonic anhydrase IX and is a marker of hypoxia and poor prognosis. Breast Cancer Res Treat 81(2):137–147. doi:10.1023/A:1025702330207
28. Nielsen TO, Hsu FD, Jensen K et al (2004) Immunohistochemical and clinical characterization of the basal-like subtype of invasive breast carcinoma. Clin Cancer Res 10(16):5367–5374. doi:10.1158/1078-0432.CCR-04-0220
29. Raghunand N, He X, van Sluis R et al (1999) Enhancement of chemotherapy by manipulation of tumour pH. Br J Cancer 80(7):1005–1011. doi:10.1038/sj.bjc.6690455
30. van Berkel M, van der Groep P, Schvarts A et al (2002) HIF-1alpha/CAIX coexpression in invasive human breast cancer. Breast Cancer Res Treat 76(Suppl 1):S146
31. Saarnio J, Parkkila S, Parkkila AK et al (1998) Immunohistochemical study of colorectal tumors for expression of a novel transmembrane carbonic anhydrase, MN/CA IX, with potential value as a marker of cell proliferation. Am J Pathol 153(1):279–285
32. Liao SY, Brewer C, Zavada J et al (1994) Identification of the MN antigen as a diagnostic biomarker of cervical intraepithelial squamous and glandular neoplasia and cervical carcinomas. Am J Pathol 145(3):598–609
33. Hockel M, Schlenger K, Aral B et al (1996) Association between tumor hypoxia and malignant progression in advanced cancer of the uterine cervix. Cancer Res 56(19):4509–4515
34. Brizel DM, Scully SP, Harrelson JM et al (1996) Tumor oxygenation predicts for the likelihood of distant metastases in human soft tissue sarcoma. Cancer Res 56(5):941–943
35. Brennan DJ, Jirstrom K, Kronblad A et al (2006) CA IX is an independent prognostic marker in premenopausal breast cancer patients with one to three positive lymph nodes and a putative marker of radiation resistance. Clin Cancer Res 12(21):6421–6431. doi:10.1158/1078-0432.CCR-06-0480
36. Chia SK, Wykoff CC, Watson PH et al (2001) Prognostic significance of a novel hypoxia-regulated marker, carbonic anhydrase IX, in invasive breast carcinoma. J Clin Oncol 19(16):3660–3668
37. Tomes L, Emberley E, Niu Y et al (2003) Necrosis and hypoxia in invasive breast carcinoma. Breast Cancer Res Treat 81(1):61–69. doi:10.1023/A:1025476722493
38. Trastour C, Benizri E, Ettore F et al (2007) HIF-1alpha and CA IX staining in invasive breast carcinomas: prognosis and treatment outcome. Int J Cancer 120(7):1451–1458. doi:10.1002/ijc.22436
39. Kronblad A, Hedenfalk I, Nilsson E et al (2005) ERK1/2 inhibition increases antiestrogen treatment efficacy by interfering with hypoxia-induced downregulation of ERalpha: a combination therapy potentially targeting hypoxic and dormant tumor cells. Oncogene 24(45):6835–6841. doi:10.1038/sj.onc.1208830
40. Storci G, Sansone P, Trere D et al (2008) The basal-like breast carcinoma phenotype is regulated by SLUG gene expression. J Pathol 214(1):25–37. doi:10.1002/path.2254
41. Fadare O, Tavassoli FA (2007) The phenotypic spectrum of basal-like breast cancers: a critical appraisal. Adv Anat Pathol 14(5):358–373. doi:10.1097/PAP.0b013e31814b26fe
42. Van den Eynden GG, Smid M, Van Laere SJ et al (2008) Gene expression profiles associated with the presence of a fibrotic focus and the growth pattern in lymph node-negative breast cancer. Clin Cancer Res 14(10):2944–2952. doi:10.1158/1078-0432.CCR-07-4397
43. Cheang MC, Voduc D, Bajdik C et al (2008) Basal-like breast cancer defined by five biomarkers has superior prognostic value than triple-negative phenotype. Clin Cancer Res 14(5):1368–1376. doi:10.1158/1078-0432.CCR-07-1658
44. Turner NC, Reis-Filho JS (2006) Basal-like breast cancer and the BRCA1 phenotype. Oncogene 25(43):5846–5853. doi:10.1038/sj.onc.1209876
45. Vleugel MM, Greijer AE, Shvarts A et al (2005) Differential prognostic impact of hypoxia induced and diffuse HIF-1alpha expression in invasive breast cancer. J Clin Pathol 58(2):172–177. doi:10.1136/jcp.2004.019885
46. Gery S, Sawyers CL, Agus DB et al (2002) TMEFF2 is an androgen-regulated gene exhibiting antiproliferative effects in prostate cancer cells. Oncogene 21(31):4739–4746. doi:10.1038/sj.onc.1205142
47. Gery S, Koeffler HP (2003) Repression of the TMEFF2 promoter by c-Myc. J Mol Biol 328(5):977–983. doi:10.1016/S0022-2836(03)00404-2
48. Glinsky GV, Berezovska O, Glinskii AB (2005) Microarray analysis identifies a death-from-cancer signature predicting therapy failure in patients with multiple types of cancer. J Clin Invest 115(6):1503–1521. doi:10.1172/JCI23412
49. Hayama S, Daigo Y, Kato T et al (2006) Activation of CDCA1-KNTC2, members of centromere protein complex, involved in pulmonary carcinogenesis. Cancer Res 66(21):10339–10348. doi:10.1158/0008-5472.CAN-06-2137
50. Gurzov EN, Izquierdo M (2006) RNA interference against Hec1 inhibits tumor growth in vivo. Gene Ther 13(1):1–7. doi:10.1038/sj.gt.3302595