Priority Report

Cancer Research

Detection of Redundant Fusion Transcripts as Biomarkers or Disease-Specific Therapeutic Targets in Breast Cancer

Yan W. Asmann[1], Brian M. Necela[2], Krishna R. Kalari[2], Asif Hossain[1], Tiffany R. Baker[2], Jennifer M. Carr[2], Caroline Davis[2], Julie E. Getz[5], Galen Hostetter[5], Xing Li[1], Sarah A. McLaughlin[3], Derek C. Radisky[2], Gary P. Schroth[6], Heather E. Cunliffe[5], Edith A. Perez[4], and E. Aubrey Thompson[2]

Abstract

Fusion genes and fusion gene products are widely employed as biomarkers and therapeutic targets in hematopoietic cancers, but their applications have yet to be appreciated in solid tumors. Here, we report the use of SnowShoes-FTD, a powerful new analytic pipeline that can identify fusion transcripts and assess their redundancy and tumor subtype-specific distribution in primary tumors. In a study of primary breast tumors, SnowShoes-FTD was used to analyze paired-end mRNA-Seq data from a panel of estrogen receptor (ER)$^+$, HER2$^+$, and triple-negative primary breast tumors, identifying tumor-specific fusion transcripts by comparison with mRNA-Seq data from nontransformed human mammary epithelial cell cultures plus the Illumina Body Map data from normal tissues. We found that every primary breast tumor that was analyzed expressed one or more fusion transcripts. Of the 131 tumor-specific fusion transcripts identified, 86 were "private" (restricted to a single tumor) and 45 were "redundant" (distributed among multiple tumors). Among the redundant fusion transcripts, 7 were unique to ER$^+$ tumors and 8 were unique to triple-negative tumors. In contrast, none of the redundant fusion transcripts were unique to HER2$^+$ tumors. Both private and redundant fusion transcripts were widely expressed in primary breast tumors, with many mapping to genomic loci implicated in breast carcinogenesis and/or risk. Our finding that some fusion transcripts are tumor subtype-specific suggests that these entities may be critical determinants in the etiology of breast cancer subtypes, useful as biomarkers for tumor stratification, or exploitable as cancer-specific therapeutic targets. *Cancer Res; 72(8); 1921–8. ©2012 AACR.*

Introduction

Widely used as both biomarkers and therapeutic targets in hematopoietic malignancies, genomic rearrangements and resultant fusion gene products have only recently begun to be appreciated in solid tumors (1, 2). The development of paired-end RNA sequencing technologies (3) combined with effective analytic pipelines that reduce false discovery rates makes it possible to interrogate the transcriptome of tumor cells and identify fusion transcripts with high confidence. A number of laboratories, including our own, have applied this technology to detection of novel fusion transcripts in established breast cancer cell lines (4–6). Fusion transcripts have been reported in a small number of primary breast tumors. The de Bellis group discovered a novel fusion transcript from a single primary breast tumor (7). Edwin Liu's group analyzed 5 breast tumors using a combination of mate-pair genomic (8) and paired-end RNA sequencing (4). Although no recurrent fusion transcripts were detected in this study, one fusion transcript (RPS6KB→VMPI) was subsequently detected in about 30% of a larger cohort of 70 breast tumors from Asian women. However, the corresponding RPSK6KB→VMPI fusion gene was not detected by Stephens and colleagues (9), who used paired-end genomic sequencing to survey genomic rearrangements in 15 primary breast tumors, of which 5 were further analyzed for fusion transcripts. Whereas these data unambiguously indicate that primary breast tumors express fusion transcripts, the redundancy of such chimeric RNAs in breast cancer remains an open question; and no studies published to date have undertaken to survey the distribution of fusion transcripts in breast tumor subtypes.

We developed a novel analytic pipeline, SnowShoes-FTD, which facilitates identification of high confidence fusion transcripts from paired-end mRNA-Seq data (6). This pipeline has now been applied to detection of fusion transcripts in RNA sequence data from a panel of 8 each estrogen receptor alpha (ESR1) positive (ER$^+$), HER2-enriched (HER2$^+$), and triple negative (TN) primary breast tumors (24 tumors total). In addition, we analyzed mRNA sequence data from 8

Authors' Affiliations: [1]Division of Biomedical Statistics and Bioinformatics, Mayo Clinic Rochester, Rochester, Minnesota; Departments of [2]Cancer Biology, [3]Surgery, and [4]Medicine, Mayo Clinic Florida, Jacksonville, Florida; [5]Translational Genomics Research Institute (TGEN), Phoenix, Arizona; and [6]Illumina, Inc., Hayward, California

Note: Supplementary data for this article are available at Cancer Research Online (http://cancerres.aacrjournals.org/).

Corresponding Author: E. Aubrey Thompson, Department of Cancer Biology, Griffin Cancer Research Bld., Room 214, 4500 San Pablo Road South, Jacksonville, FL 32224. Phone: 904-953-6226; Fax: 904-953-0277; E-mail: thompson.aubrey@mayo.edu

doi: 10.1158/0008-5472.CAN-11-3142

Table 1. Distribution of fusion transcripts among tumors subtypes

Tumor subtype	Private fusions	Range private fusions per tumor	Number of genes in private fusions	Redundant fusions	Range redundant fusions per tumor	Number of genes in redundant fusions	Subtype specific redundant fusions	Fusions with multiple isoforms
All tumors	86	0–12	149	45	1–13	76	—	6
HER2 tumors[a]	17	0–5[b]	34	18	1–9[c]	33	0	1
ER^+ tumors[d]	30	0–9	51	32	2–12	55	7	2
TN tumors	39	2–12	68	32	3–13	53	8	3

NOTE: Tumor subtype-specific incidence was abstracted from Supplementary Table S2. Statistical analyses were conducted by Mann–Whitney.

[a]Total fusions: $HER2^+$ versus ER^+, $P = 0.0055$; $HER2^+$ versus TN, $P = 0.0054$.

[b]Private fusions: $HER2^+$ versus ER^+, $P = 0.1697$; $HER2^+$ versus TN, $P = 0.0179$.

[c]Redundant fusions: $HER2^+$ versus ER^+, $P = 0.0051$; $HER2^+$ versus TN, $P = 0.0122$.

[d]No significant differences between ER^+ and TN for total, private, or redundant fusions.

nontransformed human mammary epithelial cell cultures (HMEC) as well as 16 normal human tissues (including breast) from the Body Map study. Every breast tumor expressed at least 1 tumor-specific fusion transcript. A large number of redundant fusion transcripts were identified; a subset of these seems to be tumor subtype–specific. The identification of redundant, tumor subtype–specific fusion transcripts indicates that the role of such entities may not be restricted to rare breast cancer subtypes (10). Our results suggest that fusion transcripts may be potentially useful as biomarkers to stratify breast cancer subtypes, may mark regions of localized chromosomal instability that are linked to the natural history of ER^+, $HER2^+$, or TN breast cancer, and may ultimately emerge as therapeutic targets in breast cancer.

Materials and Methods

Paired-end RNA-seq analysis

Total RNA was prepared from 8 each fresh-frozen ER^+, ERBB2-enriched ($HER2^+$), and TN breast tumors. All tumors were collected under Institutional Review Board (IRB) protocol 09-001909 "MC083H, Breast Center Blood and Tissue Bank" and analyzed under IRB protocol 954-00 "Utilization of Gene Expression Profiling to Identify Candidate Markers of Breast Cancer Aggressiveness." All tumors were T1/T2,N0,M0. Tumors were macrodissected to remove normal tissue. RNA quality was determined with an Agilent Bioanalyzer (RIN > 7.9 for all samples), and cDNA libraries were prepared and sequenced (50 nt paired-end) on the Illumina GAIIx, as previously described (11) to a depth of 20–50M end pairs per sample (Supplementary Table S1). The quantities of the fusion transcripts were calculated as the number of fusion encompassing reads per million aligned reads. Normal tissue mRNA-Seq data (50 base paired-end, 73–80 million read pairs per sample) from the Body Map 2.0 project were obtained from ArrayExpress (12). Paired-end sequence data from nontransformed human mammary epithelial cells (6) were reanalyzed, as described below.

Identification of fusion transcripts

End pairs were aligned to human genome build 36 using Burrows-Wheeler Aligner (13). The aligned SAM files were sorted according to read IDs using the SAMtools (Sequence Alignment Map tools; ref. 14). The fusion transcripts were identified with SnowShoes-FTD (6) version 2.0, which has higher sensitivity without increasing false discovery rate, compared with version 1.0. The SnowShoes analytic pipeline incorporates a number of analytic tools into a coherent workflow so as to facilitate data processing.

Fusion encompassing versus fusion spanning reads

Fusion encompassing reads (15) contain 50 nucleotides from each end which map to different fusion partners. Fusion spanning reads include one end that maps within one of the 2 fusion partners and a second end that spans the junction between the 2 different fusion partners. Sentinel fusion transcripts are defined as those detected in a single tumor with 3 or more unique, tiling fusion encompassing read pairs plus 2 or more unique, tiling fusion spanning reads. Moreover, alignment of these reads must allow unambiguous assignment of directionality (5′–3′) of the 2 fusion partners. Our initial analysis of fusion transcripts in breast cancer cell lines indicates that sentinel transcripts are predicted with very high accuracy (6). A select subset of sentinel transcripts from the breast tumors was validated, as shown in Supplementary Figure S1.

Private versus redundant fusion transcripts

A private fusion transcript is detected in only 1 tumor sample. All private transcripts, by definition, have sentinel properties. Redundant transcripts are detected in 2 or more tumors. A redundant transcript must exhibit sentinel properties in at least 1 tumor.

Tumor-specific fusion transcripts

Fusion transcripts in breast tumors were filtered to remove all candidates that were also detected in either one of the

Table 2. Multiple fusion transcripts are expressed in breast tumors of different subtypes

Fusion gene directional	S	R	I	Fusion product
HER2+ tumor fusions				
ANP32E→MYST4	Y			CIF
BAT2L2→COL3A1		Y		3′UTR
CALR→ACASA	Y			CTT
CAPN1→ARL2	Y			CIF
CD6B→NEAT1		Y		3′UTR
CD6B→PSAP		Y		3′UTR
CWC25→ROBO2	Y			CTT
DIDO1→REPS1	Y		Y	5′UTR
ELAC1→SMAD4		Y		CTT
EPN1→COL1A1		Y		CTT
FTL→ADD3		Y		3′UTR
GLI3→FAM3B	Y			CTT
GOLPH3L→CTSS	Y			CTT
GPATCH→C8orf48	Y			CIF
H1F0→ACTB		Y		3′UTR
HNRNPH1→VAPA	Y			CIF
JOSD1→RPS19BP1	Y			CTT
KCTD3→TXNDC16	Y			CTT
LGMN→NAP1L1		Y		5′UTR
LOC728606→KCTD1		Y		3′UTR
LOC96610→IGLL5		Y		5′UTR
MALAT1→IGF2		Y		5′UTR
MTF2→ARL3	Y	Y		CTT
OGT→ACTB		Y		3′UTR
OLA1→ORMDL3	Y			CTT
RALGPS2→LAMB3	Y			CTT
RBM6→SLC38A3	Y			5′UTR
RPL19→RPS16		Y		CIF
RPL23→MUCL1		Y		CTT
SPARC→TRPS1		Y	Y	CTT
TEP1→RNASE1	Y			CIF
TFG→GPR12B	Y			CIF
TMSB10→RPS16		Y		CIF
TP53113→ABCA10	Y			CTT
VPS35→DCN		Y		3′UTR
ER+ tumor fusion				
ACTG1→PPP1R12C		Y		CTT
AEBP1→THRA	Y		Y	CTT
APOOL→DCAF8	Y			3′UTR
ASAP1→MALAT1	Y			3′UTR
BAT2L2→COL3A1	Y	Y		3′UTR
CD6B→NEAT1		Y		3′UTR
CD6B→PSAP		Y		3′UTR
COL1A1→BASP1		Y		3′UTR
COL1A1→FMNL3		Y		3′UTR
COL1A1→GORASP2	Y			CTT
COL1A2→LAMP2	Y	Y		3′UTR
COL3A1→COL16A1	Y			CIF

(Continued on the following page)

Table 2. Multiple fusion transcripts are expressed in breast tumors of different subtypes (Cont'd)

Fusion gene directional	S	R	I	Fusion product
COL3A1→ZNF43	Y			3′UTR
CTTN→NCRNAD0201	Y			UNK
CYB5R3→TXNIP		Y		3′UTR
DCLK1→COL3A1		Y		UNK
DNAJA2→COL14A1	Y			3′UTR
DNM2→P1N1		Y		CTT
EIF4G1→ABCC5	Y			CTT
ELAC1→SMAD4		Y		CTT
ELF3→SLC39A6	Y	Y		3′UTR
EPN1→COL1A1		Y		CTT
FLNA→ABCA2		Y		CTT
FTL→ADD3	Y	Y		3′UTR
H1F0→ACTB		Y		3′UTR
HEATR5A→COL1A1	Y			3′UTR
HMGN3→PAQR8	Y			5′UTR
HSP90AB1→PCGF2	Y			CIF
MALAT1→IGF2		Y		5′UTR
IGFBP5→AMD1	Y			3′UTR
LOC728606→KCTD1		Y		3′UTR
LOC96610→IGLL5		Y		5′UTR
MAF→IGFBP7	Y			3′UTR
MAPK1IP1L→XP01	Y			3′UTR
MGP→NCRNA00188	Y	Y		3′UTR
MGP→REPS2	Y			3′UTR
MRPL52→USP22	Y			3′UTR
NDUFS6→ACTB	Y	Y		CTT
PLXNA1→CTSD		Y		3′UTR
POLD3→COL3A1	Y			3′UTR
POSTN→TM9SF3	Y			3′UTR
POSTN→TRIM33	Y			3′UTR
PTP4A2→MALAT1	Y	Y		UNK
RAB3IP→IGFBP5	Y			3′UTR
RAB8A→EIF4G2	Y	Y		3′UTR
RHOB→GATA3		Y		3′UTR
RHOBTB3→CRNKL1	Y			3′UTR
RPL23→MUCL1		Y		CTT
SERPINA1→K1AA1217	Y			3′UTR
SFI1→YPEL1	Y			CTT
SLC39A6→LRIG1	Y			CTT
SPATS2L→COL3A1		Y		3′UTR
STC2→RNF11	Y	Y		CTT
TAX1BP1→MALAT1	Y	Y		3′UTR
TES→HNRNPU	Y			3′UTR
THSD4→PAQR5	Y			CIF
TMEM119→ARIH2	Y			3′UTR
TTC7A→SOCS5		Y		3′UTR
UBR2→SRPK1		Y		CIF
VPS35→DCN	Y	Y		3′UTR
YWHAG→CYB561	Y			3′UTR

(Continued on the following page)

Table 2. Multiple fusion transcripts are expressed in breast tumors of different subtypes (Cont'd)

Fusion gene directional	S	R	I	Fusion product
TN tumor fusions				
AATK→USP32		Y		CIF
ACTB→C20orf112		Y	Y	3′UTR
ACTG1→PPP1R12C	Y	Y		CTT
ADCY9→C16orf5	Y			CTT
APOL1→ACTB		Y		3′UTR
BAT2L2→COL3A1		Y		3′UTR
C2orf56→SAMD4B	Y			3′UTR
CD68→PSAP		Y		3′UTR
CD74→MBD6	Y	Y		3′UTR
CDK4→UBA1	Y			3′UTR
CIRBP→UGP2	Y			3′UTR
COL1A1→BASP1		Y		3′UTR
COL1A1→FGD2	Y			CTT
COL1A2→LAMP2		Y		3′UTR
CTSD→PRKAR1B	Y			CIF
DNM2→PIN1	Y	Y		CTT
EPHA2→CTSD	Y			3′UTR
EPN1→COL1A1		Y		CTT
FLNA→ABCA2	Y	Y		CTT
GAPDH→KRT13	Y			3′UTR
GAPDH→MRPS18B	Y			3′UTR
GEMIN7→SLC39A14	Y		Y	5′UTR
GNB1→TRH	Y			3′UTR
GNB2→CTSD	Y	Y		3′UTR
GPAA1→CD24	Y		Y	CTT
H1F0→ACTB		Y		3′UTR
IFI27→CPNE3	Y			3′UTR
ITGA3→KHK	Y			3′UTR
ITGAV→ANKHD1	Y			CIF
KRT18→PLEC	Y	Y		CIF
KRT81→EMP2	Y			3′UTR
LGMN→NAP1L1	Y	Y		5′UTR
LOC72B606→KCTD1	Y	Y		3′UTR
LOC96610→IGLL5	Y	Y		5′UTR
LTBP4→CTSD	Y			3′UTR
MTF2→ARL3	Y	Y		CTT
NAV2→WDFY1	Y			3′UTR
NCOR2→ELN	Y			CTT
NDUFS6→ACTB		Y		CTT
NPLOC4→PDE6G	Y			CTT
NTN1→HDLBP	Y			3′UTR
OGT→ACTB	Y	Y		3′UTR
PACSIN3→CTSD	Y			3′UTR
PCNX→MKKS	Y			CTT
PDHX→CAT	Y			CIF
PIKFYVE→TMEM119	Y			3′UTR
PLEC→PLEKHM2	Y			CIT
PLXNA1→CTSD	Y	Y		3′UTR
PROM1→TAPT1	Y			CIF

(Continued on the following page)

Table 2. Multiple fusion transcripts are expressed in breast tumors of different subtypes (Cont'd)

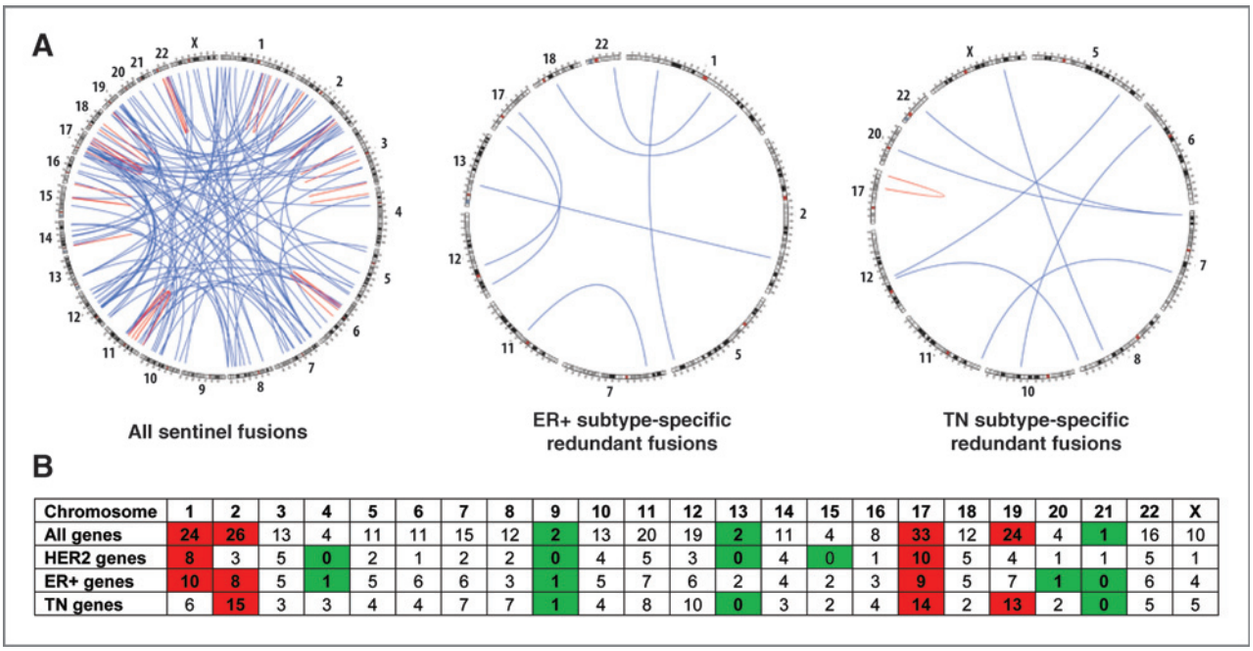
Fusion gene directional	S	R	I	Fusion product
PTM4→GNB4	Y			3′UTR
PTP4A2→MALAT1		Y		UNK
RAB8A→EIF4G2		Y		3′UTR
RHOB→GATA3		Y		3′UTR
RPL19→RPS16	Y	Y		CIF
RPL23→MUCL1		Y		CTT
RPL8→KRT4	Y			CIF
RPS15→PLEC	Y			CIF
SBF1→FLNA	Y			3′UTR
SEMA4C→PKM2	Y			CIF
SFTPC→IGLL5	Y			CTT
SLC16A3→MRPL4	Y			CTT
SLC34A2→ACTB	Y			CIF
SPATS2L→COL3A1		Y		3′UTR
TMEM109→CTSD	Y			3′UTR
TMSB10→RPS16	Y	Y		CIF
TNRC18→SLC9A3R1	Y			CIF
TSPAN14→HLA-E	Y	Y		3′UTR
TTC7A→SOCS5	Y	Y		CTT
USF2→IRX3	Y			CIF
YWHAG→PDIA3	Y			3′UTR
YWHAZ→ZBTB33	Y	Y		3′UTR

NOTE: Subtype-specific fusion transcripts are identified in bold blue font. All fusion transcripts are given according to orientation 5 fusion partner → 3′ fusion partner. Transcripts are further identified according to sentinel status in each tumor subtype (S), redundancy in each subtype (R), and fusion transcript isoforms detection in each subtype (I). Fusion products are identified as follows: 3′UTR, fusion that changes 3′UTR of 5′ fusion partner; 5′UTR, fusion in 5′UTR of 5′ fusion partner; CIF, coding inframe fusion to produce a chimeric protein; CTT, C-terminal truncation of 5′ fusion partner resulting from frameshift.

control data sets: the HMEC or Body Map data. This approach was based on the assumption that such candidates represent either annotation or alignment errors or arise from germ line rearrangement polymorphisms (8).

Results and Discussion

We detected 131 sentinel fusion transcripts in 24 tumors (Supplementary Table S2). The majority of the fusion transcripts arose from interchromosomal fusions (104 of 131). We had previously identified fusion transcripts that were expressed in multiple isoforms in breast cancer cell lines (6). Six fusion transcripts were expressed as multiple isoforms in tumors (yellow highlight Supplementary Table S2.) The majority of the fusion transcripts were "private," expressed in only one tumor sample. However, 45 sentinel transcripts were



Chromosome	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	X
All genes	24	26	13	4	11	11	15	12	2	13	20	19	2	11	4	8	33	12	24	4	1	16	10
HER2 genes	8	3	5	0	2	1	2	2	0	4	5	3	0	4	0	1	10	5	4	1	1	5	1
ER+ genes	10	8	5	1	5	6	6	3	1	5	7	6	2	4	2	3	9	5	7	1	0	6	4
TN genes	6	15	3	3	4	4	7	7	1	4	8	10	0	3	2	4	14	2	13	2	0	5	5

Figure 1. Chromosomal distribution of fusion transcripts and fusion partner genes is nonrandom. Connection between the chromosomal loci of fusion transcripts in shown in A for all sentinel fusions as well as for tumor subtype–specific fusion transcripts. The chromosomal "heat map" (B) shows the top 4 (red) and bottom 4 (green) chromosomes, identified by the genomic coordinates of fusion partner genes.

redundant, as evidenced by detection in 2 or more tumors. (highlighted pink and blue Supplementary Table S2.)

Tumor subtype distribution of fusion transcripts

Every tumor expressed at least one redundant fusion transcript, with a range of 1 to 13 redundant transcripts per tumor (Table 1). Among the redundant transcripts, 7 were uniquely expressed in ER^+ tumors and 8 in TN tumors (bold blue in Table 2), but no redundant transcript was exclusively expressed in $HER2^+$ tumors. Private transcripts were detected at a range of 0 to 12 per tumor (Table 1). ER^+ and TN tumors expressed similar numbers of fusion transcripts per tumor, whereas $HER2^+$ tumors expressed significantly fewer fusions per tumor (Table 1). Depth of sequence analysis is likely to contribute to fusion transcript detection, and significantly greater depth of sequencing was obtained with the ER^+ tumors (~40–50M) than with the TN (~20–28M) or $HER2^+$ (17–20M) tumors. However, we did not observe a significance difference in fusion transcripts per tumor between ER^+ and TN tumors, which differ in depth of sequence; whereas the TN and $HER2^+$ tumors, which were sequenced to similar depths, were significantly different in this respect. We conclude on this basis that the differences in depth of sequence among our tumors do not have a major impact on detection of the number of fusion transcripts per tumor.

Although $HER2^+$ tumors generally have fewer fusion transcripts, we noted that a few $HER2^+$ tumors expressed levels of fusions that were comparable with those observed in ER^+ or TN tumors. (see $HER2^+$ tumors_29 in (Supplementary Table S2.) It will be of interest to determine whether expression of large numbers of fusion transcripts in a subset of $HER2^+$ tumors has therapeutic implications. We conclude that fusion transcripts represent a heretofore underappreciated class of genomic features that may have considerable potential as biomarkers or therapeutic targets in breast cancer, and it will be interesting to determine whether the number or characteristics of fusion transcripts correlates with clinical outcome in $HER2^+$ or other tumor subtypes.

Chromosomal distribution of fusion transcript partners

The chromosomal mapping distribution of the sentinel fusions is clearly nonrandom (Fig. 1A). A disproportionately large number of fusion transcript partners are located on chromosomes 1, 2, 17, and 19 (Fig. 1B), whereas relatively few fusion transcript partners are located on chromosomes 4, 9, 13, 15, 20, and 21. It is difficult, because of the relatively small numbers, to make any rigorous conclusions with respect to tumor subtype–specific distribution of fusion transcripts. However, chromosome 19 seems to be a "hot spot" for TN tumors. Circos plots of ER^+-specific and TN-specific redundant fusion gene partners (Fig. 1A) indicate that there is a subtype-specific fusion transcript geography, suggesting a functional link between breast tumor subtype and formation of fusion transcripts. The observation that $HER2^+$ tumors, as a group, express significantly fewer fusion transcripts is consistent with this hypothesis.

A number of distinct clusters emerged when the fusion partner genes were mapped to genomic loci (Fig. 2). Two major clusters were observed on chromosome 17, mapping to 17q21–q23, and 17q25. Both of these regions are well known to undergo copy number variation in breast cancer. All of the

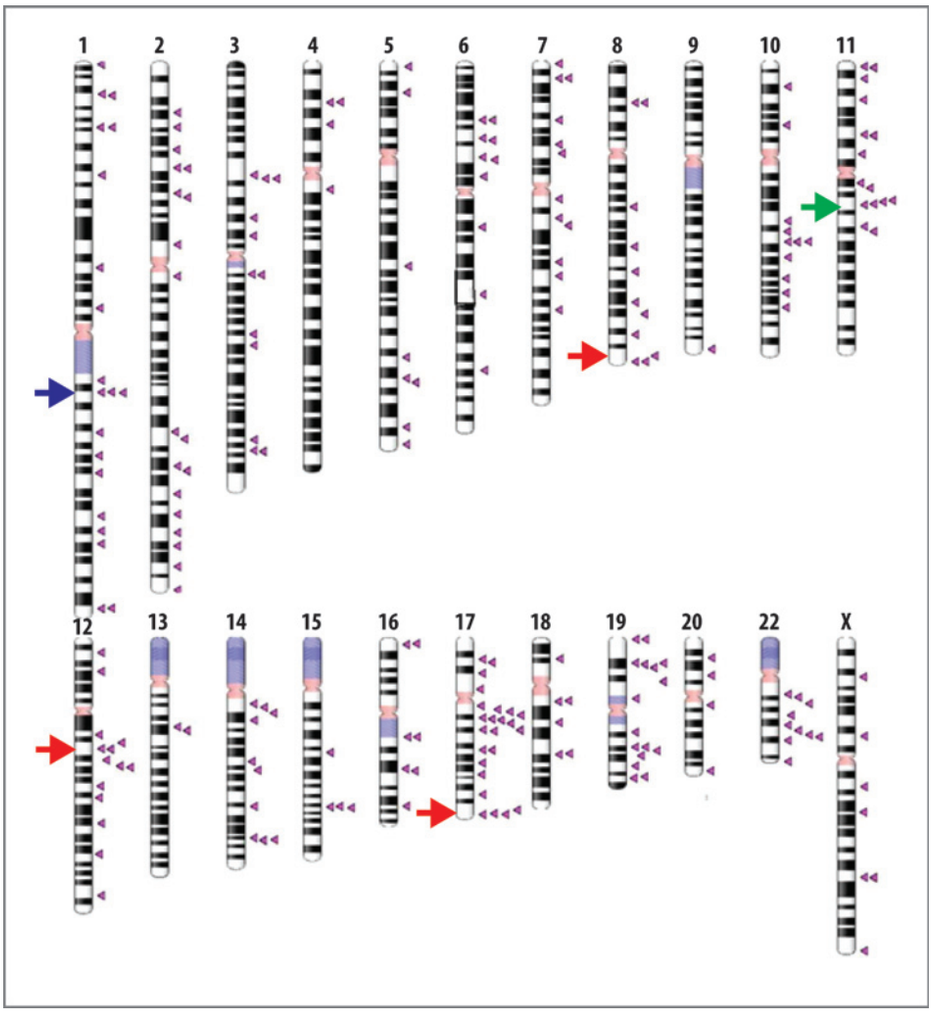


Figure 2. Chromosomal mapping of fusion partner genes reveals tumor suptype–specific clusters. Chromosomal mapping was carried out using PheGen (National Center for Biotechnology Information) to assign chromosomal coordinates of all fusion gene partners. Clusters that are uniquely associated with $HER2^+$ tumors are designated by the blue arrow (Chr1q21.22–21.3), whereas a green arrow designates a large ER^+ cluster at chr11q13.1–q13.3, and red arrows identify TN clusters at chr8q24.3, chr12q13.13, and chr17q25.1–25.3. Pink areas denote centromeric regions, whereas blue, gray, and black areas denote cytobands. Note that chr21, which contains only a single fusion gene partner, was omitted from this representation.

chromosome 19 fusion partners in TN tumors mapped to clusters located in the vicinity of 19p13 or 19q13. One large cluster of genes at 11q13.1–q13.4 was restricted to ER^+ tumors (green arrow in Fig. 2), a small cluster of genes at 1q21.2–q21.3 was restricted to $HER2^+$ tumors (blue arrow in Fig. 2), and genes that clustered at 8q24.3, 12q13.13, and 17q25.1–q25.3 were restricted to TN tumors (red arrows in Fig. 2).

Limited data from genomic analysis of both breast cancer cell lines (5) and tumors (4, 9) indicate that genomic rearrangement is the primary mechanism whereby most fusion transcripts are generated. Furthermore, review of the array comparative genomic hybridization (aCGH) data on breast cancer reveals that many of the fusion partners that we have identified map to regions that are known to undergo copy number gain or loss in breast tumors. This correlation is most obvious when one considers chromosome 17, which contains 33 genes that contribute to fusion transcripts. Among these genes, 6 map to a cluster at 17q12, 5 to 17q21, and 6 to 17q25. All 3 of these loci are known to undergo copy number variation in breast cancer (9, 16–18). The distribution of fusion partners on chromosome 19 is even more striking. All of the genes map to either 19p12–p13 or 19q13. Both aCGH and genome wide association data indicate that these 2 regions are important in breast cancer, particularly the TN subtype (19, 20). Based on these considerations, we posit that most of the fusion transcripts arise due to chromosomal rearrangements and therefore mark areas of local chromosomal instability.

Structure and potential functional significance of predicted fusion transcript products

SnowShoes_FTD assembles the predicted nucleotide sequences of the candidate fusion transcripts and translates that sequence into the predicted amino acid sequences of the putative fusion proteins (Supplementary Table S3). We and others have shown that fusion transcripts in breast cancer cell lines fall into several broad categories based on the location within the transcription unit wherein the fusion occurs (5, 6). A small number of fusions occur in 5′ untranslated region (UTR) regions (Table 2), placing the coding sequence of the 3′ fusion partner under the control of the promoter from the 5′ fusion partner. We have reported that a "promoter swap" event of this sort is associated with ERBB2 overexpression in a breast cancer cell line derived from a $HER2^+$ tumor (6).

The most common class of fusion transcripts in cell lines occur within 3′UTR). A similar distribution prevailed in primary breast tumors (Table 2). Such fusions result in the generation of full-length coding sequences of the 5′ fusion partner, but alter the 3′ UTR sequence of such transcripts, with potential effects on stability and/or translational efficiency of the fusion transcript (21).

The second broad class of chimeric transcripts involves fusion within the coding regions. Some of these transcripts contain precise exon/exon junctions (column H Supplementary Table S2) and are assumed to be processed. However, our data do not discriminate between tumor-specific trans-splicing events and processing of a primary transcript that arises due to genomic rearrangement. The fusion junctions of many chimeric transcripts do not correspond to known exon/exon boundaries. These may arise due to trans-splicing at cryptic sites or, more likely, represent novel exonic sequences derived from transcription of rearranged genes.

Coding sequence fusions fall into 2 classes. We identified 25 fusion transcripts that are predicted to give rise to chimeric proteins, many of which contain functional domains from both fusion partners and might therefore be expected to have novel properties [coding inframe fusion (CIF) in Table 2]. The deduced sequence and functional domains of all predicted fusion products is given in Supplementary Table S4. By way of example, the TFG→GPR128 fusion transcript is predicted to encode a novel 848 amino acid protein in which the PB1 protein–protein interaction domain of TFG (also known as the TRKT3 oncogene) is fused to the 7 transmembrane spanning domain of GPR128, with loss of the serine/threonine-rich N-terminal domain that is characteristic of this subclass of G-protein–coupled receptors. The potential regulatory effects of such a chimeric protein might be considerable, and the fact that these hypothetical signaling changes might devolve from a G-protein–coupled receptor makes this a potentially druggable target.

About half of the coding to coding fusions were predicted to result in frame shifts and carboxy-terminal truncation of the 5′ fusion partner (CTT in Table 2). To the extent to which such transcripts escape nonsense-mediated degradation mechanisms, they would be predicted to encode N-terminal polypeptides that are deleted of C-terminal functional domains. For example, the ADCY9→C16orf5 fusion transcript is predicted to encode a polypeptide of 585 amino acids that includes the N-terminal nucleotide binding domain of adenylylate cyclase 9, but is deleted of the C-terminal nucleotide cyclase domain and therefore unlikely to have catalytic activity. However, the N-terminal fragment contains the intact dimerization domain of ADCY9 and might therefore function as a dominant-negative inhibitor.

The functional significance of fusion transcripts in breast cancer is an area of very active investigation. The objective is to identify tumor specific, druggable fusion transcripts that are required for establishment and maintenance of the transformed phenotype (e.g., BCR-ABL1). The existence of driver fusion transcripts of this sort remains to be established in breast cancer. However, the potential significance of fusion transcripts as biomarkers does not depend upon such entities serving as driver mutations; and we are confident that a more comprehensive analysis of fusion transcripts in breast cancer will provide a set of novel biomarkers that will be useful for stratification of tumor subtypes, prediction, and prognosis. Ultimately, we anticipate that some of these fusion transcripts will emerge as therapeutic targets for treatment of the more challenging breast tumor subtypes, including $HER2^+$ tumors with both *de novo* and acquired resistance to targeted therapy, hormone-resistant ER^+ tumors, and TN tumors.

Disclosure of Potential Conflicts of Interest

All authors confirm that the information reported above is accurate and understand that this information will be disclosed publicly. The AACR reserves the right to decline to publish their work if the Association believes a serious conflict of interest exists and they also understand that failure to complete this form will disqualify their manuscript from consideration for publication. No potential conflicts of interest were disclosed.

Acknowledgments

The authors thank the expert assistance of Mr. Bruce W. Eckloff (Mayo Clinic Advanced Genomics Technology Center), who sequenced the tumor libraries, and Ms. Shujun Luo (Illumina, Inc.), who sequenced the HMEC libraries.

Grant Support

This work was supported in part by grants from the State of Florida Bankhead-Coley program (1BG12) and the Breast Cancer Research Foundation (BCRF 21J). Sequencing of HMEC libraries was supported in part by CA155129. Additional resources were provided by the 26.2 with Donna Foundation (National Marathon to Fight Breast Cancer), the Carmichael Family Foundation, and the Mayo Foundation.

Received September 26, 2011; revised January 4, 2012; accepted February 16, 2012; published online April 16, 2012.

References

1. Soda M, Choi YL, Enomoto M, Takada S, Yamashita Y, Ishikawa S, et al. Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer. Nature 2007;448:561–6.
2. Kumar-Sinha C, Tomlins SA, Chinnaiyan AM. Recurrent gene fusions in prostate cancer. Nat Rev Cancer 2008;8:497–511.
3. Maher CA, Palanisamy N, Brenner JC, Cao X, Kalyana-Sundaram S, Luo S, et al. Chimeric transcript discovery by paired-end transcriptome sequencing. Proc Natl Acad Sci U S A 2009;106:12353–8.
4. Inaki K, Hillmer AM, Ukil L, Yao F, Woo XY, Vardy LA, et al. Transcriptional consequences of genomic structural aberrations in breast cancer. Genome Res 2011;21:676–87.
5. Edgren H, Murumagi A, Kangaspeska S, Nicorici D, Hongisto V, Kleivi K, et al. Identification of fusion genes in breast cancer by paired-end RNA-sequencing. Genome Biol 2011;12:R6.
6. Asmann YW, Hossain A, Necela BM, Middha S, Kalari KR, Sun Z, et al. A novel bioinformatics pipeline for identification and characterization of fusion transcripts in breast cancer and normal cell lines. Nucleic Acids Res 2011;39:e100.
7. Guffanti A, Iacono M, Pelucchi P, Kim N, Solda G, Croft LJ, et al. A transcriptional sketch of a primary human breast cancer by 454 deep sequencing. BMC Genomics 2009;10:163.
8. Hillmer AM, Yao F, Inaki K, Lee WH, Ariyaratne PN, Teo AS, et al. Comprehensive long-span paired-end-tag mapping reveals

characteristic patterns of structural variations in epithelial cancer genomes. Genome Res 2011;21:665–75.
9. Stephens PJ, McBride DJ, Lin ML, Varela I, Pleasance ED, Simpson JT, et al. Complex landscapes of somatic rearrangement in human breast cancer genomes. Nature 2009;462:1005–10.
10. Tognon C, Knezevich SR, Huntsman D, Roskelley CD, Melnyk N, Mathers JA, et al. Expression of the ETV6-NTRK3 gene fusion as a primary event in human secretory breast carcinoma. Cancer Cell 2002;2:367–76.
11. Sun Z, Asmann YW, Kalari KR, Bot B, Eckel-Passow JE, Baker TR, et al. Integrated analysis of gene expression, CpG island methylation, and gene copy number in breast cancer cells by deep sequencing. PLoS ONE 2011;6:e17490.
12. ArrayExpress. Body Map 2.0, query ID: E-MTAB-513. [cited]. Available from: http://www.ebi.ac.uk/arrayexpress.
13. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009;25:1754–60.
14. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics 2009;25:2078–9.
15. Maher CA, Kumar-Sinha C, Cao X, Kalyana-Sundaram S, Han B, Jing X, et al. Transcriptome sequencing to detect gene fusions in cancer. Nature 2009;458:97–101.
16. Adelaide J, Finetti P, Bekhouche I, Repellini L, Geneix J, Sircoulomb F, et al. Integrated profiling of basal and luminal breast cancers. Cancer Res 2007;67:11565–75.
17. Andre F, Job B, Dessen P, Tordai A, Michiels S, Liedtke C, et al. Molecular characterization of breast cancer with high-resolution oligonucleotide comparative genomic hybridization array. Clin Cancer Res 2009;15:441–51.
18. Bae JS, Choi JS, Baik SH, Park WC, Song BJ, Kim JS, et al. Genomic alterations of primary tumor and blood in invasive ductal carcinoma of breast. World J Surg Oncol 2010;8:32.
19. Antoniou AC, Wang X, Fredericksen ZS, McGuffog L, Tarrell R, Sinilnikova OM, et al. A locus on 19p13 modifies risk of breast cancer in BRCA1 mutation carriers and is associated with hormone receptor-negative breast cancer in the general population. Nat Genet 2010;42:885–92.
20. Yang TL, Su YR, Huang CS, Yu JC, Lo YL, Wu PE, et al. High-resolution 19p13.2-13.3 allelotyping of breast carcinomas demonstrates frequent loss of heterozygosity. Genes Chromosomes Cancer 2004;41:250–6.
21. Mayr C, Hemann MT, Bartel DP. Disrupting the pairing between let-7 and Hmga2 enhances oncogenic transformation. Science 2007;315:1576–9.