

Transcriptome sequencing of human breast cancer reveals aberrant intronic transcription in amplicons and dysregulation of alternative splicing with major therapeutic implications

SHIVA SEYED FOROOTAN^{1*}, JOE M. BUTLER^{1*}, DEREK GARDENER¹, ALISON E. BAIRD¹, ANDREW DODSON¹, ALISTAIR DARBY², JOHN KENNY², NEIL HALL², ANDREW R. COSSINS², CHRISTOPHER S. FOSTER^{1,3} and CHRISTINE M. GOSDEN¹

¹Molecular Pathology Laboratory, Department of Molecular and Clinical Cancer Medicine, Institute of Translational Medicine, Faculty of Health and Life Sciences, University of Liverpool, Royal Liverpool University Hospital, Liverpool L69 3GA;

²Center for Genomic Research, University of Liverpool, Liverpool L69 7ZB; ³HCA International, Pathology Laboratories, London WC1E 6JA, UK

Received June 30, 2015; Accepted July 31, 2015

DOI: 10.3892/ijo.2015.3222

Abstract. Advances in genomic and transcriptome sequencing are revealing the massive scale of previously unrecognised alterations occurring during neoplastic transformation. Breast cancers are genetically and phenotypically heterogeneous. Each of the three major subtypes [ERBB2 amplified, estrogen receptor (ESR)-positive and triple-negative] poses diagnostic and therapeutic challenges. Here we show, using high-resolution next-generation transcriptome sequencing, that in all three breast cancer subtypes, but not matched controls, there was significant overexpression of transcripts from intronic and untranslated regions in addition to exons from specific genes, particularly amplified oncogenes and hormone receptors. For key genes *ERBB2* and *ESR1*, we demonstrate that overexpression is linked to the production of highly modified and truncated splice variants in tumours, but not controls, correlated with tumour subtype. Translation of these tumour-specific splice variants generates truncated proteins with altered subcellular locations and functions, modifying the phenotype, affecting tumour biology, and targeted antitumour therapies. In contrast, tumour suppressors *TP53*, *BRCA1/2* and *NFI* did not show intronic overexpression or truncated splice variants in cancers. These findings emphasize the detection

of intronic as well as exonic changes in the transcriptional landscapes of cancers have profound therapeutic implications.

Introduction

Breast cancer is the leading cause of cancer deaths in women. In the UK, between 1979 and 2011, the annual incidence of breast cancers increased from 23,876 to 49,936 (1) emphasizing the necessity to understand contemporary cancer biology to develop effective diagnosis and management. Of the three major subtypes, targeted therapies have been designed only for those with amplified *ERBB2* and estrogen receptor-1 (*ESR1*) expression. The majority of breast cancers, including those involving familial *BRCA1/2* mutations, are not *ERBB2* amplified and are *ESR1* and progesterone receptor (*PR*)-negative. For triple-negative breast cancers (TNC) (2-4) with five year survival rate of 77%, compared with 93% for other subtypes (5,6) treatment options are restricted to conventional chemotherapy.

Genomic alterations that define *ERBB2* or *ESR1* amplified breast cancer subtypes comprise massive oncogene or hormone receptor amplifications (7,8) but not mutations or rearrangements in RefSeq exonic sequences. These are characterised by detection of amplification/overexpression, not genetic analysis of mutations or rearrangements revealed by exome sequencing (9-14). Recent demonstrations of the expression of novel (non-RefSeq) splice variants in cell lines (15-19), suggest amplification may not simply involve overexpression of exonic sequences alone but also non-exonic sequences. Furthermore, genome-wide association studies (GWAS) of single nucleotide polymorphisms (SNPs) in breast cancer indicate only 5% of high risk SNPs are exonic, the remaining 95% localised to introns, UTRs and intergenic sequences suggesting that these intronic sequences may have functional roles (20-23). In this study next generation whole transcriptome sequencing has been employed to test the hypothesis that structural genomic changes occurring in human breast cancers are not random and stochastic but occur in combinations characteristic of specific sub-types and hence may predict phenotypic behaviour, including response to therapy.

Correspondence to: Dr Shiva Seyed Forootan, Molecular Pathology Laboratory, Department of Molecular and Clinical Cancer Medicine, Institute of Translational Medicine, Faculty of Health and Life Sciences, University of Liverpool, 5/6th floor, Duncan Building, Daulby Street, Liverpool, L69 3GA, UK
E-mail: shifor@liverpool.ac.uk

*Contributed equally

Key words: breast cancer, intronic transcription, splice variants, therapeutic target

Materials and methods

Tissue samples and patient data. This study was approved by North West 12 Research Ethics Committee (Lancaster) in accordance with the Medical Research Council guidelines (project reference no. Gosden, 10/H1015/2). From 20 primary breast cancer specimens, frozen tissue sections were cut and stained with H&E to assess morphology. Human breast tissues were obtained from Liverpool Tissue Bank as frozen tissues with the full record of the grade of carcinoma, age and hormonal status (ER, PR and HER2) of the samples (Table I).

RNA, DNA and protein extraction and quality assessment. Frozen tissues (30 mg) from each of the 20 specimens were extracted using RLT buffer containing 2-ME (1 ml buffer and 10 μ l of 2-ME) added to the tissues following evaporation of liquid nitrogen.

After centrifugation, RNA, DNA and protein were extracted from the supernatant. The quantity of RNA and DNA was measured by NanoDrop (Labtech, Ringmer, UK). The quality of RNA for each sample was assessed by 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA) using Agilent RNA 6000 Nano kit. The RNA integrity number (RIN) for all the samples was >7.0. The quality of DNA and protein was assessed using gel electrophoresis and Bradford assay, respectively.

Total RNA was further cleaned using the RiboMinus concentration modules (cat. no. K1550-05; Invitrogen) according to manufacturer's instructions. Depletion of rRNA from each sample was performed using the RiboMinus Eukaryote kit for RNA-Seq (cat. no. A1083708; Invitrogen) according to manufacturer's instructions. Successful removal of rRNA was confirmed using the Bioanalyzer.

cDNA library preparation and sequencing. Libraries of rRNA-depleted RNA suitable for sequencing using the SOLiD platform were created using the SOLiD Total RNA-Seq kit (part no. 4445374; Applied Biosystems). In each instance, 100 ng of RNase III digested rRNA depleted RNA was used as input into the library creation and 15 cycles of amplification were employed to produce the final libraries. These libraries were sequenced on the SOLiD platform from Applied Biosystems. Initially sequencing of the 10 bp library barcode was performed followed by 35 and 50 bp paired-end sequencing. Fragments were subjected to paired-end sequencing using 35 and 50 bp reads.

Mapping of RNA-Seq reads using TopHat. The sequenced paired-end tags were mapped to the reference human genome (hg19) using TopHat (24). This mapping approach ensures that a read, which spans an exon/exon boundary on the mRNA, is mapped to the genome such that it flanks the intervening intronic region (called junctions by TopHat). After alignment, two rounds of filtering were performed, the first to remove any low quality reads (mapping quality <10) and the second to remove any reads that mapped to ribosomal genes. Mappings were examined visually using SAMtools (25) and IGV (26). FPKM measures (gene expression quantified using the fragments per kilobase of exon per million reads mapped-FPKM) were calculated using the following definition: 'a fragment was counted if both paired

end tags of the fragment F3 and F5 are observed or if only one of them (F3 or F5) is observed'. RefSeq co-ordinates were used to define exonic and intronic regions of the gene. Thus two FPKM values were calculated for each gene to measure exonic and intronic expression.

Genes were screened for differential exonic expression between a tumour sample and the paired normal tissue. Normal tissue samples were grouped together to calculate a standard deviation of the FPKM for each gene. A one-sampled t-test was applied for each gene, comparing its expression in a single cancer against the pooled normal tissue. Genes with low p-values indicated a high differential expression. The q-value package in R was applied to take account of multiple testing. The false discovery rate (FDR) was set to 0.001. This was applied to each tumour sample to give a profile of the differentially expressed genes. Of the genes with significantly different exonic expression we distinguished whether they are up or down regulated in the tumour. This entire process was repeated to screen for genes with significant differential intronic expression.

In certain cancer specimens, significantly upregulated genes were clustered together at certain chromosomal locations. These were hypothesised to be the consequence of amplicons. The length of these amplicons was quantified using a method we devised that involves the minimization of a binomial probability. For a total of X genes on a chromosome, then parameter p is defined as Y/X where Y is the number of genes on that chromosome with significantly upregulated expression (intronic, exonic or both). Then, we consider n consecutive genes encompassing the cluster, and count how many of those are significantly upregulated (intronic or exonic or both), let this number be k .

$$f(k; n, p) = \binom{n}{k} p^k (1-p)^{n-k}$$

We calculate the binomial probability in the usual way with n trials and k successes. This is performed for different sets of consecutive genes encompassing the cluster; the set with the lowest binomial probability is chosen to represent the amplicon.

Validation of NGS by quantitative PCR (qPCR) and fluorescent in situ hybridization (FISH)

qPCR. For gene expression profiling, qPCR was used on a selection of upregulated or downregulated genes. The expression level was confirmed in *ERBB2*, *ERBB3*, *VEGFA*, *FLT1*, *TP53*, *BRCA1*, *BRCA2*, *ESR1* and *PGR* genes. The primers and amplicon size are addressed in Table II. The first strand cDNA was synthesized from 100-500 ng/ μ l total RNA using AffinityScript™ multiple temperature cDNA synthesis kits (Stratagene, La Jolla, CA, USA) according to manufacturer's protocol. MX3305P real-time PCR machine (Stratagene) was used. Reaction volume was 10 μ l containing 5 μ l Brilliant SYBR-Green Master Mix (2X) (Stratagene), 1 μ l forward and reverse primers (5 μ M), 1 μ l cDNA and 2 μ l water. Primers were all optimized for 60°C. Cycling conditions were: 95°C for 15 min, 40 cycles at 94°C for 15 sec, 60°C for 30 sec, plate read and 72°C for 30 sec with final extension of 72°C for 10 min. Gene expression level was normalized relative to human β -actin expression.

Table I. Clinical information for 20 breast tumours and paired normal controls, tumour grade, estrogen and progesterone receptor and ERBB2 (HER2) status.

LTBRC no. tumour/ control status	Age (years)	Grade	Estrogen receptor (<i>ESR1</i>)	Progesterone receptor (PGR)	<i>ERBB2</i> (<i>HER2</i>)
Tumour samples <i>ERBB2</i> (Her2)-positive ± estrogen receptor-positive					
1 221T tumour	53	3	+	-	3 ⁺
2 221N paired control					
3 286T	76	3	+	+	3 ⁺
4 326T	27	3	+	+	3 ⁺
5 043T	84	3	-	-	3 ⁺
6 121T	50	3	+	-	3 ⁺
7 265T	48	2	+	-	2 ⁺
8 193T	44	3	-	-	2 ⁺
Triple-negative samples (estrogen and progesterone receptor-negative and HER2-negative)					
9 194T	57	3	-	-	-
10 271T	67	3	-	-	-
11 013 tumour	32	3	-	-	-
12 013N	Normal pair control				
13 082T	37	3	-	-	-
14 160T	69	3	-	-	-
15 190T	77	3	-	-	-
16 233T	61	3	-	-	-
17 250T	67	3	-	-	-
18 287T	55	3	-	-	-
19 085T	86	3	-	-	-
20 086T	50	3	-	-	-
21 081 tumour	43	3	-	-	-
22 081N control	Normal pair control				
23 010 tumour	34	3	-	-	-
24 010N control	Normal pair control				

All breast cancers were invasive ductal carcinoma. The table shows the clinical information on the 12 breast tumours with the patient age, tumour stage and hormone receptor status for ESR1, PGR and ERBB2 (HER2). Tumours are designated with the suffix T and for tumours that have matched adjacent paired non-involved breast tissue these control samples have the same number but suffix N.

FISH. FISH performed using the Kreatech *ERBB2*, Her2/SE17 probe kit (Kreatech) for detecting *ERBB2* gene and ZytoLight SPEC *ESR1*/CEN 6 Dual Color Probe kit (Zytovision, Bremerhaven, Germany) for detecting *ESR1* gene, according to the manufacturer's instruction. Briefly, slides were deparaffinized in xylene, incubated in heat pretreatment solution, EDTA buffer, using pressure cooker for 1 min. Slides were incubated with proteinase K for 10 min at 37°C followed by wash with dH₂O for 5 min, dehydrated and air dried. Ten microliters of ZytoLight SPEC *ESR1*/CEN 6 Dual Color Probe was applied to the slides in the dark followed by denaturation at 75°C for 10 min and overnight incubation at 37°C in the Hybrid OmniSlide hybridizer (Fisher Scientific). After hybridization, coverslips were removed using wash buffer (20X SSC and NP-40), washed at 72°C for 3 min in wash buffer and air dried in the dark. Subsequently, one drop of DAPI (Vector Laboratories) was applied to the slides and slides were covered with the coverslips. Slides were viewed and scored using the

Olympus BX61 cytovision fluorescence microscope. To calculate and record the results, the number of chromosome and gene signal was counted in 20-60 non-overlapping invasive cancer cell nuclei in 3 or more distinct fields. The gene to chromosome ratio was calculated. Gene amplification was defined as a ratio of >2.2. Values of 1.8-2.2 implied borderline amplification and below 1.8, not amplified.

Immunohistochemistry (IHC). Tissues were labelled with antibodies against p53 (clone DO-7; Dako), p63 (clone 4A4; NeoMarkers) and *ESR1* (clone 1D5; Dako) as previously described (27).

Western blotting. The protein level of *ERBB2* was evaluated with western blot analysis using two *ERBB2* antibodies: mouse monoclonal antibody (ab16901) against 1242-1255 aa in C-terminus and rabbit polyclonal antibody (ab11717) (both from Abcam) against 651-660 aa. Western blotting was performed as described previously (27). The dilution of

Table II. Primer sequence for qPCR identification and validation of NGS.

Primer	Direction	Sequence	Amplicon size
ERBB2	Forward	TGCTGGACATTGACGAGACAGAGT	147
	Reverse	AAGTCATCAGCTCCCACACAGTCA	
ERBB3	Forward	GAAAGTGTCAATGTGTAGGAGC	243
	Reverse	TCTTCTTCAGTACCCAGGACAG	
VEGFA	Forward	CTACTGCCATCCAATCGAGAC	255
	Reverse	TCTTTCTTTGGTCTGCATTAC	
FLT1	Forward	TGACACTTTGATCCCTGATGGA	217
	Reverse	ATGGCCTCTAAGTAATTTGACTGG	
TP53	Forward	GCCAAAGAAGAAACCACTGGA	111
	Reverse	TGGGCATCCTTGAGTTCCAA	
BRCA1	Forward	AGAAAGAGGAACGGGCTTGGAAGA	115
	Reverse	TGAGCAGTCTTCAGAGACGCTTGT	
BRCA2	Forward	TGAACTGACAGATTCTAAACTGCC	135
	Reverse	GGTTCTCCCACTAAGATAAGGG	
ESR1	Forward	TCACAGACACTTTGATCCACC	293
	Reverse	CAAGGAATGCGATGAAGTAGAG	
PGR	Forward	ATTCCTTTGGAAGGGCTACGA	284
	Reverse	ATCCCTGCCAATATCTTGGGT	
β -actin	Forward	AGCCTCGCCTTTGCCGA	174
	Reverse	CTGGTGCCTGGGGCG	

antibodies are, 1/1,000 for polyclonal and 1/50 for monoclonal antibody. To confirm the presence of different splice variants, we employed PCR (FGFR2), IHC (p53) and western blotting (*ERBB2*). The presence of small numbers of mutations in our genes of interest such as p53 and BRCA2 was validated using Sanger sequencing.

Results

Deep transcriptome sequencing was undertaken in 20 primary breast cancer, including *ERBB2*-amplified, *ESR1* amplified/overexpressed and triple-negative tumours and compared with matched morphologically non-neoplastic breast tissue controls. Analysis of the 3' regions of AceView (Human 2010) variants of key breast cancer genes (*ERBB2*, *TP53* and *ESR1*) revealed that 16/23, 12/17 and 4/8 respectively lacked poly(A) tails (data not shown). Whole transcriptome sequencing included exons, introns and UTRs to define gene expression in relation to breast cancer subtypes and across regions of interest including *ERBB2* (chromosome 17), *ESR1* (chromosome 6), 8q24 and *MYC* (chromosome 8) together with other oncogenes and tumour suppressor genes (TSG) including *TP53/63/73*, *BRCA1/2*, *NF1*, *CDKN2A/B* implicated in breast cancer.

ERBB2 and detection of novel intronic tumour-specific transcript overexpression. *ERBB2* an oncogenic receptor tyrosine kinase (RTK), is amplified in ~20% of breast cancers (28). Of the 20 breast cancers examined, 5 contained *ERBB2* amplification identified by FISH or IHC. The transcriptional

consequences of *ERBB2* amplification were assessed using the FPKM measure defined as 'fragments per kilobase of exon per million reads mapped'. Stringent statistical analysis was employed to compare transcriptional expression in exons, introns, UTRs and overlaps on positive and negative strands in cancers with amplified and non-amplified *ERBB2* and 835 genes covering the entire chromosome 17 in cancers of all three tumour sub-types and controls. Major differences were identified between the tumours with *ERBB2* amplification at 17q11.2-12 and matched controls for *ERBB2* expression (Fig. 1). Those with *ERBB2* amplification exhibited high levels of intronic as well as exonic expression, whereas all controls were negative. Intronic expression in tumours was not limited to *ERBB2*. Detailed analysis of 835 genes located on whole chromosome 17 showed intronic tumour-specific sequence overexpression to occur 3.7 times more frequently than that for exons (Table III). Rigorous statistical analysis showed the mean intronic expression between tumours compared with controls was 48.7 (range, 1-208) (FDR 0.001 and >2-fold-change). The mean number of genes with FDR <0.001 exonic expression of the 835 genes on chromosome 17 in each of the tumour subtypes, was 12.7 (range, 0-41). Comparatively few genes had concurrent exonic and intronic expression, and for those genes with both exonic and intronic overexpression the mean was 5.1 (range, 0-32) (Table III).

To investigate whether intronic expression was specific to chromosome 17, or occurred in other breast tumours (including those with *ESR1* amplification at 6q25 or 8q24), intronic and exonic gene expression was assessed for chromosomes 6

Table III. Tumour rankings indicating significant exonic and intronic gene expression of 2120 genes of chromosomes 6, 8 and 17.

	No. of genes with (exclusively) exonic significance				Rank
	Chr 6 (Total genes=797) n (%)	Chr 8 (Total genes=488) n (%)	Chr 17 (Total genes=835) n (%)	Sum (Total genes=2120) n (%)	
221T	4 (0.50)	17 (3.48)	6 (0.72)	27 (1.27)	5
121T	1 (1.13)	4 (0.82)	4 (0.48)	9 (0.42)	7
286T	32 (4.02)	41 (8.40)	41 (4.91)	114 (5.38)	1
043T	11 (1.38)	18 (3.69)	14 (1.68)	43 (2.03)	4
326T	0 (0.00)	1 (0.20)	2 (0.24)	3 (0.14)	9
010T	6 (0.75)	1 (0.20)	0 (0.00)	7 (0.33)	8
013T	36 (4.52)	24 (4.92)	41 (4.91)	101 (4.76)	3
160T	16 (2.01)	87 (17.8)	3 (0.36)	106 (5.00)	2
081T	1 (1.13)	12 (2.46)	4 (0.48)	17 (0.80)	6
Mean	11.9 (1.58)	22.8 (4.66)	12.8 (1.53)	47.5 (2.24)	

	No. of genes with (exclusively) intronic significance				Rank
	Chr 6 (Total genes=797) n (%)	Chr 8 (Total genes=488) n (%)	Chr 17 (Total genes=835) n (%)	Sum (Total genes=2120) n (%)	
221T	18 (2.26)	17 (3.48)	6 (0.72)	41 (1.93)	6
121T	0 (0.00)	4 (0.82)	4 (0.48)	8 (0.38)	8
286T	10 (1.25)	41 (8.40)	41 (4.91)	92 (4.34)	4
043T	1 (1.13)	18 (3.69)	14 (1.68)	33 (1.56)	7
326T	0 (0.00)	1 (0.20)	2 (0.24)	3 (0.14)	9
010T	491 (61.7)	1 (0.20)	0 (0.00)	492 (23.2)	1
013T	67 (8.41)	24 (4.92)	41 (4.91)	132 (6.23)	3
160T	1 (1.13)	87 (17.8)	3 (0.36)	91 (4.29)	5
081T	297 (37.3)	12 (2.46)	4 (0.48)	313 (14.8)	2
Mean	98.3 (12.58)	22.8 (4.66)	12.8 (1.53)	133.9 (6.32)	

	No. of genes with both exonic and intronic significance				Rank
	Chr 6 (Total genes=797) n (%)	Chr 8 (Total genes=488) n (%)	Chr 17 (Total genes=835) n (%)	Sum (Total genes=2120) n (%)	
221T	0 (0.00)	49 (10.0)	32 (3.83)	81 (3.82)	2
121T	0 (0.00)	0 (0.00)	13 (1.56)	13 (0.61)	7
286T	9 (1.13)	25 (5.12)	0 (0.00)	34 (1.60)	6
043T	2 (0.25)	0 (0.00)	0 (0.00)	2 (0.09)	8
326T	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	9
010T	25 (3.14)	11 (2.25)	0 (0.00)	36 (1.70)	5
013T	24 (3.01)	25 (5.12)	0 (0.00)	49 (2.31)	3
160T	5 (0.63)	98 (20.1)	0 (0.00)	93 (4.86)	1
081T	14 (1.76)	28 (5.74)	1 (0.12)	43 (2.03)	4
Mean	9.3 (1.1)	26.2 (5.37)	5.1 (0.58)	39 (1.89)	

The table shows the number of genes for each of the chromosomes 6, 8 and 17 that have significant exonic, intronic or both exonic and intronic gene expression. The table indicates the findings for 797 genes of chr 6, 488 genes of chr 8 and 836 genes of chr 17 and from this derive the tumour rankings.

and 8 in each breast cancer subtype (Table III). Analysis of 2,119 genes spanning the whole of chromosomes 6, 8 and 17 revealed >2-fold-changes for intronic and/or exonic expres-

sion, in each of the different cancers and subtypes, but not controls (Table III). Intronic overexpression also occurred in TNCs and was not restricted to *ERBB2* and *ESR1* amplified

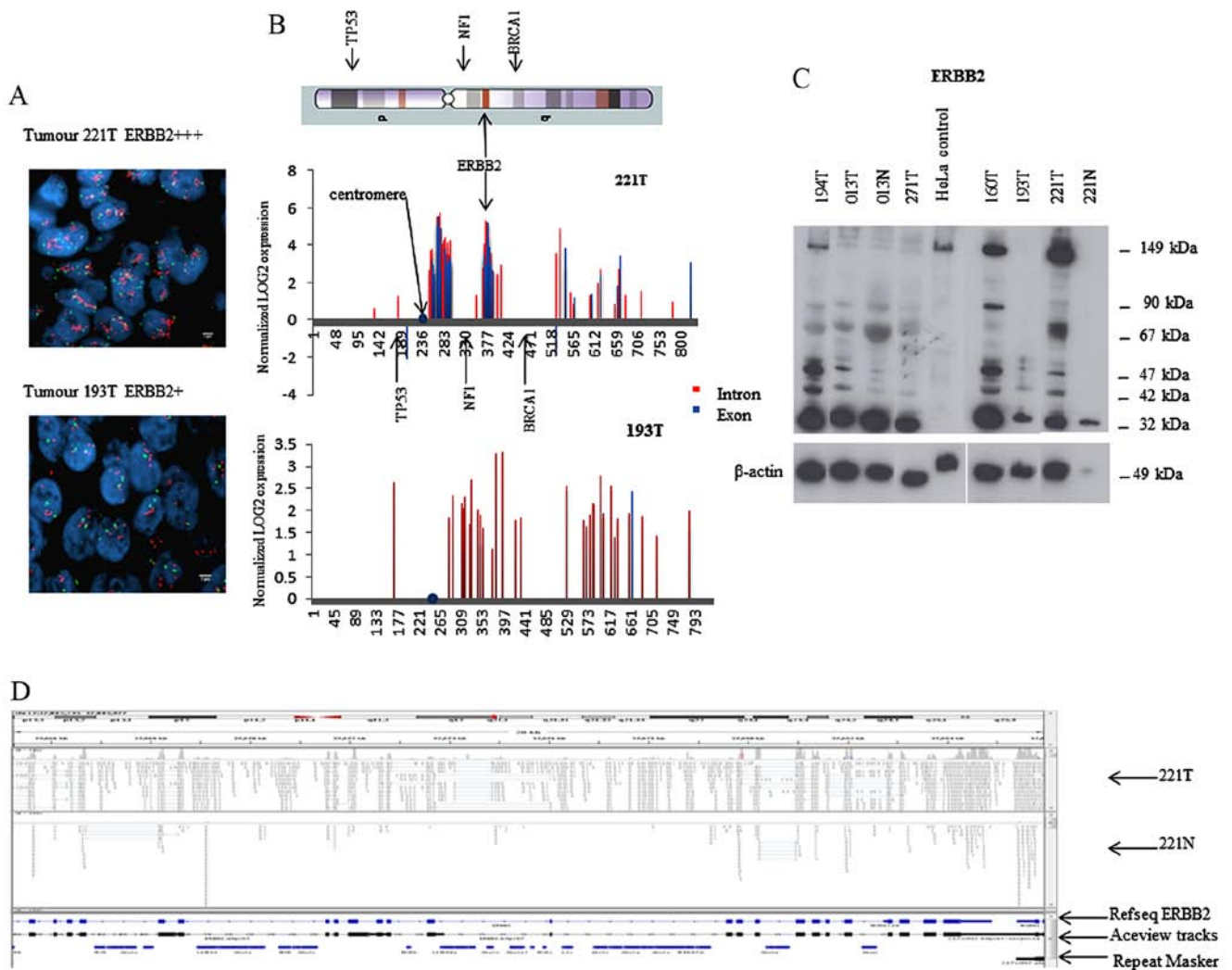


Figure 1. Transcriptional architecture of chromosome 17 in breast tumours; intronic overexpression in *ERBB2* and other genes, but not tumour suppressors *p53*, *NFI* or *BRCA1*. (A) FISH of *ERBB2* (red fluorescence) showing high level *ERBB2* amplification in tumour 221T compared with non-amplified tumour 193T. Green fluorescent centromeric marker identifies chromosome copy number enabling calculation of *ERBB2* signals to chromosome number (>5:1 in 221T vs. 1.5:1 for 193T) (magnification, $\times 100$). (B) Normalized Log₂ expression (x-axis) for exonic (red) and intronic (blue) overexpression of 835 genes covering entire chromosome 17 (y-axis), demonstrating intense intronic overexpression in genes involved in 17q amplicons of tumour 221T, but not tumour 193T without 17q amplification. Tumour 221T has two 17q amplicons with high exonic and intronic overexpression in the pericentromeric amplicon [genes 261 (*NLK*) to 302 (*CRLF3*)] and *ERBB2* amplicon [375 (*MEDI1*) to 391 (*CASC3*)]; *ERBB2* is at location 380. There was no overexpression of intronic sequences of tumour suppressors *TP53* (at 143) or *BRCA1* (455) in any tumours, not even for *NFI* (at 302) located between the two 17q amplicons. (C) Western blot analyses showing the presence of multiple shorter protein isoforms of *ERBB2* in breast tumours 194T, 013T, 271T, 160T, 193T and 221T compared with matched normal tissue controls 013N and 221N. These result from translation of the multiple alternatively spliced transcripts of *ERBB2* that include intronic and UTR sequences. HeLa cell control (lane 5) and β -actin used to normalise expression levels. Western blotting confirms truncated splice variant isoforms of *ERBB2* from alternatively spliced transcripts: 19 kDa: variant a; 67 kDa: variants f and g; 47 kDa: variant i; 42 kDa: variant k, (AceView NCBI human 2010 nomenclature). (D) IGV snapshot demonstrating exonic and intronic expression of *ERBB2* occurs in tumour 221T, whereas paired normal breast control 221N, lacks intronic expression demonstrating intronic overexpression is tumour-specific.

tumours (Table III). There was overlap of transcripts between genes on the sense and antisense strands in the *ERBB2* amplicon, but occurred only in tumours with 17q amplification. Paired end analysis of sequence overlap between *ERBB2* and adjacent gene *C17ORF37* on the opposite strand revealed transcriptional overlaps between sense and antisense strands when alternatively spliced *ERBB2* and *C17ORF37* variants with longer 5' and 3'UTRs were expressed. These cancer-specific variants extended beyond normal RefSeq variants, creating sequence overlaps that did not occur in controls or tumours without *ERBB2* amplification (Fig. 1). The amplicons and breakpoints involving *ERBB2* at chromosome 17q11-12 were identified by stringent analysis of exonic and intronic

expression of 835 genes on chromosome 17 in all three cancer subtypes and controls (Fig. 1).

ERBB2 amplified tumours were characterised by >2-fold intronic and exonic overexpression of *ERBB2*; neighbouring genes including *MEDI1*, *CDK12*, *STARD3*, *PGAP3*, *ERBB2*, *C17orf37*, *GRB7*, *GSDMB*, *ORMDL3*, *PSMD3*, *MED24*, *MSL1* and *CASC3* also showed intronic and exonic overexpression (Fig. 1). Most *ERBB2* amplified tumours contained only one amplicon that included *ERBB2* at 17q11.2-q12. However, two cancers contained both an *ERBB2*-amplicon and a second amplicon adjacent to the centromere (221T and 121T). Both amplicons were characterised by genes with >2-fold exonic and intronic expression. Between the two amplicons was a

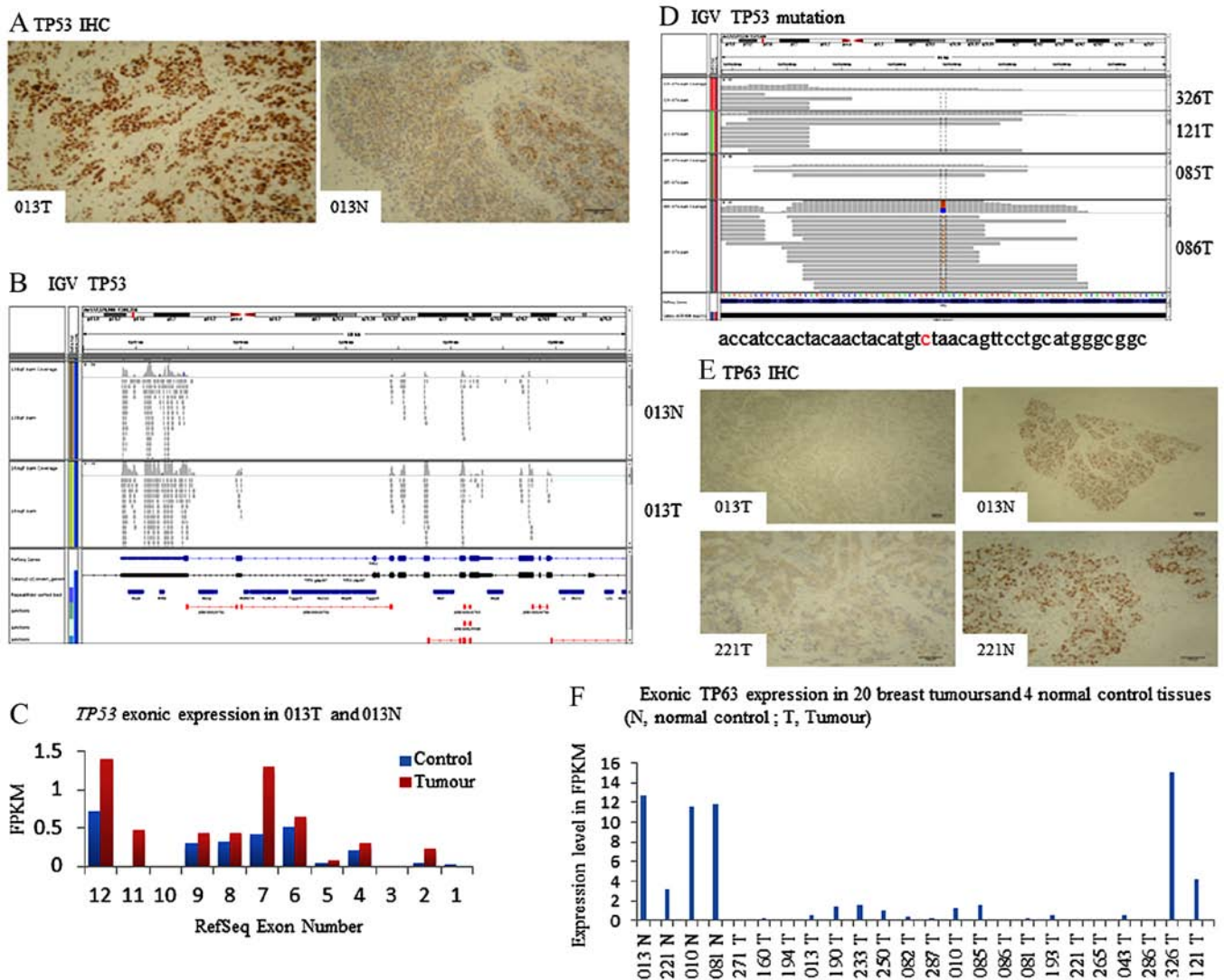


Figure 2. Tumour suppressor genes *TP53*, *TP63* show no intronic overexpression in tumours; alternative splicing determines nuclear localisation. (A) Immunohistochemistry for *TP53* demonstrating tumour 013T (x20) has nuclear localisation of *TP53*; in contrast, paired normal control 013N (x10) lacks nuclear localisation, is excluded from the nucleus and localised to the cytoplasm. (B) Nuclear localisation determined by alternatively splicing in *TP53* and *TP63*. IGV screenshot shows *TP53* in tumour 013T (lower panel) expressing of exon 11 encoding the nuclear localisation domain, thus conferring nuclear localisation for *TP53* in tumour 013T. In contrast, paired normal control 013N (upper panel) does not express exon 11 and is thus excluded from the nucleus. Nuclear *TP53* expression thus occurs only for tumour 013T not the paired control 013N. This is confirmed by (A) IHC showing respective subcellular localisations of 013T to nucleus and 013N to cytoplasm. (C) Comparison of expression of each of the 12 individual exons of *TP53* in tumour 013T in contrast to normal control 013N, showing exon 11 is expressed in tumour 013T but not in 013N providing quantitative information in relation to (A) IHC and (B) IGV results. (D) *TP53* transcriptome sequencing identifies mutations in transcripts: IGV demonstrates detection of mutation (C->G change) at position 7,577,568 chromosome 17 in tumour (lower panel), but not in control, (upper panel) confirmed by Sanger sequencing of *TP53*. (E) Immunohistochemistry of *TP63* in tumour samples 013T and 221T in comparison with their paired normal control samples 013N and 221N. This demonstrates greater expression in controls than the paired tumour samples, consistent with loss of TSG expression in tumourigenesis (013T and 013N, magnification, x10; 221T and 221N, magnification, x20). (F) Exonic *TP63* expression levels in 20 breast tumour samples and 4 controls determined by FPKM analysis shows *TP63* expression is significantly higher in controls than tumours ($p < 0.001$).

non-amplified region in which none of the genes including the TSG, *NFI* overexpressed or amplified (Fig. 1).

Tumour suppressor genes (TSG) including TP53, BRCA1/2 show no intronic overexpression in cancers. Our data revealed 17q amplification to be cancer-specific, occurring only in tumours but not matched controls (Fig. 1). The amplified 17q amplicons varied in size, involved genes with elevated (>2-fold) levels of intronic/exonic expression. Each individual amplicon comprised different breakpoints and genes, indicating a predisposition to amplicon formation in proximal 17q, although this was not site-specific (Fig. 1). In contrast, none of

the TSG on chromosome 17 (*TP53* at 17p13, *NFI* at 17q11.2), nor *BRCA1* at 17q21 showed >2-fold exonic or intronic expression in any tumour or control (Fig. 2). The most extreme case was that for the TSG and *NFI* located in the region between the two chromosome 17q amplicons. Analysis of exonic and intronic expression in the two tumours with two 17q amplicons revealed that although genes proximal and distal to *NFI* in the pericentromeric and *ERBB2* amplicons showed >2-fold intronic or exonic overexpression, *NFI* was not overexpressed in any tumour or control. Mutation analysis, SNP expression, alternative splicing and subcellular localisation assessed for each of the three TSGs *TP53*, *BRCA1* and *NFI* located on

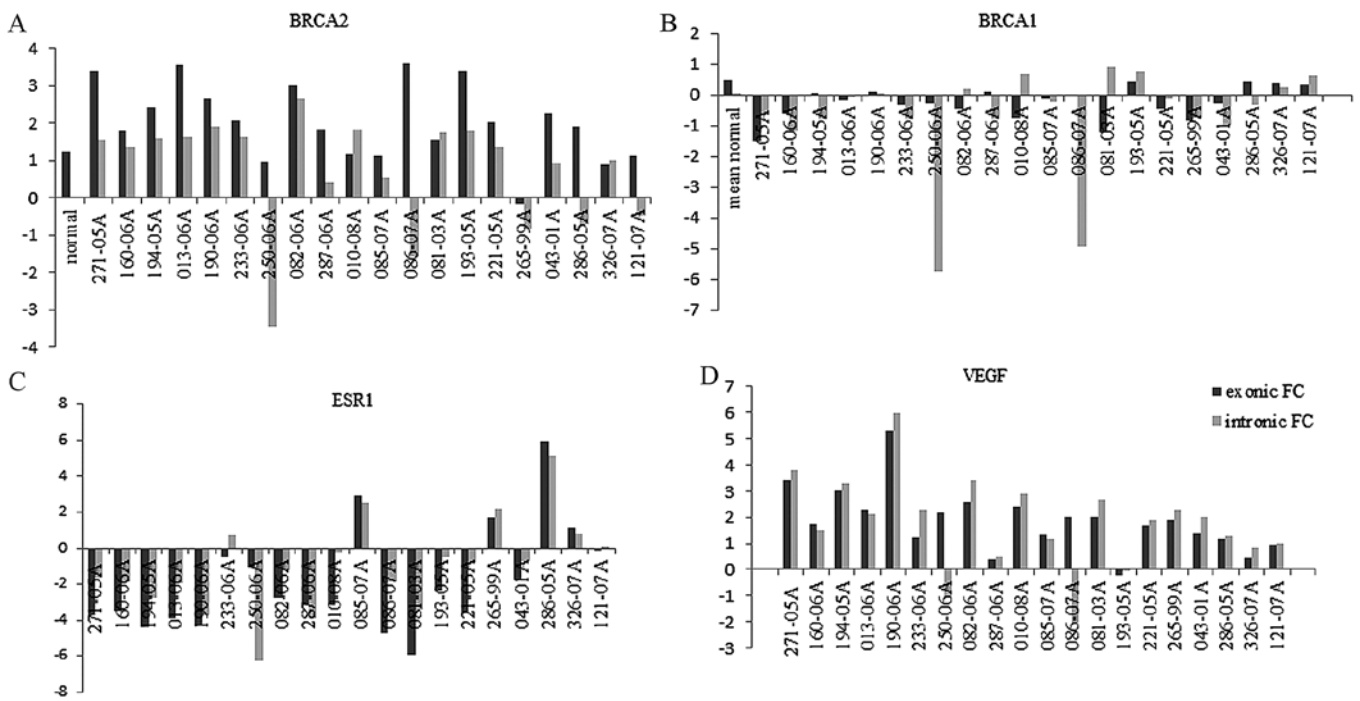


Figure 3. Exonic and intronic expression of four genes in 20 breast tumour samples and 4 controls determined by FPKM analysis. (A) *BRCA2* exonic expression is significantly higher in 11/20 carcinomas compared to controls. Intronic expression level was higher in 15/20 cancers compare to normal and it was lower than control in 4/20 cancers. (B) *BRCA1*, the exonic expression level was lower than control in 19/20 tumours. The intronic expression level was higher in 6/20 tumour compare to control. (C) *ESR1* expression is significantly higher in 4/20 and 5/20 for exonic and intronic expression respectively when normalized with controls. (D) The exonic and intronic VEGF expression is high in 19/20 and 17/20 tumours compare to control, respectively.

chromosome 17 (Fig. 2) revealed that alternative splicing of *TP53* generated only 1-2 RefSeq splice variants. The findings for *TP53*, *NF1* and *BRCA1* reveal no >2-fold intronic or exonic expression and only RefSeq isoforms were detected in any tumour or control. This was consistent in other major TSGs including *TP63/73*, *BRCA1/2*, *CDKN2A/2B* and *RB1/RBL1/RBL2* and it was in marked contrast to amplified oncogenes including *ERBB2* and *ESR1* that generated multiple truncated tumour-specific isoforms. Using mutation analysis, a non-synonymous *TP53* mutation was detected resulting in the amino acid change Cys106Ser in cancer 086T which was validated and confirmed using Sanger sequencing (Fig. 2). *TP63* showed higher expression in controls than tumours, the expected pattern for TSGs (Fig. 2). *TP63* expressed RefSeq variant *b*, an isoform with nuclear localisation and tetramerisation domains conferring nuclear localisation and confirmed by IHC (Fig. 2). *TP73* was not expressed in these breast carcinomas.

In normal tissues, *BRCA2* expression was significantly lower than in the malignant counterparts (Table IV and Fig. 3). Although no intronic or exonic expression >2-fold was found for *BRCA2*, FPKM values were higher for exon 11, involved in DNA repair (29), than other exons. *BRCA1*, expression was significantly lower in 16/20 carcinomas than the controls, with low intronic and exonic expression. Mutations were detected in *BRCA2*, in contrast to *BRCA1* where no mutations were identified (Table IV). Thus a novel and important finding of this study is that TSGs express no more than two RefSeq splice variants in any cancer or control, in contrast to the multiple alternatively spliced transcripts in oncogenes and hormone receptors, suggesting that the transcriptional integrity of TSGs

is tightly regulated in cancers, although this offers no protection against mutations or other changes.

Functional consequences of intronic overexpression: alternative splicing and truncated proteins. Transcriptome analyses for *ERBB2* were validated using qPCR to confirm transcript sequences, FISH to identify amplification, IHC to determine subcellular localisations and western blotting to confirm translation of alternatively spliced transcripts into proteins and assess the numbers and sizes of truncated protein isoforms in tumours and controls (Fig. 1). Several carcinomas expressed >30 copies of *ERBB2* when compared with matched controls and non-amplified tumours containing only two copies (Fig. 1). Chromosome 17 copy numbers in each cell were characterized using a concurrent second FISH probe to detect chromosome 17 centromeric sequences and hence identify chromosomal losses, gains and ploidy changes. In tumours with both *ERBB2* and pericentromeric amplicons, FISH revealed chromosome 17 centromeric regions in tumours to be larger than controls (Fig. 1), indicating amplification of the satellite and pericentromeric sequences, in agreement with studies demonstrating transcription of satellite, repeat, retrotransposon and other RNAs in tumorigenesis (30-33). We investigated intronic overexpression in relation to alternative splicing (AS) by calculating FPKM values for each exon, alternative exon and intron, aligning these with AceView (NCBI HG 19 Human AceView 2010) IGV and UCSC Blat. In carcinomas, when *ERBB2* was amplified, multiple highly truncated alternatively spliced variants are transcribed, resulting in the expression of a larger proportion of the genome, particularly intronic sequences not

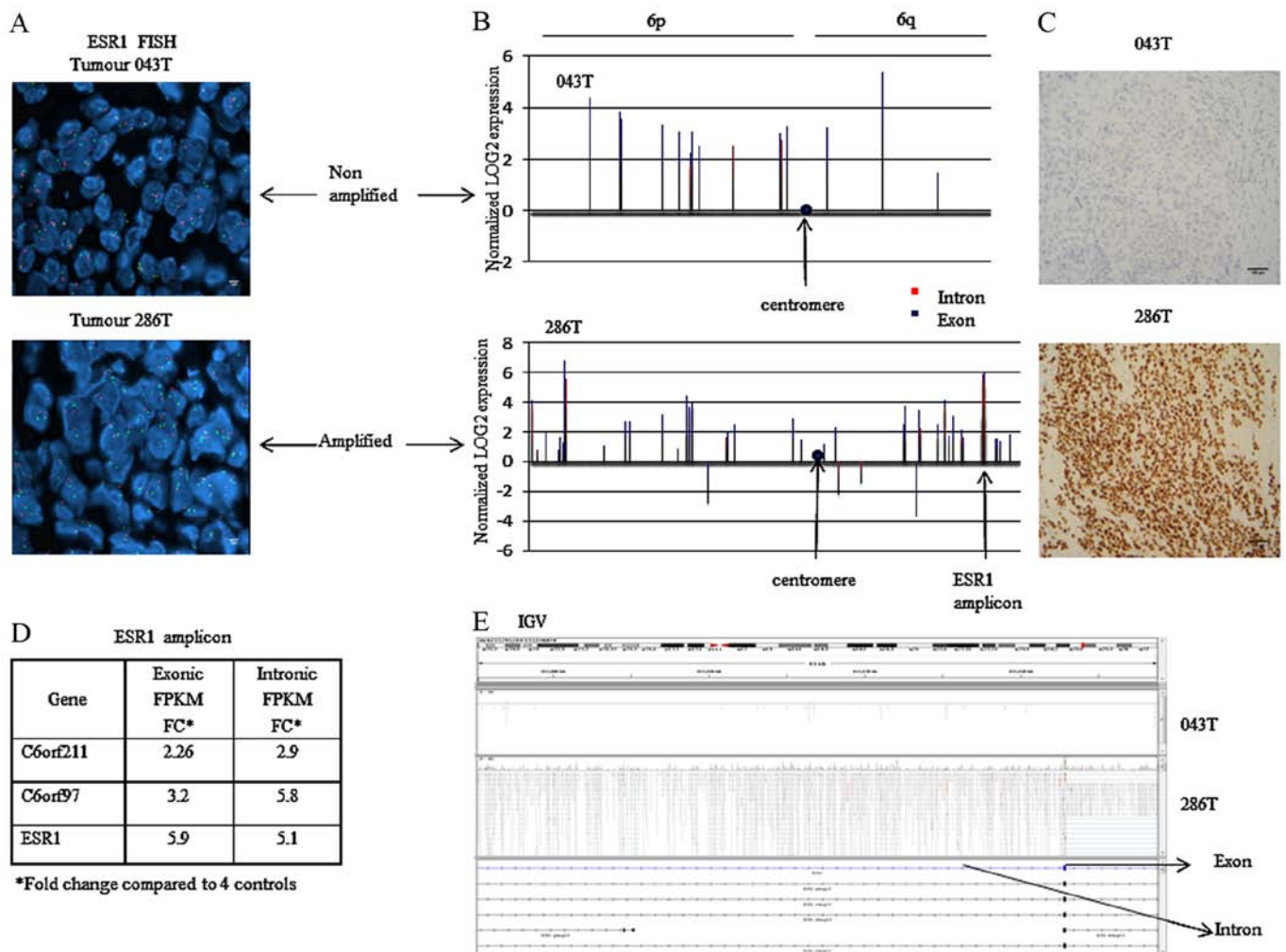


Figure 4. Tumour-specific amplification of hormone receptor *ESR1* at chromosome 6q26 associated with intronic overexpression. (A) FISH analysis of *ESR1* in *ESR1*. Tumour sample 086T shows amplification of *ESR1* by FISH analysis (*ESR1* shown by fluorescent green signals) compared with non-amplified (043T) samples. Chromosome 6 centromeric probe (orange fluorescent signals) indicates the numbers of chromosome 6 present in relation to *ESR1* signals (magnification, $\times 100$). (B) Normalized LOG2 expression of exons and introns over the 796 genes over the whole of chromosome 6, demonstrating the three genes, *ESR1*, *C6ORF211* and *C6ORF97* comprising the *ESR1* amplicon at 6q26 in breast tumour 286T, but there is no *ESR1* amplicon in the *ESR1*-negative sample 043T. (C) *ESR1* immunohistochemistry demonstrates intense staining for *ESR1* in the *ESR1* amplified tumour 286T, compared with *ESR1*-negative tumour 043T in line with the results for IHC and normalised LOG2 expression in (A and B). (D) Exonic and intronic FPKM fold-changes showing the very high levels of expression in the three genes *ESR1*, *C6ORF97* and *C6ORF211* that comprise the *ESR1* amplicon in tumour 286T compared with pooled control samples. (E) IGV transcriptome screen shot demonstrating the specific high levels of exonic and intronic expression in *ESR1* (lower panel) in *ESR1* amplified breast tumour 286T compared with the lack of intronic or expression in *ESR1*-negative tumour 043T.

expressed in normal controls. The splice variants contained initiation sites, exons, introns, 5' and 3'UTRs, differing from the normal RefSeq variants of control samples, many being highly expressed. The transcriptional start and end points were validated using RT-PCR. To demonstrate that NMD pathways were not activated in these shortened transcripts, we used western blot analysis and subcellular localisation analysis to validate that in *ERBB2* amplified breast tumours, multiple alternatively spliced truncated transcripts were translated into highly truncated proteins, preponderantly smaller and varying in size, corresponding to the alternatively spliced *ERBB2* transcripts identified according to AceView (Fig. 1). In contrast, in normal breast and non-*ERBB2* amplified tumours, the full length *ERBB2* isoform (149 kDa) was predominant (Fig. 1).

ESR1, intronic transcription and overexpression affect function. Five of the 20 breast carcinomas expressed *ESR1* protein

identified by IHC, although only three showed >2 exonic fold-change (Fig. 3). FPKM analysis of exonic and intronic expression of *ESR1* at 6q25 and a further 794 genes on chromosome 6 detected a small amplicon at 6q25 comprising the three genes *ESR1*, *C6orf211* and *C6orf9* in an *ESR1*-positive tumour sample, associated with both intronic and exonic overexpression in *ESR1* (Figs. 3 and 4). *ESR1* amplification was found to be associated with intronic overexpression and alternative splicing (Fig. 5). In the absence of estrogen, *ESR1* functions as a nuclear transcription factor that promotes breast cell growth. The receptors are inactive and cytoplasmic. However, on activation by estrogen, they dimerise and translocate to the nucleus. Thus, changes in subcellular localisation of *ESR1* from cytoplasm to the nucleus will effect its function.

We also analysed the transcriptional profiles of the other three major hormone receptors *ESR2*, progesterone receptor (*PGR*) and androgen receptor (*AR*). For *PGR*,

Table IV. Base pair changes in BRCA2 in all cases and base pair changes in the 8 repeats of BRCA2 gene.

A, Base pair changes in BRCA2 in all cases				
	Base pair change	Position	Syn/non syn	Rs
193T	GTA→GCA	32,929,387	Non syn (exon 14)	Rs169547
	ATAA→ATCA	32,914,895	exon 11	Rs80359583
	CTG→CTC	32,915,005	Syn (exon 11)	Rs206076
221T	GTA→GCA	32,929,387	Non syn (exon 14)	Rs169547
	ATAA→ATCA	32,914,895	exon 11	Rs80359583
	CTG→CTC	32,915,005	Syn (exon 11)	Rs206076
271T	GTA→GCA	32,929,387	Non syn (exon 14)	Rs169547
	CTG→CTC	32,915,005	Syn (exon 11)	Rs206076
194T	GTA→GCA	32,929,387	Non syn (exon 14)	Rs169547
	CTG→CTC	32,915,005	Syn (exon 11)	Rs206076
013T	GTA→GCA	32,929,387	Non syn (exon 14)	Rs169547
	CAG→CAT	32,911,936	Non syn (exon 11)	-
	CTG→CTC	32,915,005	Syn (exon 11)	Rs206076
190T	GTA→GCA	32,929,387	Non syn (exon 14)	Rs169547
	CTG→CTC	32,915,005	Syn (exon 11)	Rs206076
233T	G→T	32,890,572	5' UTR	Rs1799943
	GTA→GCA	32,929,387	Non syn (exon 14)	Rs169547
	CTG→CTC	32,915,005	Syn (exon 11)	Rs206076
250T	G→T	32,890,572	5' UTR	Rs1799943
	GTA→GCA	32,929,387	Non syn (exon 14)	Rs169547
	GTA→GCA	32,929,387	Non syn (exon 14)	Rs169547
043T	CTG→CTC	32,915,005	Syn (exon 11)	Rs206076
	AAT→CAT	32,906,729	Non syn (exon 10)	-
	GTA→GCA	32,929,387	Non syn (exon 14)	Rs169547
286T	CTG→CTC	32,915,005	Syn (exon 11)	Rs206076
	AAT→CAT	32,906,729	Non syn (exon 10)	-
	GTA→GCA	32,929,387	Non syn (exon 14)	Rs169547
082T	CTG→CTC	32,915,005	Syn (exon 11)	Rs206076
	GTA→GCA	32,929,387	Non syn (exon 14)	Rs169547
287T	CTG→CTC	32,915,005	Syn (exon 11)	Rs206076
	G→T	32,890,227	Intron 1	Rs9567552
	TCA→TCG	32,929,232	Syn (exon 14)	-
010T	GTA→GCA	32,929,387	Non syn (exon 14)	Rs169547
085T	CTG→CTC	32,915,005	Syn (exon 11)	Rs206076
086T	CTG→CTC	32,915,005	Syn (exon 11)	Rs206076
	AAT→CAT	32,906,729	Non syn (exon 10)	-
081T	GTA→GCA	32,929,387	Non syn (exon 14)	Rs169547

B, Base pair changes in the 8 repeats of BRCA2 gene			
	Base pair change	Position	Syn/non syn
193-05A	GUU→GUC	32,912,299	SYN (V)
271-05A	AAA→AAG	32,911,888	SYN (K)
013-06A	CAG→CAU	32,911,936	NONSYN (Q→H)
233-06A	GUU→GUC	32,912,299	SYN (V)
043-01A	GUU→GUC	32,912,299	SYN (V)
286-05A	GAA→UAA	32,914,118	NONSYN (E→STOP CODON)
287-06A	AAA→AAG	32,911,888	SYN (K)
121-07A	AGU→AUU	329,144,098	NONSYN (S→I)
081-03A	AAC→ACC	32,911,932	NONSYN (N→T)
	GUU→GUC	32,912,299	SYN (V)
081-03B	GUU→GUC	32,912,299	SYN (V)

The mutations and SNPs are shown with their genomic positions and details of whether these changes are synonymous or non-synonymous.

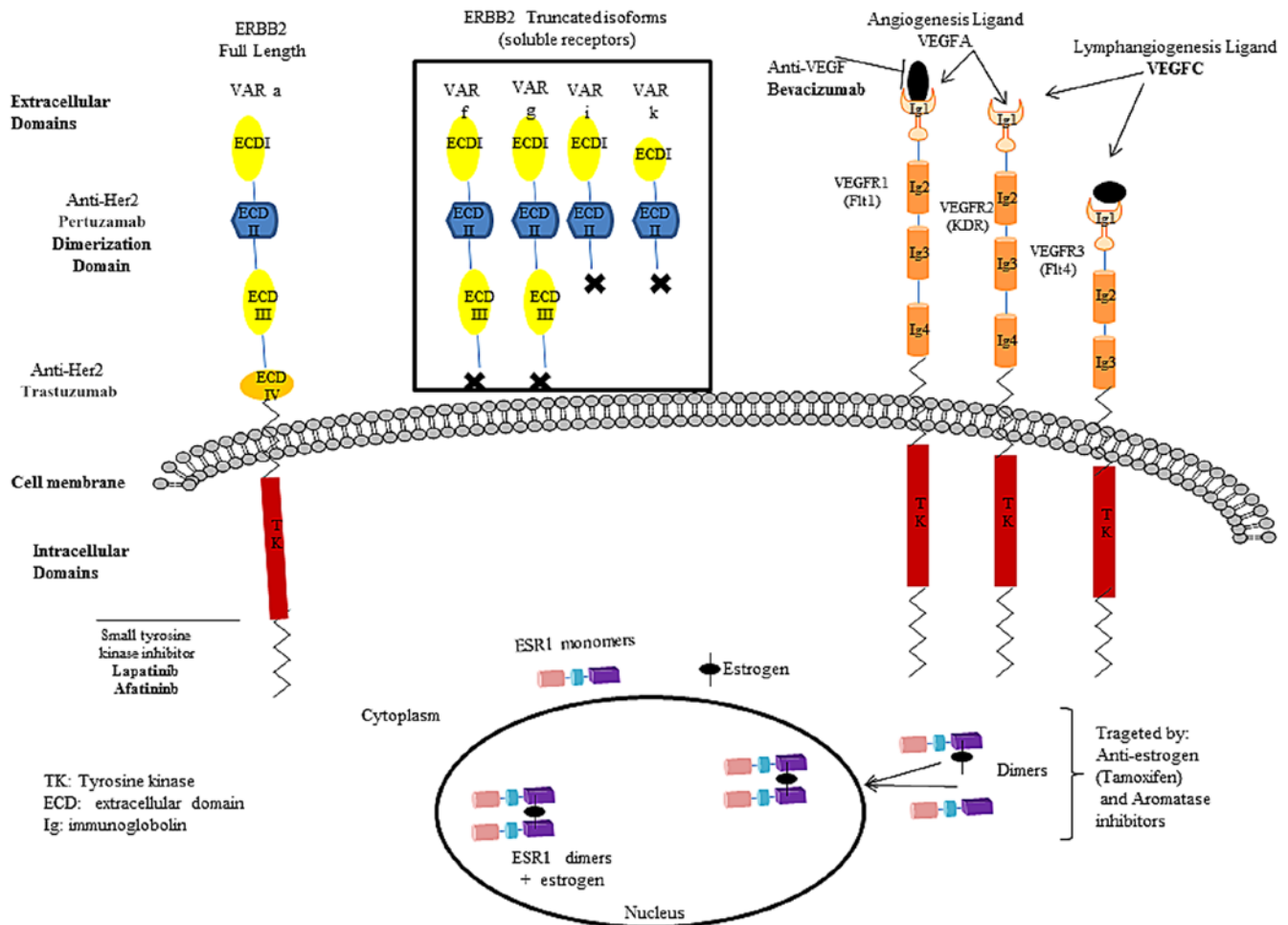


Figure 5. Schematic view of the consequences of tumour-specific intronic overexpression generating multiple alternatively spliced variants giving rise to short protein isoforms with altered structures, functions and subcellular localisations affecting targeted drug therapies. Figure shows a diagrammatic representation of a cell and its subcellular compartments with cell membrane, cytoplasm and nucleus. Upper left, full length RefSeq *ERBB2* localised to cell membrane with extracellular domains targeted by monoclonal antibodies trastuzumab (Herceptin) and pertuzumab and intracellular TK domain targeted by small molecule tyrosine kinase inhibitors (lapatinib, erlotinib and gefitinib) indicated. The central box shows the shorter protein isoforms of *ERBB2* (variants f, g, i and k) identified by western blotting after exonic and intronic expression analysis and IGV. Unlike full length *ERBB2* isoforms the short variants f, g, i and k lack ECD IV, the ECD targeted by Herceptin and intracellular TKI domains. These short isoforms also lack transmembrane and intracellular domains, resulting in 'soluble receptors' potentially carrying drugs into serum rather increasing levels in tumour cells. Upper right, multiple alternative splicing for angiogenic and lymphangiogenic receptors *VEGFR1-3* (*FLT1*, *KDR* and *FLT4*, respectively) and their ligands *VEGFA* and *VEGFC*. As for the transmembrane receptor *ERBB2*, the generation of multiple proteins shorter than full length RefSeq forms may affect whether receptors and ligands such as *VEGFA*, are inhibitory, stimulatory and signal transduction. Lower figure monomers of *ESR1* have cytoplasmic localisation but full length *ESR1* is able to bind to estrogen allowing it to dimerise and move to the nucleus where it acts as a transcription factor. Our results (Fig. 4) demonstrate that tumours with amplified *ESR1* or *ESR1* intronic overexpression have only cytoplasmic, not nuclear *ESR1* expression, which impairs its nuclear transcription factor function and responses to drug therapy.

18/20 cancers showed much lower exonic and intronic expression than the controls. The remaining 2 positive cancers were also positive by IHC. AR expression was negative in 12/20 samples but positive in 8/20. All 12 samples negative for AR were triple-negative subtypes.

Oncogene expression in amplicons: 8q24 amplification without MYC overexpression. The 8q24 region is frequently amplified in breast cancers and may be associated with MYC oncogene overexpression (34). In a tumour with 8q24 amplification, we studied intronic and exonic expression in 488 genes over the whole of chromosome 8, comparing tumours in each of the three subtypes with controls. Many genes contained exonic or intronic overexpression (Table III). Surprisingly, no exonic or intronic overexpression (>2-fold) was detected for MYC in the 8q24 amplified tumour. This finding was striking,

because the adjacent gene, *MTSS1* (at 8p22), showed >2-fold exonic and intronic expression of 4.1 and 3.7, respectively, whereas *MYC* did not (Table III and Fig. 4). It was assumed axiomatic that 8q24 amplification would involve *MYC* overexpression. However, our studies show that although 2.5% (range, 0.2-17.8%) of the 488 genes on chromosome 8 demonstrated exonic overexpression, 15.4% (range, 0-24.8%) intronic overexpression and 5.4% (range, 0-20.1%) both exonic and intronic overexpression, *MYC* showed no overexpression. To further characterise oncogene expression, we investigated whether overexpression occurred for other oncogenes including *EGFR* family members (*EGFR*, *ERBB3* and *ERBB4*) in each tumour subgroup and controls. No tumour or control showed *EGFR* amplification. For *ERBB3*, one sample with overexpressed *ERBB2* (286T) showed a >2-fold-change of both exonic and intronic expressions. For *ERBB4*, two *ERBB2* 3⁺ patients

Table V. Mutations, SNPs and base pair changes in *VEGFA* gene.

	Base pair change	Position	Syn/non syn	Rs
271T, 190T, 10T, 10B	A→G	43,753,212	3'UTR	Rs10434
043T, 085T, 086T, 081T	TCC→TCT	43,738,977		Rs25648
082T, 010T, 265T	C→T CTG→CTC	43,746,169	Intron 3	Rs3025000
081T	ATG→AGC G→T	43,738,985 43,752,287	3' exon	
086T	G→A	43,753,325	3'UTR	Rs3025023
250T, 286T, 082T, 043T, 326T, 121T, 085T, 086T, 010T, 081T, 265T, 190T, 233T, 287T	G→T	43,746,276	Intron 4	Unknown
221T, 286T	C→T	43,752,536	3'UTR	Rs3025039
221T, 194T	C→T	43,753,051	3'UTR	Rs3025040

showed >2-fold-changes (221T entrance and 286T exonic) and one triple-negative patient (085T) showed >2-fold exonic and intronic overexpression. Other major oncogene families investigated included epidermal, fibroblast, vascular endothelial and platelet derived growth factor receptors.

Specificity of intronic transcription and alternatively spliced transcripts. To investigate whether intronic overexpression was associated with particular introns and sequences, we calculated FPKM values for each exon, alternative exon and intron. We aligned these with the AceView alternative splicing database, using IGV, UCSC Blat and Repeat Masker to identify locations of repeat sequences, retrotransposons and other repeat elements (NCBI HG19 UCSC Blat, IGV and Repeat Masker). The upregulated intronic sequence expression found in cancers was not random, with high specificity for genes, introns and sequences involved associated with retrotransposons and other repeat sequences (data not shown). For *ERBB2*, the two introns with highest FPKM values were those immediately before the transmembrane domain. Predictions indicated that these would give rise to shorter, alternatively spliced variants and truncated protein isoforms lacking transmembrane and intracellular domains, no longer localised to the cell membrane, and instead generating soluble receptors (Fig. 5). Intronic LINES, *Alus* and other retroelements (enigmatic dark matter) can induce alternative splicing by mechanisms including exonisation of cryptic splice sites, forcing exonisation and cryptic polyadenylation or exon skipping (35-39). In this way, intronic or intergenic sequence expression may contribute to the acquisition of new exons and alternatively spliced variants.

TNCs, intronic overexpression and identification of drug targets. We confirmed that TNCs contained no *ERBB2* or *ESR1* amplification, but for TNCs our results demonstrate that for *VEGFA*, 10/13 exhibited >2-fold exonic or intronic expression, but 3/13 (23%) showed no overexpression, emphasising the need for personalised profiling to increase effectiveness and reduce

toxicity (Figs. 3 and 5). *VEGFA* undergoes alternative splicing, producing variants having either stimulatory or inhibitory functions (40), highlighting that transcriptome profiling may enable targeting of stimulatory isoforms. We have also shown a number of synonymous or non-synonymous mutation in *VEGFA* (Table V). In this study, TNC produced only inhibitory isoforms. VEGF receptors are also potential therapeutic targets, but our results showed no significant elevated expression for *VEGFR1* (*FLT1*) or *VEGFR2* (*KDR*) (Fig. 5).

Tumour-specific transcription, alternative splicing, protein isoforms and targeted cancer therapies. For *ERBB2* amplified tumours, intronic overexpression, multiple alternatively spliced transcripts and protein isoforms of tumours differ markedly from those of non-malignant samples and have implications for diagnosis, prognosis, targeted therapies and biomarkers. Although full length RefSeq *ERBB2* found in normal cells is localised to cell membrane, in tumours, truncated *ERBB2* isoforms have differing sizes, and subcellular localisations. Some isoforms may lack transmembrane and intracellular domains and no longer localise to cell membrane, becoming soluble receptors, but still have extracellular domains that bind to targeted drugs, forming soluble-receptor drug complexes, potentially partitioning toxic drug-soluble receptor complexes into serum (Fig. 5). Oncogenic proteins with altered or ablated drug target domains with altered monoclonal antibody or TK inhibitor binding, may render therapies ineffective or tumours unresponsive (Figs. 1 and 5). The estrogen receptor *ESR1* normally localises to the nucleus, functioning as a nuclear transcription factor. However, our results demonstrate that *ESR1* amplification may result in cytoplasmic rather than nuclear localisation, affecting nuclear transcription factor function (Figs. 3 and 5). The results suggest transcription of intronic 'dark matter' in cancers plays a key role in generating novel transcriptional landscapes, that provides insights into tumour biology and potential pitfalls and new opportunities for personalised cancer therapies (41).

Discussion

Functional implications of tumour-specific intronic transcription and alternative splicing. This study revealed the extent of genomic structural alterations occurring during the evolution of human breast cancers and has identified some of the key changes to be subtype-specific and mutually exclusive. The findings have fundamental implications for tumour-specific breast cancer therapy. In non-neoplastic cells, ~2-3% of genomic 'exonic' DNA primes synthesis of protein-coding mRNA and transcription is tightly regulated (42). Here, we show that in cancers, massive genomic rearrangements occur with pervasive cancer-specific transcriptional changes that include intronic sequence overexpression generating alternatively spliced transcripts and truncated protein isoforms cannot be revealed by exome analysis. First, RefSeq exome sequencing cannot detect intronic overexpression and tumour-specific alternative splicing. Second, the relevance of these findings to current breast cancer treatment is that they show the functional correlates of tumour-specific alterations, revealing new biologically relevant therapeutic targets. Third, alternatively spliced tumour-specific transcripts and truncated proteins are likely to alter drug target domains. These structural changes within the genome modify the therapeutic effectiveness of targeted drugs such as Herceptin (trastuzumab) or small molecule TK inhibitors, designed to target external or TK domains of full length RefSeq *ERBB2* (43). Truncation may ablate or alter drug target domains and explain lack of response, ineffectiveness or resistance to targeted therapies. Finally, GWAS have identified genetic risk loci for breast cancers and SNPs are a key resource in cancer genetics. However, many high-risk SNPs are intronic, suggesting the functional involvement of intronic sequences in cancers. More than 95% of breast cancer susceptibility variants are found in non-exonic regions previously thought to be non-coding. Our studies provide a framework for tumour-specific analysis of introns and other non-exonic regions to assess whether high risk SNPs may be associated with tumour-specific transcription.

Massive genomic rearrangements, oncogenes and tumour suppressors in breast cancer. The data herein suggested that some chromosomal regions including 17q, 6q and 8q24 appear predisposed to amplicon formation. However, each amplicon has differing breakpoints, indicating the uniqueness and complexities of individual amplicon formation. Relationships between chromosomal amplification and oncogene overexpression may also be complex. Prior to these studies, it appeared axiomatic that 8q24 amplification inevitably involved *MYC* overexpression. However, for 8q24 amplification and *MYC* our data show that oncogene overexpression and chromosomal amplification are not invariably linked. Absence of *MYC* overexpression in 8q24 amplification was surprising given that 15.4% of chromosome 8 genes showed intronic overexpression, 2.5%, showed exonic overexpression and 5.4% exhibited both. The findings illustrate the necessity for detailed gene expression studies to analyse functional correlates of chromosome amplification, rather than relying solely on surrogates such as FISH.

In contrast to oncogene/hormone receptor amplification and intronic overexpression, no TSG showed intronic or exonic

overexpression or the generation of truncated alternatively spliced variants. The comparisons were stark, given that in some tumours, TSGs including *BRCA1* and *NFI* were located immediately adjacent to 17q amplicons in which oncogenes including *ERBB2* showed extensive intronic overexpression, alternative splicing and generated truncated protein isoforms. The expression profiles of TSGs suggest that, by their resistance to transcriptional disruption and intronic overexpression, these genes retain their normal function and highlight the abnormalities and extent of tumour-specific changes in amplified oncogenes. TSGs play key roles to prevent tumourigenesis in breast and other epithelia. The implications of our findings are that in tumours there are either mechanisms leading to intronic sequences overexpression and triggering tumour specific alternative splicing or mechanisms that prevent transcription of intronic sequences, in TSGs even when closely mapping oncogenes are amplified and overexpressed. Suggestions that mutations in *TP53* may be linked with catastrophic chromosomal tumour-specific amplifications and rearrangements are intriguing and bring new perspectives in relation to TSGs and mechanisms of amplification (44). Analysis of *TP53* mutations in relation to 17q, 6q and 8p amplification revealed no direct links between *TP53* mutations and amplicons, although other loci, amplicons and mutations in other types of tumour should be analysed.

The human genome contains 40-50% of repetitive sequences derived from retrotransposable elements, with multiple copies of long and short interspersed nucleotide elements (LINEs and SINES), LTRs and other repeat sequences (31,45-47). This study provides additional new information about the extent of intronic transcription and alternative splicing that illuminates the accumulating evidence in cancer, transcription of intronic, UTR satellite and transposon sequences occurs and extends the repertoire of alternatively spliced transcripts (31). Intronic L1s, *Alus* and other retroelements (enigmatic dark matter) affect transcription and gene expression via exon skipping and exonisation using cryptic splice sites, forcing exonisation and cryptic polyadenylation so genes acquire new exons from intronic or intergenic sequences thereby generating novel alternatively spliced variants (41,48). Recently, we described intronic exonisation in human prostate cancer, suggesting that this phenomenon is likely to be common to many different malignancies (49). In normal somatic cells, this sequence compartment is attenuated by epigenetic silencing mechanisms involving DNA methylation and chromatin-mediated repression to maintain genomic integrity (45,50,51). In tumours, there are suggestions that transcription from retrotransposons occurs subsequent to loss of methylation (50). Loss or inactivation of TSGs in cancers, has led to the suggestion that demethylating agents may be utilised to promote expression of lost or inactivated TSGs. While demethylating drugs may enhance TSG expression, there are potential risks of exacerbating or increasing retrotransposon transcription, that may potentially result in oncogene or other tumour promoting gene upregulation or overexpression, accelerating tumourigenesis rather than suppressing growth (52).

Effects of tumour-specific transcription and translation on targeted cancer therapies. In this study, we demonstrate that intronic regions harbour sequences transcribed only in

cancers and not in normal tissues. Since such tumour-specific sequences encompass transcripts not included in RefSeq exons or revealed by exome analysis, studies should be extended beyond the limits of individual exomes if the full extent of cancer transcription and exploitation of intronic and intergenic sequences is to be uncovered. Such previously unrecognised transcription and translation from amplified or disrupted oncogenes will generate RNA transcripts and proteins with altered size structures, functions and subcellular localisations similar to those detected in prostate cancer (49). Drug development has been predicated on targeting normal full-length RefSeq proteins with domains such as external, transmembrane and TK domains in RTKs. Divergent protein isoforms could potentially affect function through diverse mechanisms ranging from ablation of drug target domains to alterations such as those to soluble rather than transmembrane receptors for RTKs (53,54), nuclear transcription factors localised to the cytoplasm for *ESR1* (55), mesenchymal rather than epithelial signalling for *FRGR2* (56,57) and inhibitory rather than stimulatory functions for *VEGFA* (58). For amplified *ERBB2* these findings have major implications for targeted therapies.

Despite the promise of personalized cancer medicine, therapeutic options are available only for breast cancers with amplified/overexpressed *ERBB2* or *ESR1*. TNCs are the most aggressive tumours but lack therapeutic options, having no targetable *ERBB2* or *ESR1* overexpression. Our findings suggest transcriptome profiling may identify alternative physiological mechanisms as therapeutic targets. Non-selective genome-wide demethylation is unlikely to be therapeutically advantageous while anti-angiogenic therapies may be ineffective, or potentially toxic, without therapeutic benefit, thus explaining the disappointing results of the anti-*VEGFA* monoclonal bevacizumab and FDA withdrawal of approval for use in breast cancer (FDA 2011). This study suggests transcriptional profiling may be an appropriate approach to individual cancers by providing a comprehensive identification of new therapeutic targets for effective personalised cancer therapies.

References

- Office for National Statistics: Office for National Statistics, Cancer Statistics Registrations: Registrations of Cancer Diagnosed in 2008, England. Series MB1. National Statistics, London, p43, 2012.
- Sorlie T, Tibshirani R, Parker J, Hastie T, Marron JS, Nobel A, Deng S, Johnsen H, Pesich R, Geisler S, *et al*: Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc Natl Acad Sci USA* 100: 8418-8423, 2003.
- Baldassarre G, Battista S, Belletti B, Thakur S, Pentimalli F, Trapasso F, Fedele M, Pierantoni G, Croce CM and Fusco A: Negative regulation of BRCA1 gene expression by HMGAI1 proteins accounts for the reduced BRCA1 protein levels in sporadic breast carcinoma. *Mol Cell Biol* 23: 2225-2238, 2003.
- O'Shaughnessy J, Osborne C, Pippen JE, Yoffe M, Patt D, Rocha C, Koo IC, Sherman BM and Bradley C: Iniparib plus chemotherapy in metastatic triple-negative breast cancer. *N Engl J Med* 364: 205-214, 2011.
- Dent R, Hanna WM, Trudeau M, Rawlinson E, Sun P and Narod SA: Pattern of metastatic spread in triple-negative breast cancer. *Breast Cancer Res Treat* 115: 423-428, 2009.
- Rodríguez-Pinilla SM, Sarrío D, Honrado E, Hardisson D, Calero F, Benitez J and Palacios J: Prognostic significance of basal-like phenotype and fascin expression in node-negative invasive breast carcinomas. *Clin Cancer Res* 12: 1533-1539, 2006.
- Hicks J, Krasnitz A, Lakshmi B, Navin NE, Riggs M, Leibu E, Esposito D, Alexander J, Troge J, Grubor V, *et al*: Novel patterns of genome rearrangement and their association with survival in breast cancer. *Genome Res* 16: 1465-1479, 2006.
- Ficz G, Branco MR, Seisenberger S, Santos F, Krueger F, Hore TA, Marques CJ, Andrews S and Reik W: Dynamic regulation of 5-hydroxymethylcytosine in mouse ES cells and during differentiation. *Nature* 473: 398-402, 2011.
- Hudis CA: Trastuzumab - mechanism of action and use in clinical practice. *N Engl J Med* 357: 39-51, 2007.
- Giampaglia M, Chiuri VE, Tinelli A, De Laurentiis M, Silvestris N and Lorusso V: Lapatinib in breast cancer: Clinical experiences and future perspectives. *Cancer Treat Rev* 36 (Suppl 3): S72-S79, 2010.
- Folkman J: Angiogenesis: An organizing principle for drug discovery? *Nat Rev Drug Discov* 6: 273-286, 2007.
- Carey LA and Sharpless NE: PARP and cancer - if it's broke, don't fix it. *N Engl J Med* 364: 277-279, 2011.
- Leslie M: Brothers in arms against cancer. *Science* 331: 1551-1552, 2011.
- Tao Y, Liu S, Briones V, Geiman TM and Muegge K: Treatment of breast cancer cells with DNA demethylating agents leads to a release of Pol II stalling at genes with DNA-hypermethylated regions upstream of TSS. *Nucleic Acids Res* 39: 9508-9520, 2011.
- Jones S, Zhang X, Parsons DW, Lin JC, Leary RJ, Angenendt P, Mankoo P, Carter H, Kamiyama H, Jimeno A, *et al*: Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science* 321: 1801-1806, 2008.
- Sugarbaker DJ, Richards WG, Gordon GJ, Dong L, De Rienzo A, Maulik G, Glickman JN, Chirieac LR, Hartman ML, Taillon BE, *et al*: Transcriptome sequencing of malignant pleural mesothelioma tumors. *Proc Natl Acad Sci USA* 105: 3521-3526, 2008.
- Bueno R, De Rienzo A, Dong L, Gordon GJ, Hercus CF, Richards WG, Jensen RV, Anwar A, Maulik G, Chirieac LR, *et al*: Second generation sequencing of the mesothelioma tumor genome. *PLoS One* 5: e10612, 2010.
- Chen W, Kalscheuer V, Tzschach A, Menzel C, Ullmann R, Schulz MH, Erdogan F, Li N, Kijas Z, Arkesteijn G, *et al*: Mapping translocation breakpoints by next-generation sequencing. *Genome Res* 18: 1143-1149, 2008.
- Rothberg JM and Leamon JH: The development and impact of 454 sequencing. *Nat Biotechnol* 26: 1117-1124, 2008.
- Kwan T, Benovoy D, Dias C, Gurd S, Provencher C, Beaulieu P, Hudson TJ, Sladek R and Majewski J: Genome-wide analysis of transcript isoform variation in humans. *Nat Genet* 40: 225-231, 2008.
- Lee YH, Kim JH and Song GG: Genome-wide pathway analysis of breast cancer. *Tumour Biol* 35: 7699-7705, 2014.
- Hajiloo M, Damavandi B, Hooshadad M, Sangi F, Mackey JR, Cass CE, Greiner R and Damaraju S: Breast cancer prediction using genome wide single nucleotide polymorphism data. *BMC Bioinformatics* 14 (Suppl 13): S3, 2013.
- Gascard P, Bilenky M, Sigaroudinia M, Zhao J, Li L, Carles A, Delaney A, Tam A, Kamoh B, Cho S, *et al*: Epigenetic and transcriptional determinants of the human breast. *Nat Commun* 6: 6351, 2015.
- Trapnell C, Pachter L and Salzberg SL: TopHat: Discovering splice junctions with RNA-Seq. *Bioinformatics* 25: 1105-1111, 2009.
- Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G and Durbin R; 1000 Genome Project Data Processing Subgroup: The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25: 2078-2079, 2009.
- Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G and Mesirov JP: Integrative genomics viewer. *Nat Biotechnol* 29: 24-26, 2011.
- Forootan SS, Wong YC, Dodson A, Wang X, Lin K, Smith PH, Foster CS and Ke Y: Increased Id-1 expression is significantly associated with poor survival of patients with prostate cancer. *Hum Pathol* 38: 1321-1329, 2007.
- Ro JS, el-Naggar A, Ro JY, Blick M, Frye D, Fraschini G, Fritsche H and Hortobagyi G: c-erbB-2 amplification in node-negative human breast cancer. *Cancer Res* 49: 6941-6944, 1989.
- Jensen RB, Carreira A and Kowalczykowski SC: Purified human BRCA2 stimulates RAD51-mediated recombination. *Nature* 467: 678-683, 2010.
- Faulkner GJ, Kimura Y, Daub CO, Wani S, Plessy C, Irvine KM, Schroder K, Cloonan N, Steptoe AL, Lassmann T, *et al*: The regulated retrotransposon transcriptome of mammalian cells. *Nat Genet* 41: 563-571, 2009.

31. Faulkner GJ: Retrotransposons: Mobile and mutagenic from conception to death. *FEBS Lett* 585: 1589-1594, 2011.
32. Parenteau J, Durand M, Morin G, Gagnon J, Lucier JF, Wellinger RJ, Chabot B and Elela SA: Introns within ribosomal protein genes regulate the production and function of yeast ribosomes. *Cell* 147: 320-331, 2011.
33. Baillie JK, Barnett MW, Upton KR, Gerhardt DJ, Richmond TA, De Sapio F, Brennan PM, Rizzu P, Smith S, Fell M, *et al*: Somatic retrotransposition alters the genetic landscape of the human brain. *Nature* 479: 534-537, 2011.
34. Xu J, Chen Y and Olopade OI: MYC and Breast Cancer. *Genes Cancer* 1: 629-640, 2010.
35. Pennisi E: Shining a light on the genome's 'dark matter'. *Science* 330: 1614, 2010.
36. Gal-Mark N, Schwartz S and Ast G: Alternative splicing of Alu exons - two arms are better than one. *Nucleic Acids Res* 36: 2012-2023, 2008.
37. Bennett EA, Keller H, Mills RE, Schmidt S, Moran JV, Weichenrieder O and Devine SE: Active Alu retrotransposons in the human genome. *Genome Res* 18: 1875-1883, 2008.
38. Hormozdiari F, Alkan C, Ventura M, Hajirasouliha I, Malig M, Hach F, Yorukoglu D, Dao P, Bakshi M, Sahinalp SC, *et al*: Alu repeat discovery and characterization within human genomes. *Genome Res* 21: 840-849, 2011.
39. Bazak L, Levanon EY and Eisenberg E: Genome-wide analysis of Alu editability. *Nucleic Acids Res* 42: 6876-6884, 2014.
40. Bates DO, Cui TG, Doughty JM, Winkler M, Sugiono M, Shields JD, Peat D, Gillatt D and Harper SJ: VEGF165b, an inhibitory splice variant of vascular endothelial growth factor, is down-regulated in renal cell carcinoma. *Cancer Res* 62: 4123-4131, 2002.
41. Blaxter M: Genetics. Revealing the dark matter of the genome. *Science* 330: 1758-1759, 2010.
42. Alberts B, Johnson A, Roberts K, Raff M, Lewis J and Bray D: *Molecular Biology of the Cell*. Garland Publishing, Inc., New York, 1983.
43. Jackson C, Browell D, Gautrey H and Tyson-Capper A: Clinical significance of HER-2 splice variants in breast cancer progression and drug resistance. *Int J Cell Biol* 2013: 973584, 2013.
44. Rausch T, Jones DT, Zapatka M, Stütz AM, Zichner T, Weischenfeldt J, Jäger N, Remke M, Shih D, Northcott PA, *et al*: Genome sequencing of pediatric medulloblastoma links catastrophic DNA rearrangements with TP53 mutations. *Cell* 148: 59-71, 2012.
45. Goodier JL and Kazazian HH Jr: Retrotransposons revisited: The restraint and rehabilitation of parasites. *Cell* 135: 23-35, 2008.
46. Häsler J and Strub K: Alu elements as regulators of gene expression. *Nucleic Acids Res* 34: 5491-5497, 2006.
47. Mätlik K, Redik K and Speek M: L1 antisense promoter drives tissue-specific transcription of human genes. *J Biomed Biotechnol* 2006: 71753, 2006.
48. Jurka J, Kohany O, Pavlicek A, Kapitonov VV and Jurka MV: Duplication, coclustering, and selection of human Alu retrotransposons. *Proc Natl Acad Sci USA* 101: 1268-1272, 2004.
49. Yao S, Ireland SJ, Bee A, Beesley C, Forootan SS, Dodson A, Dickinson T, Gerard P, Lian LY, Risk JM, *et al*: Splice variant PRKC- ζ (-PrC) is a novel biomarker of human prostate cancer. *Br J Cancer* 107: 388-399, 2012.
50. Szpakowski S, Sun X, Lage JM, Dyer A, Rubinstein J, Kowalski D, Sasaki C, Costa J and Lizardi PM: Loss of epigenetic silencing in tumors preferentially affects primate-specific retroelements. *Gene* 448: 151-167, 2009.
51. Polak P, Karličić R, Koren A, Thurman R, Sandstrom R, Lawrence MS, Reynolds A, Rynes E, Vlahoviček K, Stamatoyannopoulos JA, *et al*: Cell-of-origin chromatin organization shapes the mutational landscape of cancer. *Nature* 518: 360-364, 2015.
52. Ying J, Li H, Seng TJ, Langford C, Srivastava G, Tsao SW, Putti T, Murray P, Chan AT and Tao Q: Functional epigenetics identifies a protocadherin PCDH10 as a candidate tumor suppressor for nasopharyngeal, esophageal and multiple other carcinomas with frequent methylation. *Oncogene* 25: 1070-1080, 2006.
53. Hynes NE: Tyrosine kinase signalling in breast cancer. *Breast Cancer Res* 2: 154-157, 2000.
54. Vorlová S, Rocco G, Lefave CV, Jodelka FM, Hess K, Hastings ML, Henke E and Cartegni L: Induction of antagonistic soluble decoy receptor tyrosine kinases by intronic polyA activation. *Mol Cell* 43: 927-939, 2011.
55. Holst F, Stahl PR, Ruiz C, Hellwinkel O, Jehan Z, Wendland M, Lebeau A, Terracciano L, Al-Kuraya K, Jänicke F, *et al*: Estrogen receptor alpha (ESR1) gene amplification is frequent in breast cancer. *Nat Genet* 39: 655-660, 2007.
56. Martin AJ, Grant A, Ashfield AM, Palmer CN, Baker L, Quinlan PR, Purdie CA, Thompson AM, Jordan LB and Berg JN: FGFR2 protein expression in breast cancer: Nuclear localisation and correlation with patient genotype. *BMC Res Notes* 4: 72, 2011.
57. Cha JY, Lambert QT, Reuther GW and Der CJ: Involvement of fibroblast growth factor receptor 2 isoform switching in mammary oncogenesis. *Mol Cancer Res* 6: 435-445, 2008.
58. Bates DO and Harper SJ: Therapeutic potential of inhibitory VEGF splice variants. *Future Oncol* 1: 467-473, 2005.