Supplementary Information

Fusion Transcripts and Transcribed Retrotransposed Loci Discovered through Comprehensive Transcriptome Analysis using Paired-End diTagging (PET)

Yijun Ruan, Hong Sain Ooi, Siew Woh Choo, Kuo Ping Chiu, Xiao Dong Zhao, K.G. Srinivasan, Fei Yao, Chiou Yu Choo, Jun Liu, Pramila Nuwantha, Wilson G. W. Bin, Vladimir A. Kuznetsov, Atif Shahab, Wing-Kin Sung, Guillaume Bourque, Nallasivam Palanisamy, Chia-Lin Wei

TABLE OF CONTENT **Page**

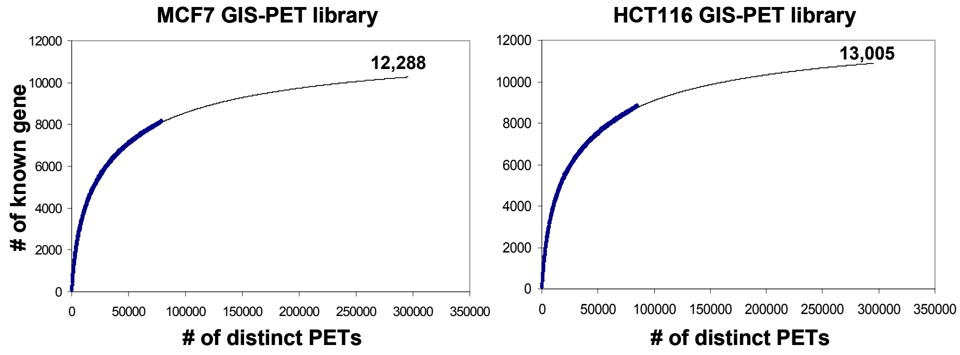
Supplemental Figures: S1-S10

Supplemental Tables: S1, S2 and S8

Supplemental Tables S 3-7 are listed separately and not included here.

I. Statistic Analysis of PET Reproducibility and Saturation of GIS-PET libraries

To estimate the level of coverage and saturation of these two GIS-PET libraries, we use Monte-Carlo-base sampling of distinct PET sequences (PET-1) to generate sub-libraries of different sizes (Kuznetsov et al. 2002). This approach allows recovering the kinetics curve of the number of the distinct transcripts in the library when the library size becomes randomly large. Figure S1 shows that the both of the MCF7 (left panel) and HCT116 (right panel) libraries are quite saturated. Using extrapolation of best-fit Hill function (see parameters of the function in the tables below), we estimate that 13,005 genes and 12,288 genes could be identified at the detectable level when the sizes of the libraries for MCF7 and HCT116 cells increase to 300,000 and more distinct PET sequences. With the numbers of known genes identified so far in these two libraries (9,240 in MCF7 and 8,923 in HCT116 cells), we estimate 75% and 69% of saturation in these two libraries, respectively. The number of genes estimated here agree to the numbers obtained base on MPSS and SAGE datasets analysis for breast and colon cancer cells (Grigoriadis et al. 2006; Jongeneel et al. 2003; Kuznetsov, 2002).



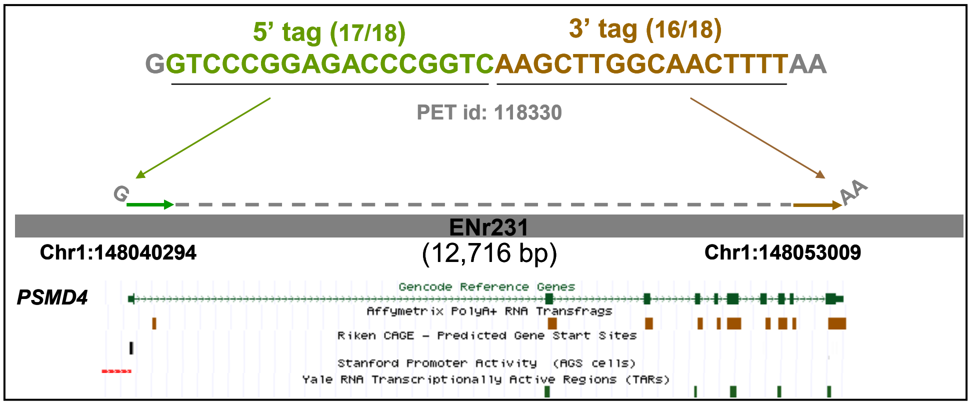
f=Ntot*x^b/(c^b+x^b)	Coefficient	Std. Error	t	P
Ntot	12288	42.3	290.4	<0.0001
b	0.7323	0.0023	316.1	<0.0001
c	32140.668	321.4819	100	<0.0001

f=Ntot*x^b/(c^b+x^b)	Coefficient	Std. Error	t	P
Ntot	13005	46.5	280.2	<0.0001
b	0.735	0.002	290.4	<0.0001
c	31856	331.3	96.2	<0.0001

Figure S1. Estimation of the number of expressed genes can be found in HCT116 and MSF7 cells. N_{tot}, b and c are the parameters of the Hill function. N_{tot}: an estimate of the total number of expressed genes

II. Alignment analysis to recover unmappable PET sequences (PET-0)

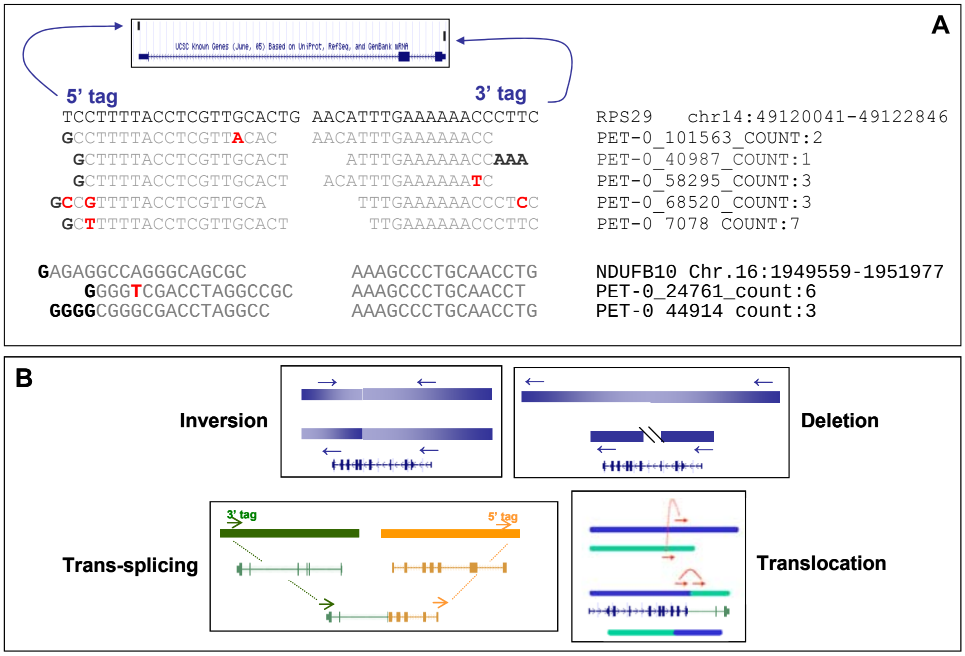
In this GIS-PET analysis, PET sequences derived from conventional transcripts are mapped to reference genome using standard mapping criteria (Ng, et al 2005). An example is shown in Figure S2.


5’ tag (17/18)
3’ tag (16/18)
GGTCCCGGAGACCCGGTCAAGCTTGGCAACTTTTAA
PET id: 118330
G
AA
ENr231
Chr1:148040294
(12,716 bp)
Chr1:148053009
PSMD4
Gencode Reference Genes
Affymetrix PolyA+ RNA Transfrags
Riken CAGE - Predicted Gene Start Sites
Stanford Promoter Activity (AGS cells)
Yale RNA Transcriptionally Active Regions (TARs)

Supplemental Figure S2. Example of standard mapping of PET sequences to the reference genome.

A typical structure of a PET sequence contains a 5’ tag of 18bp and a 3’ tag of 18bp. In the 5’ tag, often time the first nucleotide is a G resulted by the untemplated activity of RNA reverse transcriptase during cDNA synthesis. The 3’ tag, by design, contains AA, residual of the polyA tail of mRNA. The residual AA provides the orientation of PET sequence corresponding to transcript. In this example, a 36 bp GIS-PET mapped to ENCODE region ENr231. 17 out of 18 bp 5’ tag and 16 out of 18 bp 3’ tag mapped to chromosome 1 between 148040294 – 148053009 with span 12,716 bp. This region encodes the *PSMD4* gene as annotated by Gencode reference gene track. The corresponding exons were detected by Transfrag (Affymetrix) and TARs (Yale) were shown in the tracks below. Its 5’ transcription start site has also been mapped by Riken’s CAGE tag, and its promoter activity has been registered by Stanford promoter assay.

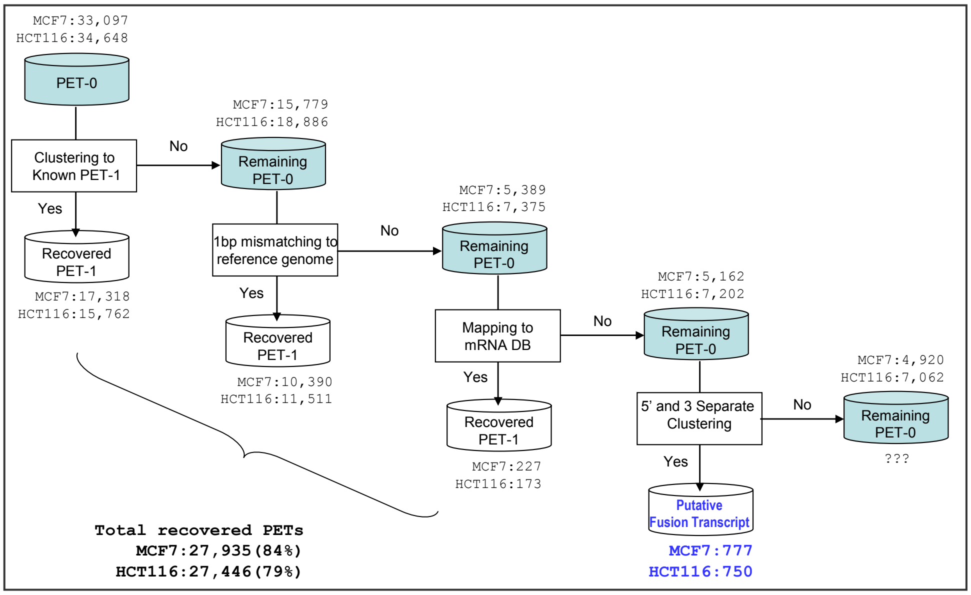
However, 24% of total PET sequences could not be mapped to the reference genome using the standard mapping criteria. The main causes for the lack of mapping are sequencing errors in PET tags, nucleotide polymorphisms between the reference genome and the target genome and probably unconventional fusion transcripts derived from either trans-splicing or translocation (Supplemental Figure S3), as well as imperfection of the assembled reference genome as we described previously (Ng et al. 2005).


A
UCSC Known Genes (June, 05) Based on UniProt, RefSeq, and GenBank mRNA
5' tag
3' tag
TCCTTTTACCTCGTTGCACTG AACATTTGAAAAAACCCTTC RPS29 chr14:49120041-49122846
GCCTTTTACCTCGTTACAC AACATTTGAAAAAACC PET-0_101563_COUNT:2
GCTTTTACCTCGTTGCACT ATTTGAAAAAACCAAA PET-0_40987_COUNT:1
GCTTTTACCTCGTTGCACT ACATTTGAAAAAATC PET-0_58295_COUNT:3
GCCGTTTACCTCGTTGCA TTTGAAAAAACCCTCC PET-0_68520_COUNT:3
GCTTTTTACCTCGTTGCACT TTGAAAAAACCCTTC PET-0_7078_COUNT:7
GAGAGGCCAGGGCAGCGC AAAGCCCTGCAACCTG NDUFB10 Chr.16:1949559-1951977
GGGGTCGACCTAGGCCGC AAAGCCCTGCAACCT PET-0_24761_count:6
GGGGCGGGCGACCTAGGCC AAAGCCCTGCAACCTG PET-0_44914 count:3
B
Inversion
Deletion
3' tag
5' tag
Trans-splicing
Translocation

Supplemental Figure S3. Categories of PET-0

(A). Alignment of unmappable PETs to the reference 5' and 3' regions of *RPS29* gene sequence to show different types of errors occurred during the PET cloning and sequencing steps. Nucleotides highlighted in red showed the sequence errors and nucleotides highlighted in bold are either resulted from non template G nucleotide incorporation in the 5' end or residual poly As in the 3' end. **(B).** Types of PET-0 created by fusion transcripts resulted from mRNA trans-splicing or genomic rearrangements as shown by 5' and 3' tags mapping in reverse orientation, ultra long span or on different chromosomes.

To recover the PET-0 sequences caused by potential sequence errors and variation between the reference genome and the target genome for analysis, we implemented a PET-0 processing pipeline as outlined in Figure S4.



Supplemental Figure S4. The outline of PET-0 sequence recovering pipeline

Three major steps were taken to remap PET-0s and identify putative fusion transcripts. The first step is the clustering of all PET-0 sequences with the PET-1s. This alignment will correct most of sequencing errors in PET-0 sequences. The second step is to allow one nucleotide mismatch for 5' or 3' tag during the mapping to genome sequences. The third step is mapping the remaining PET-0 to the mRNA sequences in Genbank database. PET clusters mapped independently on two different chromosomes were then considered as putative unconventional fusion transcripts.

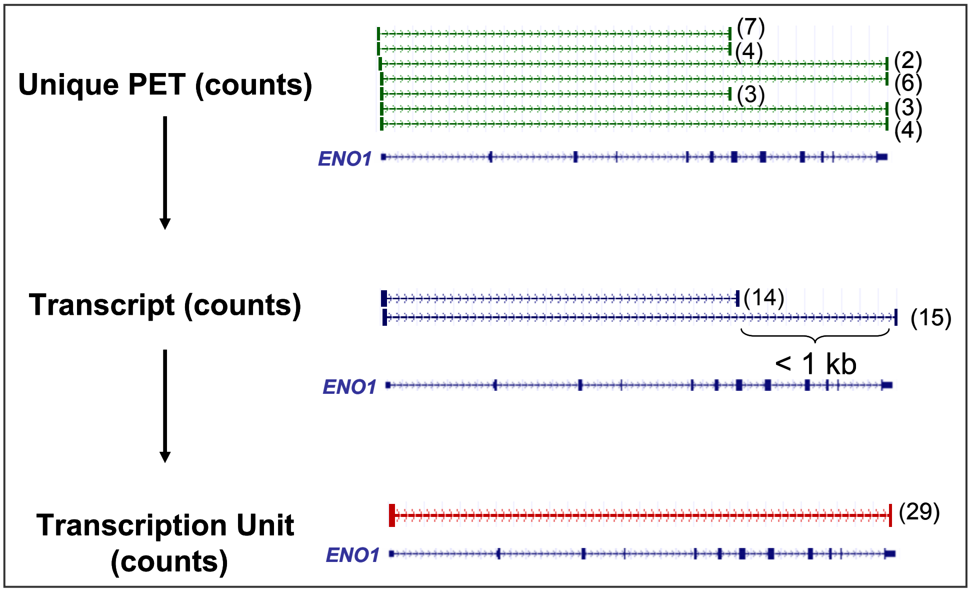
First, we assessed how many of the PET-0 sequences were due to sequencing error. An incorrect base calling in PET sequences can be readily identified by aligning the PET-0 sequences to the PET sequences that had been correctly mapped to the reference genome sequences. By aligning the 33,097 and 34,648 PET-0 sequences to PET-1 sequences, 17,318 (52.3%) from MCF7 and 15,762 (45.5%) from HCT116 cells could be aligned to PET-1 using a relaxed matching criteria. We tested various alignment criteria and determined that 5' minimal 12 bp and 3' minimal 10 bp match are the optimal criteria for aligning PET-0 sequences to PET-1 sequences with the highest percentage of supports by known gene evidence. Over 99% (17,178 and 15,570) of these reclaimed PET-0s were supported with known genes, most of which are abundantly expressed. A large proportion of these PET-0 sequences (63%) contain one mismatched nucleotide to the correct PET sequences (Supplemental Figure S3A), which reflects the 0.2% error rate in sequencing (equivalent to the phred QV 26). The remaining PET-0 sequences are probably due to nucleotide polymorphism between the transcript sequences from these cancer cell lines and the reference human genome sequences or other reasons. Therefore, by allowing a single nucleotide mismatch when mapping the remaining PET-0 sequences to the reference genome, the PET sequences with SNPs should become mappable. Indeed, 10,390 (31.4%) PET-0s from MCF7 and 11,511 (33.2%) PET-0s from HCT116 were remapped by allowing one nucleotide mismatch at either 5' or 3' tag sequences (Figure S4). The coordinates of 78.4% (8,147/10,390) and 72.8% (8,381/11,511) of these PETs were unique (PET-1) and supported by known genes, refseq and Genbank mRNA.

To search for reliable transcript evidence independently of the reference genome sequences, the remaining PET-0s were BLASTed to the first and last 100 bases of a collection of mRNA sequences which included Genbank mRNA (113,552 entries) and Refseq genes (41,730 entries) generated from a variety of genomic backgrounds. Only 227 and 173 PET-0s could be mapped. When we examined the alignments between these PETs and their corresponding mRNA by aligning them back to the hg17 genomic sequence, we found that most of them contained 2 independent SNPs in the tag sequences. For example, *IFITM2* mapped by 3 PETs, the 5' tag sequence is GGTAACCCGA**C**C**G**CCGCT whereas the sequence on the reference genome is GGTAACCCGA**T**C**A**CCGCT. The other cases were tags from either end mapped between 2 exons. *RPS15a* mapped by 7 copies of PETs, of which the 5' tags mapped to mRNA from the 12th to 29th nucleotides bridging the first exon (nucleotide 1st -25th) and second exon (nucleotide 26th -138th). Together, by combining the recoverable PET-0 sequences using one nucleotide mismatch to the reference genome and alignment to mRNA databases, we estimated that there is a variation of approximately 6 nucleotides in every 10,000 bp of gene coding sequences between the two cancer cell genomes and the reference genome sequences.

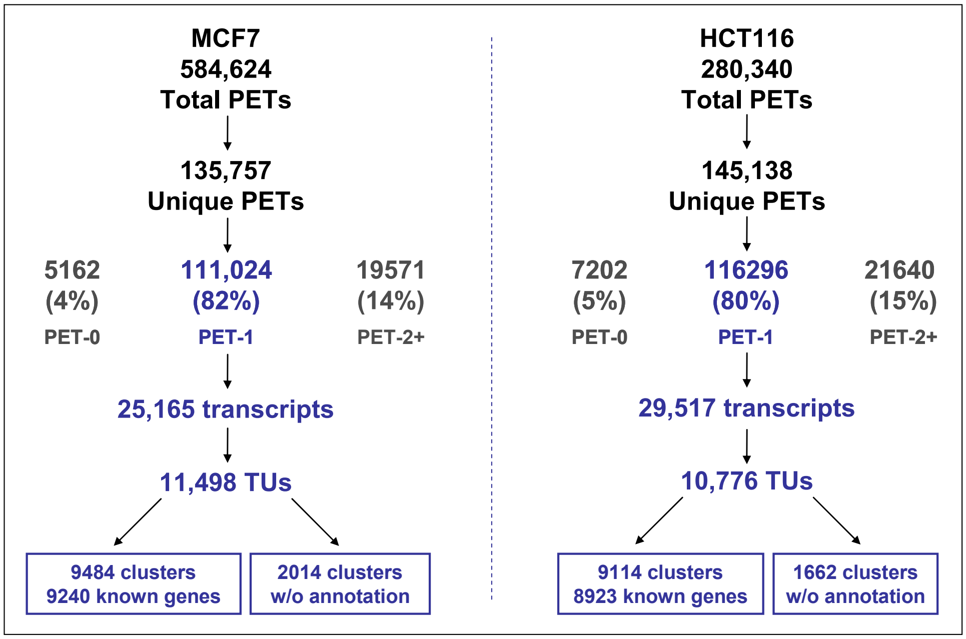
In summary, the PET-0 processing pipeline reclaimed 84.4% and 79.2% of total PET-0s from these two libraries to be mapped the human reference genome. Furthermore, ~90% of these reclaimed mappable PETs (previously PET-0s) had mRNA or known gene support, indicating that this PET-0 analysis pipeline is accurate and effective. The total numbers of unmappable PETs remaining after this processing pipeline were only 5,162 in MCF7 and 7,202 in HCT116 cells, representing only 3.8% and 4.96% out of total unique PETs analyzed. In other words, more than 95% of GIS-PET sequences could be mapped and assigned to appropriate references based on the current genome and mRNA databases.

III. Transcriptome analysis by uniquely mapped PET sequences (PET-1)

PET sequences derived from same transcript species often have slight variations of nucleotide sequences due to enzymatic slippery during PET construction. The exact transcript molecules of same species can also vary in term of length due to minor inconsistency of transcription initiation and stability. PET sequences with such minor variations will map closely along the genome. If the two ends of PET sequences directly overlap to each other, these PETs will be collapsed together to represent a specific transcript variant. If the PETs overlap to each other at one end but not at the other end, these PETs are considered to represent alternative transcripts of the same gene, or belong to the same transcription unit.



Supplemental Figure S5. Transcript and transcription unit (TU) defined by GIS-PET Group of unique PETs are clustered to define transcripts and transcription units based on the proximity of their mapping locations. As an example shown here, 7 unique PETs derived from ENO1 were grouped into 2 distinct transcripts using 1 bp overlap of their 5' and 3' tag mapping locations as criteria. These transcripts were further clustered to represent transcription unit if either their 5' or 3' ends were mapped within 1 kb distance. Total of 29 copies of PETs were generated from this TU.



Supplemental Figure S6. Transcripts and TUs expressed in MCF7 and HCT116
Total GIS-PETs from MCF7 and HCT116 cells were generated and the unique PETs were mapped to the hg17 reference genome. PETs were separated into 3 categories based on the number of their mapping coordinates and PETs mapped to single locations were further clustered to form transcripts and transcription units. Based on the public annotation information, these gene clusters were annotated to either known genes or clusters without annotation.

Supplemental Table S1

Transcript variants uncovered by GIS-PET in the ENCODE regions

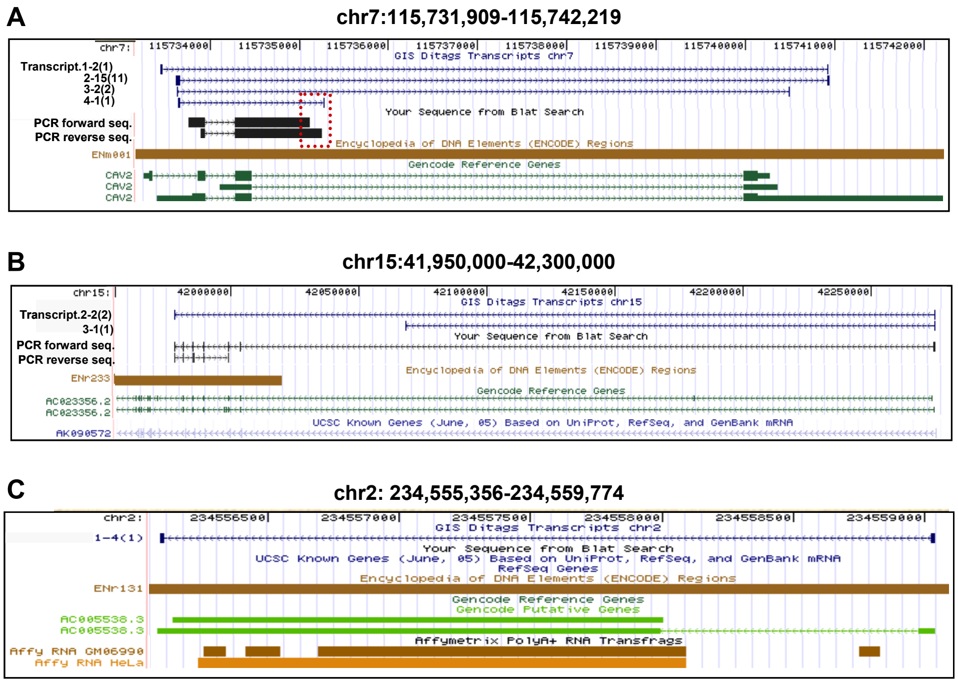
Transcripts	5' Alternative TSS	3' Alternative PAS	Novel
Supported by	**137**	**183**	**99**
Gencode	64	103	7
Transfrag	94	114	23
TARs	78	105	13
Total supported	**119 (86.9%)**	**142 (77.6%)**	**43 (43.4%)**
No support	18	41	56
Candidates for PCR validation	11	19	17
PCR/sequence confirmed	**10 (90%)**	**17 (90%)**	**13 (76%)**

Supplemental Table S2. PET sequence locations relative to known 5' TSS and 3' PAS PET sequences mapped to top 20 most abundant transcripts are used to evaluate the intactness of transcripts represented by PET sequences.

		Non-specific mapping	≥50 bp shorter than known TSS	± 50bp around known TSS	≥50 bp longer than known TSS	Longer than known TSS
MCF7						
Total	**61569**					
5'	# of PETs	7	251	39815	21496	**61143**
	%	0.01%	0.4%	64.67%	34.91%	**99.3%**
			≥50 bp shorter than known PAS	**± 50bp around known PAs**	**≥50 bp longer than known PAS**	**Longer than known PAS**
3'	# of PETs	89	4496	56728	256	**56880**
	%	0.14%	7.30%	92.14%	0.42%	**92.38%**
		Non-specific mapping	**≥50 bp shorter than known TSS**	**± 50bp around known TSS**	**≥50 bp longer than known TSS**	**Longer than known TSS**
HCT116						
Total	**23704**					
5'	# of PETs	0	318	16570	6816	**23116**
	%	0	1.30%	70%	28.80%	**97.50%**
			≥50 bp shorter than known PAS	**± 50bp around known PAs**	**≥50 bp longer than known PAS**	**Longer than known PAS**
3'	# of PETs	33	1260	22387	24	**22349**
	%	0.14%	5.32%	94.44%	0.10%	**94.28%**

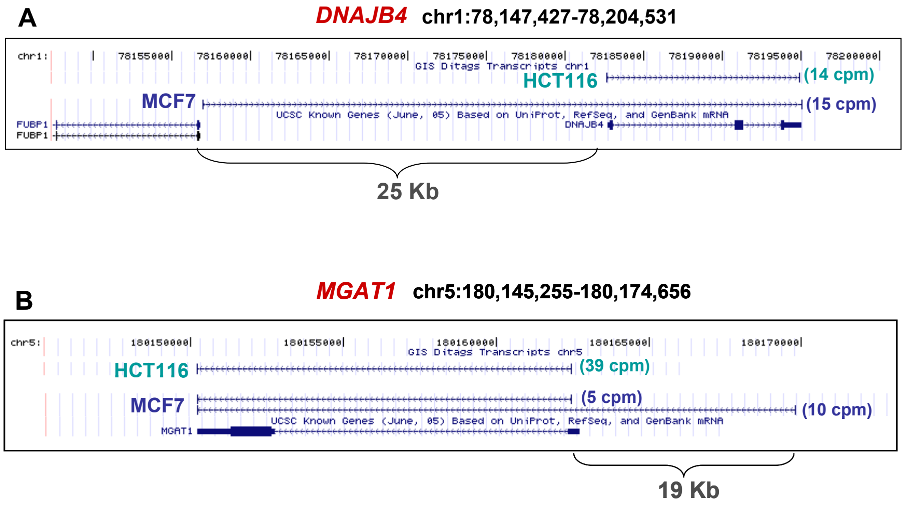
Identification of alternative TSS in MCF7 and HCT116 cells

1703 and 2274 5' alternative TSS were found that were located other than the 1st exon for MCF7 and HCT116 cells, respectively. 315 and 396 of them were overlapped by extending 250 bp to both ends when compared with 30,964 representative alternative promoters defined by 5' end of oligo-cap cDNAs (http://dbtss.hgc.jp). The complete list of TSS for each cell type and the overlapping loci are listed as Supplemental Table S3 (MCF7) and Table S4 (HCT116).


A
chr7:115,731,909-115,742,219
Transcript.1-2(1)
2-15(11)
3-2(2)
4-1(1)
PCR forward seq.
PCR reverse seq.
GIS Ditags Transcripts chr7
Your Sequence from Blat Search
Encyclopedia of DNA Elements (ENCODE) Regions
ENm001
Gencode Reference Genes
CAV2
CAV2
CAV2
B
chr15:41,950,000-42,300,000
Transcript.2-2(2)
3-1(1)
PCR forward seq.
PCR reverse seq.
GIS Ditags Transcripts chr15
Your Sequence from Blat Search
Encyclopedia of DNA Elements (ENCODE) Regions
ENr233
Gencode Reference Genes
AC023356.2
AC023356.2
UCSC Known Genes (June, 05) Based on UniProt, RefSeq, and GenBank mRNA
AK090572
C
chr2: 234,555,356-234,559,774
1-4(1)
GIS Ditags Transcripts chr2
Your Sequence from Blat Search
UCSC Known Genes (June, 05) Based on UniProt, RefSeq, and GenBank mRNA
RefSeq Genes
Encyclopedia of DNA Elements (ENCODE) Regions
ENr131
Gencode Reference Genes
Gencode Putative Genes
AC005538.3
AC005538.3
Affymetrix PolyA+ RNA Transfrags
Affy RNA GM06990
Affy RNA HeLa

Supplemental Figure S7. Validation of new transcripts and alternative PAS identified in ENCODE regions

(A) and **(B)** Examples of transcripts with new 3' alternative polyA site identified by GIS-PETs (blue track), which was subsequently validated by PCR and sequence analysis (black track). The PCR cDNA sequences aligned with GIS-PET and Gencode reference genes. **(C).** New transcript detected by GIS-PET (blue track) and supported by Gencode putative gene prediction (green track).


A
DNAJB4 chr1:78,147,427-78,204,531
GIS Ditags Transcripts chr1
HCT116
(14 cpm)
MCF7
(15 cpm)
UCSC Known Genes (June, 05) Based on UniProt, RefSeq, and GenBank mRNA
DNAJB4
FUBP1
FUBP1
25 Kb
B
MGAT1 chr5:180,145,255-180,174,656
GIS Ditags Transcripts chr5
HCT116
(39 cpm)
MCF7
(5 cpm)
(10 cpm)
UCSC Known Genes (June, 05) Based on UniProt, RefSeq, and GenBank mRNA
MGAT1
19 Kb

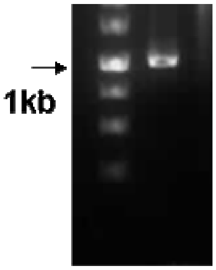
Supplemental Figure S8. Differential promoter usage indicated by alternative TSS
Two genes; *DNAJB4* **(A)** and *MGAT1* **(B)** with novel alternative TSS differentially expressed between cancer cells MCF7 and HCT116 are shown. The distances between the start sites of two transcript variants are marked and expression levels of individual variants are represented by cpm (copies per million transcripts).

IV. Validation of fusion transcripts

In total, there were 70 fusion transcripts candidates identified in this study (Supplemental table ST5). The top 20 fusion transcript PET clusters were validated by RT-PCR. PCR forward and reverse primers were designed based on the PET consensus sequences and were used to amplify the fusion cDNAs from amplified cDNA libraries. Out of 20 PCR reactions, 17 (85%) were positive and subjected to sequencing analysis, and 7 (4 from MCF7 cells and 3 from HCT116 cells) were confirmed to be derived from predicted fusion genes. The PCR primers, conditions, completed sequences and Blat analysis are described below.

1. *CXorf15/SYAP1*

```
5′ primer: GTCACCTCATGGCGACGCGGGTAGAGGA
5′ nested primer: GGCAGCGCGGGGAAGAGGCGGCGGCGCC
3′ primer: TAAATTCTAGATGTTGACAATTACTGAA
3′ nested primer: TAATCCTTAACAGTCTGCAAACTGACAT
```



The PCR product is shown on the side.

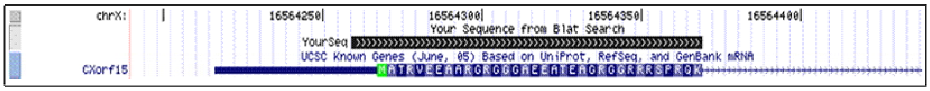
The *CXorf15/SYAP1* fusion transcript sequence:

GTCACCTCATGGCGACGCGGGTAGAGGAGGCAGCGCGGGGAAGAGGCGGCGGCGCCGAAGAGGCGACT
GAGGCCGGACGGGGCGGACGGCGACGCAGCCCGCGGCAGAAGACTATTTAT(T)TAACTTTGCATCTG
CTGCCACAAAAAAAGATAACTGAATCAGTTGCTGAAACAGCACAAACAATAAAGAAATCCCGTAGAAG
AAGGAAAAAATAGATGGCATCATTGACAAGACAATTATAGGAGATTTTCAGAAGGAACAGAAAAAATTT
GTTGAAGAGCAACATACAAAGAAGTCAGAAGCAGCTGTGCCCCCATGGGTTGACACTAACGATGAAGA
AACAATTCAACAACAAATTTTGGCCCTATCAGCTGACAAGAGGAATTTCCTTCGTGACCCTCCGGCTG
GCGTGCAATTTAATTTCGACTTTGATCAGATGTACCCCGTGGCCCTGGTCATGCTCCAGGAGGATGAG
CTGCTAAGCAAGATGAGATTTGCCCTCGTTCCTAAACTTGTGAAGGAAGAAGTGTTCTGGAGGAACTA
CTTTTACCGCGTCTCCCTGATTAAGCAGTCAGCCCAGCTCACGGCCCTGGCTGCCCAACAGCAGGCCG
CAGGGAAGGAGGAGAAGAGCAATGGCAGAGAGCAAGATTTGCCGCTGGCAGAGGCAGTACGGCCCAAA
ACGCCACCCGTTGTAATCAAATCTCAGCTTAAAACTCAAGAGGATGAGGAAGAAATTTCTACTAGCCC
AGGTGTTTCTGAGTTTGTCAGTGATGCCTTCGATGCCTGTAACCTAAATCAGGAAGATCTAAGGAAAG
AAATGGAGCAACTAGTGCTTGACAAAAAGCAAGAGGAGACAGCCGTACTGGAAGAGGATTCTGCAGAT
TGGGAAAAAGAACTGCAGCAGGAACTTCAAGAATATGAAGTGGTGACAGAATCTGAAAAACGAGATGA
AAACTGGGATAAGGAAATAGAGAAAATGCTTCAAGAGGAAAATTAGCTGTTCCTGAAATAGAAGAATA
ATCCTTAACAGTCTGCAAACTGACAT**TAAATTCTAGATGTTGACAATTACTGAA**

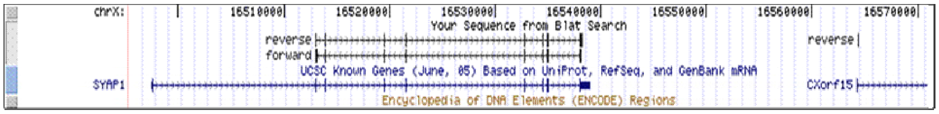
Note: Letters in red are PCR primers

Blat analysis:
The obtained cDNA sequence was BLATed to human genoem sequence (hg17) at the UCSC genome browser site (http://genome.ucsc.edu/). The screen shots of the BLAT result is shown below.

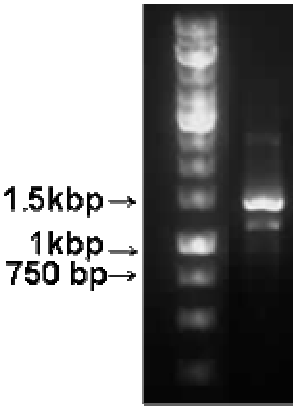
5' chrX:16,564,191-16,564,438


chrX:
16564250|
16564300|
16564350|
16564400|
Your Sequence from Blat Search
YourSeq
UCSC Known Genes (June, 05) Based on UniProt, RefSeq, and GenBank mRNA
CXorf15
MATRVEEAARGRGGGAEEATEAGRGGRRRSPRQK

3' chrX:16,491,598-16,576,592


chrX:
16510000|
16520000|
16530000|
16540000|
16550000|
16560000|
16570000|
Your Sequence from Blat Search
reverse
forward
reverse
UCSC Known Genes (June, 05) Based on UniProt, RefSeq, and GenBank mRNA
SYAP1
CXorf15
Encyclopedia of DNA Elements (ENCODE) Regions

2. *RPS6KB1/TMEM49* fusion



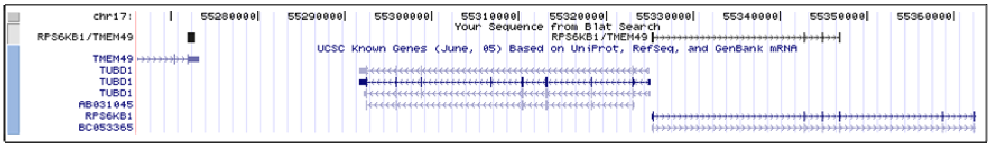
```
5′ primer: CACTGCCCACTGTTTGGCTTCACGGAAC
5′ nested primer: CCTGTACGCATGCTCCTACGCTGAACTT
3′ primer: TTTATAAAGACATCTTTAATCATTCCAA
3′ nested primer: GTATTTAGAATGAAGTCTTGAAAAAAAC
```

The *RPS6KB1/TMEM49* fusion transcript sequence:

```
CACTGCCCACTGTTTGGCTTCACGGAACCCTGTACGCATGCTCCTA
CGCTGAACTTTAGGAGCCAGTCTAAGGCCTAGGCGCAGACGCACTG
AGCCTAAGCAGCCGGTGATGGCGGCAGCGGCTGTGGTGGCTGCGGC
GGGTCCGGGCCCATGAGGCGACGAAGGAGGCGGGACGGCTTTTACC
CAGCCCCGGACTTCCGAGACAGGGAAGCTGAGGACATGGCAGGAGT
GTTTGACATAGACCTGGACCAGCCAGAGGACGCGGGCTCTGAGGATGAGCTGGAGGAGGGGGTCAGT
TAAATGAAAGCATGGACCATGGGGGAGTTGGACCATATGAACTTGGCATGGAACATTGTGAGAAATTT
GAAATCTCAGAAACTAGTGTGAACAGAGGGCCAGAAAAAATCAGACCAGAATGTTTTGAGCTACTTCG
GGTACTTGGTAAAGGGGGCTATGGAAAGGTTTTTCAAGTACGAAAAGTAACAGGAGCAAATACTGGGA
AAATATTTGCCATGAAGGTGCTTAAAAAGGGAGAAAACTGGTTGTCCTGGATGTTTGAAAAGTTGGTC
GTTGTCATGGTGTGTTACTTCATCCTATCTATCATTAACTCCATGGCACAAAGTTATGCCAAACGAAT
CCAGCAGCGGTTGAACTCAGAGGAGAAAACTAAATAAGTAGAGAAAGTTTTAAACTGCAGAAATTGGA
GTGGATGGGTTCTGCCTTAAATTGGGAGGACTCCAAGCTGGGAAGGAAAATTCCCTTTTCCAACCTGT
ATCAATTTTTACAACTTTTTTCCTGAAAGCAGTTTAGTCCATACTTTGCACTGACATACTTTTTCCTT
CTGTGCTAAGGTAAGGTATCCACCCTCGATGCAATCCACCTTGTGTTTTCTTAGGGTGGAATGTGATG
TTCAGCAGCAAACTTGCAACAGACTGGCCTTCTGTTTGTTACTTTCAAAAGGCCCACATGATACAATT
AGAGAATTCCCACCGCACAAAAAAAGTTCCTAAGTATGTTAAATATGTCAAGCTTTTTAGGCTTGTCA
CAAATGATTGCTTTGTTTTCCTAAGTCATCAAAATGTATATAAATTATCTAGATTGGATAACAGTCTT
GCATGTTTATCATGTTACAATTTAATATTCCATCCTGCCCAACCCTTCCTCTCCCATCCTCAAAAAAG
GGCCATTTTATGATGCATTGCACACCCTCTGGGGAAATTGATCTTTAAATTTTGAGACAGTATAAGGA
AAATCTGGTTGGTGTCTTACAAGTGAGCTGACACCATTTTTTATTCTGTGTATTTAGAATGAAGTCTT
GAAAAAAACTTTATAAAGACATCTTTAATCATTCCAA
```

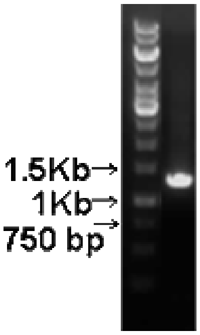
Blat analysis:

chr17:55,266,188-55,362,498



3. *CXorf53/FUNDC2*

5′ primer: AAGGGTCGGGCCAAGATGGCGGTGCAGG
5′ nested primer: TGGTGCAGGCGGTGCAGGCGGTTCATCT
3′ primer: TGAATACATTTGCCTTTAAGCATGAAAG
3′ nested primer: CCTTATTGTGTGCTGCATGGAAAGGAAC

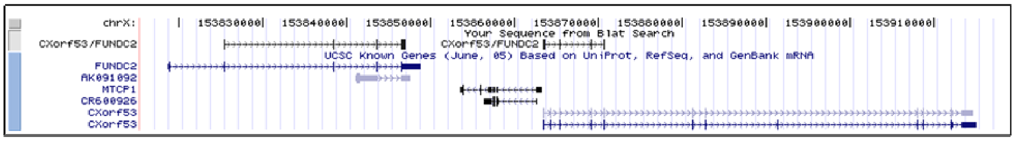


The *CXorf53/FUNDC2* fusion transcript sequence:

AAGGGTCGGGCCAAGATGGCGGTGCAGGTGGGCAGGCGGTGCAGGCGGTTCAT
CTCGAGTCTGACCTTTCCTCGTTTGTCTCAACCACGCTCTGAGCACAGAGAAG
GAGGAAGTAATGGGGCTGTGCATAGGGGAGTTGAACGATGATACAAGAGTGAC
TCCAAATTTGCATATACTGGAACTGAAATGCGCACAGTTGCTGAAAAGGTTGATGCCGTCAGAATTGT
TCACATTCATTCTGTCATCATCTTACGACGTTCTGATAAGAGGAAGGACCGAGTAGAAATTTCTCCAG
AGCAGCTGTCTGCAGCTTCAACAGAGGCAGAGAGGTTGGCTGAACTGACAGGCCGCCCCATGAGAGTT
GTGGGCTGGTATCATTCCCATCCTCATATAACTGTTTGGCCTTCACATGTTGGAAACTTTGAGGGAAA
TTTTGAGTCACTGGACCTTGCGGAATTTGCTAAGAAGCAGCCATGGTGGCGTAAGCTGTTCGGGCAGG
AATCTGGACCTTCAGCAGAAAAGTATAGCGTGGCAACCCAGCTGTTCATTGGAGGTGTCACTGGATGG
TGCACAGGTTTCATATTCCAGAAGGTTGGAAAGTTGGCTGCAACAGCTGTGGGAGGTGGATTTTTTCT
CCTTCAGCTTGCAAACCATACTGGGTACATCAAAGTTGACTGGCAACGAGTGGAGAAGGACATGAGGA
AAGCCAAAGAGCAGCTGAAGATCCGTAAGAGCAATCAGATACCTACTGAGGTCAGGAGCAAAGCTGAG
GAGGTGGTGTCATTTGTGAAGAAGAATGTTCTAGTAACTGGGGGATTTTTCGGAGGCTTTCTGCTTGG
CATGGCATCCTAAGGAAGATGACCTCATGTTCATTGTTCCTGGTTTTTTCCAGCCAGCAGCCTCTACA
CTCCATCATAGGACATCGAGTCCCTCCTCCTCTTCTCCCATGCCTTCTTCCCTGCCATGGCAAATCTG
AGTGGCTTCTCTAAGCATCTGCTGGTACAAGTCAATGTGGCACCATGAGCTTCATGGTGGCAGAAGAG
ACAATAGTCCTTAGCTCTCCTCCCAGTACACCCCCTACTTGGCCAGTCTGTAGGCCAACAAGAAGGTT
CCTTTACCCCCATGCAAGACACTTATGAGAACACATTACAAGATGGCTGACCGTGGAGGATGAGTGGA
TCCTGAAAGGTTGTCCCAAACTGTTGATTTGGAAAAGAAATAAGCACATAGATAACCTTATTGTGTGC
TGCATGGAAAGGAAC**TGAATACATTTGCCTTTAAGCATGAAAG**

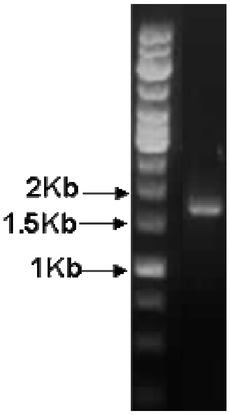
Blat analysis

chrX:153,815,552-153,917,431



4. *SFPQ/EIF5A*

5′ primer: GCCTGTGTCATCCGCCATTTTGTGAGAA
5′ nested primer: GCAAGGTGGCCTCCACGTTTCCTGAGCG
3′ primer: ACCTGAATAAAACTACAAGTTTAATATG
3′ nested primer: CTCTCCCCGACACATTTGTTAAAATCAA



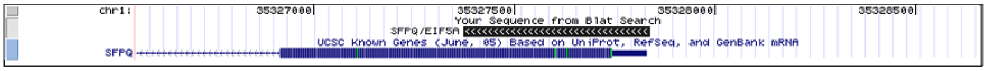
The *SFPQ/EIF5A* fusion transcript sequence:

GCCTGTGTCATCCGCCATTTTGTGAGAA
GCAAGGTGGCCTCCACGTTTCCTGAGCGTCTTCTTCGCTTTTGCCTCGACC
GCCCCTTGACCACAGACATGTCTCGGGATCGGTTCCGGAGTCGTGGCGGTG
GCGGTGGTGGCTTCCACAGGCGTGGAGGAGGCGGCGGCCGCGGCGGCCTCC
ACGACTTCCGTTCCCCGCCGCCCGGCATGGGCCTCAATCAGAATCGCGGCC

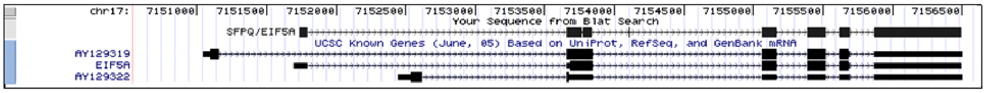
CCATGGGTCCTGGCCCGGGCCAGAGCGTCCAAAACCGTCCGATACTGCCACCGCCTCCACACCAACAG
CAGCAACAGCCACCACCGCAGCAGCCACCGCCGCAGCAGCCGCCACCGCATCAGCCGCCGCCGCATCC
ACAGCCGCATCAGCAGCAGCAGCCGCCGCCACCGCCGCAGGACTCTTCCAAGCCCGTCGTTGCTCAGG
GACCCGGCCCCGCTCCCGGAGTAGGCAGCGGTTGGGCTCGCGGCGAGCGGACGGGGTCGAGTCAGTGC
GTTCGCGCGAGTTGGAATCGAAGCCTCTTAAAATGGCAGATGACTTGGACTTCGAGACAGGAGATGCA
GGGGCCTCAGCCACCTTCCCAATGCAGTGCTCAGCATTACGTAAGAATGGCTTGTGTGCTCAAAGGCC
GGCCATGTAGATCGTCGAGATGTCTACTTCGAAGACTGGCAAGCACGGTCACGTCAAAGGTCCATCTG
GTTTGGTATTGACATCTTTACTGGGAAGAAATATGAAGATATCTGCCCGTCAACTCATAATATGGATG
TCCCCAACATCAAAAGGAATGACTTCCAGCTGATTGGCATCCAGGATGGGTACCTATCACTGCTCCAG
GACAGCGGGGAGGTACGAGAGGACCTTCGTCTCCCTGAGGGAGACCTTGGCAAGGAGATTGAGCAGAA
GTACGACTGTGGAGAAGAGATCCTGATCACGGTGCTGTCTGCCATGACAGAGGAGGCAGCTGTTGCAA
TCAAGGCCATGGCAAAATAACTGGCCCCCAGGGTGGCGGTGGTGGCAGCAGTGATCCTCTGAACCTGC
AGAGGCCCCCTCCCCGAGCCTGGCCTGGCTCTGGCCCGGTCCTAAGCTGGACTCCTCCTACACAATTT
ATTTGACGTTTTATTTTGGTTTTCCCCACCCCCTCAATCTGTCGGGGAGCCCCTGCCCTTCACCTAGC
TCCCTTGGCCAGGAGCGAGCGAAGCTGTGGCCTTGGTGAAGGTGCCCTCCTCTTCTCCCCTCACACTA
CAGCCCTGGTGGGGGAGAAGGGGGTGGGTGCTGCTTGTGGTTTAGTCTTTTTTTTTTTTTTTTTAATT
CAATCTGGAATCAGAAAGCGGTGGATTCTGGCAAATGGTCCTTGTGCCCTCCCCACTCATCCCTGGTC
TGGTCCCCTGTTGCCTATAGCCCTTTACCCTGAGCACCACCCCAACAGACTGGGGACCAGCCCCCTCG
CCTGCCTGTGTCTCTCCCCAAACCCCTTTAGATGGGGAGGGAAGAGGAGGAGAGGGGAGGGGACCTGC
CCCCTCCTCAGGCATCTGGGAGGGCCCTGCCCCCATGGGCTTTACCCTTCCCTGCGGGCTCTCTCCCC
GACACATTTGTTAAAATCAA**ACCTGAATAAAACTACAAGTTTAATATG**

Blat analysis:

chr1:35,326,559-35,328,655

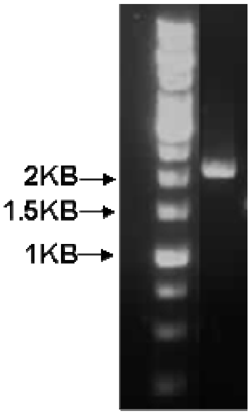

chr1:
35327000
35327500
35328000
35328500
Your Sequence from Blat Search
SFPQ/EIF5A
UCSC Known Genes (June, 05) Based on UniProt, RefSeq, and GenBank mRNA
SFPQ

chr17:7,150,550-7,156,600



5. *SRP9/RPS8*

5′ primer: TGGGACTCGCGTCGGTTGGCGACTCCCG
5′ nested primer: GACGTAGGTAGTTTGTTGGGCCGGGTTC
3′ primer: TAATAAAGGTGTTTATTGTTTTGTTCCC
3′ nested primer: ATAAATCCTTGTTTTGTCTTCACCCATG



The *SRP9/RPS8* fusion transcript sequence:

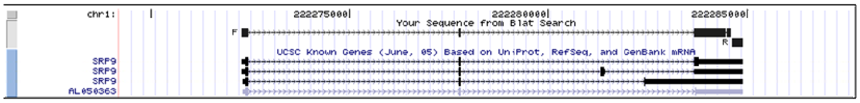
TGGGACTCGCGTCGGTTGGCGACTCCCGGACGTAGGTAGTTTGTTGGGC
CGGGTTCTGAGGCCTTGCTTCTCTTTACTTTTCCACTCTAGGCCACGAT
GCCGCAGTACCAGACCTGGGAGGAGTTCAGCCGCGCTGCCGAGAAGCTT
TACCTCGCTGACCCTATGAAGGCACGTGTGGTTCTCAAATATAGGCATT
CTGATGGGAACTTGTGTGTTAAAGTAACAGATGATTTAGTTTGTTTGGT
GTATAAAACAGACCAAGCTCAAGATGTAAAGAAGATTGAGAAATTCCACAGTCAACTAATGCGACTTA
TGGCAGCCAAGGAAGCCCGCAATGTTACCATGGAAACTGAGTGAATGGTTTGAAATGAAGACTTTGTC
GTGTACTTAGGAAGTAAATATCTTTTGAATTAGAGAAAGTGTTGGGACAGAAAGTACTTTATGTAACT
AAGTGGGCTGTTCAGAAGCTTAGAGGTCATTTTTTGTAATTTTCTTTTTAATTACTTTAGAGAGCTAG
GGATGCAAATGTTTTCAGTTAGAAAGCCTTTATTTACTTTTGGAAATTGAACAAGAAATGCATCTGTC

TTAGAAACTGGAGATTATTTGATGTTAGGTAAAACATGTAATTGTTTCTCTGGCAAATTTGTATCAGT
AATTTGAAAATGAGATATTAGGAAAAACCATCCTTCTTAAATTTAGTTCATCTTTCTTTAAAAGAACA
TTAAATGTAACCATTTTGTCAGATCCATGTATTTTGGAGCATAAAATGTATGCTGTTGTGACCAATAA
ATATAAAATATGGTAATTGGAATTAACTCCACACCATAGTATGCATTGTTATACATACTGTGTACCTA
ATTATGTATAGCAGTGTAGTCTCAATTATATCTGAAGTAATTGTGACTAACAAGTATGCTTTGCCTTA
TTTCCACATTTAACTACCTGTTAATATACGATTTGTAGTATCAGCTTGTTGAGCATGACTTTGATCTA
GTTTCAGTGATCAGAGCAGCAGTTATTTGAGTGTATGGATGCATGATGATCACTGTGGCTATAATGTA
ACTGAAACCACCATATTTACAGAATTTACTTACATATTTTCCCAATCTGAGTT

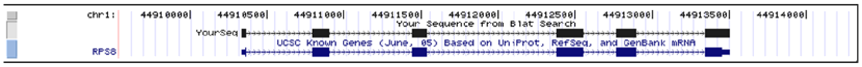
CTGCCCATTTCTGTCTATATCAGTGTTAGTGTGACTGTCACACTGTTATGTTTCAGTAACTAGAAGTA
TGATATTGATATATATTGTATTCACCACTAAATGTAATGTTGATCTCAAGAATGAAATGAAGGCACTA
CATTGAAATATGTTTTGTATAAATTTGTCATGTTGAACAGCATTTTAGCATGGTAAGTTCCCTTAGCT
ATATGAATTTTGGCATGTTTCAGAGAGATCAGTAAATAAAATATTAGATAAAATAAAAAAAAAAAAAA
AAAAAAAAAACCAAAAAAAAAAAAAAAAAACTCCAAACTCGAGGCGGCCGCGGATCCGACGCTCTTTCC
AGCCAGCGCCGAGCGATGGGCATCTCTCGGGACAACTGGCACAAGCGCCGCAAAACCGGGGGCAAGAG
AAAGCCCTACCACAAGAAGCGGAAGTATGAGTTGGGGCGCCCAGCTGCCAACACCAAGATTGGCCCCC
GCCGCATCCACACAGTCCGTGTGCGGGGAGGTAACAAGAAATACCGTGCCCTGAGGTTGGACGTGGGG
AATTTCTCCTGGGGCTCAGAGTGTTGTACTCGTAAAACAAGGATCATCGATGTTGTCTACAATGCATC
TAATAACGAGCTGGTTCGTACCAAGACCCTGGTGAAGAATTGCATCGTGCTCATCGACAGCACACCGT
ACCGACAGTGGTACGAGTCCCACTATGCGCTGCCCCTGGGCCGCAAGAAGGGAGCCAAGCTGACTCCT
GAGGAAGAAGAGATTTTAAACAAAAAACGATCTAAAAAAATTCAGAAGAAATATGATGAAAGGAAAAA
GAATGCCAAAATCAGCAGTCTCCTGGAGGAGCAGTTCCAGCAGGGCAAGCTTCTTGCGTGCATCGCTT
CAAGGCCGGGACAGTGTGGCCGAGCAGATGGCTATGTGCTAGAGGGCAAAGAGTTGGAGTTCTAT
CTTAGGAAAATCAAGGCCCGCAAAGGCAAATAAATCCTTGTTTTGTCTTC
ACCCATA**TAATAAAGGTGTTTATTGTTTTGTTCCC**

Blat analysis:

chr1:222,269,212-222,287,631

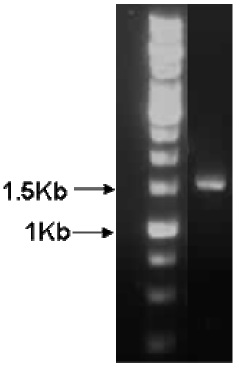

chr1:
222275000
222280000
222285000
Your Sequence from Blat Search
F
R
UCSC Known Genes (June, 05) Based on UniProt, RefSeq, and GenBank mRNA
SRP9
SRP9
SRP9
AL050363

chr1:44,909,548-44,914,293


chr1:
44910000
44910500
44911000
44911500
44912000
44912500
44913000
44913500
44914000
Your Sequence from Blat Search
YourSeq
UCSC Known Genes (June, 05) Based on UniProt, RefSeq, and GenBank mRNA
RPS8

6. *AMD1/GAPDH*

5′ primer: GCTTACACAGTATGGCCGGCGACATTAG
5′ nested primer: CTAGCGCTCGCTCTACTCTCTCTAACGG
3′ primer: CCATCAATAAAGTACCCTGTGCTCAACC
3′ nested primer: GGCCTAGGGAGCCGCACCTTGTCATGTA



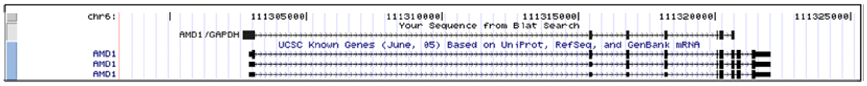
The *AMD1/GAPDH* fusion transcript sequence:

GCTTACACAGTATGGCCGGCGACATTAGCTAGCGCTCGCTCTACTCTCTCT
AACGGGAAAGCAGCGGAATACAAGAGACTGAACTGTATCTGCCTCTATTTC
CAAAAGACTCACGTTCAACTTTCGCTCACACAAAGCCGGGAAAATTTTATT
AGTCCTTTTTTTAAAAAAAGTTAATATAAAATTATAGCAAAAAAAAAAGGA
ACCTGAACTTTAGTAACACAGCTGGAACAATCCGCAGCGGCGGCGGCAGCGGCGGGAGAAGAGGTTTA
ATTTAGTTGATTTTCTGTGGTTGTTGGTTGTTCGCTAGTCTCACGGTGATGGAAGCTGCACATTTTTT
CGAAGGGACCGAGAAGCTGCTGGAGGTTTGGTTCTCCCGGCAGCAGCCCGACGCAAACCAAGGATCTG
GGGATCTTCGCACTATCCCAAGATCTGAGTGGGACATACTTTTGAAGGATGTGCAATGTTCAATCATA

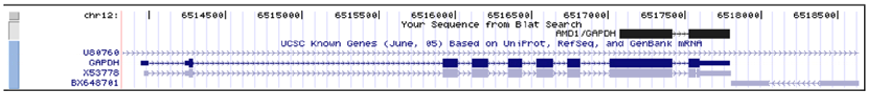
AGTGTGACAAAAACTGACAAGCAGGAAGCTTATGTACTCAGTGAGAGTAGCATGTTTGTCTCCAAGAG
ACGTTTCATTTTGAAGACATGTGGTACCACCCTCTTGCTGAAAGCACTGGTTCCCCTGTTGAAGCTTG
CTAGGGATTACAGTGGGTTTGACTCAATTCAAAGCTTCTTTTATTCTCGTAAGAATTTCATGAAGCCT
TCTCACCAAGGGTACCCACACCGGAATTTCCAGGAAGAAATAGAGTTTCTTAATGCAATTTTCCCAAA
TGGAGCAGCATATTGTATGGGACGTATGAATTCTGACTGTTGGTACTTATATACTCTGGATTTCCCAG
AGAGTCGGGTAATCAGTCAGCCAGATCAAACCTTGGAAATTCTGATGAGTGAGCTTGACCCAGCAGTT
ATGGACCAGTTCTACATGAAAGATGGTGTTACTGCAAAGGATGTCACTCGTGAGAGTGGAATTCGTGA
CCTGATACCAGGTTCTGTCATGATGCCACAATGTCATCCTTGTGGGTATTCGATGAATGGAATGAAAT
CGGATGGAACTTATTGATGGCCGCGGGGCTCTCCAGAACATCATCCCTGCCTCTACTGGCGCTGCCAA
GGCTGTGGGCAAGGTCATCCCTGAGCTGAACGGGAAGCTCACTGGCATGGCCTTCCGTGTCCCCACTG
CCAACGTGTCAGTGGTGGACCTGACCTGCCGTCTAGAAAAACCTGCCAAATATGATGACATCAAGAAG
GTGGTGAAGCAGGCGTCGGAGGGCCCCCTCAAGGGCATCCTGGGCTACACTGAGCACCAGGTGGTCTC
CTCTGACTTCAACAGCGACACCCACTCCTCCACCTTCGACGCTGGGGCTGGCATTGCCCTCAACGACC
ACTTTGTCAAGCTCATTTCCTGGTATGACAACGAATTTGGCTACAGCAACAGGGTGGTGGACCTCATG
GCCCACATGGCCTCCAAGGAGTAAGACCCCTGGACCACCAGCCCCAGCAAGAGCACAAGAGGAAGAGA
GAGACCCTCACTGCTGGGGAGTCCCTGCCACACTCAGTCCCCCACCACACTGAATCTCCCCTCCTCAC
AGTTTCCATGTAGACCCCTTGAAGAGGGGAGGGGCCTAGGGAGCCGCACCTTGTCATGTA**CCATCAAT**
AAAGTACCCTGTGCTCAACC

Blat analysis:

chr6:111,298,170-111,325,235


chr6:
111305000|
111310000|
111315000|
111320000|
111325000|
Your Sequence from Blat Search
AMD1/GAPDH
UCSC Known Genes (June, 05) Based on UniProt, RefSeq, and GenBank mRNA
AMD1
AMD1
AMD1

chr12:6,513,828-6,518,640



7. *BCAS4/BCAS3*

Molecular characterization of fusion gene

BCAS4 forward primer: 5’CCTCCTGATGCTGCTCGT

BCAS3 reverse primer: 5’CTGCAGCGTGATTTATTGGA

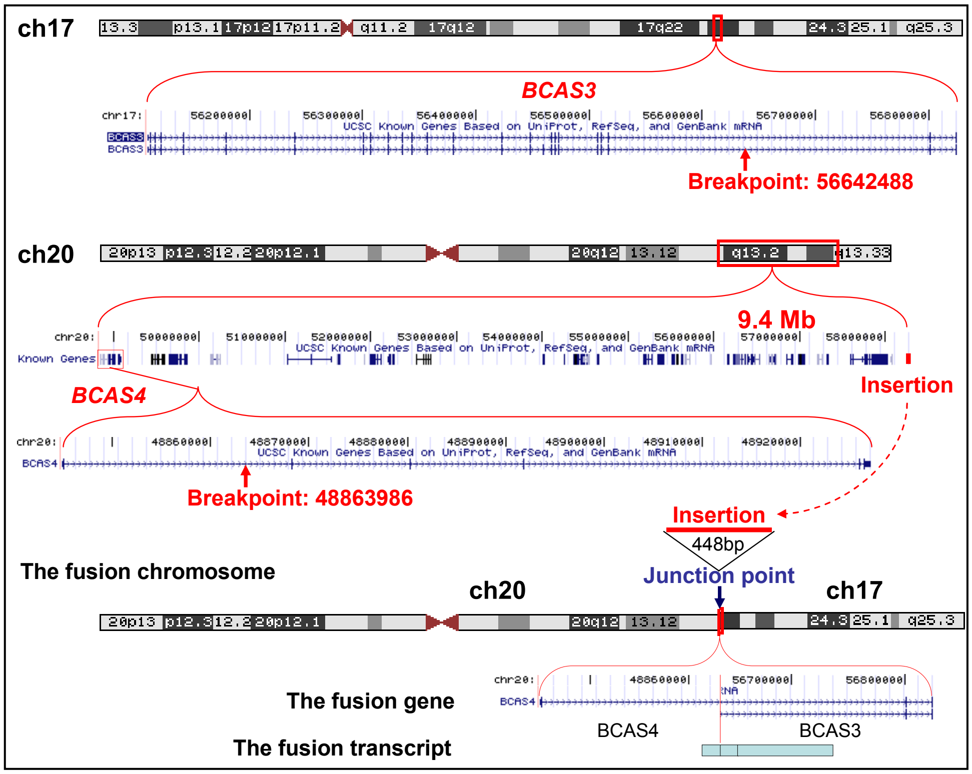
The *BCAS4/BCAS3* fusion transcript sequence

BCAS4 gene

TCGTTATGTGACCCGCTCGATACAC**GCCGGAGCCCATGCGCAGCGGGCGCGCGAGCTCGCGCTCTTC**
CTGACCCCCGATCCTGGGGCCGAG↑GTACCTTTGACAGGAGCGTGACCCTGCTGGAGGTGTGCGGGAG
CTGGCCTGAGGGCTTCGGGCTGCGGCACATGTCCTCCATGGAGCACACGGAGGAGGGCCTCCGGGAGC
GACTTGCCGACGCCATGGCCGAGTCACCTAGCCGGGACGTCGTGGGATCCGGAACAGAACTTCAGCGA
GAGGGAAGCATCGAGACTCTGAGTAACAGCTCAGGCTCCACCAGCGGCAGCATACCAAGAAACTTTGA
TGGCTACCGATCTCCGCTGCCCACCAATGAGAGCCAGCCCCTCAGCCTCTTCCCGACTGGCTTCCCGT
AGGTACCAGCAACCTGCTTCTGACTGGCCAGCCCCCTCCCCTGCTGGAGGAGGGGAGAAGCCCCGCTC
TGGTCCTACCCTTCAGTCTCTGCTCTTCCTTCATCAACCACCTTCCCCAAGCTTAGTGACAGCAGCCG
CCCATCCTACCTGGATGGAGAGGAGACCCTTCTCCAAGCACCTCAGCGCACTTGCCCTCTGCCACACC
TGTCGGTGGAGGCTGTGGCCAGGAGAGACTGTAGAAGCTCGGTCCCTGTGTATGTTTGCATATGACAT
CCTGCATTGGATCCGCTTTTGTATTTTTTAACCATACCCACGGTGGGGCGGGTGGGGGGAGCCTGGAA
CAGTGACCAGATCTGGCGGCCTGAGTGGGGACAGAGTTGATCGTCCACCTGGCCATTTTGACCCTGAG

```
TGGACAGTCACAGCCTCAGCTCATGTCTGGCTGTGACACACACTGCCCCCAGCTTGCCTTGGTCAGCC
CCACTCCAGCACGGGGTGAACGGAGGCCCAGAGTACTAGGGAATGAGGAAGGGAGGACATGCCTCTTC
TTCCTCCTTTCTTTCCCCATCTGTTCCTGGGAAGAGTTTGTCTTTCTTAATCTTTAAGCCCTGTTACC
CTGTTCCTGTACTGATCAATGGAAGGAAACCGGTGGTTACTGAAAGCCCTGTTGAAAAGGTGAACGGT
GTGGTCCAATAAATCACGGCTGCAAAATCGAATTCCCGCGGTCGCCATGGTCGGGCCGGGAGCCTGCG
AACGTCGGGCCCAATTCGCCCTATAGGTGAGTCGTATATACAATTTCACTGTGGCGGTCGTTTAACAC
CGCTCTGTGAACTGGAAACCCTTGGCGTAACCCACTAATTCGCCGTGGGGACAAATCCCCCTTTGCGC
CACGGGTGGGGTATAATACCTAAAAGAGCTCA       BCAS3 gene
```

The blue arrow indicates the junction of first exon of BCAS4 and 23rd exon of BCAS3. Presence of complete exon sequence suggest the breakpoint in the intronic regions.



Supplemental Figure S9. Molecular characterization of fusion gene *BCAS4/3*

Top: The genomic structures of *BCAS3* and *BCAS4* are shown on chromosome 17 and 20, respectively. The two red arrows indicate the breakpoints where the two chromosomes joint together. **Bottom:** The fusion chromosome, gene structure and resulted fusion transcript are displayed together with the 448 bp insertion 9.4 Mb 3' from the *BCAS4* genomic location. The junction point is indicated by a blue arrow.

V. Comparison of fusion transcripts identified by this and other studies in MCF7 transcriptome

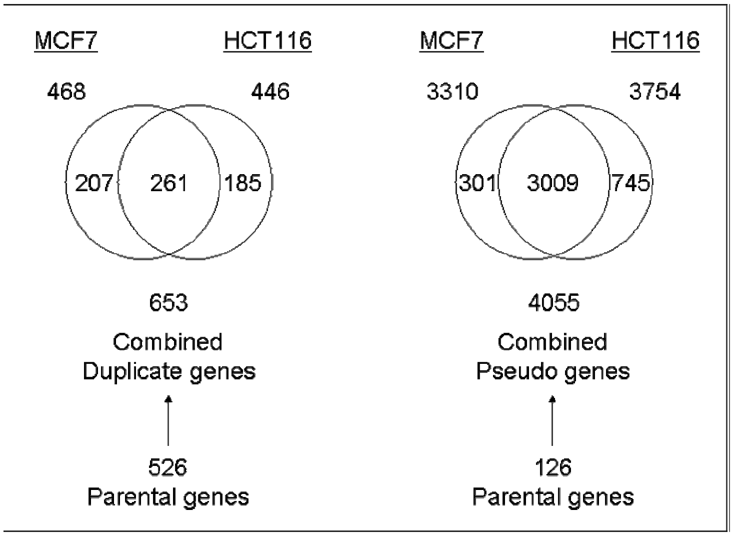
There were two studies specifically attempted to identify fusion genes at large scale effort. Hahn et al (Hahn et al. 2004) took the EST computation mapping approach and reported 237 fusion genes expressed among normal and malignant tissues. They went further and validated 2 of them (*BCAS4/BCAS3* and *IRA1/RGS17*) through RT-PCR approach in MCF7 cells. We compared the 237 fusion genes from Hahn's list with the 7202 PET-0 sequences. Only the *BCAS4/BCAS3* fusion transcript was matched. To test whether the *IRA1/RGS17* fusion can be found in our cDNA library, we carried out PCR with reported primers and obtained correct size fragment which confirmed the presence of this fusion gene in our MCF cells. Further examining all the PET mapping coordinates, we found 2 PETs (total 35 counts) in the PET11+ category (more than 11 paired mapping locations on hg17 genome) can be mapped close to *IRA1/RGS17* when the numbers of 5' and 3' genomic mapping position reported were expanded up to 2,000.

1. PETid: SHC012-U_47939_COUNT:31
GGTAATCCCAGCACTTTGGAAATAAAACTTTACAG
5' mapped at chr3 178401645 + TBL1XR1 chr3 178397742
3' mapped at chr6 153419593 + RGS17 chr6 153424148

2. PETid: SHC012-U_49305_COUNT:4
GGTAATCCCAGCACTTTGGAAATAAAACTTTACAGG
5' mapped at chr3 178401645 + TBL1XR1 chr3 178397742
3' mapped at chr6 153419593 + RGS17 chr6 153424148

The other study has revealed 66 fusion clones through end pair sequencing of full length cDNA clones from MCF7 cells and 4 of them were confirmed by PCR (Volik et al. 2006). Intriguingly, the previously characterized *BCAS4/BCAS3* and *IRA1/RGS17* are not included in this list of 66 candidates. We conducted PCR using the reported primers to validate these 4 fusion hits that were reportedly confirmed by Volik et al. However, our experiment did not yield any specific fragments from the mRNA of MCF7 cells and even the MCF7 cDNA library. We also did not find any of our 7202 PET-0 mapping locations in the close proximity (up to 5 Kb) with any of the 66 candidate fusion gene locations reported by Volik et al.

VI. GIS-PET identified pseudogenes and transcribed pseudogenes



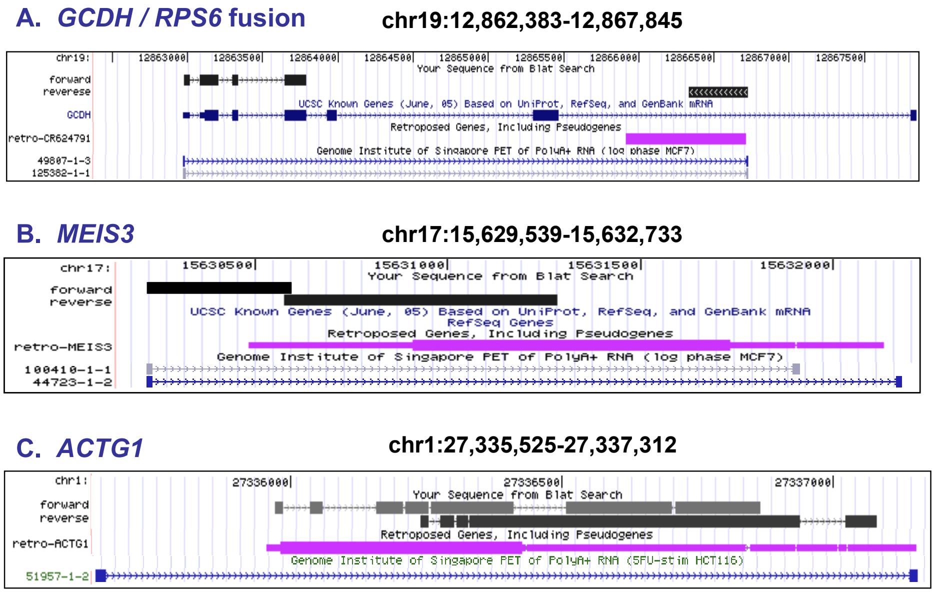
Supplemental Figure S10. Venn diagram displays the numbers of duplicated genes and pseudogenes loci identified in MCF7 and HCT116 cell genomes.

Supplemental Table S8.

Overlaps between GIS-PET identified psedugogene loci and the retrotransposed loci from datasets

		GIS-PET	Yale	Sanger_Vega*	UCSC_retro
		4055	**7368**	**4416**	**11306**
GIS-PET	**4055**	x	2624	1213	3573
Yale	**7368**	2624	x	2048	5076
Sanger_Vega*	**4416**	1213	2048	x	2644
UCSC_retro	**11306**	3573	5076	2644	x

* Only 9 chromosomes are present in Vega pseudogene dataset. Therefore, 2719 loci from Yale pseudogenes, 4061 from UCSC retrogenes and 1618 pseudogenes from GIS-PET were used for comparison in this category.


A. GCDH / RPS6 fusion
chr19:12,862,383-12,867,845
chr19:
12863000
12863500
12864000
12864500
12865000
12865500
12866000
12866500
12867000
12867500
Your Sequence from Blat Search
forward
reverese
UCSC Known Genes (June, 05) Based on UniProt, RefSeq, and GenBank mRNA
GCDH
Retroposed Genes, Including Pseudogenes
retro-CR624791
Genome Institute of Singapore PET of PolyA+ RNA (log phase MCF7)
49807-1-3
125382-1-1
B. MEIS3
chr17:15,629,539-15,632,733
chr17:
15630500
15631000
15631500
15632000
Your Sequence from Blat Search
forward
reverse
UCSC Known Genes (June, 05) Based on UniProt, RefSeq, and GenBank mRNA
RefSeq Genes
Retroposed Genes, Including Pseudogenes
retro-MEIS3
Genome Institute of Singapore PET of PolyA+ RNA (log phase MCF7)
100410-1-1
44723-1-2
C. ACTG1
chr1:27,335,525-27,337,312
chr1:
27336000
27336500
27337000
Your Sequence from Blat Search
forward
reverse
Retroposed Genes, Including Pseudogenes
retro-ACTG1
Genome Institute of Singapore PET of PolyA+ RNA (5FU-stim HCT116)
51957-1-2

Supplemental Figure S11. Transcribed pseudogenes loci

(A). A fusion cDNA between *GCDH* and the *RPS6* processed pseudogene at chromosome 19 was detected by PET mapping and verified by sequencing the PCR product. Forward and reverse sequence reads of the amplified cDNA were shown. The position of pseudogene is marked in pink. **(B).** Forward and reverse sequence reads of partially PCR amplified transcribed *MEIS3* pseudogene locus were shown. **(C).** Forward and reverse sequence reads from amplified cDNA confirmed the transcribed *ACTG1* pseudogene locus.

VII. References cited in Supplementary Information

Grigoriadis, A., A. Mackay, J.S. Reis-Filho, D. Steele, C. Iseli, B.J. Stevenson, C.V. Jongeneel, H. Valgeirsson, K. Fenwick, M. Iravani, M. Leao, A.J. Simpson, R.L. Strausberg, P.S. Jat, A. Ashworth, A.M. Neville, and M.J. O'Hare. 2006. Establishment of the epithelial-specific transcriptome of normal and malignant human breast cells based on MPSS and array expression data. *Breast Cancer Res* **8:** R56.

Hahn, Y., T.K. Bera, K. Gehlhaus, I.R. Kirsch, I.H. Pastan, and B. Lee. 2004. Finding fusion genes resulting from chromosome rearrangement by analyzing the expressed sequence databases. *Proc Natl Acad Sci U S A* **101:** 13257-13261.

Jongeneel, C.V., C. Iseli, B.J. Stevenson, G.J. Riggins, A. Lal, A. Mackay, R.A. Harris, M.J. O'Hare, A.M. Neville, A.J. Simpson, and R.L. Strausberg. 2003. Comprehensive sampling of gene expression in human cell lines with massively parallel signature sequencing. *Proc Natl Acad Sci U S A* **100:** 4702-4705.

Kuznetsov, V.A. (2002) Statistics of the numbers of transcripts and protein sequences encoded in the genome. In*: Computational and Statistical Methods to Genomics.* (W. Zhang, and I. Shmulevish, Eds.; 1-st Ed.) Kluwer: Boston-Dordrecht, pp. 125-171.

Kuznetsov, V.A., G.D. Knott, and R.F. Bonner. 2002. General statistics of stochastic process of gene expression in eukaryotic cells. *Genetics* **161:** 1321-1332.

Ng, P., C.L. Wei, W.K. Sung, K.P. Chiu, L. Lipovich, C.C. Ang, S. Gupta, A. Shahab, A. Ridwan, C.H. Wong, E.T. Liu, and Y. Ruan. 2005. Gene identification signature (GIS) analysis for transcriptome characterization and genome annotation. *Nat Methods* **2:** 105-111.

Volik, S., B.J. Raphael, G. Huang, M.R. Stratton, G. Bignel, J. Murnane, J.H. Brebner, K. Bajsarowicz, P.L. Paris, Q. Tao, D. Kowbel, A. Lapuk, D.A. Shagin, I.A. Shagina, J.W. Gray, J.F. Cheng, P.J. de Jong, P. Pevzner, and C. Collins. 2006. Decoding the fine-scale structure of a breast cancer genome and transcriptome. *Genome Res* **16:** 394-404.