Abstract

Purpose: Although fusion genes serve as an effective target in hematologic malignancies, recurrent gene fusion events are relatively rare in solid tumors. To detect recurrent novel fusion transcripts we performed whole transcriptome sequencing primary breast cancer samples.

Experimental Design: Whole-transcriptome sequencing of 120 fresh-frozen primary breast cancer samples and five adjacent normal breast tissues using the Illumina HiSeq2000 platform was performed. Three different fusion-detecting tools (deFuse, Chimerascan, and TopHat) were used, and the results were compared.

Results: These tools detected 3831, 6630 and 516 fusion transcripts (FTs) overall. We primarily focused on the results obtained using the deFuse software. More FTs were identified from HER2 subtype breast cancer samples than from the luminal or triple-negative subtypes ($p < 0.05$). Seventy fusion candidates were selected for validation, and 32 (45.7%) were confirmed by RT-PCR and Sanger sequencing. Of the validated fusions, six were recurrent (found in 2 or more samples), three were in-frame (PRDX1-AKR1A1, TACSTD2-OMA1 and C2CD2-TFF1) and three were off-frame (CEACAM7-CEACAM6, CYP4X1-CYP4Z2P, and EEF1DP3-FRY). Notably, the novel read-through fusion, EEF1DP3-FRY, was identified and validated in 6.7% (8/120) of the breast cancer samples. This off-frame fusion results in early truncation of the FRY gene, which plays a key role in the structural integrity

during mitosis. Three previously reported fusions, PPP1R1B-STARD3, MFGE8-HAPL, and ETV6-NTRK3, were detected in 8.3%, 3.3% and 0.8% of the 120 samples, respectively, by both deFuse and Chimerascan. The recently reported MAGI3-AKT3 fusion was not detected in our analysis.

Conclusion: Among the 3800s fusions detected by mRNA Sequencing, 32 fusions were validated using Sanger sequencing, including the novel recurrent read-through fusion, EEF1DP3-FRY. Previously reported FTs, such as PPP1R1B-STARD3, MFGE8-HAPLN3 and ETV6-NTRK3, were also identified. Future work is warranted to elucidate the biological significance of these fusions.

Key words: Fusion gene, Fusion transcript, Whole transcriptome sequencing, RNA-Seq, Breast cancer

Student Number: 2012-30493

Contents

List of Tables

List of Figures

Introduction

A fusion gene is a hybrid formed from two distinct genes that undergo chromosomal rearrangement. Gene fusion events are observed at a significantly higher rate in cancer samples compared to benign samples, and are thought to play critical carcinogenic roles via various mechanisms such as oncogene activation, tumor suppressor deletion/downregulation, and the creation of novel proteins capable of altering cellular pathways.(Kannan et al., 2011) The BCR-ABL1 fusion gene, which is a prototypic fusion oncogene associated with leukemia, has proven to be an efficient therapeutic target for tyrosine kinase inhibitors which yield significant survival gains in leukemia patients (Rowley, 1973; de Klein et al., 1982). The identification of recurrent fusion genes in lung adenocarcinoma and prostate cancer suggest that gene fusion events may also be biologically relevant in solid tumors (Tomlins et al., 2005, 2007; Soda et al., 2007; Ju et al., 2012). For example, 79% of prostate tumors harbor ETS family translocations, and up to 8.6% of non-small cell lung cancers were reportedly positive for the EML4-ALK fusion gene (Kangaspeska et al., 2012; Scagliotti et al., 2012). Thus, the identification of gene fusions in cancer has greatly changed our clinical practice in terms of cancer diagnosis, therapy, and prognosis.

Breast cancer is a heterogeneous disease that has been associated with various genetic alterations. New advances in sequencing technology have enabled the

increasing identification of novel fusion genes, including the ETV6-NTRK3 fusion in secretory breast ductal carcinoma and the MYB-NFIB fusion in adenoid cystic carcinoma of breast. Both are recurrent fusions, but they are found only in very uncommon histologic subtypes of breast cancer (Knezevich et al., 1998; Stenman et al., 2010). Banerji et al. recently reported the MAGI3-AKT3 fusion oncogene in almost 7% of triple-negative breast cancer, and further indicated that this oncogene could be therapeutically targetable.(Banerji et al., 2012) The identification of additional (hopefully more common) biologically functional fusion genes is critical, as these could potentially serve as biomarkers and/or therapeutic targets.

In an effort to identify novel recurrent fusion transcripts in breast cancer, we performed whole-transcriptome sequencing (RNA-seq) on 120 primary breast cancer samples (Table 1) and five matched normal samples. PCR and Sanger sequencing was used to confirm the fusion junction sequences of selected candidates. An overview of our study design is presented in Figure 1.

Materials and methods

1. Study Population

Patients diagnosed with pathologically confirmed primary invasive carcinoma of the breast and with available clinical and pathological data were selected for this study. Patients who underwent neoadjuvant systemic therapy or those with stage lV disease at diagnosis were excluded. All patients underwent appropriate adjuvant treatment after curative resection according to institutional guidelines. All patients were followed for more than 5 years or until the first systemic recurrence. Immunohistochemical assessments of estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2) were performed by an experienced pathologist, using antibodies against ER (1DO5; DAKO, Denmark), PR (PgR626, DAKO), and HER2 (CB11, diluted 1:200; Novocastra Laboratories, Newcastle-Upon-Tyne, UK).

2. Primary Tumor and Normal Mammary Tissue Samples

One-hundred-and-twenty primary tumor samples were obtained from surgical specimens immediately after the resection of tumors in the operating room. Tumor blocks (5-7 mm3) were harvested grossly from the tumor cores and/or normal

mammary tissues at a distance from the primary tumor, and stored in nitrogen tanks at a tissue bank (Laboratory of Breast Cancer Biology, Seoul National University College of Medicine, Seoul Korea). Informed consent was obtained prior to sampling, and the study was approved by the Institutional Review Board of Seoul National University Hospital (IRB No 1109-007-376).

3. Paired-End Massively Parallel RNA Sequencing

Total RNA was obtained from archived tumor blocks and from five normal samples. 1-mg aliquots were used to construct cDNA libraries with the TruSeq RNA kit. The kit protocol included polyA-selected RNA extraction, RNA fragmentation, random-hexamer-primed reverse transcription, and 101 nucleotide paired-end sequencing, which was performed on an Illumina HiSeq2000. The obtained libraries were quantified using qPCR, which was performed according to the qPCR Quantification Protocol Guide with the results quantified using an Agilent Technologies 2100 Bioanalyzer. To estimate the relevant expression levels and identify alternatively spliced transcripts, the RNA-Seq reads were mapped to the human genome using TopHat (version 1.3.3), which is capable of reporting split-read alignments across splice junctions, and determined using Cufflinks (version 1.2.1) under the default options.(Trapnell et al., 2009; Trapnell et al., 2010) The reference genome sequence (hg19, Genome Reference Consortium Homo_Sapiens. GRCh37.62) and annotation

data were downloaded from the UCSC website (http://genome.uscs.edu). The transcript counts were calculated at the isoform and gene levels, and the relative transcript abundances were measured in FPKM (fragments per kilobase of exon per million fragments mapped) using Cufflinks. To identify potential gene fusions from the RNA-seq data, we used the deFuse program (version 0.4.3).(McPherson et al., 2011) Additionally, TopHatFusion (version 2.0.9) and Chimerascan (version 0.4.5) fusion detecting algorithms were applied. The filtering parameters are as described in the following; clustering_precision = 0.95, span_count_threshold = 5, split_count_threshold = 3, percent_identity_threshold = 0.90, max_dist_pos = 600, num_dist_genes = 500, split_min_anchor = 4, max_concordant_ratio = 0.1, splice_bias = 10, denovo_assembly = no, probability_threshold = 0.50..

4. Validation by Reverse-Transcription PCR (RT-PCR) and Sanger Sequencing

RT-PCR and Sanger sequencing were used to validate the 70 selected FTs in the index cases. For RT-PCR, an oligo (dT) primer and a PrimeScript RT-PCR kit (TAKARA, Kyoto) were used to synthesize cDNA from 2 μg of total RNA, according to the manufacturer's protocol. PCR was performed with forward and reverse primers designed to encompass the identified fusion junctions

(Supplementary Table 1) plus TAKARA Ex-Taq Polymerase (Takara, Kyoto). The amplification conditions were as follows: 94°C for 5 min; followed by 35 cycles of 94°C for 30 sec, the appropriate annealing temperature for 30 sec, and 72°C for 40 sec, and then a final soak at 72°C for 7 min. The obtained PCR products were purified using EXOSAP-IT (USB, Ohio, USA) and sequenced using a BigDye Terminator v3.1 sequencing kit and a 3730xl automated sequencer (both from Applied Biosystems, Foster City, CA).

Results

1. Identification of Fusion Transcripts by the deFuse Algorithm

We performed high-throughput whole-transcriptome sequencing using the Illumina HiSeq2000 platform on 120 primary breast cancer samples and five matched normal samples. RNAs were extracted from fresh-frozen surgical specimens harvested immediately after surgical resection. Of the tumors, 49.2% (59/120) were luminal subtype (hormone-receptor-positive), 10.8% (13/120) were HER2 subtype (HER2-amplified and hormone-receptor-negative), and 30% were triple-negative subtype. At the 5-year follow-up, 30% (84/120) of the patients showed distant metastases (Table 1).

To detect fusion transcripts (FTs), we initially applied the deFuse algorithm. This method yielded a total of 3831 FTs.(McPherson et al., 2011) Of them, 29.5% (1130/3831) were inter-chromosomal FTs (occurring between two genes located on different chromosomes) while 70.5% (2701/3831) were intra-chromosomal FTs (Figure 2a). Among the 2701 intra-chromosomal FTs, 51.9% (1401/2701) were adjacent fusions (i.e., fusion of two adjacently located genes on the genome without other genes in between) (Figure 2a). Overall, 34.2% of the fusions (1310/3831) had appropriate donor (5')-acceptor (3') relationships. FTs with an inappropriate donor-acceptor relationship (donor-donor or acceptor-acceptor) are theoretically unable to

undergo translation (Figure 2b). In terms of distribution, 62.6% of the potential FTs (2399/3831) were expressed in two or more samples (recurrent FTs), while 38.4% (1432/3831) were identified in a single sample (private FTs) (Figure 2c). Fewer than twenty of the 3831 FTs were detected in the RNA-Seq results obtained from the five normal samples.

HER2-subtype tumors accounted for more of the total FTs compared to the luminal and triple-negative subtypes (Figure 3a). On an individual level, the HER2-positive tumors harbored significantly more FTs than the HER2-negative tumors (mean number of FTs per tumor, 24.20 ± 13.38 vs. 16.65 ± 14.35; $p = 0.036$) (Figure 3b). The number of FTs did not differ according to the age at diagnosis or distant-metastasis-free survival status. HER2 type tumors showed higher number of fusion transcripts which were both detected by deFUSE and Chimerascan.

2. Fusion Junction Validation by RT-PCR And Sanger Sequencing

To confirm that we had identified true fusion events rather than false positives arising from errors in our RNA-seq or fusion-detecting strategies, we performed Sanger sequencing of high-priority candidates using cDNA extracted from the index tumor samples. From among the 3831 identified FTs, high-priority candidate FTs

were selected using the following criteria: 1) the presence of an appropriate donor (5')-acceptor (3') relationship; 2) recurrent expression in multiple samples; 3) high probability scores derived from the deFuse algorithms; 4) the involvement of a known tumor-associated gene, oncogene or tumor suppressor gene, or a gene with a known targeting drug. The forward and reverse primers were designed to encompass the identified fusion junctions (Table 2), and the obtained Sanger sequencing data were compared with the sequences obtained from RNA-Seq. Seventy FTs were nominated for validation. Among them, 32 (45.7%) FTs were successfully validated by RT-PCR and Sanger sequencing (Table 3, Figure 4). There were no significant characteristics of the successfully validated fusions. When we subsequently applied two different fusion detecting algorithms, the validation rate were different and whether the fusion was detected by single or multiple methods, didn't affect the validation rate. Likewise, the validated fusions were not identified by all fusion detecting algorithms. Among the 32 validated fusions, 15 were identified by Chimerascan and 4 by TopHatFusion. The validation rate differed between the three algorithms. The well-known fusion ETV6-NTKR3 was found by the deFuse and Chimerascan algorithms. Also, deFuse and Chimerascan both detected the previously reported fusions PPP1R1B-STARD3 and MFGE8-HAPLN3, along with the novel recurrent fusion EEF1DP3-FRY while this was not found among TopHatFusion calls. This suggests the diversity of fusion detecting

algorithms and the heterogeneous results which may contribute to the paucity of replicated results from different datasets (Hofvander et al., 2015).

3. Novel Validated Recurrent EEF1DP3-FRY Fusion Transcript

EEF1DP3-FRY is a fusion between eukaryotic translation elongation factor 1 delta pseudogene 3 (EEF1DP3) and furry homolog (FRY), both of which are located on the long arm of chromosome 13. This novel read-through fusion was found in 6.7% (8/120) of the studied tumor samples. While EEF1DP3 is a pseudogene with no known function, the Fry protein is known to play a crucial role in the structural integrity of mitotic centrosomes.(Nagai et al., 2013) The off-frame fusion of these two adjacent genes resulted in the early-truncation, loss-of-function mutation of the large FRY gene (267,000 bps, 61 exons, and 3,013 amino acids). The confirmed breakpoint was located upstream of exon 2; it caused a frameshift that created a TAG stop at codon 33 of exon 2 (Figure 5). Chimerascan analysis of the isoform ratio (i.e., the proportion of reads covering the fusion junction among all reads that map to each of the involved genes) revealed that the mean isoform ratios of the EEF1DP3 and FRY genes in the fusion-positive samples were 0.93 and 0.72, respectively.

5. Confirmation Of Previously Reported FTs

We tried to confirm the previously reported fusions in our data. The PPP1R1B-STARD3 chimeric fusion transcript was previously described in breast and gastric cancer(Robinson et al., 2011; Yun et al., 2013). In our dataset, 8.3% (10/120) of the tested tumor tissues were positive for this fusion gene, whereas none of the five normal samples showed such expression. The fusion junction was located between exon 6 of PPP1R1B and exon 2 of STARD3, which is consistent with the previous reports. Nine of the 10 tumors that expressed this fusion were ER-negative. Another known read-through chimeric transcript, MFGE8-HAPLN3(Varley et al., 2014), was detected in four of the tested cancer samples: three triple-negative samples and one luminal sample. ETV6-NTRK3, a prototype fusion gene of secretory breast carcinoma, was identified in one of our tumor samples (a secretory carcinoma sample), as assessed by both deFuse and Chimerascan, and confirmed by immunohistochemical staining. In contrast, the previously reported recurrent rearrangements involving genes encoding microtubule-associated serine-threonine kinase (MAST) and members of the Notch family (Robinson et al., 2011) were not identified in the 120 breast cancer samples tested in the present work.

Table1. Demographics of 120 primary tumor samples

Sample number	Age(yr)	Distant metastasis	ER	PR	HER2	Subtype
9827	51	Yes	Positive	Negative	Negative	Luminal
2224	51	No	Positive	Positive	Negative	Luminal
2006-53	60	Yes	Positive	Negative	Negative	Luminal
2008-010	55	No	Negative	Negative	Negative	Triple negative
9857	71	Yes	Positive	Positive	Unknown	Luminal
2003-200	64	Yes	Negative	Negative	Negative	Triple negative
2214	54	No	Positive	Positive	Negative	Luminal
2105	67	No	Positive	Positive	Negative	Luminal
2386	59	No	Negative	Negative	Positive	HER2
9970	58	Yes	Negative	Negative	Unknown	Non-luminal
2123	41	No	Negative	Negative	Positive	HER2
2188	38	No	Positive	Negative	Negative	Luminal
99143	63	Yes	Positive	Positive	Unknown	Luminal
2189	52	No	Negative	Negative	Negative	Triple negative
2083	65	No	Positive	Positive	Positive	Luminal
9653	53	Yes	Positive	Negative	Unknown	Luminal
2107	54	No	Positive	Positive	Negative	Luminal
2006-21	71	No	Negative	Negative	Negative	Triple negative
2115	42	No	Negative	Positive	Negative	Luminal
2009-050	70	Yes	Negative	Negative	Positive	HER2
2006-56	62	Yes	Negative	Negative	Negative	Triple negative
2003-237	61	No	Positive	Positive	Negative	Luminal
2096	58	No	Negative	Negative	Negative	Triple negative
2062	51	No	Positive	Positive	Positive	Luminal
9829	59	Yes	Negative	Negative	Negative	Triple negative
9964	46	Yes	Negative	Negative	Unknown	Non-luminal
9984	55	Yes	Negative	Negative	Unknown	Non-luminal
99124	49	Yes	Negative	Negative	Unknown	Non-luminal
99150	51	Yes	Positive	Positive	Unknown	Luminal

2003-265	40	Yes	Negative	Negative	Unknown	Non-luminal
2006-316	30	No	Negative	Negative	Negative	Triple negative
2014	58	No	Positive	Negative	Positive	Luminal
2027	53	No	Positive	Positive	Positive	Luminal
2051	45	No	Negative	Negative	Negative	Triple negative
2013	41	No	Positive	Positive	Unknown	Luminal
2012	37	Yes	Positive	Positive	Unknown	Luminal
2007	36	No	Negative	Negative	Unknown	Non-luminal
2053	32	No	Negative	Negative	Negative	Triple negative
2060	51	No	Negative	Positive	Negative	Luminal
2129	32	Yes	Negative	Negative	Negative	Triple negative
2150	45	No	Negative	Negative	Unknown	Non-luminal
2008-118	39	No	Positive	Negative	Negative	Luminal
2142	43	No	Negative	Negative	Negative	Triple negative
2146	69	No	Negative	Negative	Positive	HER2
2172	37	No	Negative	Negative	Negative	Triple negative
2155	41	No	Negative	Negative	Negative	Triple negative
2216	59	No	Positive	Negative	Unknown	Luminal
2003-169	49	No	Negative	Negative	Negative	Triple negative
2222	51	No	Positive	Negative	Unknown	Luminal
2234	39	No	Positive	Negative	Negative	Luminal
2236	54	No	Negative	Negative	Negative	Triple negative
2192	52	No	Negative	Negative	Positive	HER2
2006-173	58	No	Negative	Negative	Unknown	Non-luminal
2195	53	No	Positive	Negative	Negative	Luminal
2268	34	No	Negative	Negative	Negative	Triple negative
2186	63	Yes	Positive	Positive	Negative	Luminal
2194	56	No	Positive	Negative	Negative	Luminal
2243	41	No	Positive	Positive	Unknown	Luminal
2303	52	No	Negative	Negative	Negative	Triple negative
2242	56	No	Positive	Positive	Negative	Luminal
2204	32	No	Negative	Negative	Negative	Triple negative

2202	37	No	Negative	Negative	Positive	HER2
2183	39	No	Positive	Positive	Negative	Luminal
2198	36	No	Negative	Negative	Positive	HER2
2247	53	No	Negative	Negative	Unknown	Non-luminal
2304	39	No	Positive	Positive	Negative	Luminal
2003-159	43	Yes	Negative	Negative	Negative	Triple negative
2281	40	No	Positive	Positive	Unknown	Luminal
2008-134	67	Yes	Negative	Negative	Negative	Triple negative
2317	39	No	Positive	Positive	Negative	Luminal
2332	56	Yes	Negative	Positive	Negative	Luminal
2384	51	No	Negative	Negative	Negative	Triple negative
2358	63	No	Positive	Positive	Negative	Luminal
2354	46	No	Negative	Positive	Negative	Luminal
2370	37	Yes	Positive	Positive	Negative	Luminal
2359	45	No	Negative	Negative	Negative	Triple negative
2357	39	No	Negative	Negative	Unknown	Non-luminal
2379	36	No	Negative	Negative	Positive	HER2
2377	47	No	Negative	Positive	Negative	Luminal
2378	41	No	Negative	Negative	Negative	Triple negative
2006-84	40	No	Negative	Negative	Negative	Triple negative
2006-315	69	No	Negative	Negative	Negative	Triple negative
2003-47	68	Yes	Positive	Negative	Negative	Luminal
2006-187	53	No	Negative	Negative	Negative	Triple negative
2006-313	63	No	Negative	Negative	Positive	HER2
2006-285	43	No	Negative	Negative	Positive	HER2
2006-270	42	No	Positive	Negative	Negative	Luminal
2006-139	43	No	Positive	Positive	Negative	Luminal
2003-148	40	No	Positive	Positive	Negative	Luminal
2006-223	62	Yes	Positive	Positive	Positive	Luminal
2006-174	42	No	Negative	Negative	Negative	Triple negative
2003-186	51	Yes	Negative	Negative	Negative	Triple negative
2006-327	37	Yes	Positive	Positive	Negative	Luminal

2003-191	40	No	Negative	Negative	Negative	Triple negative
2006-311	44	No	Negative	Positive	Negative	Luminal
2006-241	52	No	Negative	Positive	Negative	Luminal
2006-147	39	Yes	Positive	Positive	Positive	Luminal
2006-225	51	No	Negative	Positive	Negative	Luminal
2006-149	73	No	Positive	Positive	Negative	Luminal
2003-267	38	No	Positive	Positive	Negative	Luminal
2003-256	55	Yes	Negative	Negative	Negative	Triple negative
2006-197	71	No	Negative	Negative	Negative	Triple negative
2003-296	77	No	Positive	Positive	Negative	Luminal
2003-303	43	No	Negative	Negative	Unknown	Non-luminal
2003-269	45	No	Negative	Positive	Negative	Luminal
2006-266	49	No	Negative	Negative	Unknown	Non-luminal
2006-267	60	Yes	Positive	Negative	Positive	Luminal
2006-194	52	No	Negative	Negative	Positive	HER2
2006-281	36	No	Negative	Negative	Negative	Triple negative
2006-16	38	No	Positive	Positive	Negative	Luminal
2006-38	64	No	Positive	Positive	Unknown	Luminal
2006-235	36	Yes	Positive	Positive	Negative	Luminal
2006-49	37	Yes	Positive	Positive	Negative	Luminal
2006-82	43	No	Positive	Positive	Negative	Luminal
2006-72	39	Yes	Negative	Negative	Negative	Triple negative
2008-56	72	Yes	Negative	Negative	Positive	HER2
2008-86	44	Yes	Negative	Negative	Negative	Triple negative
2008-112	35	No	Negative	Negative	Negative	Triple negative
2008-217	49	Yes	Negative	Negative	Positive	HER2
2010-050	36	Yes	Positive	Positive	Negative	Luminal

Table 2. Primer sequence of the validated fusion transcripts

Fusion No	gene1	gene2	OLIGO	sequence
1	MAP2K6	AARSD1	LEFT PRIMER	AGGTTCTGGTCCGGGTAGAT
			RIGHT PRIMER	TCACCTCCCAAGTGTCAGTG
			PRODUCT SIZE: 176, PAIR ANY COMPL: 4.00, PAIR 3' COMPL: 0.00	
2	C20orf195	PTK6	LEFT PRIMER	CAGCGTGGCTGTAAAGAACA
			RIGHT PRIMER	AACACCCTCTGCAAAGTTGG
			PRODUCT SIZE: 176, PAIR ANY COMPL: 4.00, PAIR 3' COMPL: 1.00	
3	SUPT6H	CASC3	LEFT PRIMER	TGAAGATGGCATTGAAGGTG
			RIGHT PRIMER	TTACTGGATGGGGTGAGGAG
			PRODUCT SIZE: 235, PAIR ANY COMPL: 3.00, PAIR 3' COMPL: 0.00	
4	CASC3	RGS9	LEFT PRIMER	TGAAGATGGCATTGAAGGTG
			RIGHT PRIMER	GTGGAGATCCCTCCAGTGAA
			PRODUCT SIZE: 192, PAIR ANY COMPL: 5.00, PAIR 3' COMPL: 2.00	
5	CASC3	CCDC46	LEFT PRIMER	TGAAGATGGCATTGAAGGTG
			RIGHT PRIMER	TGGCGTAGGACACAACAAAG
			PRODUCT SIZE: 157, PAIR ANY COMPL: 3.00, PAIR 3' COMPL: 2.00	
6	MSL1	CDK12	LEFT PRIMER	TGTGATTGCCTTTGGTTCTG
			RIGHT PRIMER	CATTACGGGAATCCTCTCCA
			PRODUCT SIZE: 187, PAIR ANY COMPL: 5.00, PAIR 3' COMPL: 2.00	
7	CDK13	RALA	LEFT PRIMER	GACAGCTCATGGAGGGTCTG
			RIGHT PRIMER	TCCATCTTTCTCGCTCGAAT
			PRODUCT SIZE: 189, PAIR ANY COMPL: 5.00, PAIR 3' COMPL: 0.00	
8	CEP72	FNTA	LEFT PRIMER	TGCATCAGAGGACTCACTCG

			RIGHT PRIMER	ACTCGCCTATGATGCCAAAC
			PRODUCT SIZE: 180, PAIR ANY COMPL: 8.00, PAIR 3' COMPL: 1.00	
9	RAB2A	CHD7	LEFT PRIMER	ATCTACCCCACCAGGGAAGT
			RIGHT PRIMER	ACATCATAATCGGCGACACA
			PRODUCT SIZE: 231, PAIR ANY COMPL: 3.00, PAIR 3' COMPL: 1.00	
10	TNFRSF11B	COLEC10	LEFT PRIMER	CTTGCCAATATTGCCCTGAT
			RIGHT PRIMER	CTGTGCTTGTGTCTCCTCCA
			PRODUCT SIZE: 222, PAIR ANY COMPL: 3.00, PAIR 3' COMPL: 1.00	
11	BCAS3	DPY19L2P1	LEFT PRIMER	GTTGCCAGGCTGGAGTGT
			RIGHT PRIMER	GGACAGATGTTGCCAGGATT
			PRODUCT SIZE: 203, PAIR ANY COMPL: 4.00, PAIR 3' COMPL: 0.00	
12	ATP6V1C1	VEGFB	LEFT PRIMER	CGAGAGGTGAGGGCAGTTAC
			RIGHT PRIMER	GGTAGTCCCAGCAACTGAGG
			PRODUCT SIZE: 188, PAIR ANY COMPL: 5.00, PAIR 3' COMPL: 2.00	
13	ERBB2	SDK2	LEFT PRIMER	ACAGTGGCATCTGTGAGCTG
			RIGHT PRIMER	CGTCTTCACTCGCTCTGTGA
			PRODUCT SIZE: 174, PAIR ANY COMPL: 4.00, PAIR 3' COMPL: 2.00	
14	IGHG1	ERBB2	LEFT PRIMER	TCCCGGACATGGTCTAAGAG
			RIGHT PRIMER	CAAGTGCAAGGTCTCCAACA
			PRODUCT SIZE: 207, PAIR ANY COMPL: 3.00, PAIR 3' COMPL: 2.00	
15	RGPD2	FASN	LEFT PRIMER	AGGAGGACAAGCCTGAGGAG
			RIGHT PRIMER	TTTTATGGGCTTTGGTGAGG
			PRODUCT SIZE: 183, PAIR ANY COMPL: 5.00, PAIR 3' COMPL: 1.00	
16	FBXL20	TOM1L1	LEFT PRIMER	AACGGAGTGACCAAGAGCAG

			RIGHT PRIMER	ACCATGGCTTTGACATGTTG
			PRODUCT SIZE: 205, PAIR ANY COMPL: 4.00, PAIR 3' COMPL: 1.00	
17	IFT20	ERBB2	LEFT PRIMER	CAGGAGAAGCTGCCTGTCTA
			RIGHT PRIMER	AGCAGAGGTGGGTGTTATGG
			PRODUCT SIZE: 159, PAIR ANY COMPL: 4.00, PAIR 3' COMPL: 1.00	
18	MED1	CCL2	LEFT PRIMER	CCTGTGAATGACTCCCTGGT
			RIGHT PRIMER	TGCCAAGTCACTTCCCTTCT
			PRODUCT SIZE: 169, PAIR ANY COMPL: 6.00, PAIR 3' COMPL: 0.00	
19	NLK	PPP1R1B	LEFT PRIMER	GACAGGAGCCCCTACTTTCA
			RIGHT PRIMER	TTGCCTGCCTGTTAAATTCC
			PRODUCT SIZE: 220, PAIR ANY COMPL: 5.00, PAIR 3' COMPL: 1.00	
20	TMEM199	NLK	LEFT PRIMER	TTTCCAATTTCATTGCACCA
			RIGHT PRIMER	TGCTATACCTGCCTCCATCC
			PRODUCT SIZE: 228, PAIR ANY COMPL: 4.00, PAIR 3' COMPL: 0.00	
21	NOS2	ERBB2	LEFT PRIMER	ACCAGCATTTTGCCCTGTAG
			RIGHT PRIMER	TGGCCCCAATTCATTTCTTA
			PRODUCT SIZE: 184, PAIR ANY COMPL: 3.00, PAIR 3' COMPL: 1.00	
22	ETV6	NTRK3	LEFT PRIMER	CCTCCCTCTGGAAATCCTTC
			RIGHT PRIMER	CATGCCCATTGGGAGAATAG
			PRODUCT SIZE: 203, PAIR ANY COMPL: 6.00, PAIR 3' COMPL: 1.00	
23	TACSTD2	OMA1	LEFT PRIMER	CACCACAAAGAGCAATCCAA
			RIGHT PRIMER	GGGGAGTTGAGAAAGGAACC
			PRODUCT SIZE: 150, PAIR ANY COMPL: 4.00, PAIR 3' COMPL: 0.00	
24	PLAT	CCND1	LEFT PRIMER	AGGGCTGGAGAGAAAACCTC

			RIGHT PRIMER	TGAGGCGGTAGTAGGACAGG
			PRODUCT SIZE: 162, PAIR ANY COMPL: 4.00, PAIR 3' COMPL: 2.00	
25	PLCB3	VEGFB	LEFT PRIMER	TCAGTTCCATCAGGGACACA
			RIGHT PRIMER	AGTGGGATGGGTGATGTCAG
			PRODUCT SIZE: 183, PAIR ANY COMPL: 5.00, PAIR 3' COMPL: 2.00	
26	PPP1R9B	CLTC	LEFT PRIMER	CGGAAGATCCATTTCAGCAC
			RIGHT PRIMER	TTGACTCGCTCAGGATCGTA
			PRODUCT SIZE: 184, PAIR ANY COMPL: 5.00, PAIR 3' COMPL: 1.00	
27	RAB26	E4F1	LEFT PRIMER	TTTCCTGACATGGCTTTTGA
			RIGHT PRIMER	GTCACCTGGGAGCATGAACT
			PRODUCT SIZE: 243, PAIR ANY COMPL: 4.00, PAIR 3' COMPL: 1.00	
28	RAB9A	EGFL6	LEFT PRIMER	CGACCCTCTCTCTGTCCTCA
			RIGHT PRIMER	AGGGCTGCAAGTTCACCTTA
			PRODUCT SIZE: 244, PAIR ANY COMPL: 4.00, PAIR 3' COMPL: 1.00	
29	CDK19	REV3L	LEFT PRIMER	CACTGGTTTTCCCAATCAGG
			RIGHT PRIMER	CTCTATGGGGGAAGCAGACA
			PRODUCT SIZE: 234, PAIR ANY COMPL: 6.00, PAIR 3' COMPL: 2.00	
30	CASC3	SPOP	LEFT PRIMER	TACACACCGACACACACCAG
			RIGHT PRIMER	GAAGCACCAGATTCCTCGTC
			PRODUCT SIZE: 200, PAIR ANY COMPL: 3.00, PAIR 3' COMPL: 2.00	
31	GRB7	STAT3	LEFT PRIMER	TCCAGGTACCGTGTGTCAAG
			RIGHT PRIMER	CTGGAGGGGGACAGGTAGTT
			PRODUCT SIZE: 156, PAIR ANY COMPL: 5.00, PAIR 3' COMPL: 1.00	
32	FOXN1	TNFAIP1	LEFT PRIMER	CATGGAGATGCCAAGCACTA

			RIGHT PRIMER	GGGGATAATTTGACGCTTGA
			PRODUCT SIZE: 154, PAIR ANY COMPL: 5.00, PAIR 3' COMPL: 2.00	
33	TRAF4	WSB1	LEFT PRIMER	CATCTCCCACCAGGACATTC
			RIGHT PRIMER	GGTGCCAGCATAACCTCAAT
			PRODUCT SIZE: 247, PAIR ANY COMPL: 4.00, PAIR 3' COMPL: 2.00	
34	CASC3	WIPF2	LEFT PRIMER	GCTCTGTGTTTGCCTGATGA
			RIGHT PRIMER	GAAGCACCAGATTCCTCGTC
			PRODUCT SIZE: 235, PAIR ANY COMPL: 5.00, PAIR 3' COMPL: 3.00	
35	COL1A1	ACTB	LEFT PRIMER	GTGGCTTTTAGGATGGCAAG
			RIGHT PRIMER	ATGGGTCTTCAAGCAAGTGG
			PRODUCT SIZE: 245, PAIR ANY COMPL: 4.00, PAIR 3' COMPL: 0.00	
36	CEACAM7	CEACAM6	LEFT PRIMER	CCTGTTGTCAGTGGAGAGCA
			RIGHT PRIMER	AGCCTGGTGGTCTGCAGTAT
			PRODUCT SIZE: 228, PAIR ANY COMPL: 5.00, PAIR 3' COMPL: 3.00	
37	COL1A1	COL1A2	LEFT PRIMER	ATTGGTAACCCTGGCAGAGA
			RIGHT PRIMER	GCCTTTAGCACCAGCATCAC
			PRODUCT SIZE: 231, PAIR ANY COMPL: 4.00, PAIR 3' COMPL: 2.00	
38	CSMD3	SAMD12	LEFT PRIMER	TAGCCATCAAGTGCAGCAAC
			RIGHT PRIMER	CCTTCACGTCTCTCCAAAGC
			PRODUCT SIZE: 237, PAIR ANY COMPL: 4.00, PAIR 3' COMPL: 0.00	
39	CYP4X1	CYP4Z2P	LEFT PRIMER	GGACTTGGGATCTGTTCTGC
			RIGHT PRIMER	CCCAGAACTTTGGAAATTGG
			PRODUCT SIZE: 192, PAIR ANY COMPL: 6.00, PAIR 3' COMPL: 1.00	
40	EEF1DP3	FRY	LEFT PRIMER	GAACCGGAGGCTTGATGTAA

			RIGHT PRIMER	CCACAATGGCAGTTGATGAC
			PRODUCT SIZE: 249, PAIR ANY COMPL: 3.00, PAIR 3' COMPL: 2.00	
41	PCBD2	HNRNPA2B1	LEFT PRIMER	CAAAGTTGCCTCCTCTTCCA
			RIGHT PRIMER	GAGAAGGGTCAGGGCTGATT
			PRODUCT SIZE: 228, PAIR ANY COMPL: 5.00, PAIR 3' COMPL: 1.00	
42	RGPD2	MALAT1	LEFT PRIMER	TCCAAAAGCCTTCTGCCTTA
			RIGHT PRIMER	TTTTATGGGCTTTGGTGAGG
			PRODUCT SIZE: 224, PAIR ANY COMPL: 6.00, PAIR 3' COMPL: 3.00	
43	PCBD2	MUC6	LEFT PRIMER	GGAACGTGAGTGGGAAGTGT
			RIGHT PRIMER	TAGTAGGCTCCCTCCCCCTA
			PRODUCT SIZE: 214, PAIR ANY COMPL: 5.00, PAIR 3' COMPL: 2.00	
44	NEAT1	RGPD2	LEFT PRIMER	CTACCGGTGTACCCACCATT
			RIGHT PRIMER	TTTTATGGGCTTTGGTGAGG
			PRODUCT SIZE: 181, PAIR ANY COMPL: 5.00, PAIR 3' COMPL: 3.00	
45	PRDX1	AKR1A1	LEFT PRIMER	TGACCATCTGGCATAACAGC
			RIGHT PRIMER	GCCTCAGCCTTGACAGAGTT
			PRODUCT SIZE: 155, PAIR ANY COMPL: 5.00, PAIR 3' COMPL: 1.00	
46	ERBB2	MUC6	LEFT PRIMER	AAAGGCCCAAGACTCTCTCC
			RIGHT PRIMER	TTAATGAGCTCAGGGCTTGG
			PRODUCT SIZE: 170, PAIR ANY COMPL: 5.00, PAIR 3' COMPL: 3.00	
47	RUFY1	FLT4	LEFT PRIMER	ATCTTGATGGTGGCAAGGAG
			RIGHT PRIMER	CTCCCCATACTCGCTGTTGT
			PRODUCT SIZE: 209, PAIR ANY COMPL: 4.00, PAIR 3' COMPL: 1.00	
48	VTCN1	NRG1	LEFT PRIMER	TACCCAGATACGCTGGGAAC

			RIGHT PRIMER	AACTGGTTTCACACCGAAGG
			PRODUCT SIZE: 209, PAIR ANY COMPL: 5.00, PAIR 3' COMPL: 2.00	
49	BRIP1	KIAA1468	LEFT PRIMER	GGTTTGGGTTGGTACCATTG
			RIGHT PRIMER	GCTCTTGGACGGTCTTGTTT
			PRODUCT SIZE: 227, PAIR ANY COMPL: 3.00, PAIR 3' COMPL: 0.00	
50	MYO1D	CLTC	LEFT PRIMER	TTGAACGGTGTGGTTTTCAG
			RIGHT PRIMER	GAGCACCAAATCATGGACAA
			PRODUCT SIZE: 209, PAIR ANY COMPL: 4.00, PAIR 3' COMPL: 2.00	
51	TMEM49	PDK2	LEFT PRIMER	CGGGTGTGAGAGTCCGTAAG
			RIGHT PRIMER	GGCTCTGGACATACCAGCTC
			PRODUCT SIZE: 224, PAIR ANY COMPL: 4.00, PAIR 3' COMPL: 3.00	
52	HDLBP	HJURP	LEFT PRIMER	AATCGACCTTCCAGCAGAGA
			RIGHT PRIMER	GACCTCACCGCTTTTTGAAT
			PRODUCT SIZE: 226, PAIR ANY COMPL: 4.00, PAIR 3' COMPL: 1.00	
53	COL1A2	COL1A1	LEFT PRIMER	ATTGCAAGGTCTGCCTGGTA
			RIGHT PRIMER	CAGGGAAACCAGTAGCACCA
			PRODUCT SIZE: 242, PAIR ANY COMPL: 6.00, PAIR 3' COMPL: 3.00	
54	SLC39A9	MAP3K9	LEFT PRIMER	TGTTGGTGGGATGTTACGTG
			RIGHT PRIMER	GATGTCCTCATCAGGGTCGT
			PRODUCT SIZE: 193, PAIR ANY COMPL: 4.00, PAIR 3' COMPL: 1.00	
55	COL3A1	COL1A1	LEFT PRIMER	CAGCCTGGAGATAAGGGTGA
			RIGHT PRIMER	GCCTTTAGCACCAGCATCA
			PRODUCT SIZE: 153, PAIR ANY COMPL: 7.00, PAIR 3' COMPL: 3.00	
56	PITPNC1	PRKCA	LEFT PRIMER	CGGGTGTATCTCAACAGCAA

			RIGHT PRIMER	ACAACATTGTCTGGGCATCA
			PRODUCT SIZE: 300, PAIR ANY COMPL: 4.00, PAIR 3' COMPL: 2.00	
57	TOM1L1	HLF	LEFT PRIMER	TTTTGGTGTTTTTGGTTTTGG
			RIGHT PRIMER	CTTCCAATCTGTGGCTTGCT
			PRODUCT SIZE: 243, PAIR ANY COMPL: 4.00, PAIR 3' COMPL: 0.00	
58	RNF19A	STK3	LEFT PRIMER	CTTGACCGGATTCCTTGTGT
			RIGHT PRIMER	ATCCCCATGATATTCGCTTG
			PRODUCT SIZE: 252, PAIR ANY COMPL: 4.00, PAIR 3' COMPL: 0.00	
59	TBL1XR1	NRG1	LEFT PRIMER	AGGAGGGGAATTTCCTTGTG
			RIGHT PRIMER	CGATTCAATTCATTCCCATTC
			PRODUCT SIZE: 204, PAIR ANY COMPL: 6.00, PAIR 3' COMPL: 2.00	
60	UBE3C	BCAS3	LEFT PRIMER	CTGTAAGCCTGTGGCACAAC
			RIGHT PRIMER	CTGCCAGGATGTTCAGCTTC
			PRODUCT SIZE: 225, PAIR ANY COMPL: 5.00, PAIR 3' COMPL: 2.00	
61	LRP5	GAL	LEFT PRIMER	GCAGCCTTTCTTCCACACTC
			RIGHT PRIMER	AGGACCGCTCGATGTCTTCT
			PRODUCT SIZE: 290, PAIR ANY COMPL: 3.00, PAIR 3' COMPL: 0.00	
62	LTF	CD74	LEFT PRIMER	GATAAGGTGGAACGCCTGAA
			RIGHT PRIMER	GGGTCTCATGGGATGAGGTA
			PRODUCT SIZE: 267, PAIR ANY COMPL: 3.00, PAIR 3' COMPL: 0.00	
63	ATG5	EIF3E	LEFT PRIMER	ATGAAGGCACACCACTGAAA
			RIGHT PRIMER	GAGTCATTTGGCCGTTGATT
			PRODUCT SIZE: 293, PAIR ANY COMPL: 4.00, PAIR 3' COMPL: 2.00	
			PRODUCT SIZE: 293, PAIR ANY COMPL: 4.00, PAIR 3' COMPL: 2.00	
64	SYNJ2	DCC	LEFT PRIMER	ACTGTAGCGTGCTGCTGGAG

			RIGHT PRIMER	GCCAGATGAATGCCATCTTT
			PRODUCT SIZE: 229, PAIR ANY COMPL: 4.00, PAIR 3' COMPL: 1.00	
65	C2CD2	TFF1	LEFT PRIMER	GCGATGTCTGACGTTCTCAA
			RIGHT PRIMER	TGACACCAGGAAAACCACAA
			PRODUCT SIZE: 280, PAIR ANY COMPL: 4.00, PAIR 3' COMPL: 1.00	
66	PCBD2	COX6C	LEFT PRIMER	GCACGAATGCTACAGCCATA
			RIGHT PRIMER	GGCTGTGAGTGGTTGTGTTG
			PRODUCT SIZE: 236, PAIR ANY COMPL: 6.00, PAIR 3' COMPL: 2.00	
67	PCBD2	TPM3	LEFT PRIMER	GCGTTCCAAGTCTCCTTCAA
			RIGHT PRIMER	TAGGGGGAGGGAGCCTACTA
			PRODUCT SIZE: 204, PAIR ANY COMPL: 6.00, PAIR 3' COMPL: 2.00	
68	CBFA2T2	CHMP4B	LEFT PRIMER	GCTGGAACTCGATGGTTGAT
			RIGHT PRIMER	GGTGGTGGTGTCTGGTTAGC
			PRODUCT SIZE: 266, PAIR ANY COMPL: 3.00, PAIR 3' COMPL: 1.00	
69	LPP	BNC2	LEFT PRIMER	TGCATCATCTGCAACAACAA
			RIGHT PRIMER	ATAAGGCAGCCGGAAAGAAT
			PRODUCT SIZE: 255, PAIR ANY COMPL: 4.00, PAIR 3' COMPL: 1.00	
70	RNF213	ABL1	LEFT PRIMER	TCCTGCTCTTGCTTCTGGAT
			RIGHT PRIMER	CTGCACCAGGTTAGGGTGTT
			PRODUCT SIZE: 220, PAIR ANY COMPL: 4.00, PAIR 3' COMPL: 0.00	

Table 3. Characteristics of the 32 validated, 70 nominated fusion transcripts

Sample Number	Donor gene (5')	Acceptor gene (3')	Chr (5')	Chr (3')	Donor-Acceptor	In-frame vs Off-frame	Detected sample number	Validated	10 FTs for screening	Splitr count	Span count
2268	ATG5	EIF3E	6	8	D-A	In	1	YES		5	6
2006-241	ATP6V1C1	VEGFB	8	17	D-D	Off	1			5	5
2083	BCAS3	DPY19L2P1	17	7	A-A	In	1	YES		18	12
2123	BRIP1	KIAA1468	17	18	D-A	In	1	YES		29	7
2014	C20orf195	PTK6	20	20	D-D	In	1	YES		32	18
9827	C2CD2	TFF1	21	21	D-A	In	2	YES	YES	62	27
2006-241	CASC3	SPOP	17	17	A-A	In	1			3	6
2146	CASC3	WIPF2	17	17	A-A	Off	1			1	6
2006-147	CASC3	CCDC46	17	17	D-A	In	1	YES		16	13
2006-147	CASC3	RGS9	17	17	D-A	Off	1	YES		71	49
2003-159	CBFA2T2	CHMP4B	20	20	D-A	In	1			18	55
2006-311	CDK13	RALA	7	7	D-A	Off	1	YES		35	37
2006-281	CDK19	REV3L	6	6	D-A	In	1	YES	YES	47	42
2236	CEACAM7	CEACAM6	19	19	A-A	Off	3	YES		3	10
2188	CEP72	FNTA	5	8	D-A	In	1	YES		5	6
2003-148	COL1A1	ACTB	17	7	A-A	In	2			10	9
2242	COL1A1	COL1A2	17	7	D-A	In	7			2	5
2006-16	COL1A2	COL1A1	7	17	D-A	Off	2			11	6
2194	COL3A1	COL1A1	2	17	D-A	In	4			3	5
2357	CSMD3	SAMD12	8	8	D-A	In	1			5	9
2006-313	CYP4X1	CYP4Z2P	1	1	A-A	Off	2	YES		8	3
2083	E2F6	GREB1	2	2	A-A	In	3			2	13
2386	EEF1DP3	FRY	13	13	D-A	Off	15	YES	YES	10	5
2003-200	ERBB2	MUC6	17	11	D-A	In	1			122	5
2192	ERBB2	SDK2	17	17	D-A	Off	1			9	6
2234	ETV6	NTRK3	14	15	D-A	In	1	YES		8	15
2006-267	FBXL20	TOM1L1	17	17	D-A	Off	1			22	20
2006-173	FOXN1	TNFAIP1	17	17	D-D	In	1			16	22
2146	GRB7	STAT3	17	17	D-A	In	1			149	12
2150	HDLBP	HJURP	2	2	D-A	In	1			6	9
2006-173	IFT20	ERBB2	17	17	D-A	Off	1			22	14
2006-267	IGHG1	ERBB2	14	17	D-A	In	3			5	6
2003-159	LPP	BNC2	3	9	D-A	In	1			16	6
2222	LRP5	GAL	11	11	D-A	In	1	YES		51	46
2234	LTF	CD74	3	5	D-A	In	4			107	7
2354	MAP2K6	AARSD1	17	17	D-D	In	1			1	5
2247	MED1	CCL2	17	17	D-D	In	1	YES		47	75
2198	MSL1	CDK12	17	17	D-A	Off	1	YES		43	59

2123	MYO1D	CLTC	17	17	D-A	In	1	YES		44	30
2006-173	NEAT1	RGPD2	2	11	D-A	Off	4			2	7
2006-173	NLK	PPP1R1B	17	17	A-A	Off	1	YES		8	19
2006-173	NOS2	ERBB2	17	17	D-D	In	1			11	37
R2008-118	PCBD2	HNRNPA2B1	5	7	D-A	In	5			4	5
2202	PCBD2	MUC6	2	11	D-D	In	1			33	37
99150	PCBD2	COX6C	5	8	D-A	In	3			55	6
99150	PCBD2	TPM3	5	1	D-A	In	1			3	11
2198	PITPNC1	PRKCA	17	17	D-A	In	1	YES	YES	15	10
2188	PLAT	CCND1	8	11	D-A	In	1			81	28
2003-191	PLCB3	VEGFB	11	11	D-A	Off	1	YES		27	21
2123	PPP1R9B	CLTC	17	17	D-A	Off	1	YES		99	80
2006-316	PRDX1	AKR1A1	1	1	A-A	In	13	YES	YES	9	6
R2216	RAB26	E4F1	16	16	D-A	Off	1	YES		12	10
2378	RAB2A	CHD7	8	8	D-A	In	1	YES		17	6
2006-197	RAB9A	EGFL6	X	X	D-A	Off	1			4	6
2006-173	RGPD2	FASN	2	17	D-D	In	1			2	7
2006-241	RGPD2	MALAT1	2	11	D-A	Off	7			3	8
2202	RNF19A	STK3	8	8	D-A	In	1	YES	YES	43	17
2003-159	RNF213	ABL1	17	9	D-A	In	1			21	31
2014	RUFY1	FLT4	5	5	D-A	In	1			13	21
2188	SLC39A9	MAP3K9	14	14	D-A	In	1	YES	YES	11	5
2006-147	SUPT6H	CASC3	17	17	D-D	In	1			8	6
2354	SYNJ2	DCC	6	18	D-A	In	1			6	6
2003-191	TACSTD2	OMA1	1	1	D-A	In	2	YES	YES	4	5
2202	TBL1XR1	NRG1	3	8	D-A	In	1			1	23
2006-173	TMEM199	NLK	17	17	D-A	Off	1			5	9
2123	TMEM49	PDK2	17	17	D-A	In	1	YES		189	30
2006-316	TNFRSF11B	COLEC10	8	8	A-A	Off	1	YES		3	11
2198	TOM1L1	HLF	17	17	D-A	In	1			54	26
2006-241	TRAF4	WSB1	17	17	D-A	Off	1	YES		18	12
2202	UBE3C	BCAS3	7	17	D-A	In	1			6	5
2062	VTCN1	NRG1	1	8	D-A	In	1	YES	YES	21	56

Figure 1. Study overview

Whole-transcriptome sequencing of 120 primary breast cancer samples was performed. Fusion transcripts (FTs) were identified by applying multiple fusion-detecting tools: deFuse, followed by Chimerascan and TopHatFusion. Highly nominated candidates were confirmed by RT-PCR and Sanger sequencing.

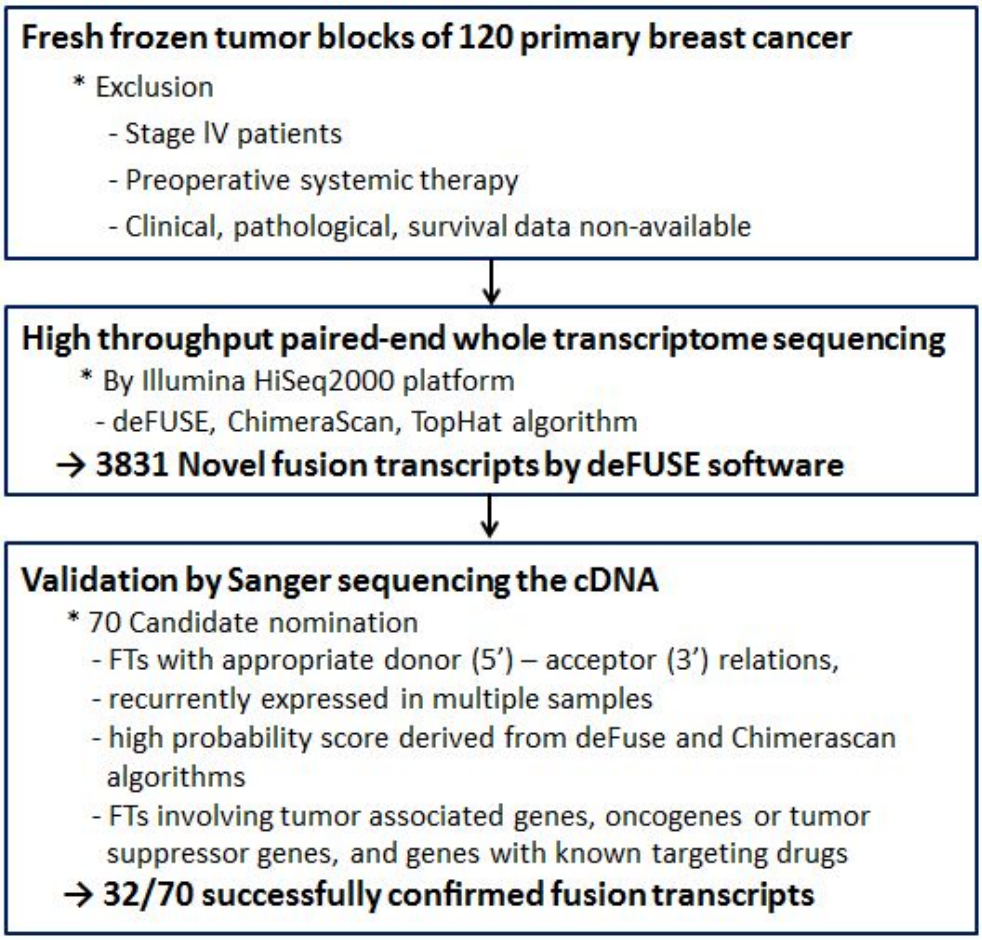


Figure 2. Fusion transcripts detected by deFuse

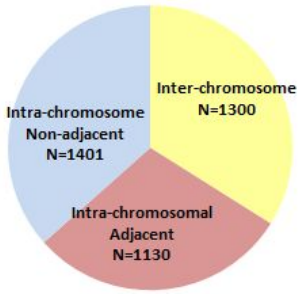


Figure 2a

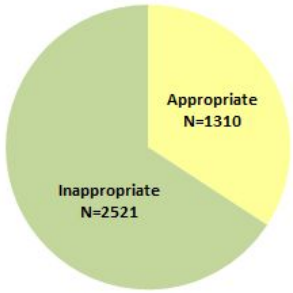


Figure 2b

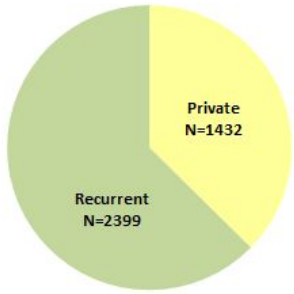


Figure 2c

Definitions: intra-chromosome, the two genes are on the same chromosome; inter-chromosome, the two genes are on different chromosomes; adjacent, the two genes are located side-by-side; non-adjacent, there is at least one gene between the two genes; appropriate, the two genes share an appropriate donor (5')-acceptor (3') relationship; inappropriate; both genes are either donors or acceptors; private, the FT was detected in only one sample; and recurrent, the FT was detected in multiple samples.

Figure 3. Number of FTs detected by deFuse

(a) HER2-positive tumors harbored the greatest numbers of FTs among the four subtypes. (b) HER2-positive breast cancer samples harbored significantly more FTs.

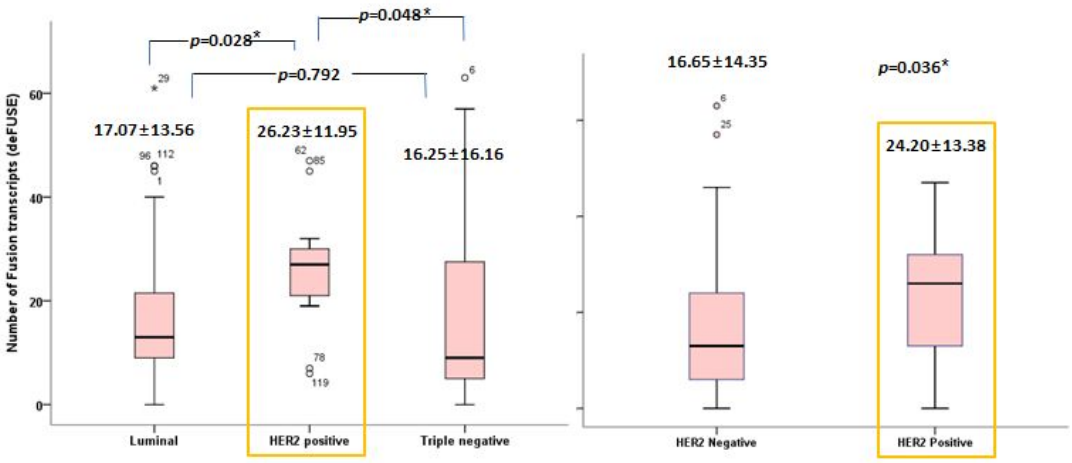
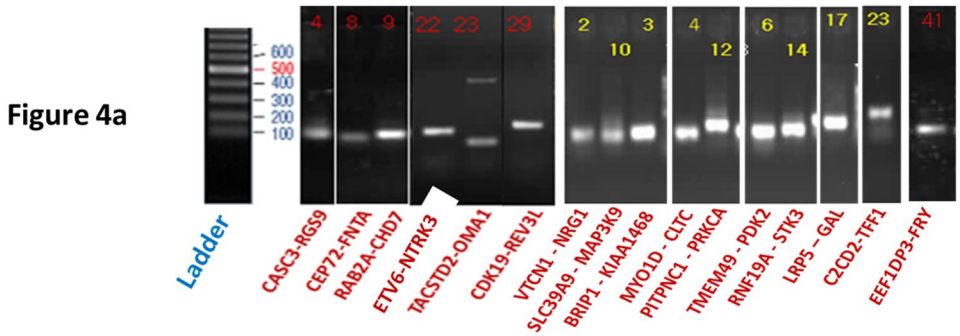


Figure 3a

Figure 3b

Figure 4. RT-PCR and Sanger sequencing of the 70 candidates

(a) PCR bands indicating an FT. (b) Sanger sequencing-based chromatogram presenting a fusion junction. (c) Circos plot of the validated FTs. The outer ring depicts the chromosomes.



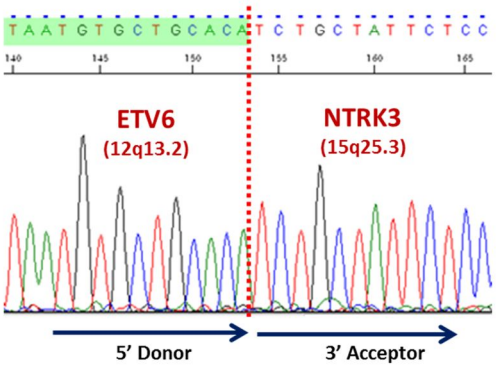


Figure 4b

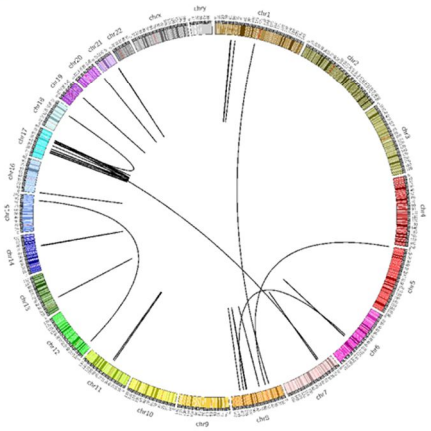


Figure 4c

Figure 5. The novel recurrent fusion gene, EEF1DP3-FRY

(a) Identification of the EEF1DP3-FRY FT by whole-transcriptome sequencing. The sequences of the fusion-junction-spanning reads are displayed. Black indicates EEF1DP3 exons 1 and 2, while red indicates exon 2 of FRY. (b) Schematic representations of the wild-type EEF1DP3 (yellow) and FRY (green) genes, with the breakpoints shown as the respective exon numbers and mRNA positions. Fusion occurred between the second exon of EEF1DP3 and the second exon of FRY. The red arrow indicates the predicted fusion protein, which involves the early truncation of FRY within exon 2.

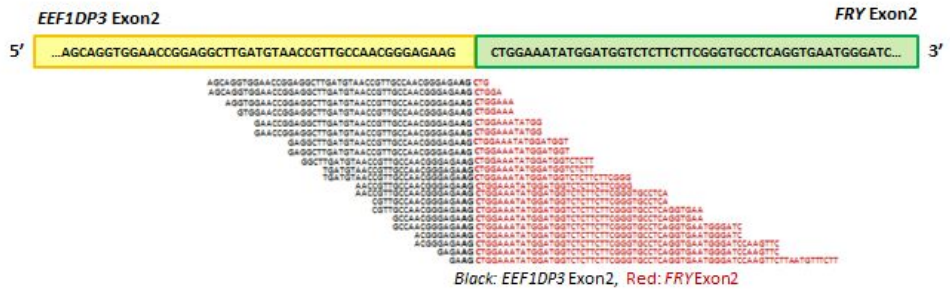


Figure 5a

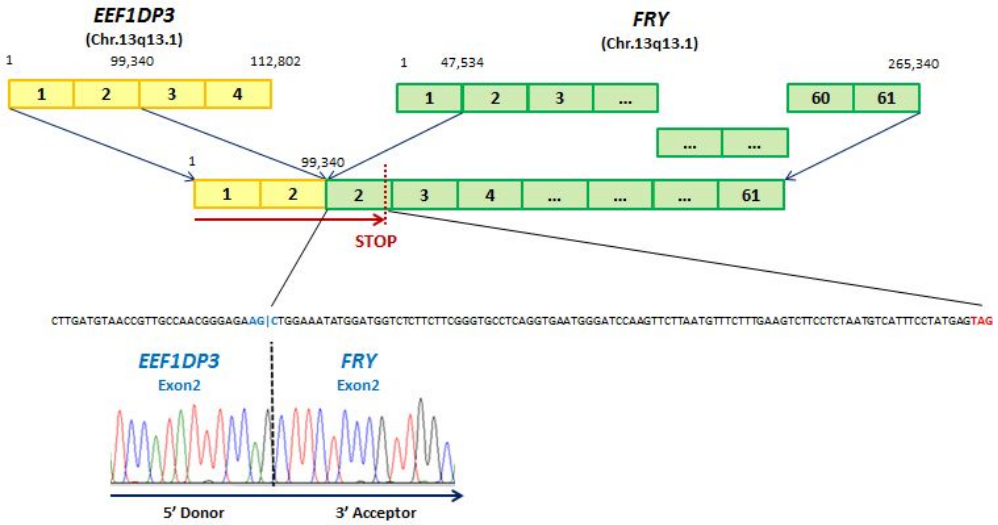

EEF1DP3
(Chr.13q13.1)
1
99,340
112,802
1
2
3
4
FRY
(Chr.13q13.1)
1
47,534
265,340
1
2
3
...
60
61
...
...
1
99,340
1
2
2
3
4
...
...
...
61
STOP
CTTGATGTAACCGTTGCCAACGGGAGAAG|CTGGAAATATGGATGGTCTCTTCTTCGGGTGCCTCAGGTGAATGGGATCCAAGTTCTTAATGTTTCTTTGAAGTCTTCCTCTAATGTCATTTCCTATGAGTAG
EEF1DP3
Exon2
FRY
Exon2
5' Donor
3' Acceptor

Figure 5b

Discussion

Unlike prostate and lung cancers, which are known to frequently express driver fusion genes, limited data are available for fusion genes in primary breast cancer(Tomlins et al., 2007; Takeuchi et al., 2009). Two recurrent fusion genes have been identified in breast cancer, ETV6-NTRK3 in secretory breast carcinoma and MYB-NFIB in adenoid cystic carcinoma, but these are very uncommon histologic types of breast malignancy (Kangaspeska et al., 2012). Recent massive parallel sequencing studies revealed that all classes of genetic alterations may be detected throughout various types of cancer. However, relatively few studies have focused on the comprehensive detection of novel fusion genes in breast cancer. Our work is the first large-scale RNA-Seq study of breast cancer among Asian women.

Our RNA-Seq study of 120 primary breast cancer tissues identified numerous FTs, with each tumor sample having from 0 to 36 FTs. HER2 amplification/overexpression-positive tumors experienced more fusion events compared to their negative counterparts. Kalyana-Sundaram and colleagues previously sequenced 14 breast cancer cell lines matched with array comparative genomic hybridization (aCGH), and found that 60% of the identified fusion events occurred within chromosome arms 17q12, 17q23, 20q13, and 8q (Kalyana-Sundaram et al., 2012). These findings and those from other studies showing an association between gene fusions and recurrent amplicons (Robinson et al., 2011)

support our finding that HER2-positive breast cancers harbor a higher level of fusion transcripts. Relatively low isoform ratio of these 'amplicon-associated' gene fusions also implicates the by-product nature that most transcripts are full length and only minor proportion exists as a fusion transcript. Although functional studies are needed, it is more likely that these amplicon-associated fusions are passenger aberrations rather than a biologically relevant driver fusion (Yoshihara et al., 2014; Mertens et al., 2015). As our NGS work were matched sequencing of whole exome and transcriptome, we tried to see the copy number alterations and number of fusions but fusion events were not clustered around those sites showing copy number variations. This may be because of the limited nature of whole exome sequencing method for identifying accurate copy number variations unlike aCGH.

The prevalence of a certain genetic alteration among a certain tumor type is important for its clinical implications. Among the 3831 FTs identified in our initial screen, 62.6% were recurrent fusions expressed in more than one sample. The novel recurrent fusion gene, EEF1DP3-FRY, was relatively common, being detected in eight of the 120 samples (6.7%). This fusion event creates a frameshift and introduces a stop codon, resulting in early truncation of the FRY protein. The fusion junction was identical in all eight samples positive for this particular FT. The isoform ratios were found to be 0.93 and 0.75 for EEF1DP3 and FRY, respectively, indicating that most of the gene transcripts found in the tumors were the fused form, with little transcription seen from wild-type genes. The FRY protein interacts with

the nuclear Dbf2-related (NDR) family of Ser/Thr kinases, and has been implicated in regulating cell division, morphogenesis and growth (Cong et al., 2001; Hirata et al., 2002; Emoto et al., 2004; Hergovich et al., 2006; Nagai et al., 2013). Although future work is needed to investigate the biological relevance of this novel recurrent fusion transcript, it seems likely that the loss of the FRY protein would decrease cellular integrity.

The reliability of our data is supported by our detection of the previously reported recurrent fusions, PPP1R1B-STARD3, MFGE8-HAPLN3, and ETV6-NTRK3. Of them, PPP1R1B-STARD3 is an interesting read-through fusion that causes a frameshift and early truncation of STARD3. Yun et al. found this fusion in 21.3% of their gastric cancer patients, and suggested that it might increase cancer cell proliferation by activating phosphatidylinositol-3-kinase (PI3K)/AKT signaling (Yun et al., 2013). We identified this FT in 8.3% of our breast cancer samples, and found that the breakage point and the predicted fusion protein structure were the same as those reported by Yun et al. in gastric cancer. We then performed validation of PPP1R1B-STARD3 fusion among samples representing it using the same primers and applying same method as shown, but failed in all samples. Validation in SKBR3 – HER2 overexpressing breast cancer cell line – from which the fusion had been detected in previous work had failed likewise. This may be due to the intra-tumoral heterogeneity of breast cancer. One of the possibility is modification during reverse-transcription process as it is a read-through fusion. Also this read-through fusion

transcript may be unstable than the wild type transcript easier to degrade in certain condition during the validation process.

We did not identify all of the FTs that had been previously reported in breast cancer. For example, in a recent next-generation sequencing analysis of breast cancer, Banerji et al. identified the recurrent fusion gene MAGI3-AKT3 (3.4% of all validation samples) as being enriched in triple-negative breast cancer, and suggested that this fusion gene could be a driver mutation that might be targeted by existing biological agents (Banerji et al., 2012). However, this FT was not found in our RNA-seq data from 120 breast cancer samples, regardless of which fusion-detecting tool we used (deFuse, Chimerascan or TopHatFusion). We further used Sanger sequencing to screen another 90 primary breast cancer samples, but failed to detect this FT (data not shown). A recent analysis of kinase fusions in 7,000 samples of several cancer types from TCGA RNA-Seq dataset also failed to identify this MAGI3-AKT3 fusion (Stransky et al., 2014). We have made personal contact to the corresponder and requested to re-check whether any of the breast samples represented the MAGI3-AKT3 fusion but none of the sample harbored this fusion. Also a more recently identified gene fusion, ESR1-CCDC170 (Veeraraghavan et al., 2014), was not identified in our study. It is unclear whether this apparent discrepancy reflects differences in the ethnicities of the studied patients, the utilized sequencing platform, or the fusion-detecting tools. Extensive tumor heterogeneity of breast cancer entity could be another reason as minor allelic fraction of the

passenger aberrations will be likely to be found by one algorithm and not by the other.

Read-through fusions, which take place between adjacent genes when they have the same coding orientation and are located relatively close to one another (Yun et al., 2013) occur during splicing and do not involve genomic rearrangements (Varley et al., 2014). While most of the previous fusion analyses have focused on those associated with genomic rearrangements, the existence of read-through chimeric fusions opens the door to a panoply of fusion events within the tumor, irrespective of their ability to drive tumorigenesis. In the future, integrative analyses of both genomes and transcriptomes will allow us to identify a more comprehensive repertoire of the gene fusions that exist among the various cancers. Furthermore, the recurrent fusions identified herein were relatively frequent in the tumor tissues (5-8%) and were not found the normal counterpart sample. Unlike in mammary tissue, the two read-through fusion transcripts were found in two other normal organs, lung and thyroid tissue. This suggests that this read-through fusion event is a byproduct of an alternative splicing which may exist in non-tumoral conditions. As it was free in normal mammary tissue (the counterpart), the isoform switching could perhaps be a tumor specific event in breast when other organs having different neutral isoform ratio. Thus, these fusions may prove generally useful for identifying mammary tumor cells and/or facilitating diagnosis.

In conclusion, we herein performed whole-transcriptome sequencing of 120 primary breast cancer samples, and sought to identify novel and recurrent FTs that could play potential driving roles in breast cancer. In total, we identified 516 to 6630 FTs depending on the utilized fusion-detecting tool. Thirty-two fusions were validated using Sanger sequencing, including the novel recurrent read-through fusion, EEF1DP3-FRY. Previously reported FTs, such as PPP1R1B-STARD3, MFGE8-HAPLN3 and ETV6-NTRK3, were also identified. Future work is warranted to elucidate the biological significance of these fusions.

References

Banerji S, Cibulskis K, Rangel-Escareno C, Brown KK, Carter SL, Frederick AM, Lawrence MS, Sivachenko AY, Sougnez C, Zou L, Cortes ML, Fernandez-Lopez JC, Peng S, Ardlie KG, Auclair D, Bautista-Pina V, Duke F, Francis J, Jung J, Maffuz-Aziz A, Onofrio RC, Parkin M, Pho NH, Quintanar-Jurado V, Ramos AH, Rebollar-Vega R, Rodriguez-Cuevas S, Romero-Cordoba SL, Schumacher SE, Stransky N, Thompson KM, Uribe-Figueroa L, Baselga J, Beroukhim R, Polyak K, Sgroi DC, Richardson AL, Jimenez-Sanchez G, Lander ES, Gabriel SB, Garraway LA, Golub TR, Melendez-Zajgla J, Toker A, Getz G, Hidalgo-Miranda A, Meyerson M. 2012. Sequence analysis of mutations and translocations across breast cancer subtypes. Nature 486:405-409.

Cong J, Geng W, He B, Liu J, Charlton J, Adler PN. 2001. The furry gene of Drosophila is important for maintaining the integrity of cellular extensions during morphogenesis. Development 128:2793-2802.

de Klein A, van Kessel AG, Grosveld G, Bartram CR, Hagemeijer A, Bootsma D, Spurr NK, Heisterkamp N, Groffen J, Stephenson JR. 1982. A cellular oncogene is translocated to the Philadelphia

chromosome in chronic myelocytic leukaemia. Nature 300:765-767.

Emoto K, He Y, Ye B, Grueber WB, Adler PN, Jan LY, Jan YN. 2004. Control of dendritic branching and tiling by the Tricornered-kinase/Furry signaling pathway in Drosophila sensory neurons. Cell 119:245-256.

Hergovich A, Stegert MR, Schmitz D, Hemmings BA. 2006. NDR kinases regulate essential cell processes from yeast to humans. Nat Rev Mol Cell Biol 7:253-264.

Hirata D, Kishimoto N, Suda M, Sogabe Y, Nakagawa S, Yoshida Y, Sakai K, Mizunuma M, Miyakawa T, Ishiguro J, Toda T. 2002. Fission yeast Mor2/Cps12, a protein similar to Drosophila Furry, is essential for cell morphogenesis and its mutation induces Wee1-dependent G(2) delay. EMBO J 21:4863-4874.

Hofvander J, Tayebwa J, Nilsson J, Magnusson L, Brosjo O, Larsson O, von Steyern FV, Domanski HA, Mandahl N, Mertens F. 2015. RNA sequencing of sarcomas with simple karyotypes: identification and enrichment of fusion transcripts. Lab Invest.

Ju YS, Lee WC, Shin JY, Lee S, Bleazard T, Won JK, Kim YT, Kim JI, Kang JH, Seo JS. 2012. A transforming KIF5B and RET gene fusion in

lung adenocarcinoma revealed from whole-genome and transcriptome sequencing. Genome Res 22:436-445.

Kalyana-Sundaram S, Shankar S, Deroo S, Iyer MK, Palanisamy N, Chinnaiyan AM, Kumar-Sinha C. 2012. Gene fusions associated with recurrent amplicons represent a class of passenger aberrations in breast cancer. Neoplasia 14:702-708.

Kangaspeska S, Hultsch S, Edgren H, Nicorici D, Murumagi A, Kallioniemi O. 2012. Reanalysis of RNA-sequencing data reveals several additional fusion genes with multiple isoforms. PLoS One 7:e48745.

Kannan K, Wang L, Wang J, Ittmann MM, Li W, Yen L. 2011. Recurrent chimeric RNAs enriched in human prostate cancer identified by deep sequencing. Proc Natl Acad Sci U S A 108:9172-9177.

Knezevich SR, McFadden DE, Tao W, Lim JF, Sorensen PH. 1998. A novel ETV6-NTRK3 gene fusion in congenital fibrosarcoma. Nat Genet 18:184-187.

McPherson A, Hormozdiari F, Zayed A, Giuliany R, Ha G, Sun MG, Griffith M, Heravi Moussavi A, Senz J, Melnyk N, Pacheco M, Marra MA, Hirst M, Nielsen TO, Sahinalp SC, Huntsman D, Shah SP. 2011. deFuse: an algorithm for gene fusion discovery in tumor RNA-Seq

data. PLoS Comput Biol 7:e1001138.

Mertens F, Johansson B, Fioretos T, Mitelman F. 2015. The emerging complexity of gene fusions in cancer. Nat Rev Cancer 15:371-381.

Nagai T, Ikeda M, Chiba S, Kanno S, Mizuno K. 2013. Furry promotes acetylation of microtubules in the mitotic spindle by inhibition of SIRT2 tubulin deacetylase. J Cell Sci 126:4369-4380.

Robinson DR, Kalyana-Sundaram S, Wu YM, Shankar S, Cao X, Ateeq B, Asangani IA, Iyer M, Maher CA, Grasso CS, Lonigro RJ, Quist M, Siddiqui J, Mehra R, Jing X, Giordano TJ, Sabel MS, Kleer CG, Palanisamy N, Natrajan R, Lambros MB, Reis-Filho JS, Kumar-Sinha C, Chinnaiyan AM. 2011. Functionally recurrent rearrangements of the MAST kinase and Notch gene families in breast cancer. Nat Med 17:1646-1651.

Rowley JD. 1973. Letter: A new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and Giemsa staining. Nature 243:290-293.

Scagliotti G, Stahel RA, Rosell R, Thatcher N, Soria JC. 2012. ALK translocation and crizotinib in non-small cell lung cancer: an evolving paradigm in oncology drug development. Eur J Cancer

48:961-973.

Soda M, Choi YL, Enomoto M, Takada S, Yamashita Y, Ishikawa S, Fujiwara S, Watanabe H, Kurashina K, Hatanaka H, Bando M, Ohno S, Ishikawa Y, Aburatani H, Niki T, Sohara Y, Sugiyama Y, Mano H. 2007. Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer. Nature 448:561-566.

Stenman G, Andersson MK, Andren Y. 2010. New tricks from an old oncogene: gene fusion and copy number alterations of MYB in human cancer. Cell Cycle 9:2986-2995.

Stransky N, Cerami E, Schalm S, Kim JL, Lengauer C. 2014. The landscape of kinase fusions in cancer. Nat Commun 5:4846.

Takeuchi K, Choi YL, Togashi Y, Soda M, Hatano S, Inamura K, Takada S, Ueno T, Yamashita Y, Satoh Y, Okumura S, Nakagawa K, Ishikawa Y, Mano H. 2009. KIF5B-ALK, a novel fusion oncokinase identified by an immunohistochemistry-based diagnostic system for ALK-positive lung cancer. Clin Cancer Res 15:3143-3149.

Tomlins SA, Laxman B, Dhanasekaran SM, Helgeson BE, Cao X, Morris DS, Menon A, Jing X, Cao Q, Han B, Yu J, Wang L, Montie JE, Rubin MA, Pienta KJ, Roulston D, Shah RB, Varambally S, Mehra R,

Chinnaiyan AM. 2007. Distinct classes of chromosomal rearrangements create oncogenic ETS gene fusions in prostate cancer. Nature 448:595-599.

Tomlins SA, Rhodes DR, Perner S, Dhanasekaran SM, Mehra R, Sun XW, Varambally S, Cao X, Tchinda J, Kuefer R, Lee C, Montie JE, Shah RB, Pienta KJ, Rubin MA, Chinnaiyan AM. 2005. Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. Science 310:644-648.

Trapnell C, Pachter L, Salzberg SL. 2009. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25:1105-1111.

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol 28:511-515.

Varley KE, Gertz J, Roberts BS, Davis NS, Bowling KM, Kirby MK, Nesmith AS, Oliver PG, Grizzle WE, Forero A, Buchsbaum DJ, LoBuglio AF, Myers RM. 2014. Recurrent read-through fusion transcripts in breast cancer. Breast Cancer Res Treat 146:287-297.

Veeraraghavan J, Tan Y, Cao XX, Kim JA, Wang X, Chamness GC, Maiti SN, Cooper LJ, Edwards DP, Contreras A, Hilsenbeck SG, Chang EC, Schiff R, Wang XS. 2014. Recurrent ESR1-CCDC170 rearrangements in an aggressive subset of oestrogen receptor-positive breast cancers. Nat Commun 5:4577.

Yoshihara K, Wang Q, Torres-Garcia W, Zheng S, Vegesna R, Kim H, Verhaak RG. 2014. The landscape and therapeutic relevance of cancer-associated transcript fusions. Oncogene.

Yun SM, Yoon K, Lee S, Kim E, Kong SH, Choe J, Kang JM, Han TS, Kim P, Choi Y, Jho S, Yoo H, Bhak J, Yang HK, Kim SJ. 2013. PPP1R1B-STARD3 chimeric fusion transcript in human gastric cancer promotes tumorigenesis through activation of PI3K/AKT signaling. Oncogene.

요약(국문초록)

연구목적: 융합유전자는 고전적으로 혈액암에서 효과적인 치료의 표적이 되어왔으나 고형암에서 반복적인 두 유전자의 융합은 드물다. 본 연구에서는 유방암에서의 의미 있는 융합유전자를 찾기 위해 유방암 조직에서 전장 전사체 서열 분석을 시행하였다.

연구방법: 유방암으로 진단된 120명의 환자의 수술 당시 채취되어 얼려진 유방암 조직과 5명에서의 정상조직에서 RNA를 추출하여 Illumina HiSeq 2000 (Illumina, California, USA) 을 이용하여 전장 전사체 서열 분석을 시행하였다. 세 개의 독립적인 융합 전사체를 발굴하는 알고리즘(deFuse, ChimeraScan, TopHat)을 통해 분석 비교하였다.

연구결과: 각각의 알고리즘을 통해 3831, 6630, 516 개의 융합 전사체가 발견되었고 일차적으로 deFuse 를 통해 발견된 융합전사체를 분석하였다. HER2 유전자의 증폭이 있는 유방암의 아형에서 융합유전체의 수가 호르몬수용체 양성유방암이나 삼중음성 유방암에서보다 유의하게 많았다($p<0.05$). 3800 여개의 융합전사체 중 70 개의 후보를 선별하여 RT-PCR 및 Sanger sequencing 으로 32 개(45.7%)의 융합전사체를 검증하였다. 검증이 된 융합전사체 중 6 개가 2 개 이상의 샘플에서

발견된 반복성 융합이었고 3 개가 in-frame (PRDX1-AKR1A1, TACSTD2-OMA1, C2CD2-TFF1), 3 개가 off-frame (CEACAM7-CEACAM6, CYP4X1-CYP4Z2P, EEF1DP3-FRY) 이었다. 특히, 이전에 보고된 바 없는 read-through 융합 전사체인 EEF1DP3-FRY 는 6.7%(8/120)의 샘플에서 확인이 되었다. 이 off-frame 융합으로 인해 세포분열 과정에서의 구조적 안정을 유지시키는 FRY gene 이 frameshift 로 인한 'early trunction'이 초래된다. 기존의 연구를 통해 밝혀진 세 개의 PPP1R1B-STARD3, MFGE8-HAPL, ETV6-NTRK3 의 융합전사체는 각각 8.3%, 3.3%, 0.8%에서 확인되었다. 최근에 보고된 MAGI3-AKT3 융합전사체는 본 샘플에서는 발견되지 않았다.

결론: 전장 전사체 서열 분석을 통해 발굴된 3800 여개 융합 전사체 중 32 개가 RT-PCR 및 Sanger sequencing 으로 검증되었고 이 중 기존에 보고된 바 없는 read-through 융합 EEF1DP3-FRY 도 존재하였다. 더불어 기존에 보고된 PPP1R1B-STARD3, MFGE8-HAPLN3 및 ETV6-NTRK3 또한 확인되었다. 이러한 융합 전사체들의 생물학적 기능에 대한 검증이 필요하겠다.

주요어: 융합 유전자, 융합 전사체, 전장 전사체 서열 분석, RNA-Seq, 유방암

학번: 2012-30493