

ORIGINAL PAPER

Mutation of *GATA3* in human breast tumors

Jerry Usary^{1,2}, Victor Llaca³, Gamze Karaca^{1,2}, Shafaq Presswala³, Mehmet Karaca^{1,2}, Xiaping He^{1,2}, Anita Langerød⁴, Rolf Kåresen⁵, Daniel S Oh^{1,2}, Lynn G Dressler⁶, Per E Lønning⁷, Robert L Strausberg⁸, Stephen Chanock³, Anne-Lise Børresen-Dale⁴ and Charles M Perou^{*,1,2,9}

¹Lineberger Comprehensive Cancer Center, University of North Carolina, Chapel Hill, NC 27599, USA; ²Department of Genetics, University of North Carolina, Chapel Hill, NC 27599, USA; ³Section on Genomic Variation, Pediatric Oncology Branch, Advanced Technology Center, National Cancer Institute, 8717 Grovemont Circle, Gaithersburg, MD 20892-4605, USA; ⁴Department of Genetics, The Norwegian Radium Hospital, 0310 Oslo, Norway; ⁵Department of Surgery, Ullevål University Hospital, Oslo, Norway; ⁶Department of Medicine, University of North Carolina, Chapel Hill, NC 27599, USA; ⁷Department of Medicine, Section of Oncology, Haukeland University Hospital, 5021 Bergen, Norway; ⁸The Institute for Genomic Research, 9712 Medical Drive, Rockville, MD 20850, USA; ⁹Department of Pathology and Laboratory Medicine, University of North Carolina, Chapel Hill, NC 27599, USA

GATA3 is an essential transcription factor that was first identified as a regulator of immune cell function. In recent microarray analyses of human breast tumors, both normal breast luminal epithelium and estrogen receptor (ESR1)-positive tumors showed high expression of GATA3. We sequenced genomic DNA from 111 breast tumors and three breast-tumor-derived cell lines and identified somatic mutations of *GATA3* in five tumors and the MCF-7 cell line. These mutations cluster in the vicinity of the highly conserved second zinc-finger that is required for DNA binding. In addition to these five, we identified using cDNA sequencing a unique mis-splicing variant that caused a frameshift mutation. One of the somatic mutations we identified was identical to a germline *GATA3* mutation reported in two kindreds with HDR syndrome/OMIM #146255, which is an autosomal dominant syndrome caused by the haplo-insufficiency of *GATA3*. The ectopic expression of *GATA3* in human 293T cells caused the induction of 73 genes including six cytokeratins, and inhibited cell line doubling times. These data suggest that *GATA3* is involved in growth control and the maintenance of the differentiated state in epithelial cells, and that *GATA3* variants may contribute to tumorigenesis in ESR1-positive breast tumors.

Oncogene advance online publication, 13 September 2004; doi:10.1038/sj.onc.1207966

Keywords: breast cancer; microarrays; estrogen receptor; transcription factor; *GATA3*

Introduction

The GATA family of zinc-finger transcription factors are critical for the development and differentiation of cell types in vertebrates (Ho *et al.*, 1991; Orkin, 1992;

Pandolfi *et al.*, 1995; Lim *et al.*, 2000). *GATA3* is a prototypical member of the family and was initially identified as a DNA-binding protein involved in the activation of transcription at the T-cell receptor alpha locus (Ho *et al.*, 1991). The phenotype of a knockout of *GATA3* by homologous recombination in the mouse includes significant developmental abnormalities and is lethal during the embryonic stage (Pandolfi *et al.*, 1995; Lim *et al.*, 2000). Van Esch *et al.* (2000) identified mutations in *GATA3* from kindreds notable for a rare and complex disease of hypoparathyroidism, sensorineural deafness and renal insufficiency (HDR syndrome/OMIM #146255). In addition, significant mutations in other members of the GATA family have been defined in other human diseases; for example, an acquired mutation in *GATA1* results in megakaryoblastic leukemia in children with Down's syndrome (Wechsler *et al.*, 2002; Groet *et al.*, 2003), while germline mutations of human *GATA4* have been associated with congenital heart defects (Garg *et al.*, 2003).

Gene expression profiling studies have demonstrated that *GATA3* is highly expressed in a subset of human breast tumors (Perou *et al.*, 2000; Gruberger *et al.*, 2001; West *et al.*, 2001; van 't Veer *et al.*, 2002) and that *GATA3* expression in breast tumors highly correlates with expression of the estrogen receptor alpha gene/protein (ESR1) (Hoch *et al.*, 1999; van de Rijn *et al.*, 2002). Analysis of our most recent cohort of breast tumors ($n = 115$) confirmed that expression of *GATA3* correlated with the expression of a subset of genes considered important in breast luminal epithelial cell biology, including *ESR1*, *LIV-1* (*SLC39A6*), *REGG* and *TFF3* (Sorlie *et al.*, 2003). *GATA3* is also part of a larger gene set that we have termed the breast 'intrinsic' gene set, which identifies distinct tumor subtypes that show differences in survival outcomes for our cohort, as well as other cohorts (Sørli *et al.*, 2001; Sørli *et al.*, 2003). The highest expression of *GATA3* and *ESR1* is seen in tumors of the 'Luminal A' subtype, which is a subtype that is associated with the most favorable survival outcomes. On the other hand, lower expression of

*Correspondence: CM Perou; E-mail: cperou@med.unc.edu

Received 9 February 2004; revised 28 April 2004; accepted 11 May 2004

GATA3 and *ESR1* is characteristic of the 'Luminal B' ER+ subtype, which is associated with poor survival. Still lower expression of *GATA3* is seen in 'HER2+' and 'Basal-like' tumor subtypes, which show the worst outcomes. Here we show that *GATA3* is mutated in a subset of ESR1-positive breast tumors and demonstrate that these alterations likely impair *GATA3* function.

Results

GATA3 variant identification

Based upon our DNA microarray and tissue array studies, we determined that the highest expression of *GATA3* in breast tumors portends good patient outcomes, and that, as *GATA3* levels lower, the prognosis worsens (Sørli *et al.*, 2001; van de Rijn *et al.*, 2002; Sørli *et al.*, 2003). To determine if somatic mutations of *GATA3* occur in breast tumors, bidirectional sequence analysis was performed on all exons and the adjacent intron/exon boundaries of the human *GATA3* gene. Genomic DNA from 111 grossly dissected breast tumors was analysed and five heterozygous *GATA3* variants were identified (see Supplementary Materials Figure 1 for the sequence traces and Table 1), all of which occurred in clinically defined ESR1-positive tumors. These variants clustered in a region of *GATA3* near its C-terminal second zinc-finger (Figure 1b), which is highly conserved across vertebrate GATA family members (Figure 1d). The spectrum of mutations identified included two missense, a base insertion, a nonsense and a two base pair deletion that alters a splice acceptor site. The base insertion, nonsense and splice acceptor site mutations predict truncated proteins, while the two missense mutations were observed in conserved residues (Figure 1d). In addition, tumor BR99-0207 showed multiple bands by Western blot analysis (Figure 2) and we identified an unusual splicing alteration that was only identified by the direct sequencing of cDNA from this tumor (Figure 1c). The cDNA sequence analysis of BR99-0207, shown in Supplementary Materials Figure 1m, identified two different exon 4 to exon 5 splice products; one product generated the properly spliced mRNA, while the second event utilized an alternative AG for the exon 5 splice acceptor site. The abnormal splice product resulted in a seven base deletion in the mature mRNA, which then created a frameshift alteration in exon 5. We sequenced

the 5.3 kb intron between exon 4 and exon 5 in BR99-0207 and no genomic DNA alteration was identified.

As germline genomic DNA was not available from the six patients with *GATA3* alterations, we sequenced *GATA3* in 92 healthy controls from Norway and another 102 control individuals; in the analysis of 388 chromosomes, we failed to observe any of the variants that were present in the tumors, nor did we observe that one or more of these sequence variants could be considered common or rare polymorphisms (Kruglyak and Nickerson, 2001). Several facts strongly suggest that the observed sequence variants are not polymorphic alleles, but are instead somatic mutations of *GATA3*: (1) failure to observe any of the variants in 388 normal chromosomes, (2) the variants occurred in a highly conserved and functionally critical region, (3) the spectrum of mutational events are disruptive of the protein structure and (4) none of the patients showed any evidence of HDR syndrome. An additional mutation was found in the breast-tumor-derived cell line MCF-7, in which a heterozygous guanine nucleotide was inserted at position 1566 (Supplementary Materials Figure 1f). The mutations identified in MCF-7, BR99-0348 and BR00-0587 were confirmed using RT-PCR analysis of total RNA and direct sequencing of cDNA products (Supplementary Materials Figure 1f, k and data not shown). In all the three cases, both wild-type and mutant mRNA sequences were detected.

GATA3 Western blot analysis

We performed Western blots using a monoclonal antibody specific for *GATA3* on 11 of the tumors that were sequenced for *GATA3* mutations (not all tumors had materials available for protein analysis), one normal breast sample and on MCF-7 cells (Figure 2). Previous studies have shown that *GATA3* mRNA is expressed in normal breast tissue. *GATA3* is predicted to encode a 48 kDa protein that is known to undergo alternative splicing (<http://www.ncbi.nlm.nih.gov/IEB/Research/Acembly/index.html>). By Western blot analysis, we identified the expected 48 kDa band as well as a smaller approximately 23 kDa band when we used the monoclonal antibody sc-269 (Figure 2). Both bands were also identified when we used a second monoclonal antibody (sc-268) raised against human *GATA3* (Figure 2, far right lane), and both bands were detected by Nesbit *et al.* (2004) when they performed *in vitro* transcription of *GATA3*, followed by Western blot analyses (using sc-

Table 1 *GATA3* variants in tumor samples and the MCF-7 cell line

Sample	Sample type	Location	Nucleotide change	Amino-acid change
Ull-214	Breast tumor	Exon 4-5	CA deletion	Splice site deletion
Ull-030	Breast tumor	Exon 5	A insertion	T316 frameshift
BR99-0207	Breast tumor	Exon 4-5	7 bp skip in cDNA	Exon 5 frameshift
BR00-0587	Breast tumor	Exon 5	CTC > TTC	Leu344-Phe
Ull-011	Breast tumor	Exon 6	CGA > CTA	Arg367-Leu
BR99-0348	Breast tumor	Exon 6	CGA > TGA	Arg367-Stop
MCF-7	Cell line	Exon 5	G insertion	D336 frameshift

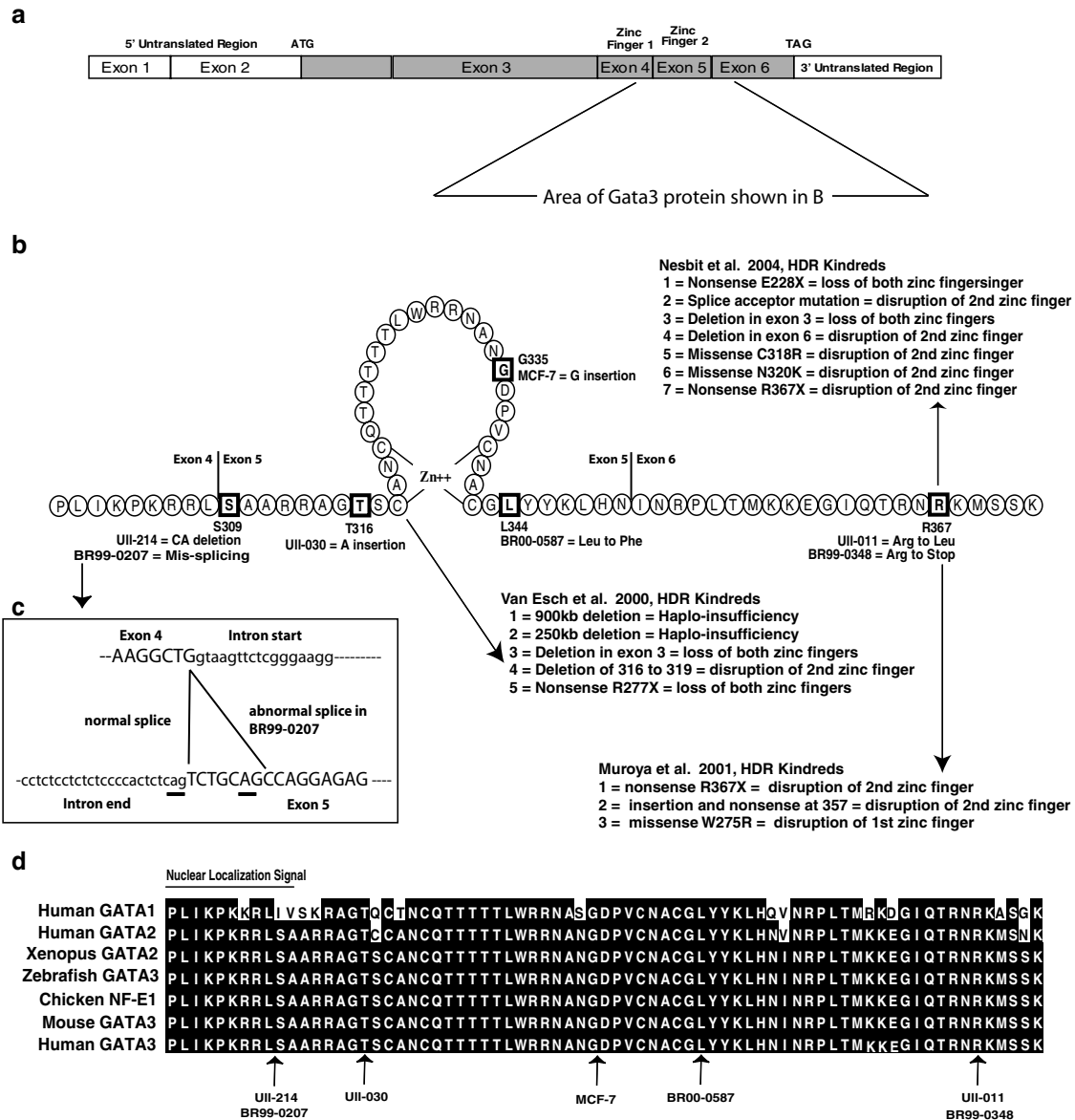


Figure 1 Human *GATA3* gene and protein overview. **(a)** Exon structure of *GATA3* showing the functional domains and region of mutations. **(b)** *GATA3* protein sequence surrounding the second zinc-finger and the location of variants identified in this study, and in three different HDR kindred studies. **(c)** Diagrammatic representation of the mis-splicing event that occurred in tumor BR99-0207. **(d)** Amino-acid identity plot of similar vertebrate *GATA* proteins and the locations of the breast tumor *GATA3* variants; identical residues are identified by black boxes

268) on wild-type and *GATA3* mutants, although the 23kDa band was weak. These data suggest that the lower band represents the *GATA3* protein and not an unknown crossreacting protein. In each breast tumor sample assayed, we detected both *GATA3* bands; however, it is likely that these products in the two *ESR1*-negative tumors were detected due to contamination by normal breast tissue because the tumor cells from these two samples were negative for *GATA3* by immunohistochemistry (data not shown).

Tumor BR99-0348 contained a nonsense mutation and showed a protein product that corresponded to the size of the predicted truncated protein (38 kDa). The mutation in the MCF-7 cell line also predicts a

truncated protein, as does the BR99-0207 splicing alteration, both of which were precisely observed by Western blot analysis (Figure 2). It is worth noting that, in BR99-0207, BR99-0348 and MCF-7, the levels of the 48 kDa wild-type *GATA3* protein were very low or not detected, even though wild-type mRNA was expressed. No *GATA3* protein reactivity was detected in the luminal-tumor-derived cell line ZR-75-1, in the basal-like immortalized human mammary epithelial cell line ME16C, or in the human embryonal kidney epithelial cell line 293T (data not shown). In Figure 2, two of the breast tumor samples (BR00-0587 and BR00-0365) showed aberrant *GATA3* protein profiles when compared to normal breast; tumor BR00-0587 showed an

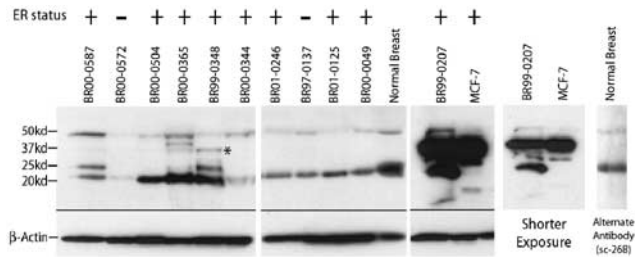


Figure 2 Western blot analysis of GATA3 protein expression. In all, 11 breast tumors, one normal breast sample and the MCF-7 cell line protein extracts were run on a 4–20% SDS-PAGE gel and assayed for expression of GATA3 protein using the sc-269 monoclonal antibody. The expected wild-type band of 48 kDa was seen in most lanes along with a smaller approximately 23 kDa band, while tumor BR99-0348 (identified by asterisk), BR99-0207 and the MCF-7 cell line showed their predicted truncated GATA3 protein products. On the right is a shorter exposure of BR99-0207 and MCF-7 to show finer detail because their expression levels were high compared to the other samples. On the far right is a Western blot of the same normal breast sample using a second monoclonal antibody (sc-268), which also identified both the 48 and 23 kDa bands

extra band while tumor BR00-0365 showed three. Tumor BR00-0587 contained a missense mutation; however, this does not explain the extra band, and in tumor BR00-0365 no alteration at the DNA level was detected. These data suggest that either we missed DNA sequence variants in these two tumors, or that these two tumors have alterations in the post-translational processing or modification of the GATA3 protein.

GATA3 immunohistochemistry

To further explore the biology of GATA3 in breast tissues, we performed immunohistochemistry (IHC) on paraffin-embedded sections from a normal breast sample for GATA3 and other markers of breast luminal (ESR1) and myoepithelial cells (CK5/6 and CK17). ESR1 staining in a terminal ductal lobular unit gave a characteristic nuclear staining of the luminal cells that line the ducts (Figure 3a), while the staining pattern for CK5/6 and CK17 (Figure 3c and d) identified the outer cell layer that likely represents the myoepithelial cells. GATA3 staining gave a very similar pattern to ESR1 and showed nuclear staining in the cells that line the ducts (Figure 3b).

Next, tumors were stained for ESR1 and GATA3; BR99-0348, BR00-0587, BR99-0207 and Ull-030 all gave strong nuclear staining for ESR1 (Figure 4a, c, e and g) and nuclear staining for GATA3 (Figure 4b, d, f and h), despite the observation that BR99-0348, BR99-0207 and Ull-030 have truncated GATA3 proteins. Tumor Ull-011 was positive for ESR1 and showed no immunohistochemical staining for GATA3 (Figure 4i, j). Finally, tumor Ull-214 is the splice site deletion mutant, which is predicted to disrupt exon 4 to exon 5 splicing. If correct, this mutation will disrupt the amino acids shown by Yang *et al.* (1994) (residues 249–311), to be required for the nuclear localization of GATA3. This

is the same region predicted by PSORT (<http://psort.nibb.ac.jp>) to contain a nuclear localization signal. Immunohistochemical analysis of Ull-214 for GATA3 revealed a cytoplasmic staining pattern in the tumor cells and a nuclear staining pattern in the adjacent normal breast ductal cells (Figure 4l); this finding demonstrates that patient Ull-214 had a somatically acquired mutation.

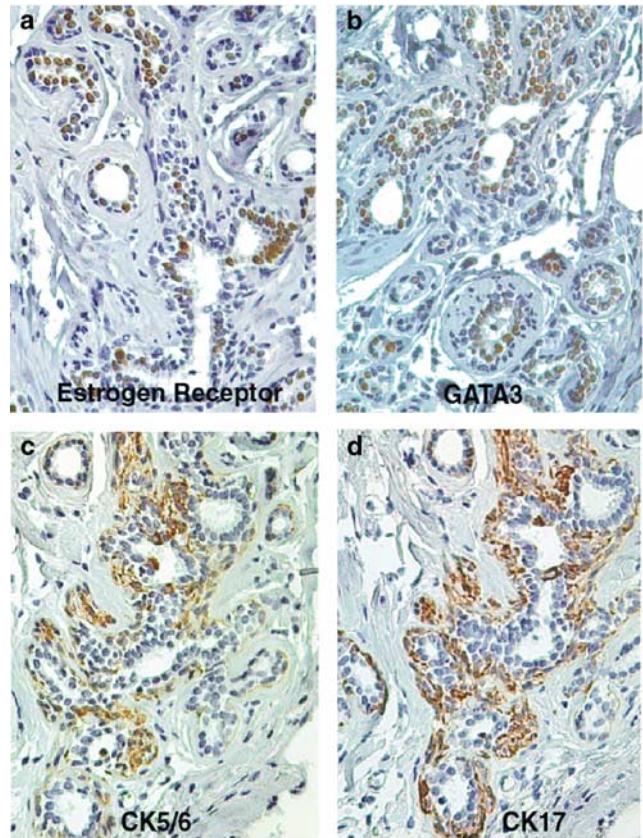


Figure 3 Immunohistochemical analysis of markers of breast luminal and myoepithelial cells in adult normal breast. Consecutive sections of the same terminal ductal lobular unit (TDLU) from an adult female were stained for (a) the estrogen receptor (ESR1), (b) GATA3, (c) cytokeratin 5/6 and (d) cytokeratin 17. Note that in panels a and b the staining is nuclear and typically stains the inner cell layer, while in c and d, the staining is cytoplasmic and stains the outer cell layer. Magnification is $\times 200$

Figure 4 Immunohistochemical analysis of the Estrogen Receptor (ESR1) and GATA3 in breast tumors. Immunohistochemistry was performed on the six *GATA3* mutated tumors for ESR1 (a, c, e, g, i, k) and for GATA3 (b, d, f, h, j, l). (a) Tumor BR99-0348 stained for the ESR1. (b) Tumor BR99-0348 stained for GATA3. (c) Tumor BR00-0587 stained for ESR1. (d) Tumor BR00-0587 stained for GATA3. (e) Tumor BR99-0207 stained for ESR1. (f) Tumor BR99-0207 stained for GATA3. (g) Tumor Ull-030 stained for ESR1. (h) Tumor Ull-030 stained for GATA3. (i) Tumor Ull-011 stained for ESR1. (j) Tumor Ull-011 stained for GATA3. (k) Tumor Ull-214 stained for ESR1. (l) Tumor Ull-214 stained for GATA3; the arrow denotes a normal breast duct that shows nuclear GATA3 staining. The magnification is $\times 150$

Ectopic expression of GATA3 in human epithelial cell lines

We cloned full-length *GATA3* into the pBabe retroviral expression vector (Morgenstern and Land, 1990) and created a site-directed mutant that corresponded to one of the mutations identified in our breast tumors (R367L). We first attempted to create stable transfectant

cell lines expressing wild-type *GATA3* (*GATA3*-WT) and the *GATA3*-R367L in the ZR-75-1 and ME16C breast epithelial cell lines (both negative for *GATA3*), and were unable to obtain any transfectants, suggesting that *GATA3* may be 'toxic' to these cells. Next, we turned to the highly transfectable human embryonal kidney epithelial cell line 293T (Graham *et al.*, 1977; Sena-Esteves *et al.*, 1999) and we were able to obtain stable transfectant cell populations expressing *GATA3*-WT and *GATA3*-R367L. We noted that the *GATA3*-WT transfectants had a slower growth rate than their transfection-matched empty vector control cell lines (Figure 5). Two independent transfectants for *GATA3*-WT and *GATA3*-R367L were created and, in both cases, the *GATA3*-WT lines had slower growth rates ($P=0.025$ and <0.001) than their transfection-matched empty vector control lines, while the *GATA3*-R367L transfectants grew at the same rate ($P=0.55$ and 0.81) as their empty vector controls.

We quantitated the expression level of *GATA3* mRNA in these transfectants using quantitative RT-PCR analysis and determined that the *GATA3* transgenes were expressed at fivefold (WT) and fourfold (R367L) above 293T baseline *GATA3* levels. The 293T transfectants did not show any *GATA3* protein by Western analysis; however, by quantitative RT-PCR analysis, *GATA3* was expressed at 25-fold lower levels than normal breast, and 400-fold less than MCF-7. Therefore, the 293T cell lines with 'overexpressed' *GATA3*-WT and *GATA3*-R367L were still 5–6-fold lower than normal breast levels, which was below the detection limit of our Western blot analysis. When we paraffin embedded these cell lines and assayed for *GATA3* expression using IHC, we detected strong nuclear staining in MCF-7 cells as expected, and nuclear staining in the *GATA3*-WT expression line (Supplementary Materials Figure 2a, b). The empty vector

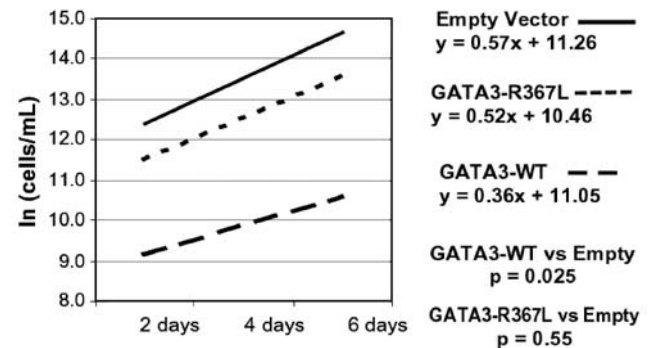
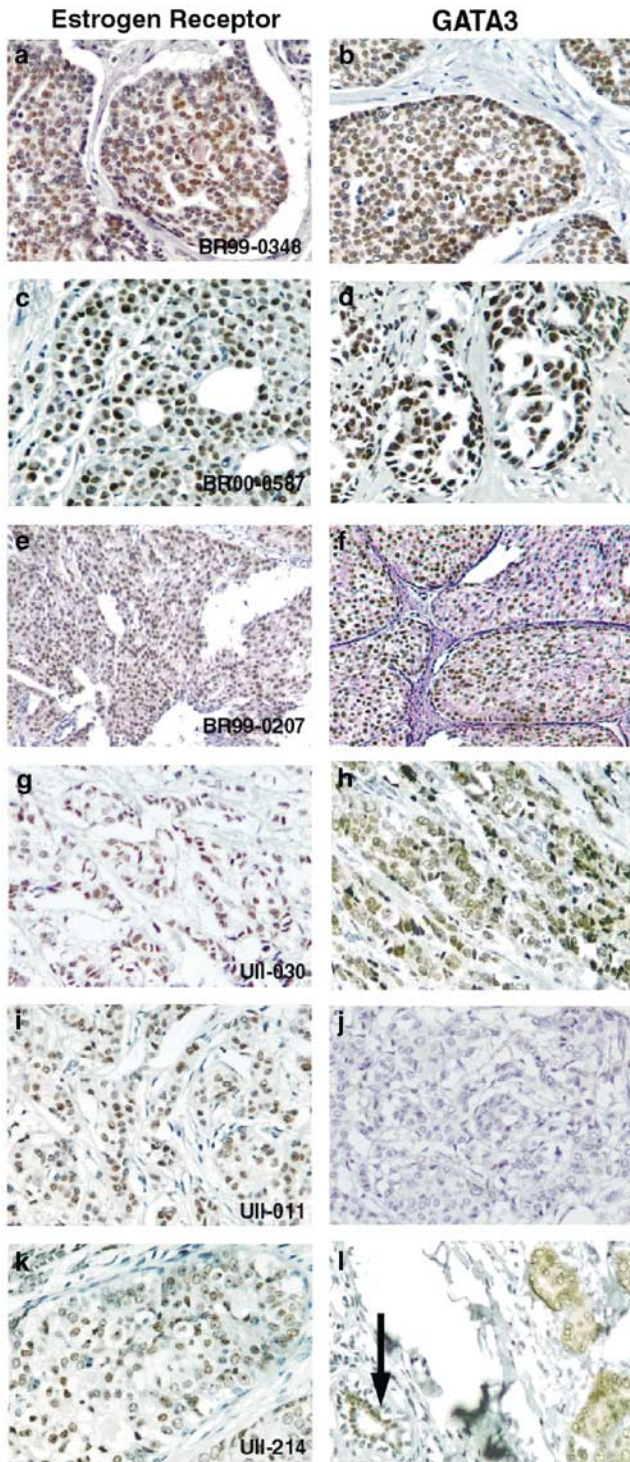


Figure 5 Log-normal growth curves for 293T transfectants. 293T cells were transfected with the pBabe empty vector, pBabe-*GATA3*-WT or pBabe-*GATA3*-R367L. We determined the PDTs from the slope of these curves (see Materials and methods) and used a multiple regression model, as described in Troester *et al.* (2002), to determine whether the slopes of the growth curves were significantly different. The growth rates for the *GATA3*-R367L and the empty vector controls were not different, while the *GATA3*-WT transfectant growth rate was different from that of the empty vector control. These are the same transfectants assayed by microarray analysis

control line and the *GATA3*-R367L mutant did not show any nuclear staining (Supplementary Materials Figure 2c, d), which suggests that the R367L mutant may be a less stable protein than WT since, by quantitative RT-PCR, the *GATA3*-WT and R367L were expressed at similar levels (five- and fourfold induced, respectively). Further support for this hypothesis comes from the IHC on tumor Ull-011 (which is the R367L mutant), which showed no staining for *GATA3* protein (Figure 4j).

To further investigate the function of *GATA3* in epithelial cells, each *GATA3* transfectant was assayed *versus* a pool of three 293T empty vector control lines using Agilent DNA microarrays containing approximately 18 000 different human genes (see Materials and methods). We performed a 'one-class' SAM supervised analysis to identify genes that differed in expression in the *GATA3*-WT and *GATA3*-R367L cell lines *versus* the empty vector controls (Tusher *et al.*, 2001). We identified 89 spots, which represented 73 induced and one repressed genes (with a false-discovery rate of 1), that showed significantly altered expression in both the *GATA3*-WT and *GATA3*-R367L transfectants (Supplementary Materials Table 1); we also performed an SAM analysis on the *GATA3*-WT line *versus* the *GATA3*-R367L line and no differentially expressed genes were detected. We next examined the promoter regions of these genes (defined as the 5 kb upstream of the ATG and 100 bp downstream of the ATG) using rVista (Loots *et al.*, 2002); rVista first uses a human-to-mouse similarity metric to identify regions of sequence conservation and then searches the conserved regions for the consensus-binding site for *GATA3*. This analysis determined that 38/61 (62%) of the genes tested by rVista contained at least one *GATA3*-binding site in their promoter region, including all five keratin genes. The analysis of a subset of genes not believed to be *GATA3* regulated because they appear in CK5/6-positive basal-like tumors (Sørli *et al.*, 2003), showed that 6/20 (30%) genes tested by rVista had *GATA3* sites; however, 2/6 genes with *GATA3* sites were also keratin genes (CK5 and CK17).

Despite the fact that 293T cells are a kidney-derived cell line, included among the 73 induced genes were many known to be involved in breast cancer biology, including the four cytokeratin genes that define cells of the luminal epithelial lineage (cytokeratins 7, 8, 18 and 19) (Dairkee *et al.*, 1988; Guelstein *et al.*, 1988; Bocker *et al.*, 1992; Boecker and Buerger, 2003) and the gene for *ERBB2*, which is critically important in a subset of breast tumors. In addition to *ERBB2*, *GRB7* was also induced, which is a gene located near *ERBB2* on chromosome 17 and that is typically co-amplified with *ERBB2* (Pollack *et al.*, 2002). Using the software program EASE available from the NIH DAVID website <http://apps1.niaid.nih.gov/david/> (Dennis *et al.*, 2003), which is a program that determines if a given Gene Ontology category is statistically over-represented, we determined that the 'intermediate filament' annotation was over-represented relative to chance (see Supplementary Materials Table 1). *GATA3* also induced the

expression of five different S100 proteins and, according to EASE, the molecular function 'calcium binding' was also significant. The gene for Hepatocyte Nuclear Factor 3 α (*FOXA1*) and *TFF3* were also induced by *GATA3*, and, importantly, these are two genes that clustered very close to *GATA3* in Supplementary Materials Figure 3, which is a close-up of the 'luminal epithelial' gene cluster taken from our previous studies on breast tumors (Sørli *et al.*, 2003). Finally, it should be noted that the gene for *ESR1* was not induced by *GATA3*, nor is *GATA3* regulated by estrogen (Hoch *et al.*, 1999; Finlin *et al.*, 2001).

One measure of the differentiation status of breast tumor cells *in vivo* is tumor grade, which is determined by a pathological assessment of tumor cell duct formation, nuclear morphology and mitotic rate (Tavassoli and Schnitt, 1992). We used SAM to identify genes that correlated with tumor grade using the 115 tumor samples described in Sørli *et al.* (2003) and visualized the data using hierarchical clustering analysis (Supplementary Materials Figure 4); this analysis showed that two dominant sets of genes were identified that include a large set of genes whose expression correlates with proliferation (Perou *et al.*, 1999, 2000; Sørli *et al.*, 2001), and a second set of genes that mostly corresponds to the 'luminal epithelial/*ESR1*+' gene set that includes *GATA3* (Perou *et al.*, 2000; Sørli *et al.*, 2001). We also determined that the high expression of *GATA3* correlates with low tumor grade and low proliferation rates (Sørli *et al.*, 2001; Sørli *et al.*, 2003). Also included (along with *GATA3*) on the *in vivo* list of grade-correlated genes were 10 genes that were induced by *GATA3* in 293T cells, including *TFF1*, *TFF3*, *APM2*, *FXD3*, *KRT7*, *LAD1*, *BACE2*, *MAL2*, *S100A11* and *TFRC*. These data demonstrate that *GATA3*, and *GATA3*-regulated genes, correlate with breast epithelial cell differentiation status *in vivo* and that the 293T experiments identified *GATA3*-regulated genes that are relevant to breast cancer.

Discussion

GATA3 is a highly conserved gene/protein whose absolute expression levels are critical for function. Our data identified multiple changes in the second zinc-finger region of *GATA3* in breast tumors and a tumor-derived cell line. Experimental evidence has shown that the second zinc-finger of *GATA3* is required for DNA binding (Yang *et al.*, 1994), and Van Esch *et al.* showed that the 316–319 deletion abolished the DNA-binding abilities of this *GATA3* variant. In addition, Nesbit *et al.* (2004) have shown that any alteration that affects the second zinc-finger, or the region immediately 3' of this region, also inhibits *GATA3* function *in vitro* and causes HDR syndrome. Smith *et al.* (1995) showed that a C-terminal deletion mutant of *GATA3* truncated after residue 370, failed to activate transcription and was not a dominant-negative interfering protein in transactivation assays using a *GATA3* reporter construct. When all of these data are considered together, in particular the

mutant that shifted the subcellular localization of *GATA3* from the nucleus to the cytoplasm, and the studies of Smith *et al.* and Nesbit *et al.*, it is probable that *GATA3* function is compromised and/or lessened in the breast tumors that have these variants. To our knowledge, none of our breast cancer patients were from pedigrees with HDR syndrome; thus, these somatic mutations in breast tumors provide further support for the functional importance of this region of *GATA3*. It is also interesting to note that, even though the wild-type mRNA was expressed in MCF-7 cells, little or no wild-type protein was produced (Figure 2), suggesting that the presence of the mutant protein may interfere with the expression or stability of the wild-type protein.

In DNA microarray and IHC analyses, *GATA3* is a highly expressed gene/protein whose expression correlates with *ESR1* and a subset of genes important for breast luminal cell biology including *LIV-1* (*SLC39A6*), *REG* and *TFF3* (Supplementary Materials Figure 3) (Finlin *et al.*, 2001; Gruvberger *et al.*, 2001; Sørli *et al.*, 2001; van de Rijn *et al.*, 2002; van't Veer *et al.*, 2002; Sørli *et al.*, 2003). In addition, high *GATA3* (and/or *ESR1*) expression levels predict favorable patient outcomes (Sørli *et al.*, 2001; van de Rijn *et al.*, 2002; Sørli *et al.*, 2003). These data raise the question as to whether the high expression of *GATA3* (and *ESR1*) contributes to tumorigenesis, or do high levels reflect the baseline state of a cell type that gives rise to this specific breast tumor subtype? Our analysis of normal breast showed that *GATA3* and *ESR1* were highly expressed in many of the normal luminal cells that line the mammary ducts (Figure 3a, b). When coupled with the observation that low *GATA3* expression correlates with worse patient outcomes and high grade, these data argue that the high expression of *GATA3* is 'normal' and that any deviation from high levels due to deletion or mutation (as is seen in HDR syndrome) or impairment through mutation (as is seen in breast) can contribute to a disease state.

In humans, the absolute levels of *GATA3* protein are critical since the haplo-insufficiency of *GATA3* is responsible for the rare autosomal dominant malformation disease called HDR syndrome (OMIM #146255) (Van Esch *et al.*, 2000; Muroya *et al.*, 2001; Nesbit *et al.*, 2004). The spectrum of *GATA3* somatic mutations found in our breast tumor samples was diverse and clustered in the vicinity of the highly conserved second zinc-finger. The mutations identified here map to the same region in which point mutations have been described in HDR patients (Van Esch *et al.*, 2000; Muroya *et al.*, 2001; Nesbit *et al.*, 2004). In fact, the nonsense mutation in tumor BR99-0348 is the exact variant found in a Japanese (Muroya *et al.*, 2001) and a Northern European kindred (Nesbit *et al.*, 2004) with HDR syndrome. In addition, one of the deletion variants of Van Esch *et al.* (2000) (deletion of residues 316–319) starts at the site of the frameshift mutation (residue 316) seen in tumor Ull-030, and Nesbit *et al.* also found a unique mis-splicing of exon 5 to exon 6 (whereas we found a mis-splicing of exon 4 to exon 5). Nesbit *et al.* also showed using *in vitro* assays that

GATA3 mutations that affected the second zinc-finger, or the region 3' to the second zinc-finger, affected DNA binding. In total, these data suggest that a functional haplo-insufficiency of *GATA3* is likely occurring in these breast tumors, which would cause a perturbation in the developmental state of these cells that contributes to tumorigenesis.

The function of *GATA3* in mammals is critical for the development and maintenance of a differentiated state. In the placenta, *GATA3* is needed for proper trophoblast-specific gene expression and function (Ma *et al.*, 1997), while in the T-cell lineage it plays a role in early T-cell development by directly inducing expression of the T-cell receptor genes, and plays a role in the transition to the differentiated Th2 effector state (Nawijn *et al.*, 2001). In breast, *GATA3* is produced by cells of luminal origin and is not produced by cytokeratin 5/6-positive myoepithelial cells (Figure 3). Recently, a role for *GATA3* in cell lineage determination in murine skin was identified and showed that *GATA3* was critical for the formation and maturation of the inner root sheath (Kaufman *et al.*, 2003); similar to what we observed in breast, *GATA3* was not expressed by the cytokeratin 5/6-positive epithelial cells of the skin.

A statistical analysis of our previous published breast tumor microarray data (Perou *et al.*, 2000; Sørli *et al.*, 2001) showed that high *GATA3* expression correlated with low tumor grade and slow proliferation rates (Supplementary Materials Figure 4 and data not shown). By analogy from other developmental systems, the role of *GATA3* in the human breast is likely to influence or drive luminal cell development and/or differentiation, and, therefore, a lessening of *GATA3* function could cause a transition to a less differentiated state (i.e. an increase in tumor grade) and increased proliferation. Support for this hypothesis comes from our studies in 293T cells in which the ectopic expression of wild-type *GATA3* caused the induction of many genes involved in luminal cell differentiation (*cytokeratins 8, 18 and 19, TFF1, TFF3*), and caused a statistically significant reduction in cell line population doubling times (PDTs). It is interesting to note that the R367L mutant had a similar transcriptional effect when compared to wild-type *GATA3*; however, the R367L mutant failed to affect PDTs, suggesting that *GATA3* may exert effects on growth-regulatory pathways that were independent of its transcriptional activation abilities.

We did not perform microarray analyses on all 111 tumors that were sequenced in this study; however, we did perform microarray analyses on five of the *GATA3* altered tumors and determined that these tumors were all of the 'luminal' subtype. These data are consistent with the involvement of *GATA3* in luminal-derived *ESR1*-positive tumors; however, it should be noted that the *GATA3* mutated tumors ranged from Stage I to Stage IIIA. In addition, all six *GATA3* altered tumors were *TP53* wild type by genomic DNA sequence analysis (data not shown). This is especially notable since the MCF-7 cell line also has a similar phenotype,

luminal-derived, ESR1-positive and *TP53* wild type. In conclusion, our data provide further evidence for an important role for *GATA3* in breast luminal cell biology and suggest that *GATA3* insufficiency in breast tumors could be important for the development of some ESR1-positive breast cancers.

Materials and methods

Breast tumor and normal blood DNA sample

We used genomic DNA from 111 grossly dissected breast tumors, 29 of which were obtained through the University of North Carolina at Chapel Hill under an IRB-approved protocol, and 82 of which were obtained from the Ullevål University Hospital, Oslo Norway (mean age at diagnosis 64, range 28–87 years old), under a separate IRB protocol. All but five tumor samples were collected from the primary tumor (five from node metastases), and all were snap frozen in dry ice and stored at -80°C . In addition, 92 genomic DNA samples were isolated from the peripheral blood of healthy Norwegian women without evidence of breast cancer (controls); these samples were obtained from postmenopausal women (55–72 years), who were part of a routine mammography screening and who had two negative mammograms at the time of sample collection. In addition, sequence analysis was performed in 102 anonymized samples from The SNP500Cancer (<http://snp500cancer.nci.nih.gov>) representing the four major US self-described ethnic groups (Caucasian, African, Hispanic and Pacific Rim ancestry) (Packer *et al.*, 2004). Genomic DNA from the UNC samples was isolated using a Qiagen Genomic DNA isolation kit, while genomic DNA from the Norwegian samples was isolated using chloroform/phenol extraction followed by ethanol precipitation (Nuclear Acid Extractor 340A; Applied Biosystems). None of the patients in this study showed the clinical manifestations of the HDR syndrome, and all stages and grades of breast tumors were represented.

Identification of GATA3 variants

Bi-directional sequence analysis was performed on genomic DNA derived from grossly dissected tumor samples and covered all exons and intron–exon junctions of *GATA3*; specifically, unique oligonucleotide primers were tagged with either the universal forward and reverse M13 sequences. PCR reactions were carried out in $10\ \mu\text{l}$ volumes containing 5–10 ng of genomic DNA, which consisted of $1\ \mu\text{l}$ of AmpliTaq Buffer, 2.5 mM MgCl_2 , 115 μM dNTPs, 0.1 U of AmpliTaq Gold DNA polymerase and PCR primers at a final concentration of 100 nM. Reactions were performed in 96-well plates on a MJ Research PTC-200 Thermal Cycler using the following parameters: a 10 min denaturation step at 95°C , followed by 38 cycles of 30 s at 94°C , 45 s at 63.6°C , 45 s at 72°C and one last 10 min hold at 72°C . Unincorporated dNTPs and primers were removed from the amplified PCR products using a Qiaquick 96-well purification kit. The purified PCR products were directly sequenced in both directions using the ABI BigDye terminator Mix (version 3, Applied Biosystems, Foster City, USA) and analysed on an ABI 3700 DNA capillary sequencer (ABI), according to the manufacturer's instructions. Cycle sequencing conditions consisted of an initial denaturation step at 98°C for 2 min, followed by 25 cycles of denaturation at 95°C for 10 s, annealing at 50°C for 10 s and extension at 60°C for 4 min. Sequence variants were identified using SEQUENCHER 4.1.1 analysis software (Gene Codes

Corp.) and manual inspection. A subset of the variants were confirmed (BR00-0587, BR99-0348 and MCF-7) or identified (BR99-0207) using reverse transcription of total RNA using Superscript (Gibco-BRL) and an oligo-dT primer, followed by PCR amplification of the region in question and direct cycle sequencing. All primer sequences are available upon request.

Cell lines

The cell lines MCF-7, ZR-75-1 and 293T were grown in RPMI + 10% FCS until they were 70–80% confluent, at which time half of the culture was harvested for genomic DNA as described above, and half harvested for protein extracts as described below. The ME16C line was obtained from Jerry Shay (UTSW) and represents an hTERT-immortalized human mammary epithelial cell line. This line was grown in fully supplemented MEGM media (Cambrex) and represents breast cells of basal epithelial-like origins (Ross and Perou, 2001). Cell line PDTs were determined by plating $\sim 5 \times 10^4$ cells in 100-mm petri dishes. Three dishes per cell line were counted at 2, 4 and 6 days after plating. The growth of the cells during the log phase can be modeled using the following equation:

$$\ln A(t) = \ln A_0 + kt, \quad (1)$$

where $A(t)$ is the number of cells per plate at time t and k represents the first-order rate constant of cell growth, with units d^{-1} . Using this regression equation, independent estimates of k were obtained for each cell line in question. The PDT for each cell line was calculated as:

$$\text{PDT} = \frac{0.693}{k} \quad (2)$$

To compare the growth rates, k , for two cell lines, a multiple regression model similar to that described in Troester *et al.* (2002) was employed.

Western blot analysis

Whole cell protein lysates were extracted from fresh frozen tumor samples, a normal breast sample and the four cell lines using the Pierce tissue (T-PER) and cell line (M-PER) protein lysate extraction kits (Pierce, Rockford, IL, USA). In all, $40\ \mu\text{g}$ of total protein from each sample was run on a 4–20% Bio-Rad Criterion precast SDS-PAGE gel and transferred to a Hybond-P nylon membrane (Amersham Pharmacia Biotech). Membranes were blocked by incubation with 5% milk in blotting buffer (Tris buffered-saline), washed with blotting buffer and incubated with a mouse monoclonal antibody specific for *GATA3* (sc-269, Santa Cruz Biotechnology, Santa Cruz, CA, USA) at a 1:500 dilution; similar results were obtained with a second antibody (sc-268) manufactured by the same company; however, sc-269 consistently gave more intense staining results on Western blot analysis. The primary antibody binding was detected using an anti-mouse IgG-horse peroxidase-linked antibody (Amersham) and was visualized with the Pierce West Pico Chemiluminescent detection kit. The Western blot was stripped and reprobed with an antibody to β -actin (Abcam #6276, Cambridge, UK) as a loading control.

Immunohistochemistry

Paraffin-embedded breast tumors, normal breast blocks and cell line blocks were cut into $5\ \mu\text{m}$ sections. Tissue sections were deparaffined with xylene, dehydrated with ethanol and endogenous peroxidase activity was blocked with a 3% hydrogen peroxide solution. The slides were incubated with

10 mM citrate buffer (pH 6.0) and microwaved for 20 min for antigen retrieval. The slides were then blocked with goat serum and incubated with *GATA3* antibodies for 30 min (Santa Cruz sc-268, 1:50 dilution), and then incubated with biotin-conjugated goat anti-mouse IgG (Vector Laboratories, CA, USA). Proteins were visualized with streptavidin-conjugated HRP (Vector Laboratories, CA, USA). The slides were counterstained with 50% hematoxylin and examined by light microscopy on a Zeiss microscope at $\times 100$ magnification. For ESR1 IHC, we used an antibody from DAKO, clone 1D5 (#M7047) at a 1:25 dilution. For CK5/6 and CK17, we performed IHC as described in van de Rijn *et al.* (2002).

Expression of *GATA3* in 293T cells

A full-length clone for *GATA3* (BC006793) was obtained from the Mammalian Gene Collection (<http://mgc.nci.nih.gov/>). The Invitrogen Gateway BP clonase enzyme (#11789013) was used to transfer the *GATA3* insert to the pDonr201 plasmid (#11798014). The pBabe-Puro plasmid (Morgenstern and Land, 1990) was modified to be Gateway compatible using the Gateway Conversion System kit (#11828019) to produce pBabe-puro-GWrfA. The LR clonase enzyme (#11991019) was used to transfer the *GATA3* insert from pDonr201 to the pBabe-puro-GWrfA plasmid. The *GATA3* insert was then sequenced and designated pBabe-Puro-*GATA3*-WT. We next created a mutated version of *GATA3* in pBabe-Puro-*GATA3* using the Quick Change site-directed mutagenesis kit from Stratagene (#200518-5). The R367L mutation found in tumor Ull-011 was created and confirmed by sequencing.

The 293T cell line was next used to produce infectious retrovirus as follows. Cells were transfected using Lipofectamine2000 (Invitrogen catalog #11668) with pVpack-GP (Stratagene catalog #217566) and pVpack-Ampho (Stratagene catalog #217568) along with either pBabe-puro-Empty, pBabe-puro-*GATA3*-WT, or pBabe-Puro-*GATA3*-R367L plasmids. Supernatants containing replication-incompetent retrovirus were harvested after 48 h and added to the cell lines for transfection. Stable clones were selected using 2 μ g/ml puromycin over a 2-week period. The harvest of mRNA for microarrays was performed using an Invitrogen Micro-FastTrak 2.0 kit.

Gene expression analysis

Cell line mRNA samples were reverse transcribed into fluorescent cDNA using 1 μ g of mRNA and the Agilent Fluorescent Direct Label Kit according to the manufacturer's protocol (<http://www.chem.agilent.com/Scripts/PDS.asp?lPage=11753>). For these experiments, three different 'empty-vector' control 293T transfectants were pooled together and assayed *versus* 293T lines that contained *GATA3*-

WT and *GATA3*-R367L, with dye-flip replicates. These four experiments were performed on Agilent custom DNA oligo microarrays that contained the exact 18 000 probes/oligos that are present on the Agilent Human A1 microarrays plus another 3000 custom oligos. All microarray data are available from our website at <http://dragon.med.unc.edu/pubsup/GATA3/>, and have been deposited into the Gene Expression Omnibus under the Accession Number GSE841 (submitter C Perou). To identify *GATA3*-regulated genes, we (1) flipped each dye-flip experiment to create four similar experiments, (2) filtered for genes that gave a signal intensity of 50 or more in both channels after LOWESS normalization, (3) used SAM analysis (Tusher *et al.*, 2001) to identify genes that changed in the R367L and *GATA3*-WT samples as a single class *versus* the empty vector controls. This resulted in the identification of 74 genes (89 spots) with a false-discovery rate of one gene. We also performed quantitative RT-PCR analysis using 'TaqMan' probes distributed by Applied Biosystems (Foster City, CA, USA) to quantitate *GATA3* mRNA expression levels. For each experimental sample (cell line, tumor or normal breast), we made cDNA as described above and then followed the manufacturer's protocol using an ABI Prism 7900. We assayed for the expression of three genes in each sample (*PUM1*, *SF3A* and *GATA3*), and normalized the expression values of *GATA3* in each sample relative to the control genes *PUM1* and *SF3A*.

To identify genes associated with tumor grade, the 122 samples profiled in Sørli *et al.* (2003) were used in a 'multi-class' SAM analysis using tumor grade (I, II and III) as the supervising parameter. The data were preprocessed as described in Sørli *et al.* (2003) and the analysis resulted in the identification of 415 genes with a false-discovery rate of 3%. These 415 genes were then used in an average linkage hierarchical clustering analysis across the 122 samples, with the genes being clustered and the order of the experimental samples being maintained according to the sample order that appears in Supplementary Materials Figure 3 (which was determined by the clustering of these samples using the breast tumor 'intrinsic' gene list).

Acknowledgements

We wish to thank Jerry Shay (UTSW) for the ME16C cell line. We also thank Phil Bernard (Univ. of Utah) and Juan Palazzo (Thomas Jefferson Univ.) for contributing breast tumor samples to this study. This work was supported by funds from the NCI Breast SPORE program to UNC-CH (C.M.P. P50-CA58223-09A1), The Norwegian Cancer Society (A.-L. B-D D99061), the Research Council of Norway (A.-L. B-D 137012/310 and 155218/30) and the NCI CGAP Project (S.C.). We also wish to acknowledge the technical assistance and support of the Tissue Procurement and Analysis Facility of UNC-CH.

References

- Bocker W, Bier B, Freytag G, Brommelkamp B, Jarasch ED, Edel G, Dockhorn-Dworniczak B and Schmid KW. (1992). *Virchows Arch. A Pathol. Anat. Histopathol.*, **421**, 315–322.
- Boecker W and Buerger H. (2003). *Cell Prolif.*, **36** (Suppl 1), 73–84.
- Dairkee SH, Puett L and Hackett AJ. (1988). *J. Natl. Cancer Inst.*, **80**, 691–695.
- Dennis Jr G, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC and Lempicki RA. (2003). *Genome Biol.*, **4**, P3.
- Finlin BS, Gau CL, Murphy GA, Shao H, Kimel T, Seitz RS, Chiu YF, Botstein D, Brown PO, Der CJ, Tamanoi F, Andres DA and Perou CM. (2001). *J. Biol. Chem.*, **31**, 31.
- Garg V, Kathiriya IS, Barnes R, Schluterman MK, King IN, Butler CA, Rothrock CR, Eapen RS, Hirayama-Yamada K, Joo K, Matsuoka R, Cohen JC and Srivastava D. (2003). *Nature*, **424**, 443–447.
- Graham FL, Smiley J, Russell WC and Nairn R. (1977). *J. Gen. Virol.*, **36**, 59–74.
- Groet J, McElwaine S, Spinelli M, Rinaldi A, Burtscher I, Mulligan C, Mensah A, Cavani S, Dagna-Bricarelli F, Basso G, Cotter FE and Nizetic D. (2003). *Lancet*, **361**, 1617–1620.
- Gruvberger S, Ringner M, Chen Y, Panavally S, Saal LH, Borg A, Ferno M, Peterson C and Meltzer PS. (2001). *Cancer Res.*, **61**, 5979–5984.

- Guelstein VI, Tchypysheva TA, Ermilova VD, Litvinova LV, Troyanovsky SM and Bannikov GA. (1988). *Int. J. Cancer*, **42**, 147–153.
- Ho IC, Vorhees P, Marin N, Oakley BK, Tsai SF, Orkin SH and Leiden JM. (1991). *EMBO J.*, **10**, 1187–1192.
- Hoch RV, Thompson DA, Baker RJ and Weigel RJ. (1999). *Int. J. Cancer*, **84**, 122–128.
- Kaufman CK, Zhou P, Pasolli HA, Rendl M, Bolotin D, Lim KC, Dai X, Alegre ML and Fuchs E. (2003). *Genes Dev.*, **17**, 2108–2122.
- Kruglyak L and Nickerson DA. (2001). *Nat. Genet.*, **27**, 234–236.
- Lim KC, Lakshmanan G, Crawford SE, Gu Y, Grosveld F and Engel JD. (2000). *Nat. Genet.*, **25**, 209–212.
- Loots GG, Ovcharenko I, Pachter L, Dubchak I and Rubin EM. (2002). *Genome Res.*, **12**, 832–839.
- Ma GT, Roth ME, Groskopf JC, Tsai FY, Orkin SH, Grosveld F, Engel JD and Linzer DI. (1997). *Development*, **124**, 907–914.
- Morgenstern JP and Land H. (1990). *Nucleic Acids Res.*, **18**, 3587–3596.
- Muroya K, Hasegawa T, Ito Y, Nagai T, Isotani H, Iwata Y, Yamamoto K, Fujimoto S, Seishu S, Fukushima Y, Hasegawa Y and Ogata T. (2001). *J. Med. Genet.*, **38**, 374–380.
- Nawijn MC, Ferreira R, Dingjan GM, Kahre O, Drabek D, Karis A, Grosveld F and Hendriks RW. (2001). *J. Immunol.*, **167**, 715–723.
- Nesbit MA, Bowl MR, Harding B, Ali A, Ayala A, Crowe C, Dobbie A, Hampson G, Holdaway I, Levine MA, McWilliams R, Rigden S, Sampson J, Williams A and Thakker RV. (2004). *J. Biol. Chem.*, **279**, 22624–22634.
- Orkin SH. (1992). *Blood*, **80**, 575–581.
- Packer BR, Yeager M, Staats B, Welch R, Crenshaw A, Kiley M, Eckert A, Beerman M, Miller E, Bergen A, Rothman N, Strausberg R and Chanock SJ. (2004). *Nucl. Acids Res.*, **32**, D528–D532.
- Pandolfi PP, Roth ME, Karis A, Leonard MW, Dzierzak E, Grosveld FG, Engel JD and Lindenbaum MH. (1995). *Nat. Genet.*, **11**, 40–44.
- Perou CM, Jeffrey SS, van de Rijn M, Rees CA, Eisen MB, Ross DT, Pergamenschikov A, Williams CF, Zhu SX, Lee JC, Lashkari D, Shalon D, Brown PO and Botstein D. (1999). *Proc. Natl. Acad. Sci. USA*, **96**, 9212–9217.
- Perou CM, Sørli T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, Fluge O, Pergamenschikov A, Williams C, Zhu SX, Lonning PE, Borresen-Dale AL, Brown PO and Botstein D. (2000). *Nature*, **406**, 747–752.
- Pollack JR, Sørli T, Perou CM, Rees CA, Jeffrey SS, Lonning PE, Tibshirani R, Botstein D, Borresen-Dale AL and Brown PO. (2002). *Proc. Natl. Acad. Sci. USA*, **20**, 12963–12968.
- Ross DT and Perou CM. (2001). *Dis. Markers*, **17**, 99–109.
- Sena-Esteves M, Saeki Y, Camp SM, Chiocca EA and Breakefield XO. (1999). *J. Virol.*, **73**, 10426–10439.
- Smith VM, Lee PP, Szychowski S and Winoto A. (1995). *J. Biol. Chem.*, **270**, 1515–1520.
- Sørli T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, Hastie T, Eisen MB, van de Rijn M, Jeffrey SS, Thorsen T, Quist H, Matese JC, Brown PO, Botstein D, Eystein Lonning P and Borresen-Dale AL. (2001). *Proc. Natl. Acad. Sci. USA*, **98**, 10869–10874.
- Sørli T, Tibshirani R, Parker J, Hastie T, Marron JS, Nobel A, Deng S, Johnsen H, Pesich R, Geisler S, Demeter J, Perou CM, Lonning PE, Brown PO, Borresen-Dale AL and Botstein D. (2003). *Proc. Natl. Acad. Sci. USA*, **100**, 8418–8423.
- Tavassoli FA and Schnitt SJ. (1992). *Pathology of the Breast*. Elsevier: New York.
- Troester MA, Lindstrom AB, Waidyanatha S, Kupper LL and Rappaport SM. (2002). *Toxicol. Sci.*, **68**, 314–321.
- Tusher VG, Tibshirani R and Chu G. (2001). *Proc. Natl. Acad. Sci. USA*, **98**, 5116–5121.
- van de Rijn M, Perou CM, Tibshirani R, Haas P, Kallioniemi O, Kononen J, Torhorst J, Sauter G, Zuber M, Kochli OR, Mross F, Dieterich H, Seitz R, Ross D, Botstein D and Brown P. (2002). *Am. J. Pathol.*, **161**, 1991–1996.
- Van Esch H, Groenen P, Nesbit MA, Schuffenhauer S, Lichtner P, Vanderlinden G, Harding B, Beetz R, Bilous RW, Holdaway I, Shaw NJ, Fryns JP, Van de Ven W, Thakker RV and Devriendt K. (2000). *Nature*, **406**, 419–422.
- 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, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R and Friend SH. (2002). *Nature*, **415**, 530–536.
- Wechsler J, Greene M, McDevitt MA, Anastasi J, Karp JE, Le Beau MM and Crispino JD. (2002). *Nat. Genet.*, **32**, 148–152.
- West M, Blanchette C, Dressman H, Huang E, Ishida S, Spang R, Zuzan H, Olson Jr JA, Marks JR and Nevins JR. (2001). *Proc. Natl. Acad. Sci. USA*, **98**, 11462–11467.
- Yang Z, Gu L, Romeo PH, Bories D, Motohashi H, Yamamoto M and Engel JD. (1994). *Mol. Cell. Biol.*, **14**, 2201–2212.

Supplementary Information accompanies the paper on Oncogene website (<http://www.nature.com/onc>).