Title

Delineating transcriptional networks of prognostic gene signatures refines treatment recommendations for lymph node-negative breast cancer patients

Journal

The FEBS Journal, 282(18)

ISSN

1742-464X

Authors

Lanigan, Fiona

Brien, Gerard L

Fan, Yue

et al.

Publication Date

2015-09-01

DOI

10.1111/febs.13354

Peer reviewed

the FEBS Journal

Delineating transcriptional networks of prognostic gene signatures refines treatment recommendations for lymph node-negative breast cancer patients

Fiona Lanigan[1,*], Gerard L. Brien[1,*], Yue Fan[2], Stephen F. Madden[2], Emilia Jerman[1], Ashwini Maratha[2], Fatima Aloraifi[1], Karsten Hokamp[1], Eiseart J. Dunne[1], Amanda J. Lohan[2], Louise Flanagan[3], James C. Garbe[4], Martha R. Stampfer[4], Marie Fridberg[5], Karin Jirstrom[5], Cecily M. Quinn[3], Brendan Loftus[2], William M. Gallagher[2], James Geraghty[3] and Adrian P. Bracken[1]

1 The Smurfit Institute of Genetics, Trinity College Dublin, Ireland
2 UCD School of Biomolecular and Biomedical Science, UCD Conway Institute, University College Dublin, Ireland
3 Department of Histopathology, St Vincent's University Hospital, Dublin, Ireland
4 Life Science Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA
5 Department of Clinical Sciences, Division of Oncology and Pathology, Lund University, Sweden

Keywords
breast cancer; cellular senescence; Master Transcriptional Regulators; OncoMasTR; proliferation

Correspondence
A. P. Bracken, Smurfit Institute of Genetics, Trinity College Dublin, Dublin 2, Ireland
Fax: +353 1 8964121
Tel: +353 1 8964121
E-mail: adrian.bracken@tcd.ie

*These authors contributed equally to this work.

(Received 8 May 2015, revised 2 June 2015, accepted 17 June 2015)

doi:10.1111/febs.13354

The majority of women diagnosed with lymph node-negative breast cancer are unnecessarily treated with damaging chemotherapeutics after surgical resection. This highlights the importance of understanding and more accurately predicting patient prognosis. In the present study, we define the transcriptional networks regulating well-established prognostic gene expression signatures. We find that the same set of transcriptional regulators consistently lie upstream of both 'prognosis' and 'proliferation' gene signatures, suggesting that a central transcriptional network underpins a shared phenotype within these signatures. Strikingly, the master transcriptional regulators within this network predict recurrence risk for lymph node-negative breast cancer better than currently used multigene prognostic assays, particularly in estrogen receptor-positive patients. Simultaneous examination of $p16^{INK4A}$ expression, which predicts tumours that have bypassed cellular senescence, revealed that intermediate levels of $p16^{INK4A}$ correlate with an intact pRB pathway and improved survival. A combination of these master transcriptional regulators and $p16^{INK4A}$, termed the OncoMasTR score, stratifies tumours based on their proliferative and senescence capacity, facilitating a clearer delineation of lymph node-negative breast cancer patients at high risk of recurrence, and thus requiring chemotherapy. Furthermore, OncoMasTR accurately classifies over 60% of patients as 'low risk', an improvement on existing prognostic assays, which has the potential to reduce overtreatment in early-stage patients. Taken together, the present study provides new insights into the transcriptional regulation of cellular proliferation in breast cancer and provides an opportunity to enhance and streamline methods of predicting breast cancer prognosis.

Abbreviations
ChIP, chromatin immunoprecipitation; DMSF, distant metastasis-free survival; ER, estrogen receptor; HMEC, human mammary epithelial cell; IHC, immunohistochemistry; LN, lymph node; MEF, mouse embryonic fibroblast; MTR, master transcriptional regulator; qPCR, quantitative real-time PCR; RFS, recurrence-free survival; ROC, receiver operating characteristic; TMA, tissue microarray.

Introduction

Accurately predicting patient prognosis in early-stage, lymph node (LN) negative breast cancers is a major challenge for clinicians. The widespread use of screening programmes means that increasing numbers of women are being diagnosed at an early stage with the disease [1]. In a positive testament to the early detection measures, only 10–30% of women with early-stage, LN-negative disease will go on to develop a recurrence after surgical resection [2,3]. However, the corollary of this is that the majority of women diagnosed with early-stage breast cancer are unnecessarily treated with chemotherapy [4]. Therefore, given the debilitating nature of chemotherapy on patient health, it is crucial to be able to more accurately delineate LN-negative patients into those that require this treatment, and the majority of patients who do not, thereby eliminating unnessary treatment burden.

In the past 10–12 years, a number of studies have defined so-called breast cancer 'poor prognosis' expression signatures, containing tens or even hundreds of individual genes [5–13]. Several of these prognosis signatures have been adopted in the clinic and proven useful in terms of stratifying patients into lower and higher risk groups [14,15]. Intriguingly, despite the ability of several different signatures to predict breast cancer outcome, there is little overlap between the genes [16,17]. This raises the question: do these apparently disparate gene sets represent a particular shared phenotypic contribution to tumour biology, and hence patient prognosis?

Previous studies have suggested that the underlying biological principal driving the prognostic performance of these markers is proliferation [16,18]. We hypothesized that identifying the upstream master transcriptional regulators (MTRs) of breast cancer poor prognosis gene expression signatures would allow us to better understand their phenotypic contribution to tumour biology, and thus improve treatment recomendations. Towards this goal, we identified a shared transcriptional network upstream of two well-validated breast cancer prognostic signatures: the 70-gene signature or 'MammaPrint' [10] and the 'Genomic Grade' signature [9]. Notably, the same MTRs were shared between both signatures, suggesting that these prognostic signatures represent a common tumour phenotype. Supporting this, we show that the normal role of the upstream MTRs is to directly regulate the promoters of a set of 'core proliferation' genes, many of which are highly enriched within breast cancer prognostic signatures. The levels of these MTRs in breast tumours, together with $p16^{INK4A}$, a key tumour suppressor deregulated in tumours that have bypassed the cellular senescence checkpoint, are strong predictors of recurrence risk, and provide an improvement upon currently used prognostics such as Ki67, and surrogate estimates of the MammaPrint [10] and Oncotype Dx [7] multigene signatures. Importantly, this combination, called the 'OncoMasTR score', accurately classifies increased numbers of patients in the 'low risk' group. Therefore, the present study provides new insights into the transcriptional regulation of cellular proliferation in breast cancer and stands to contribute to reduction of overtreatment of patients with early stage, LN-negative breast cancer.

Results

Delineating the transcriptional network of breast cancer prognostic signatures identifies upstream MTRs

To identify the MTRs upstream of breast cancer prognosis gene signatures, we used a bioinformatic approach called ARACNe [19,20]. This approach uses networks constructed from gene expression datasets to infer direct transcriptional interactions, without using survival information. We applied ARACNe to three independent, publicly available breast cancer gene-expression datasets [11,21,22], and used the resulting interaction networks to predict the upstream MTRs of two independent breast cancer prognosis gene signatures: the 'Genomic Grade [9] and '70-gene' [10] signatures. Remarkably, the results of these analyses were very similar, with FOXM1, UHRF1, PTTG1, E2F1, MYBL2 and HMGB2 among the top ten scoring MTRs (Fig. 1A and Table 1), suggesting that a central transcriptional network underpins both prognostic signatures.

Next, we aimed to explore the possibility that the levels of the prognosis-linked MTRs would provide a more accurate prediction than measuring their downstream 'passenger' genes. We examined the association of each individual MTR with patient survival in a combined dataset of three published microarray studies representing the genome-wide mRNA expression of 457 LN-negative chemotherapy-naïve breast tumours [11,21,23]. This revealed that high mRNA expression levels of any of the six MTRs in breast tumours was significantly associated with reduced recurrence-free survival (RFS) time, whereas FOXM1, MYBL2, UHRF1 and PTTG1 were better predictors of outcome than the established prognostic marker Ki67 (Figs 1B and 2). Strikingly, a combination of just six MTRs was more powerful at stratifying the patients compared to the 61 genes within the 'Poor Prognosis' signature, which constitute the MammaPrint '70-gene

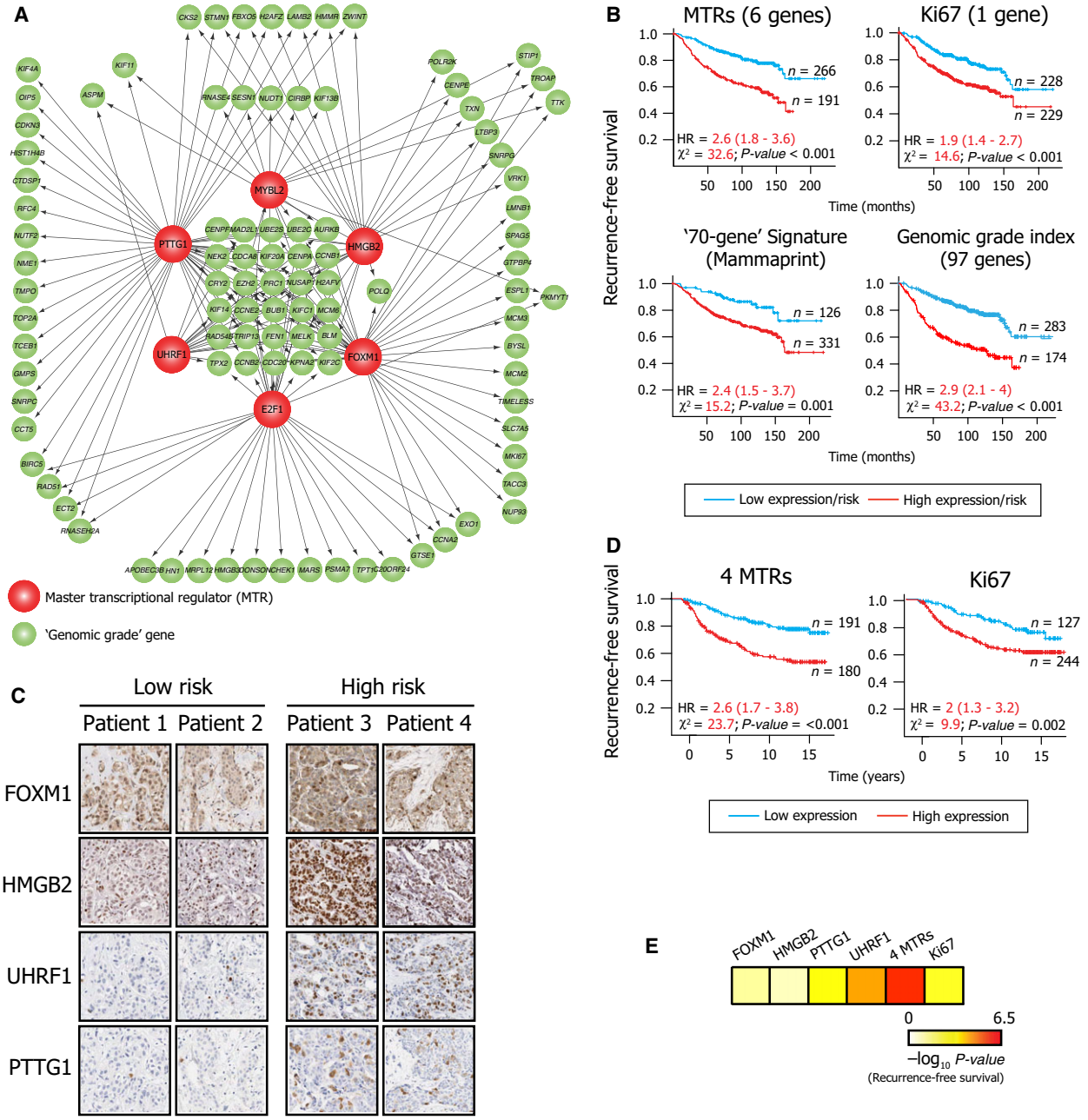

Fig. 1. The transcriptional network of breast cancer prognosis genes. (A) Representative ARACNe network for the 'Genomic Grade' signature [9] showing the predicted upstream MTRs. The upstream MTRs are shown in red and the downstream 'Genomic Grade' genes are shown in green. The Genomic grade genes predicted to be downstream of three or more MTRs are shown in the centre. (B) Kaplan–Meier analyses of a combination of the six MTRs from (A), the proliferation marker Ki67, the 61 genes that constitute the 'Mammaprint' signature and the 97 genes that constitute the 'Genomic Grade' signature. The analysis was performed on LN-negative samples without chemotherapy, in three microarray ($n = 457$) datasets combined [11,21,23]. (C) Representative immunohistochemical staining for the indicated MTRs in low and high-risk tumours on a breast cancer TMA. Low-risk tumours were defined as those that did not recur within the study timeframe, whereas high-risk tumours did recur. (D) Kaplan–Meier analyses of a combination of the four MTRs from (C), and Ki67 on TMA tumours from patients not treated with chemotherapy ($n = 371$). (E) Heat map illustrating the prognostic power of FOXM1, UHRF1, HMGB2 and PTTG1 alone, these four MTRs combined, and Ki67, as shown in (D). The scale represents $-\log_{10}$ of the P values calculated using the log-rank test.

Table 1. Top ranking MTRs of the indicated expression signatures as predicted by ARACNe analysis on three independent breast cancer datasets [11,21,22].

Rank	Poor prognosis signature [10]	Genomic grade signature [9]	Core proliferation signature (identified in-house)
1	PTTG1	PTTG1	FOXM1
2	FOXM1	FOXM1	PTTG1
3	UHRF1	UHRF1	UHRF1
4	ATAD2	MYBL2	MYBL2
5	MYBL2	ATAD2	HMGB2
6	ZNF367	HMGB2	ATAD2
7	HMGB2	ZBTB20	E2F1
8	TCF19	E2F1	E2F8
9	E2F8	E2F8	ZNF367
10	E2F1	ZNF367	TCF19

signature' (Fig. 1B). Moreover, the 97 genes within the 'Genomic Grade' signature, which constitute the 'Genomic Grade Index', had only slightly better prognostic capabilities than the six MTR genes. Taken together, these comparisons suggest that just six upstream MTRs provide superior, or at least similar, prognostic information compared to the downstream signatures they are predicted to regulate.

We next examined the protein levels of the MTRs in an independent breast cancer patient cohort via immunohistochemistry (IHC). Antibodies were screened for all six MTRs and four identified that specifically recognized FOXM1, HMGB2, PTTG1 and UHRF1 (Optimization data available on request). Tissue microarrays (TMAs) representing 498 invasive breast tumours were evaluated for the protein levels of each MTR (Fig. 1C).

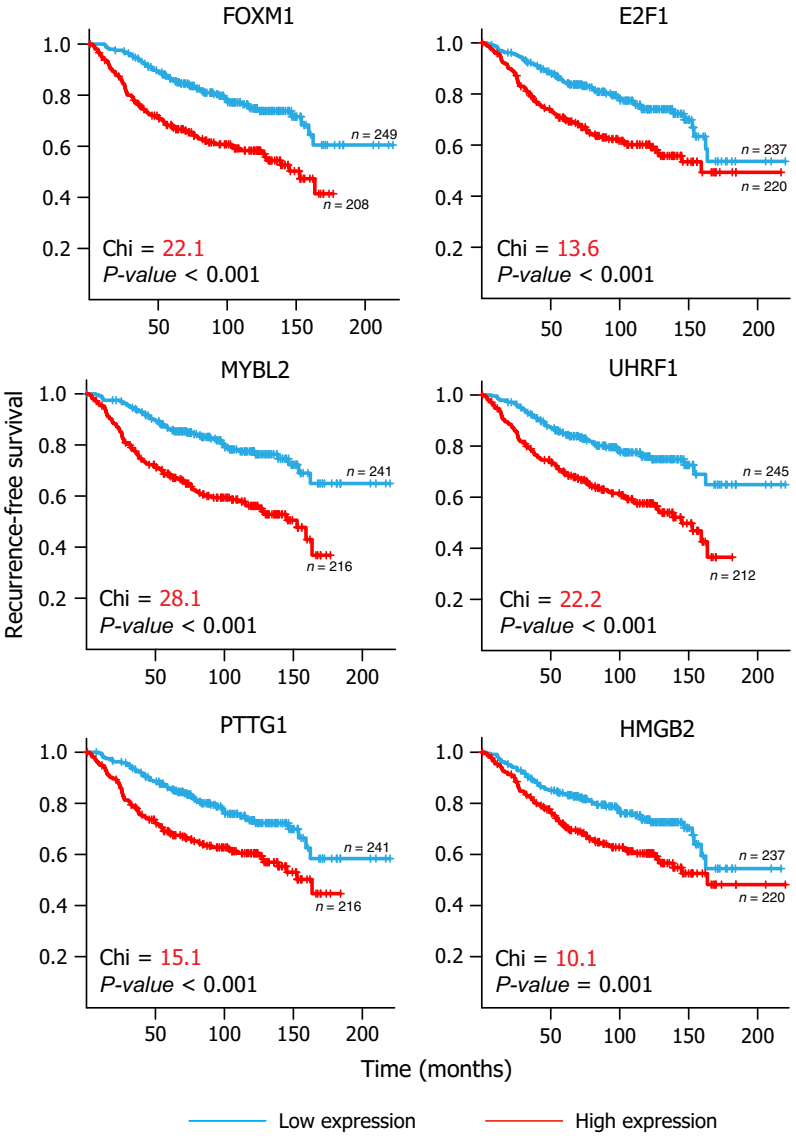

Fig. 2. MTR mRNA levels are associated with poor patient survival in breast cancer. Kaplan–Meier analysis of RFS for the six MTRs in 457 LN-negative tumours in a number of published microarray datasets [11,21,23]. Patients were stratified into two equal groups based on the expression of each MTR and analyzed using the log-rank test.

The stained TMAs were manually scored and analyzed in relation to RFS for the 371 chemotherapy-untreated tumours with information on all four MTRs (Figs 1D and 3). The combination of all four MTRs was more powerful at stratifying the patients in relation to survival compared to existing prognostic indicators such as Ki67 (Fig. 1D,E). These results indicate that both mRNA and protein levels of the prognosis-linked MTRs predict patient outcome in breast cancer.

Prognosis-linked MTRs bind the promoters of proliferation genes

We next sought to understand the phenotypic contribution of 'poor prognosis' expression signatures to tumour biology, and to relate this to MTR function. A gene ontology analysis of both the 'Genomic Grade' and '70-gene' signatures confirmed that both are significantly enriched in genes encoding regulators of cell growth and proliferation (Fig. 4A). This indicates that the primary utility of these signatures is to simply delineate the tumours with higher growth rates, as reported previously [18,24]. This led us to speculate that the prognosis-linked MTRs are simply upstream transcriptional regulators of proliferation genes, even in normal mammary epithelial cells.

We next set out to identify a cohort of 'core proliferation' genes that are highly expressed in actively growing mammary epithelial cells and whose expression is down-regulated in senescent cells. We reasoned that a strategy to combine the expression changes of serially passaged human mammary epithelial cells (HMECs) and mouse embryonic fibroblasts (MEFs) would allow us to identify a set of 'core proliferation' genes, distinct from those genes whose mRNA expression levels change as a result of shifts in cell type. Therefore, we isolated HMECs and MEFs and passaged them towards cellular senescence, as characterized by an increase in the levels of $p16^{INK4A}$ [25,26] and a decrease in the levels of the proliferative marker gene EZH2 [27]. A genome-wide mRNA expression analysis identified four differentially expressed gene clusters (Fig. 4B). The expression changes of representative genes from each cluster were validated by quantitative RT-PCR (Fig. 4C). Of particular interest were the cluster 3 genes, which were down-regulated during serial passaging of HMEC, but not MEF cells, and the cluster 4 genes, which were down-regulated in both HMEC and MEF cells. The cluster 3 genes included several genes involved in mammary epithelial cell-specific processes, such as the luminal cytokeratin *KRT19* and the tight junction protein *CLDN3*, consistent with

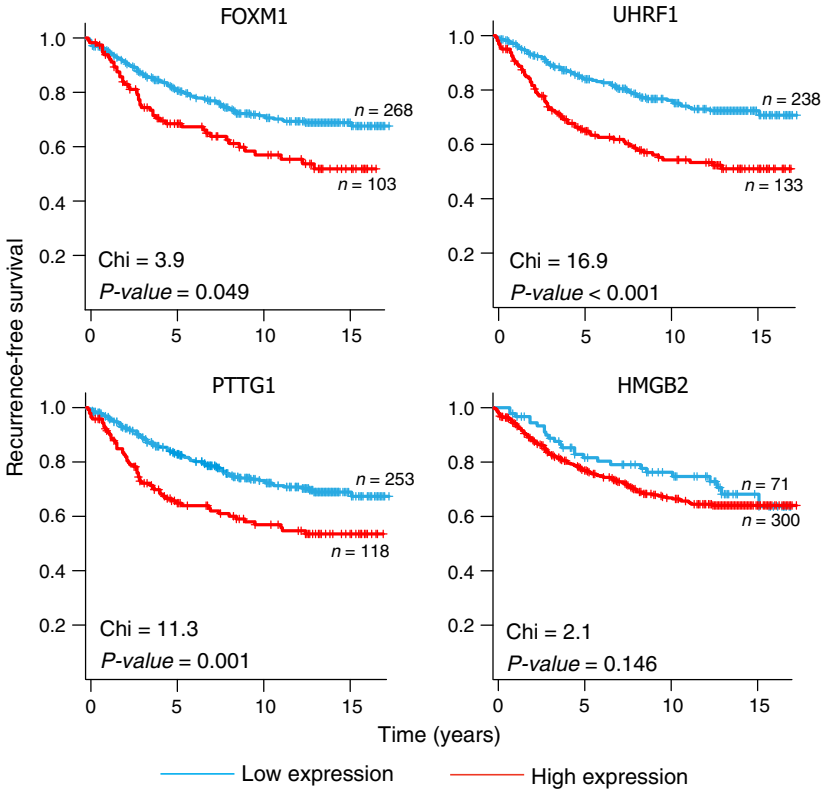

Fig. 3. Kaplan–Meier immunohistochemical analysis of FOXM1, UHRF1, PTTG1 and HMGB2 on breast cancer TMAs. Kaplan–Meier estimates of RFS for each MTR in 371 breast tumour samples. Patients were stratified into two groups using thresholds specific to each MTR. *P* values are calculated using the log-rank test. Only samples with information for all four MTRs are included in the analysis.

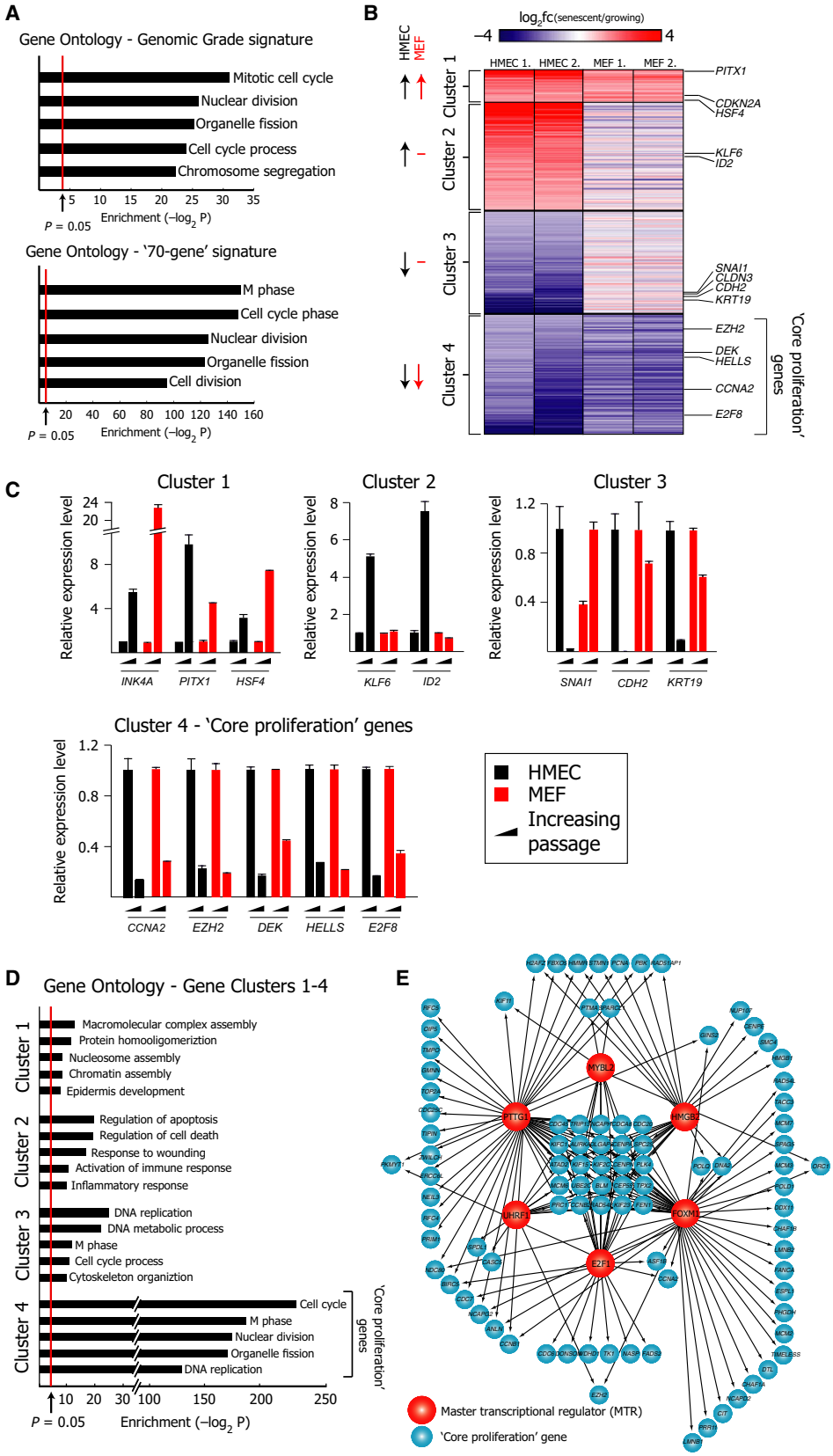
A
Gene Ontology - Genomic Grade signature
Mitotic cell cycle
Nuclear division
Organelle fission
Cell cycle process
Chromosome segregation
5 10 15 20 25 30 35
P = 0.05
Enrichment (–log2 P)
Gene Ontology - '70-gene' signature
M phase
Cell cycle phase
Nuclear division
Organelle fission
Cell division
20 40 60 80 100 120 140 160
P = 0.05
Enrichment (–log2 P)
B
log2fc(senescent/growing)
–4
4
HMEC
MEF
HMEC 1.
HMEC 2.
MEF 1.
MEF 2.
Cluster 1
Cluster 2
Cluster 3
Cluster 4
PITX1
CDKN2A
HSF4
KLF6
ID2
SNAI1
CLDN3
CDH2
KRT19
EZH2
DEK
HELLS
CCNA2
E2F8
'Core proliferation' genes
C
Cluster 1
Cluster 2
Cluster 3
Relative expression level
INK4A
PITX1
HSF4
KLF6
ID2
SNAI1
CDH2
KRT19
Cluster 4 - 'Core proliferation' genes
CCNA2
EZH2
DEK
HELLS
E2F8
HMEC
MEF
Increasing passage
D
Gene Ontology - Gene Clusters 1-4
Cluster 1
Macromolecular complex assembly
Protein homooligomeriztion
Nucleosome assembly
Chromatin assembly
Epidermis development
Cluster 2
Regulation of apoptosis
Regulation of cell death
Response to wounding
Activation of immune response
Inflammatory response
Cluster 3
DNA replication
DNA metabolic process
M phase
Cell cycle process
Cytoskeleton organiztion
Cluster 4
Cell cycle
M phase
Nuclear division
Organelle fission
DNA replication
'Core proliferation' genes
10 20 30 100 150 200 250
P = 0.05
Enrichment (–log2 P)
E
MYBL2
PTTG1
HMGB2
UHRF1
FOXM1
E2F1
Master transcriptional regulator (MTR)
'Core proliferation' gene

Fig. 4. Prognosis-linked MTRs are predicted to regulate proliferative genes. (A) The top five gene ontology terms identified for the Genomic Grade (top) and '70 gene' Poor Prognosis (bottom) gene signatures. The red line indicates a *P* value of 0.05, calculated using *t*-test statistics via the DAVID tool. (B) Transcriptomic profiling experiments in growing and senescent HMEC and MEF cultures were aligned to identify genes expressed at consistently high levels in proliferating cells from different lineage types; the so-called 'core-proliferative' genes. The heat-map analyses depict all genes up- or down-regulated by more than two-fold in growing versus senescent HMECs, and the corresponding change in MEFs (cluster 1 = 58 genes; cluster 2 = 193 genes; cluster 3 = 184 genes; cluster 4 = 214 genes). Cluster 4 represents a 'core proliferation' signature comprised of the genes whose mRNA levels are most significantly and consistently down-regulated during serial passaging of both HMECs and MEFs cells. (C) Quantitative real-time PCR validation of the mRNA expression level changes of representative genes from each of the four clusters shown in (B). (D) The top five gene ontology terms identified for each of the four gene clusters from (B). The red line indicates a *P* value of 0.05. (E) ARACNe network analysis for upstream MTRs of the 'core proliferation' signature. The predicted MTRs are shown in red and the core proliferation genes are shown in blue.

the fact that the proportion of luminal and myoepithelial cells shifts during serial passaging of HMEC cells [28]. This suggests that many of the genes within cluster 3 were down-regulated independently of the progressive decrease in proliferation rate. Consistent with this idea, a gene ontology analysis for each of the four gene clusters revealed the greatest enrichment of functional categories linked to cell cycle and proliferation in cluster 4 (Fig. 4D). Strikingly, an ARACNe analysis aimed at predicting the upstream regulators of the cluster 4 or 'core proliferation' genes identified the prognosis-linked MTRs, FOXM1, PTTG1, UHRF1, MYBL2, HMGB2 and E2F1 as key upstream regulators (Fig. 4E and Table 1). This confirms that the primary power of the 'Genomic Grade' and '70-gene' prognostic signatures is to indicate tumours with high levels of 'core proliferation' genes and thus, higher growth rates.

Next, we aimed to confirm that the normal role of the prognosis-linked MTRs is to directly bind to the promoters of the 'core proliferation' genes. Chromatin immunoprecipitation (ChIP) followed by quantitative real-time PCR (qPCR) confirmed the direct binding of four of the MTRs (FOXM1, MYBL2, E2F1 and HMGB2) to the promoters of core proliferation genes in HMEC-*TERT* cells (Fig. 5A). To gain a broader view on MTR binding throughout the genome, we performed ChIP followed by high-throughput sequencing (ChIP-seq) on HMEC-*TERT* cells for E2F1, MYBL2 and FOXM1. All three MTRs primarily associate with the promoters of the cluster 4 'core proliferation' genes and, to a lesser extent, some cluster 3 genes (Fig. 5B). The ChIP-seq tracks were aligned together with RNA-Seq performed on proliferating (passage 2) and nonproliferating (passage 14) HMECs, and three representative genes; *CCNB1*, *UBE2C* and *CENPA*, as well as a negative control, *KRT2*, are represented (Fig. 5C). The co-binding of FOXM1, E2F1 and MYBL2 on their downstream target genes is analogous to the mode of action of core pluripotency transcription factors in embryonic stem cells [29] and suggests that the prognosis-linked MTRs synergize to drive the expression of genes required for proliferation.

We were unable to assess the genome-wide binding patterns of PTTG1 or UHRF1 because of a lack of ChIP-grade antibodies. However, the fact that PTTG1 has been reported to have a role in the transcriptional activation of cell cycle genes [30,31] supports our ARACNe predictions. On the other hand, UHRF1 is generally considered to be a transcriptional repressor, being required for the maintenance of DNA methylation during cell division [32]. Therefore, UHRF1 is unlikely to directly regulate core proliferation genes, and is more likely to be a co-regulated proliferative gene. Supporting this possibility, E2F1, MYBL2 and FOXM1 also co-bind the promoter of the *UHRF1* gene in HMEC-*TERT* cells (data not shown).

Predicting tumours that have bypassed cellular senescence

We next aimed to develop a gauge of tumours that had bypassed the cellular senescence checkpoint and to determine whether this could help further refine breast cancer prognosis predictions. Accordingly, we examined whether deregulated levels of *CDKN2A* mRNA (the gene which encodes $p16^{INK4A}$) correlated with genetic perturbation of the cellular senescence checkpoint. Analysis of 463 samples in *The Cancer Genome Atlas* breast cancer dataset [33] revealed that reduced levels of *CDKN2A* mRNA correlated with deletion of the *CDKN2A* gene (Fig. 6A). However, interestingly, increased *CDKN2A* mRNA levels correlated with deletion of the *RB1* gene, as reported previously [34–37]. The fact that both very low or high levels of *CDKN2A* mRNA correlate with genetic perturbation of the pRB pathway raised the previously unanticipated possibility that cancers with intermediate $p16^{INK4A}$ levels would tend to have a better prognosis.

We next evaluated the prognostic potential of measuring intermediate or 'moderate' levels of $p16^{INK4A}$, both at the mRNA (Fig. 6B) and protein levels

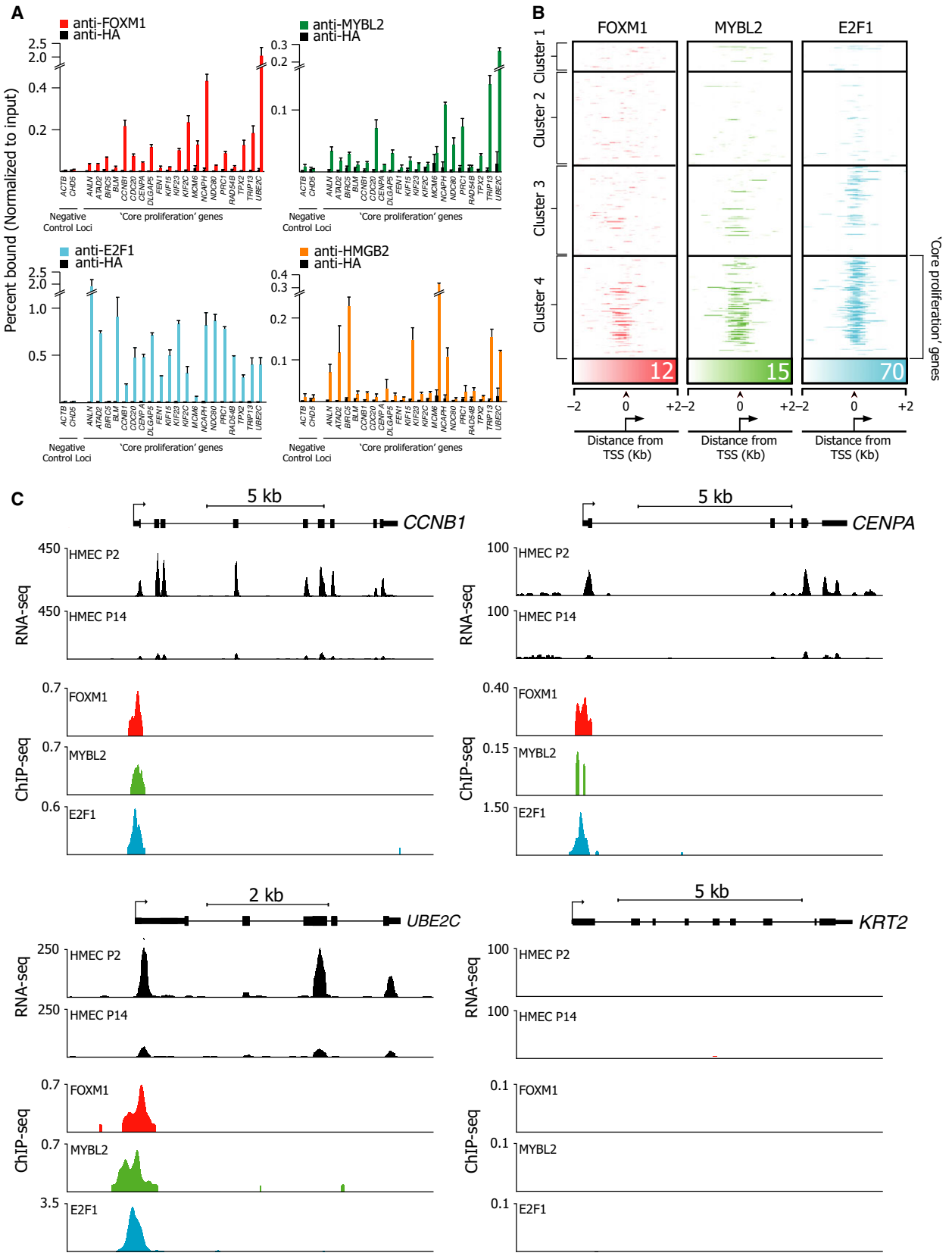
A
anti-FOXM1
anti-HA
anti-MYBL2
anti-HA
anti-E2F1
anti-HA
anti-HMGB2
anti-HA
Percent bound (Normalized to input)
Negative Control Loci
'Core proliferation' genes
B
FOXM1
MYBL2
E2F1
Cluster 1
Cluster 2
Cluster 3
Cluster 4
'Core proliferation' genes
12
15
70
Distance from TSS (Kb)
C
5 kb
CCNB1
CENPA
2 kb
UBE2C
5 kb
KRT2
RNA-seq
HMEC P2
HMEC P14
ChIP-seq
FOXM1
MYBL2
E2F1

Fig. 5. Prognosis-linked MTRs co-bind proliferative genes. (A) Validation of MTR binding to genes within the 'core proliferation' signature by ChIP-qPCR. Precipitated DNA was analyzed by qPCR using primers directed towards the promoters of the indicated genes. Anti-HA antibody was used as a negative control ChIP and the *β-ACTB* and *CHD5* promoters were included as negative control promoters. The ChIP enrichments are presented as the percentage of protein bound normalized to input. The error bars indicate the SD of three technical replicates. (B) Heat-map representations showing ChIP-seq data for FOXM1, MYBL2 and E2F1 in HMEC-*TERT* cells. The binding at the promoters of genes from cluster 4 and, to a lesser extent, cluster 3 is indicated for the FOXM1 (red), MYBL2 (green) and E2F1 (blue) MTRs. The region between –2 and +2 kb of the transcriptional start site (TSS) of all the cluster 1–4 genes is shown for comparison. (C) Representative ChIP-seq tracks at the indicated genes in HMEC-*TERT* cells showing ChIP-seq enrichment on the *y*-axis. RNA-seq data from both low and high passage HMECs is also depicted for each gene (top), showing reads per kilobase per million (RKPM). The *KRT2* gene is included as a negative control.

(Fig. 6C–E). Breast tumours with moderate *CDKN2A* mRNA levels correlated with improved RFS (Fig. 6B) in the combined microarray dataset of 457 LN-negative breast cancer patients [11,21,23]. By contrast, tumours with either very low or very high *CDKN2A* mRNA levels correlated with shorter RFS. A TMA stained for $p16^{INK4A}$ confirmed the same trend at the protein level, whereby tumours with either very high or very low $p16^{INK4A}$ correlated with both shorter recurrence-free and breast cancer-specific survival (Fig. 6C–E). Based on these observations, it is likely that the breast cancers with moderate levels of *CDKN2A* mRNA and $p16^{INK4A}$ protein are enriched in cells that have an intact pRB pathway, have not bypassed the cellular senescence checkpoint, have a lower proliferative rate and thus have a more favorable prognosis.

Development of an 'OncoMasTR score' that predicts outcome better than current prognostic tests

We next developed a scoring system combining the mRNA expression levels of prognosis-linked MTRs with those of *CDKN2A*, termed the 'OncoMasTR RNA score', and compared this with estimates of other clinically utilized multigene prognostic assays. We analyzed LN-negative, chemotherapy-untreated patients from each of three individual microarray datasets [11,21,23], as well as a combination of all three (Fig. 7A). This revealed that OncoMasTR compared favourably with surrogate estimations of both the MammaPrint and Oncotype Dx signatures, using low/high categories for comparison with the former, and low/moderate/high categories for comparison with the latter. Although the MammaPrint 70-gene signature performed best in the dataset composed of samples used in its derivation [10,11], the OncoMasTR RNA score outperformed estimates of both the MammaPrint and Oncotype Dx assays, on the other two datasets and also overall when all three datasets were combined. We also developed an 'OncoMasTR IHC score', based on the protein levels of $p16^{INK4A}$ protein and the four prognosis-linked MTRs, and tested this on the cohort of 371 chemotherapy-untreated patients described previously. The inclusion of $p16^{INK4A}$ provided a striking improvement upon prognosis-linked MTRs alone in terms of the ability to predict patient survival on all patients, as well as on a LN-negative subcohort (Fig. 7B).

For a prognostic test to be clinically useful, it should provide additional prognostic information, independent of standard clinicopathological variables. Therefore, we also performed multivariate analysis on the OncoMasTR RNA and protein scores using Cox proportional hazard models. This confirmed that the OncoMasTR score contributes significant information in predicting RFS, on top of a standard clinicopathological variable model, at both mRNA (Table 2) and protein (Table 3) levels. This was also the case in the LN-negative patient cohort. The added prognostic value of the OncoMasTR scores on top of the standard clinical model is superior to all other prognostic indicators, including Ki67, the 70-gene signature (MammaPrint) and the 21-gene signature (Oncotype Dx). Furthermore, the OncoMasTR RNA score was found to provide significant additional prognostic information to a model comprising the standard clinical variables together with the 21-gene Oncotype Dx signature (Table 4). This suggests that the OncoMasTR score can provide additional clinical utility on top of that provided by the Oncotype Dx panel.

The OncoMasTR score accurately predicts outcome for ER-positive, LN-negative patients

To further evaluate the potential clinical utility of the OncoMasTR score within the patient population most likely to require it, we examined its prognostic power on 366 estrogen receptor (ER)-positive, LN-negative patients from the combined microarray dataset described earlier [11,21,23], reflecting the inclusion criteria for the Oncotype Dx assay. Remarkably, the On-

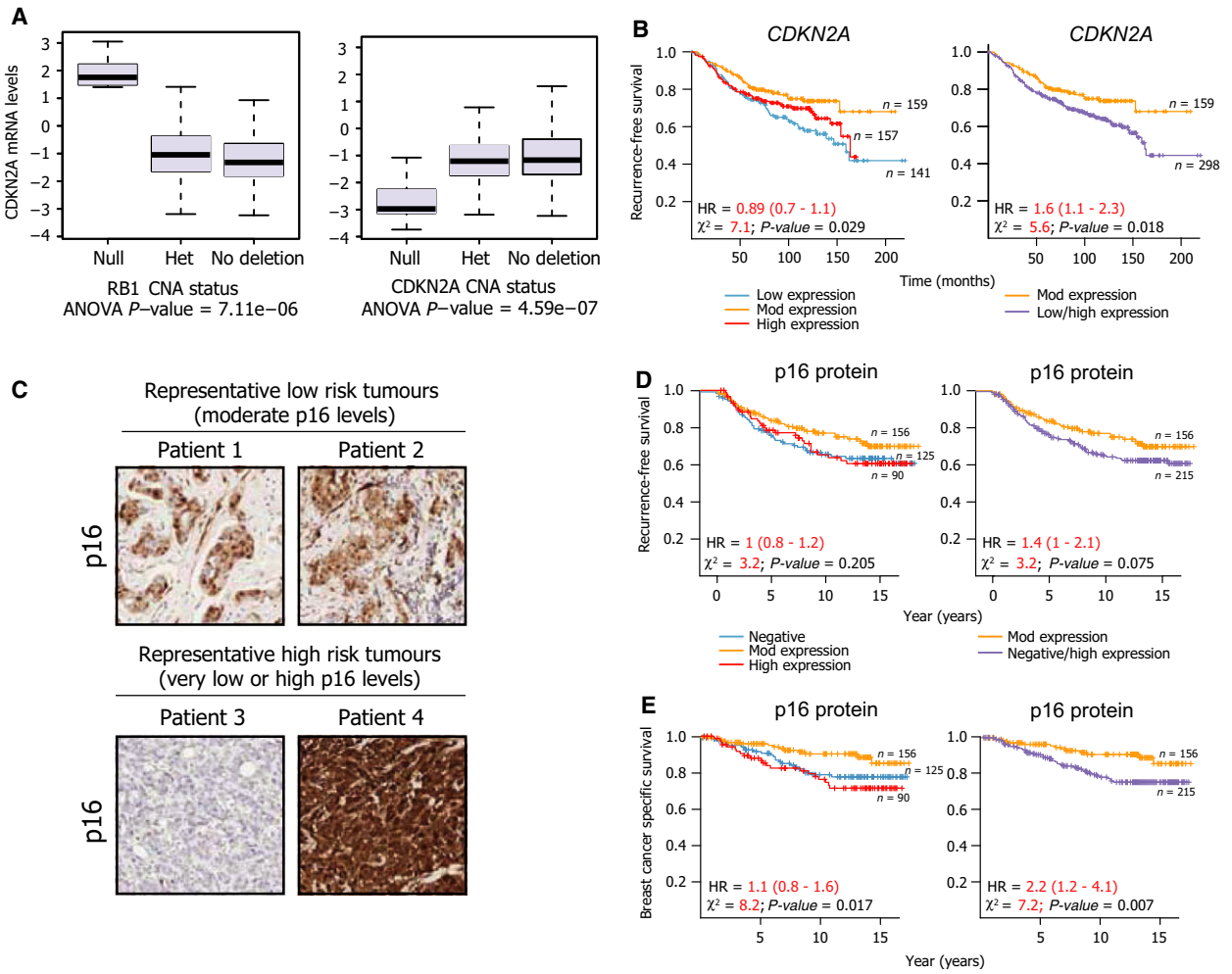

Fig. 6. p16^{INK4A} levels are indicative the genetic status of the p16-pRB pathway and predict patient outcome. (A) Correlations of the mRNA expression levels of *CDKN2A* with gene copy number alterations (CNA) in the *RB1* and *CDKN2A* gene loci using the GISTIC tool on data from 463 breast cancers from *The Cancer Genome Atlas* [33]. (B) Kaplan–Meier analyses of *CDKN2A* mRNA in LN-negative breast cancers, without adjuvant chemotherapy on the combined (n = 457) microarray dataset [11,21,23]. The samples were stratified into three groups based on *CDKN2A* mRNA expression levels, cut at the 33rd and 66th percentile (left). Additionally, the low and high expression groups were combined and compared with the moderate expression group (right). Chi-squared values and P values were calculated using a log-rank test. (C) Representative immunohistochemical stainings for p16 protein levels on low and high risk tumours (as in Fig. 1C). Note, that the tumours from patients 1–4 were stained for FOXM1, HMGB2, UHRF1 and PTTG1 in Fig. 1C. (D) Kaplan–Meier analyses based on p16 protein levels on tumours taken from patients not treated with chemotherapy (n = 371) from the TMA cohort using RFS data as the output. The samples were stratified into three groups based on p16 protein expression levels (left). Additionally, the negative and high expression groups were combined and compared with the moderate expression group (right). Chi-squared values and P values were calculated using the log-rank test. (E) Kaplan–Meier analyses as in (D), using breast cancer specific survival as the endpoint.

coMasTR RNA score outperformed surrogate estimates of both the MammaPrint (low/high groups), and Oncotype Dx (low/mod/high groups) assays in this patient cohort (Fig. 7C,D).

With the aim of addressing one of the major issues with the Oncotype Dx assay, namely the high proportion of 'intermediate risk' patients with no clear treatment guidelines [38], we performed a comparative analysis of population distribution and recurrence risk on the ER-positive, LN-negative cohort as above (n = 366), using distant metastasis-free survival (DMFS) as an endpoint (Table 5). This revealed that the OncoMasTR RNA score was capable of classifying an increased proportion of patients as low risk in this cohort: 69.3% compared to 36.2% for the Oncotype Dx 21-gene signature and 31.8% for the MammaPrint 70-gene signature. This is almost double what is classified as low risk using other platforms. Rates of

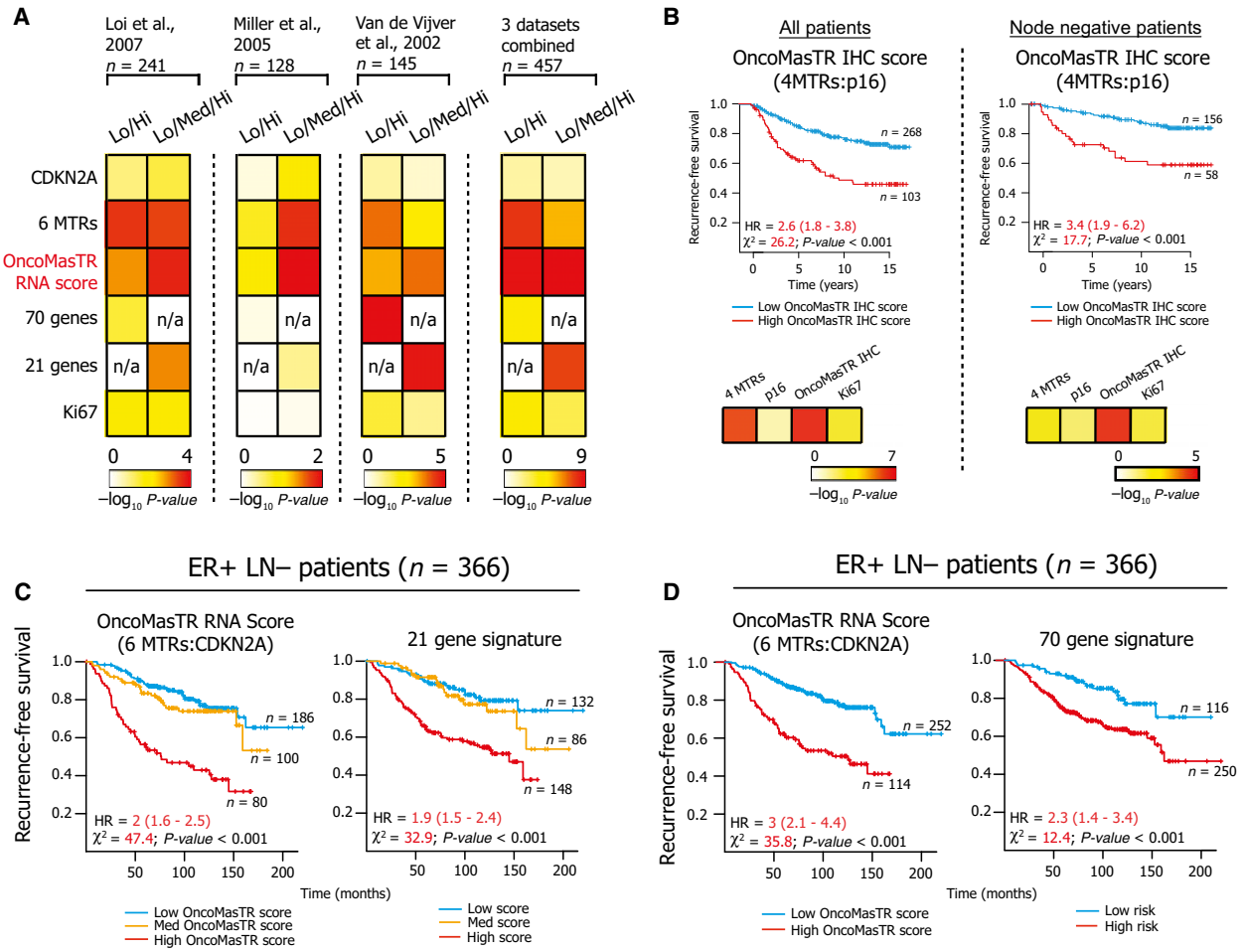

Fig. 7. The OncoMasTR score outperforms estimates of currently used multigene prognostic assays. (A) Heat maps illustrating the prognostic value of *CDKN2A* alone, six MTRs combined, OncoMasTR RNA score, 70-gene signature, 21-gene signature and Ki67 by Kaplan–Meier analysis on LN-negative, chemotherapy-untreated samples in the combined (n = 457) microarray dataset [11,21,23]. The 70-gene and 21-gene signature predicted risk groups were estimated based on gene expression data using the *genefu* package in the R-software. The scale represents $-\log_{10}$ of the P values calculated using log-rank test. Both low/moderate/high and low/high splits were used to facilitate comparison to existing the prognostic signatures. (B) Kaplan–Meier analyses of the combined score of the protein levels of four MTRs (FOXM1, UHRF1, HMGB2 and PTTG1) and p16, termed the OncoMasTR IHC score, in all samples not treated with chemotherapy (left; n = 371) and LN-negative chemotherapy-untreated samples (right; n = 214) from the TMA cohort using RFS data. The prognostic values of the four MTRs alone, p16 alone, the OncoMasTR IHC score, and Ki67 are shown as a heat map based on the $-\log_{10}$ of P values calculated using the log-rank test. (C) Kaplan–Meier analyses of the OncoMasTR RNA score (six MTRs and *CDKN2A*) in comparison with the 21-gene signature, using a low/moderate/high split, on the subset of ER-positive, LN-negative patients who did not receive adjuvant chemotherapy (n = 366), from the combined microarray dataset [11,21,23]. (D) Kaplan-Meier analyses as in C, but in comparison with the 70-gene signature, using a low/high split.

distant metastasis in the low risk groups were comparable for OncoMasTR RNA score (13.9%) and the Oncotype Dx 21-gene signature (13%) and slightly lower for the MammaPrint 70-gene signature (11.3%). However, it should be noted that this dataset contains some of the training samples for the 70-gene signature.

To further examine the sensitivity and specificity of the OncoMasTR score, and to rule out any potential issues with the microarray probe set data, we performed a TaqMan-based qRT-PCR analysis of the six MTRs and *CDKN2A*/p16 on 151 ER-positive, LN-negative tumours untreated by chemotherapy, which matched the TMA cohort used previously (Fig. 8). To identify the optimum threshold for binary classification, and to minimize any reduction in accuracy as a result of the increase in the size of the low risk group, we carried out a receiver operating characteristic (ROC) analysis, revealing that the OncoMasTR RNA

Table 2. Multivariate Cox regression analysis using a standard clinical variable model based on data from three published microarray studies [11,21,23]. Clinical variables used: age (≥ 50 years), nodal status, tumour size (≥ 2 cm), tumour grade (> 1), treatment (endocrine therapy) and ER status. LR, likelihood ratio.

	All patients (n = 567)		Node-negative patients (n = 410)	
Variable	LR-$\Delta\chi^{2a}$	P value	LR-$\Delta\chi^{2a}$	P value
Lo/Med/Hi				
FOXM1	24.14	< 0.001	26.59	< 0.001
E2F1	25.28	< 0.001	15.56	< 0.001
HMGB2	10.89	< 0.001	7.47	0.006
MYBL2	25.43	< 0.001	15.91	< 0.001
PTTG1	12.37	< 0.001	10.16	0.001
UHRF1	22.71	< 0.001	17.61	< 0.001
CDKN2A	2.23	0.135	13.82	< 0.001
6MTR	33.8	< 0.001	20.27	< 0.001
OncoMasTR RNA score	43.87	< 0.001	44.04	< 0.001
21-gene	29.02	< 0.001	38.03	< 0.001
Ki-67	8.3	0.004	7.45	0.006
Lo/Hi				
6MTR	23.82	< 0.001	29.32	< 0.001
OncoMasTR RNA score	29.62	< 0.001	32.20	< 0.001
70-gene	30.20	< 0.001	28.88	< 0.001
Ki-67	5.52	0.018	8.99	0.003

[a]Added prognostic value of each variable is represented by change in the LR-$\Delta\chi^2$ value from the model of only clinical variables to the model of clinical variable + marker in the three combined microarray datasets.

Table 3. Multivariate Cox regression analysis using a standard clinical variable model in TMAs. Clinical variables used: age (≥ 50 years), nodal status, tumour size (≥ 2 cm), tumour grade (> 1), treatment (endocrine therapy, radiotherapy), ER and HER2 status. LR, likelihood ratio.

	All patients (n = 254)		Node-negative patients (n = 166)	
Variable	LR-$\Delta\chi^{2a}$	P value	LR-$\Delta\chi^{2a}$	P value
Lo/Hi				
FOXM1	0.09	0.763	0.78	0.378
HMGB2	0.04	0.834	1.67	0.197
PTTG1	4.72	0.030	0.74	0.390
UHRF1	4.08	0.043	0.35	0.552
p16	8.02	0.005	7.62	0.006
4 MTRs	11.24	< 0.001	0.89	0.344
OncoMasTR IHC Score	23.74	< 0.001	9.92	0.002
Ki67	2.82	0.093	3.24	0.072

[a]Added prognostic value of each variable is represented by change in the LR-$\Delta\chi^2$ value from the model of only clinical variables to the model of clinical variable + marker in the TMA dataset.

score had a sensitivity/specificity of 0.67/0.74 for predicting RFS and 0.82/0.73 for predicting DMFS. The area under the curve values were 0.697 and 0.798 for RFS and DMFS, respectively, comparing favourably to published values for other prognostic tests such as Oncotype Dx [39]. Finally, using these thresholds, we carried out survival analysis on this cohort, which classified 67% of patients as 'low risk' with an impressively low recurrence rate of 3% at 5 years. No patients categorized as 'low risk' by the OncoMasTR score experienced a distant metastasis event within 5 years. Although the chosen threshold will need to be validated in an independent patient cohort, these results suggest that the OncoMasTR score holds enormous promise for improving the stratification of patients diagnosed with ER-positive LN-negative breast cancer, who need not be unnecessarily treated with damaging chemotherapeutics after surgical resection.

Discussion

The decision of whether or not to subject women diagnosed with LN-negative breast cancer to chemotherapy is a major challenge in the clinic. Indeed, in accordance with current guidelines, the majority diagnosed are unnecessarily treated with chemotherapy after surgery. In the present study, we provide a strategy to better distinguish the 10–30% of women requiring aggressive adjuvant chemotherapy from the majority who need no further treatment [2,3]. We developed a scoring system, called 'OncoMasTR', which outperforms currently used multigene prognostic assays. This scoring system is based on the expression of several prognosis-linked MTRs, which we show act upstream of proliferation genes, and $p16^{INK4A}$, for which we show that intermediate, as opposed to high or low, levels in tumours are indicative of better patient prognosis. Importantly, OncoMasTR stands to provide a better delineation of the LN-negative breast cancer patients requiring chemotherapy, and performs exceptionally well in ER-positive, LN-negative patients, a subgroup in which it is particularly difficult to identify women at high risk of recurrence, who would therefore be most likely to require chemotherapy.

The deconvolution of transcriptional networks is crucial for our understanding of normal cell biology and disease pathogenesis, which frequently results from perturbed transcriptional programming [40]. In the present study, we observed that identical transcription factors are predicted to function as upstream MTRs of independent breast cancer prognostic signatures. This suggests that these signatures represent similar tumour characteristics. Indeed, the significant enrichment of functional terms related to cell growth

Table 4. Multivariate Cox regression analysis using a clinical variable model including the 21-gene signature based on data from three published microarray studies [11,21,23]. Clinical variables used: age (≥ 50 years), nodal status, tumour size (≥ 2 cm), tumour grade (> 1), treatment (endocrine therapy), ER status and 21-gene signature. LR, likelihood ratio.

	All patients (n = 570)		Node-negative patients (n = 413)	
Variable	LR-$\Delta\chi^{2a}$	P value	LR-$\Delta\chi^{2a}$	P value
Lo/Hi				
6MTR	6.02	0.014	5.45	0.019
OncoMasTR RNA score	9.92	0.002	7.11	0.008
Lo/Med/Hi				
6MTR	12.68	< 0.001	2.28	0.131
OncoMasTR RNA score	18.56	< 0.001	14.88	< 0.001

[a]Added prognostic value of each variable is represented by change in the LR-$\Delta\chi^2$ value from the model of clinical variables + 21-gene signature to the model of clinical variable + 21-gene signature + marker in the three combined microarray datasets.

Table 5. Comparison of OncoMasTR score to estimates of the Oncotype Dx and Mammaprint prognostic assays in 366 ER-positive, LN-negative breast tumours, untreated by chemotherapy, based on data from three published microarray studies [11,21,23].

	Risk group		
Predictor	Low (%)	Intermediate (%)	High (%)
OncoMasTR			
% Population	69.3	–	30.7
% DMFS (at 5 years)	8.4	–	29.7
% DMFS (at 10 years)	13.9		36.9
21-gene (Oncotype Dx)			
% Population	36.2	23.8	40.1
% DMFS (at 5 years)	8.4	3.5	27.6
% DMFS (at 10 years)	13	10.5	34.5
70-gene (Mammaprint)			
% Population	31.8	–	68.2
% DMFS (at 5 years)	5.2	–	19.4
% DMFS (at 10 years)	11.3		25.5

and proliferation within both signatures, together with the fact that several of the MTRs co-bind the promoters of genes within our 'core proliferation' signature, indicates that the signatures mainly reflect a proliferative phenotype. Additional tumour phenotypes, such as metastasis or angiogenesis, would provide additional prognostic information [41]. Thus, identifying the MTRs of breast cancer 'metastatic' and 'angiogenic' gene signatures may further refine the OncoMasTR score.

The bypass of cellular senescence has generally been considered as a pre-requisite for the formation of invasive tumours [42]. However, evidence exists that some tumours can retain the ability to engage the cellular senescence response, subsequent to cytotoxic therapy [43–45], the inactivation of a driving oncogene such as c-Myc [46] or the reactivation of p53 [47,48]. The inference is that those tumours rendered incapable of activation of the cellular senescence checkpoint would thus confer a worse prognosis on patients. This concept is consistent with our demonstration that high levels of proliferative MTRs combined with aberrantly high or low levels of $p16^{INK4A}$, both indicative of an abrogated $p16^{INK4A}$-pRB-E2F pathway, correlate with poor outcomes for the patient. In particular, we show that high levels of $p16^{INK4A}$ correlate with deletion of the *RB1* gene, which encodes the pRB protein, whereas low levels of $p16^{INK4A}$ correlates with deletion of the *CDKN2A* gene locus. Both cases result in inactivation of the $p16^{INK4A}$-pRB-E2F pathway. Previous studies of $p16^{INK4A}$ expression in relation to breast cancer prognosis have reported conflicting results. For example, although high $p16^{INK4A}$ expression was found to be associated with poor prognosis in some patient cohorts [49–53], other studies found that it was associated with improved outcome [54]. However, these studies divided the patients into just two groups: low versus high $p16^{INK4A}$. In our analysis, we instead used a three-group stratification, providing a clear delineation of the association between $p16^{INK4A}$ expression and breast cancer prognosis; specifically, that tumours with either very high or very low $p16^{INK4A}$ are likely to reflect a poor prognosis for the patient, whereas intermediate levels correlate with better prognosis.

In conclusion, measuring the levels of prognosis-linked MTRs and $p16^{INK4A}$, known as the OncoMasTR score, is very powerful as a prognostic signature in breast cancer. This is most likely a result of its ability to identify those tumours with high growth rates that have also bypassed the cellular senescence checkpoint. Significantly, the OncoMasTR score classifies a higher proportion of ER-positive, LN-negative breast cancer patients as 'low risk' in the cohorts that we have examined, thereby reducing the number of patients in the ambiguous 'intermediate' group and potentially avoiding overtreatment of patients. Given the universal importance of proliferation and senescence regulation in cancer progression, we suggest that this approach may also prove useful in other cancer types. It will be important to evaluate the OncoMasTR score on a prospective patient cohort to determine its potential as a prognostic assay for early-stage breast cancer, and to determine whether this combination of factors could prove prognostic in other cancer types.

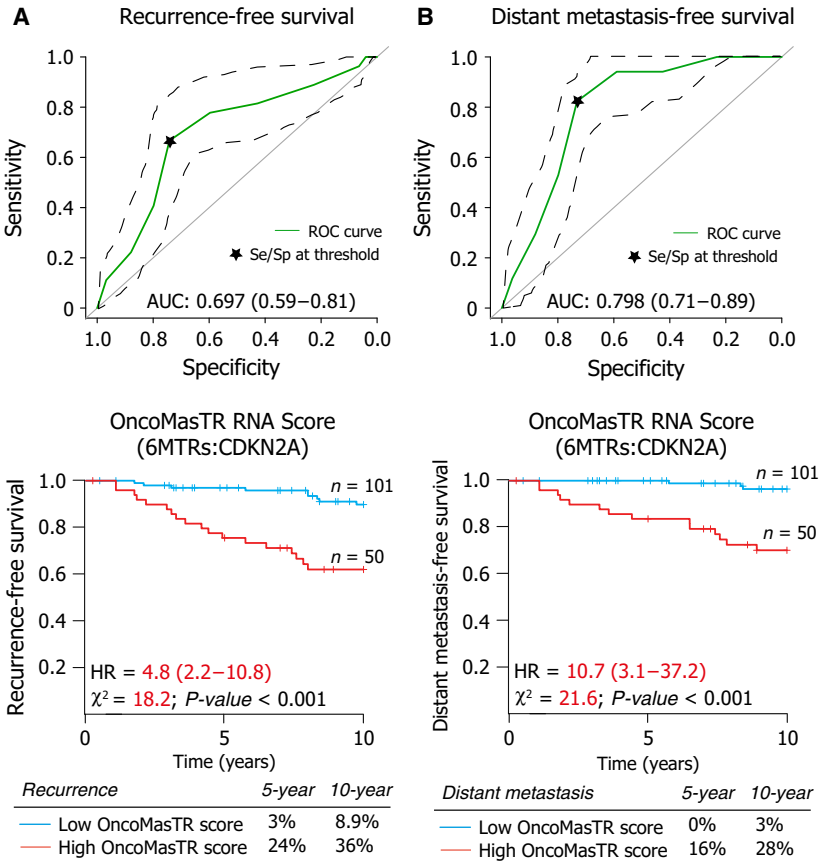

Fig. 8. Prognostic power of the OncoMasTR score as measured by TaqMan qRT-PCR in ER-positive, LN-negative patients. (A) ROC curves for the OncoMasTR RNA score (six MTRs and CDKN2A) in ER-positive, LN-negative patients who did not receive adjuvant chemotherapy, as measured by Taqman-based qRT-PCR (n = 151). RFS is used as the endpoint. Kaplan–Meier analyses for this cohort are shown beneath. (B) ROC curves and Kaplan–Meier analyses as shown in (A), with DMSF as the endpoint.

Materials and methods

Cell culture

Primary HMEC cells were grown as described previously [25]. HMEC-TERT cells were immortalized at passage 2 using a pBABE-hTERT-hygro construct. MEFs were derived from embryonic day 13.5 C57BL6 mouse embryos and maintained in DMEM media supplemented with 10% (v/v) FBS (HyClone, Logan, UT, USA), 100 U·mL^{-1} penicillin and 100 U·mL^{-1} streptomycin (Gibco, Gaithersburg, MD, USA).

RNA sequencing

Total RNA was extracted from proliferating and senescent HMECs using the RNeasy kit (Qiagen, Valancia, CA, USA). Polyadenylated RNA species were enriched from 5 μg of total RNA, and sequencing libraries were prepared from this PolyA+ RNA from two independent HMEC lines using the TruSeq Sample Prep kit (Illumina, Inc., San Diego, CA, USA). Libraries were used for cluster generation and sequencing analysis using the Genome Analyzer II (Illumina) in accordance with the manufacturer's instructions. Reads were trimmed to 42 bp to remove low-quality nucleotides and were mapped to the human genome (hg19) with the Burrows–Wheeler sequence alignment tool [55]. TOPHAT [56] was used for splice-aware mapping to the transcriptome. Reads were summarized for each transcript and fed into DESeq for normalization and calculation of mRNA fold changes.

Microarray analysis

Total RNA was extracted from proliferating and senescent MEFs using the RNeasy kit (Qiagen). For each time-point, RNA was prepared from three independent MEF lines and subsequently pooled. Cy3-labeled cRNA, for use with a custom-designed 44k microarray (Agilent Technologies Inc., Santa Clara, CA, USA), was prepared and hybridized in accordance with the manufacturer's instructions. Microarrays were scanned using Agilent's DNA microarray scanner and data were analyzed as described previously [57].

Microarray dataset mining

Gene ontology analysis was carried out using the DAVID Bioinformatics Resource (http://david.abcc.ncifcrf.gov). The processed publicly available breast cancer microarray

datasets were downloaded from Rosetta Inpharmatics [11] and GEO (GSE6532 and GSE3494) [21,23]. The datasets were normalized as described in the original publications. Within each dataset, the expression data of each gene was divided at the median into two groups, or at the 33rd and 66th percentile into three groups, for comparison with the two and three risk groups of Mammaprint and Oncotype Dx respectively. Once the samples have been dichotomized, the gene expression data are no longer used, allowing comparisons across different datasets/platforms, as described previously [58]. To generate a combined MTR score, the gene expression values for each of the six genes were divided at the median, given a score of 1 or 2 based on the expression level, and the sum of these scores was then divided, as above, to create two or three groups. *CDKN2A* gene expression was divided into three groups (low, moderate and high) at the 33rd and 66th percentile. The moderate group was given a score of 1 and the low and high groups were combined and given a score of 2, corresponding to the fact that both very low and high levels of *CDKN2A* mRNA correlate with genetic perturbation of the pRB pathway. To generate the OncoMasTR RNA score, the combined MTR score and the *CDKN2A* score were summed together and the final scores were divided into two or three groups. The genefu package in R was used to estimate the risk groups which approximate the Oncotype Dx assay (based on a 21-gene signature), the MammaPrint assay (based on a 70-gene signature) and the Genomic Grade Index (based on a 97-gene signature).

Real-time quantitative PCR

For validation of microarray experiments, total RNA was extracted from cells using the RNeasy kit (Qiagen) in accordance with the manufacturer's instructions, and 1 μg of total RNA was used to generate cDNA using the TaqMan Reverse Transcription kit (Applied Biosystems, Foster City, CA, USA). Relative mRNA expression levels were determined using the SYBR Green I detection chemistry (Applied Biosystems) on an ABI Prism 7500 Fast RT-PCR System. RPLPO was used as a control gene for normalization and the generation of ΔC_t values. For Taqman qRT-PCR analysis on FFPE breast tumours, RNA was extracted using the RNeasy FFPE kit (Qiagen), and 2 μg of total RNA was used to generate cDNA using the SuperScript VILO cDNA Synthesis Kit (Applied Biosystems). Optimized probes for each gene were assayed on TaqMan Low-Density Array cards in accordance with the manufacturer's instructions. The primer/probe sequences used are available upon request.

ChIP and ChIP-sequencing

ChIP and ChIP-seq analyses were performed as described previously [59]. Briefly, for ChIP-seq, DNA from two independent ChIP experiments was pooled and sequencing libraries generated using the ChIP-seq Sample Prep Kit (Illumina). DNA libraries were sequenced using the Genome Analyzer II (Illumina) in accordance with the manufacturer's instructions. Mapping of sequence reads to the human genome (hg19) was performed using BOWTIE [60]. Peak detection was performed using the Model-based Analysis for ChIP-Seq (MACS) software [61] and input DNA was used as a control for normalization.

ARACNe analysis

Breast cancer transcriptional networks were supplied by Andrea Califano (Columbia University, New York, NY, USA) and generated by ARACNe [19], using published breast cancer datasets [11,21,23]. The prognosis gene signatures used can be found in the relevant publications [7,9,10]. The larger 231-gene signature from which the 70-gene MammaPrint signature was derived [10], as well as the larger 207-gene list from which the 97-gene Genomic Grade Index was derived [9], were used as input for the ARACNe algorithm to identify the common transcriptional regulators driving the expression of the genes in these signatures.

Statistical analysis

Kaplan–Meier survival curves were used for survival analysis and chi-squared and P values were calculated using the log-rank test. Multivariate Cox proportional hazards analysis was used to evaluate the prognostic value of genes and combined scores and generate hazard ratios. The contribution of each marker was assessed by the change in likelihood ratio (LR-Chi, d.f. = 1). $P < 0.05$ was considered statistically significant. The primary clinical endpoints used for analysis were RFS and DMFS. The ROC curves were generated using the R-package pROC (http://web.expasy.org/pROC). The response vector was DMFS events at 10 years and the predictor was the '6MTR:CDKN2A' score. The 95% confidence intervals were computed with 2000 stratified bootstrap replicates. All statistical analysis was carried out using R (http://cran.r-project.org/; version 2.15.0). Heatmaps were created using MATRIX2PNG (http://chibi.ubc.ca/matrix2png).

TMA patient cohort

The TMA used in the present study was derived from a reference cohort of 498 consecutive invasive breast cancer cases diagnosed at the Department of Pathology, Malmo University Hospital, Malmo, Sweden, between 1988 and 1992, and has been described previously [62]. In brief, the median age was 65 years (range 27–96 years) and median follow-up time regarding disease-specific and overall survival was 11 years (range 0–17 years). Patients with

recurrent disease and previous systemic therapies were excluded. Two hundred and sixty-three patients were dead at the last follow-up, 90 of whom were classified as breast cancer-specific deaths. Of the total cohort of 498 patients, 23 patients received chemotherapy, 208 patients received radiotherapy and 161 patients received hormone therapy. Tissue cores (1 mm) from areas representative of invasive cancer were extracted from donor blocks and arrayed in duplicate. The present study was approved by the Ethics Committee at Lund University and Malmo University Hospital.

IHC and TMA analysis

TMA slides were stained using the LabVision IHC kit (LabVision, Newmarket, UK) as described previously [63]. The primary antibodies used were rabbit polyconal HMGB2 (dilution 1 : 1500; Abcam, Cambridge, MA, USA), mouse monoclonal UHRF1 (dilution 1 : 1000; Clone28/ICBP90; BD Biosciences, Clontech, Palo Alto, CA, USA), rabbit polyclonal PTTG1 (dilution 1 : 500; Invitrogen, Carlsbad, CA, USA), rabbit polyclonal FOXM1 (dilution 1 : 300; C20; Santa Cruz, CA, USA) and mouse monoclonal p16 (dilution 1 : 5000; Clone JC8). Slides were scanned at × 20 magnification using a ScanScope XT slide scanner (Aperio Technologies, Vista, CA, USA). TMA staining of tumour cells was evaluated by a pathologist on the basis of intensity as negative (0), weak (1), moderate (2) and strong (3), as well as percentage on a scale of 0–6 (0 = 0–1%; 1 = 1–10%; 2 = 10–25%: 3 = 25–50%; 4 = 50–75%; 5 = 75–90%; 6 = 90–100%). Staining for the factors HMGB2 and UHRF1 was predominantly nuclear, whereas PTTG1, FOXM1 and $p16^{INK4A}$ stained both the nuclear and cytoplasmic compartments and were scored accordingly. For UHRF1, PTTG1 and $p16^{INK4A}$, the percentage of positive tumour nuclei was the most significant variable in relation to outcome and was used in all further analysis. For HMGB2, a modified Allred score (intensity plus percentage) was used and, for FOXM1, the percentage of cytoplasmic positivity within tumour cells was the most significant variable, and was used for further analysis. For analysis of the four MTRs, a threshold for positivity was applied independently for each variable, to create a binary score with low (0) and high (1) expression. For p16, the 'negative' (0% positivity) and 'high' (> 50% positivity) expression groups were combined (score = 1) and compared with the 'moderate' group (score = 0).

To generate a combined MTR score at the protein level, the sum of the binary scores for all four MTRs was generated. Tumours with high expression of > 1 MTR were classified as having a high MTR score. To generate the combined 4MTR + p16 score (OncoMasTR IHC score), the binary 4MTR score was combined with the binary p16 score, and divided into two groups with a threshold of > 2.

Acknowledgements

We would like to thank Andrea Califano of Columbia University for kindly allowing use of the ARACNe breast cancer networks. Work in the Bracken Laboratory is supported by Science Foundation Ireland (SFI PICA SFI/10/IN.1/B3002), the Health Research Board (HRA_POR/2010/124), Enterprise Ireland (EI-CF/2012/2014A-Bracken), The Irish Research Council and St Vincent's Foundation. Work in the Gallagher Laboratory is supported by Science Foundation Ireland-funded Strategic Research Cluster, Molecular Therapeutics for Cancer Ireland (http://www.mtci.ie), the FP7 Framework Collaborative Research programme and RATHER (http://www.ratherproject.com). Both the Bracken and Gallagher laboratories are supported by the Irish Cancer Society Collaborative Cancer Research Centre, BREAST-PREDICT Grant, CCRC13GAL (http://www.breastpredict.com). MRS and JCG were supported by the Office of Science, Office of Biological and Environmental Research, of the US Department of Energy under Contract No. DE-AC02-05CH11231. Potential conflict of interest: Dr Adrian Bracken, Dr Fiona Lanigan and Professor William Gallagher are co-inventors of a patent application underlying the OncoMasTR technology. Professor Gallagher is a co-founder and Chief Scientific Officer of OncoMark Limited, which recently in-licenced the OncoMasTR technology.

Author contributions

GLB performed all of the ChIP analyses. GLB, AJL and BL performed the RNA-seq and ChIP-seq analyses. FL, GLB and EJ performed the quantitative RT-PCR analyses. FL performed the TMA staining, scoring and statistical analysis. YF, FL and SM performed the statistical and bioinformatics analysis. KH, FL and EJD performed the ARACNE analysis. FAO, CQ, LF, FL and MF performed the TMA scoring. KJ provided TMAs and clinical data. MRS and JCB provided HMEC cells, as well as advice on their culture. FL, GLB and YF prepared the artwork and text, and contributed to the study design. AB, JG and WMG contributed to study design and preparation of the manuscript. All authors approved the final manuscript submitted for publication.

References

1 Bleyer A & Welch HG (2012) Effect of three decades of screening mammography on breast-cancer incidence. *N Engl J Med* **367**, 1998–2005.

2 Fisher B, Jeong JH, Bryant J, Anderson S, Dignam J, Fisher ER & Wolmark N (2004) Treatment of lymph-node-negative, oestrogen-receptor-positive breast cancer: long-term findings from National Surgical Adjuvant Breast and Bowel Project randomised clinical trials. *Lancet* **364**, 858–868.
3 Goldhirsch A, Ingle JN, Gelber RD, Coates AS, Thurlimann B & Senn HJ (2009) Thresholds for therapies: highlights of the St Gallen International Expert Consensus on the primary therapy of early breast cancer 2009. *Ann Oncol* **20**, 1319–1329.
4 Bedard PL & Cardoso F (2011) Can some patients avoid adjuvant chemotherapy for early-stage breast cancer? *Nat Rev Clin Oncol* **8**, 272–279.
5 Chang HY, Nuyten DS, Sneddon JB, Hastie T, Tibshirani R, Sorlie T, Dai H, He YD, van't Veer LJ, Bartelink H *et al.* (2005) Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival. *Proc Natl Acad Sci USA* **102**, 3738–3743.
6 Filipits M, Rudas M, Jakesz R, Dubsky P, Fitzal F, Singer CF, Dietze O, Greil R, Jelen A, Sevelda P *et al.* (2011) A new molecular predictor of distant recurrence in ER-positive, HER2-negative breast cancer adds independent information to conventional clinical risk factors. *Clin Cancer Res* **17**, 6012–6020.
7 Paik S, Shak S, Tang G, Kim C, Baker J, Cronin M, Baehner FL, Walker MG, Watson D, Park T *et al.* (2004) A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N Engl J Med* **351**, 2817–2826.
8 Sorlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, Hastie T, Eisen MB, van de Rijn M, Jeffrey SS *et al.* (2001) Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci USA* **98**, 10869–10874.
9 Sotiriou C, Wirapati P, Loi S, Harris A, Fox S, Smeds J, Nordgren H, Farmer P, Praz V, Haibe-Kains B *et al.* (2006) Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *J Natl Cancer Inst* **98**, 262–272.
10 van't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT *et al.* (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**, 530–536.
11 van de Vijver MJ, He YD, van't Veer LJ, Dai H, Hart AA, Voskuil DW, Schreiber GJ, Peterse JL, Roberts C, Marton MJ *et al.* (2002) A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med* **347**, 1999–2009.
12 Wang Y, Klijn JG, Zhang Y, Sieuwerts AM, Look MP, Yang F, Talantov D, Timmermans M, Meijer-van Gelder ME, Yu J *et al.* (2005) Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet* **365**, 671–679.
13 Sgroi DC, Sestak I, Cuzick J, Zhang Y, Schnabel CA, Schroeder B, Erlander MG, Dunbier A, Sidhu K, Lopez-Knowles E *et al.* (2013) Prediction of late distant recurrence in patients with oestrogen-receptor-positive breast cancer: a prospective comparison of the breast-cancer index (BCI) assay, 21-gene recurrence score, and IHC4 in the TransATAC study population. *Lancet Oncol* **14**, 1067–1076.
14 Cardoso F, Van't Veer L, Rutgers E, Loi S, Mook S & Piccart-Gebhart MJ (2008) Clinical application of the 70-gene profile: the MINDACT trial. *J Clin Oncol* **26**, 729–735.
15 Sparano JA (2006) TAILORx: trial assigning individualized options for treatment (Rx). *Clin Breast Cancer* **7**, 347–350.
16 Fan C, Oh DS, Wessels L, Weigelt B, Nuyten DS, Nobel AB, van't Veer LJ & Perou CM (2006) Concordance among gene-expression-based predictors for breast cancer. *N Engl J Med* **355**, 560–569.
17 Haibe-Kains B, Desmedt C, Piette F, Buyse M, Cardoso F, Van't VeerL, Piccart M, Bontempi G & Sotiriou C (2008) Comparison of prognostic gene expression signatures for breast cancer. *BMC Genom* **9**, 394.
18 Wirapati P, Sotiriou C, Kunkel S, Farmer P, Pradervand S, Haibe-Kains B, Desmedt C, Ignatiadis M, Sengstag T, Schutz F *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**, R65.
19 Margolin AA, Wang K, Lim WK, Kustagi M, Nemenman I & Califano A (2006) Reverse engineering cellular networks. *Nat Protoc* **1**, 662–671.
20 Carro MS, Lim WK, Alvarez MJ, Bollo RJ, Zhao X, Snyder EY, Sulman EP, Anne SL, Doetsch F, Colman H *et al.* (2010) The transcriptional network for mesenchymal transformation of brain tumours. *Nature* **463**, 318–325.
21 Loi S, Haibe-Kains B, Desmedt C, Lallemand F, Tutt AM, Gillet C, Ellis P, Harris A, Bergh J, Foekens JA *et al.* (2007) Definition of clinically distinct molecular subtypes in estrogen receptor-positive breast carcinomas through genomic grade. *J Clin Oncol* **25**, 1239–1246.
22 ExPO GSE2109, http://wwwintgenorg/expo/.
23 Miller LD, Smeds J, George J, Vega VB, Vergara L, Ploner A, Pawitan Y, Hall P, Klaar S, Liu ET *et al.* (2005) An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. *Proc Natl Acad Sci USA* **102**, 13550–13555.

24 Mosley JD & Keri RA (2008) Cell cycle correlated genes dictate the prognostic power of breast cancer gene lists. *BMC Med Genomics* **1**, 11.
25 Garbe JC, Bhattacharya S, Merchant B, Bassett E, Swisshelm K, Feiler HS, Wyrobek AJ & Stampfer MR (2009) Molecular distinctions between stasis and telomere attrition senescence barriers shown by long-term culture of normal human mammary epithelial cells. *Cancer Res* **69**, 7557–7568.
26 Zindy F, Quelle DE, Roussel MF & Sherr CJ (1997) Expression of the p16INK4a tumor suppressor versus other INK4 family members during mouse development and aging. *Oncogene* **15**, 203–211.
27 Bracken AP, Pasini D, Capra M, Prosperini E, Colli E & Helin K (2003) EZH2 is downstream of the pRB-E2F pathway, essential for proliferation and amplified in cancer. *EMBO J* **22**, 5323–5335.
28 Garbe JC, Pepin F, Pelissier FA, Sputova K, Fridriksdottir AJ, Guo DE, Villadsen R, Park M, Petersen OW, Borowsky AD *et al.* (2012) Accumulation of multipotent progenitors with a basal differentiation bias during aging of human mammary epithelia. *Cancer Res* **72**, 3687–3701.
29 Boyer LA, Lee TI, Cole MF, Johnstone SE, Levine SS, Zucker JP, Guenther MG, Kumar RM, Murray HL, Jenner RG *et al.* (2005) Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell* **122**, 947–956.
30 Tong Y & Eigler T (2009) Transcriptional targets for pituitary tumor-transforming gene-1. *J Mol Endocrinol* **43**, 179–185.
31 Tong Y, Tan Y, Zhou C & Melmed S (2007) Pituitary tumor transforming gene interacts with Sp1 to modulate G1/S cell phase transition. *Oncogene* **26**, 5596–5605.
32 Bostick M, Kim JK, Esteve PO, Clark A, Pradhan S & Jacobsen SE (2007) UHRF1 plays a role in maintaining DNA methylation in mammalian cells. *Science* **317**, 1760–1764.
33 TCGA (2012) Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70.
34 Hara E, Smith R, Parry D, Tahara H, Stone S & Peters G (1996) Regulation of p16CDKN2 expression and its implications for cell immortalization and senescence. *Mol Cell Biol* **16**, 859–867.
35 Kotake Y, Cao R, Viatour P, Sage J, Zhang Y & Xiong Y (2007) pRB family proteins are required for H3K27 trimethylation and Polycomb repression complexes binding to and silencing p16INK4alpha tumor suppressor gene. *Genes Dev* **21**, 49–54.
36 Li Y, Nichols MA, Shay JW & Xiong Y (1994) Transcriptional repression of the D-type cyclin-dependent kinase inhibitor p16 by the retinoblastoma susceptibility gene product pRb. *Cancer Res* **54**, 6078–6082.
37 Tam SW, Shay JW & Pagano M (1994) Differential expression and cell cycle regulation of the cyclin-dependent kinase 4 inhibitor p16Ink4. *Cancer Res* **54**, 5816–5820.
38 Kelly CM, Krishnamurthy S, Bianchini G, Litton JK, Gonzalez-Angulo AM, Hortobagyi GN & Pusztai L (2010) Utility of oncotype DX risk estimates in clinically intermediate risk hormone receptor-positive, HER2-normal, grade II, lymph node-negative breast cancers. *Cancer* **116**, 5161–5167.
39 Goldstein LJ, Gray R, Badve S, Childs BH, Yoshizawa C, Rowley S, Shak S, Baehner FL, Ravdin PM, Davidson NE *et al.* (2008) Prognostic utility of the 21-gene assay in hormone receptor-positive operable breast cancer compared with classical clinicopathologic features. *J Clin Oncol* **26**, 4063–4071.
40 Lee TI & Young RA (2013) Transcriptional regulation and its misregulation in disease. *Cell* **152**, 1237–1251.
41 Hanahan D & Weinberg RA (2011) Hallmarks of cancer: the next generation. *Cell* **144**, 646–674.
42 Collado M & Serrano M (2010) Senescence in tumours: evidence from mice and humans. *Nat Rev Cancer* **10**, 51–57.
43 Roberson RS, Kussick SJ, Vallieres E, Chen SY & Wu DY (2005) Escape from therapy-induced accelerated cellular senescence in p53-null lung cancer cells and in human lung cancers. *Cancer Res* **65**, 2795–2803.
44 Schmitt CA, Fridman JS, Yang M, Lee S, Baranov E, Hoffman RM & Lowe SW (2002) A senescence program controlled by p53 and p16INK4a contributes to the outcome of cancer therapy. *Cell* **109**, 335–346.
45 te Poele RH, Okorokov AL, Jardine L, Cummings J & Joel SP (2002) DNA damage is able to induce senescence in tumor cells in vitro and in vivo. *Cancer Res* **62**, 1876–1883.
46 Wu CH, van Riggelen J, Yetil A, Fan AC, Bachireddy P & Felsher DW (2007) Cellular senescence is an important mechanism of tumor regression upon c-Myc inactivation. *Proc Natl Acad Sci USA* **104**, 13028–13033.
47 Ventura A, Kirsch DG, McLaughlin ME, Tuveson DA, Grimm J, Lintault L, Newman J, Reczek EE, Weissleder R & Jacks T (2007) Restoration of p53 function leads to tumour regression in vivo. *Nature* **445**, 661–665.
48 Xue W, Zender L, Miething C, Dickins RA, Hernando E, Krizhanovsky V, Cordon-Cardo C & Lowe SW (2007) Senescence and tumour clearance is triggered by p53 restoration in murine liver carcinomas. *Nature* **445**, 656–660.
49 Hui R, Macmillan RD, Kenny FS, Musgrove EA, Blamey RW, Nicholson RI, Robertson JF & Sutherland RL (2000) INK4a gene expression and methylation in primary breast cancer: overexpression of p16INK4a messenger RNA is a marker of poor prognosis. *Clin Cancer Res* **6**, 2777–2787.

50 Milde-Langosch K, Bamberger AM, Rieck G, Kelp B & Loning T (2001) Overexpression of the p16 cell cycle inhibitor in breast cancer is associated with a more malignant phenotype. *Breast Cancer Res Treat* **67**, 61–70.
51 Emig R, Magener A, Ehemann V, Meyer A, Stilgenbauer F, Volkmann M, Wallwiener D & Sinn HP (1998) Aberrant cytoplasmic expression of the p16 protein in breast cancer is associated with accelerated tumour proliferation. *Br J Cancer* **78**, 1661–1668.
52 Han S, Ahn SH, Park K, Bae BN, Kim KH, Kim HJ, Kim YD & Kim HY (2001) P16INK4a protein expression is associated with poor survival of the breast cancer patients after CMF chemotherapy. *Breast Cancer Res Treat* **70**, 205–212.
53 Singh M, Parnes MB, Spoelstra N, Bleile MJ & Robinson WA (2004) p16 expression in sentinel nodes with metastatic breast carcinoma: evaluation of its role in developing triaging strategies for axillary node dissection and a marker of poor prognosis. *Hum Pathol* **35**, 1524–1530.
54 Peurala E, Koivunen P, Haapasaari KM, Bloigu R & Jukkola-Vuorinen A (2013) The prognostic significance and value of cyclin D1, CDK4 and p16 in human breast cancer. *Breast Cancer Res* **15**, R5.
55 Li H & Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics* **25**, 1754–1760.
56 Trapnell C, Pachter L & Salzberg SL (2009) TopHat: Discovering splice junctions with RNA-seq. *Bioinformatics* **25**, 1105–1111.
57 Hokamp K, Roche FM, Acab M, Rousseau ME, Kuo B, Goode D, Aeschliman D, Bryan J, Babiuk LA, Hancock RE *et al.* (2004) ArrayPipe: a flexible processing pipeline for microarray data. *Nucleic Acids Res* **32**, W457–W459.
58 Madden SF, Clarke C, Gaule P, Aherne ST, O'Donovan N, Clynes M, Crown J & Gallagher WM (2013) BreastMark: an integrated approach to mining publicly available transcriptomic datasets relating to breast cancer outcome. *Breast Cancer Res* **15**, R52.
59 Brien GL, Gambero G, O'Connell DJ, Jerman E, Turner SA, Egan CM, Dunne EJ, Jurgens MC, Wynne K, Piao L *et al.* (2012) Polycomb PHF19 binds H3K36me3 and recruits PRC2 and demethylase NO66 to embryonic stem cell genes during differentiation. *Nat Struct Mol Biol* **19**, 1273–1281.
60 Langmead B, Trapnell C, Pop M & Salzberg SL Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**, R25.
61 Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W *et al.* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**, R137.
62 Svensson S, Jirstrom K, Ryden L, Roos G, Emdin S, Ostrowski MC & Landberg G (2005) ERK phosphorylation is linked to VEGFR2 expression and Ets-2 phosphorylation in breast cancer and is associated with tamoxifen treatment resistance and small tumours with good prognosis. *Oncogene* **24**, 4370–4379.
63 Lanigan F, McKiernan E, Brennan DJ, Hegarty S, Millikan RC, McBryan J, Jirstrom K, Landberg G, Martin F, Duffy MJ *et al.* (2009) Increased claudin-4 expression is associated with poor prognosis and high tumour grade in breast cancer. *Int J Cancer* **124**, 2088–2097.