AD_______________

Award Number: W81XWH-08-1-0110

TITLE: A Search for Gene Fusions/Translocations in Breast Cancer

PRINCIPAL INVESTIGATOR: Arul M. Chinnaiyan, M.D., Ph.D.

CONTRACTING ORGANIZATION: University of Michigan
Ann Arbor, MI 48109-1274

REPORT DATE: November 2013

TYPE OF REPORT: Final

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

# REPORT DOCUMENTATION PAGE

*Form Approved*
*OMB No. 0704-0188*

**1. REPORT DATE** *(DD-MM-YYYY)*
November 2013

**2. REPORT TYPE**
Final

**3. DATES COVERED** *(From - To)*
1 September 2008-31 August 2013

**4. TITLE AND SUBTITLE**

A Search for Gene Fusions/Translocations in Breast Cancer

**5a. CONTRACT NUMBER**

**5b. GRANT NUMBER**
W81XWH-08-1-0110

**5c. PROGRAM ELEMENT NUMBER**

**6. AUTHOR(S)**

Arul M. Chinnaiyan, M.D., Ph.D.

Email: chinnaiyangrants@umich.edu

**5d. PROJECT NUMBER**

**5e. TASK NUMBER**

**5f. WORK UNIT NUMBER**

**7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)**

University of Michigan
Ann Arbor, MI 48109-1274

**8. PERFORMING ORGANIZATION REPORT NUMBER**

**9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)**
U.S. Army Medical Research
And Material Command
Fort Detrick, Maryland
21702-5012

**10. SPONSOR/MONITOR'S ACRONYM(S)**

**11. SPONSOR/MONITOR'S REPORT NUMBER(S)**

**12. DISTRIBUTION / AVAILABILITY STATEMENT**

Approved for public release; distribution unlimited

**13. SUPPLEMENTARY NOTES**

**14. ABSTRACT**

Previously, we completed the molecular/ biochemical characterization of several shortlisted candidate gene fusions from the transcriptome sequencing of over 70 previously validated samples. From these studies, we identified two rare but recurrent gene fusions in breast cancer cell lines and tissues involving the MAST and Notch genes. Both of these fusion genes are potentially targetable and patients harboring MAST or Notch fusions may benefit from MAST or Notch inhibitors. We also describe a novel study of cancer-specific pseudogenes, including those in breast cancer. Most recently, through our clinical sequencing initiative, we discovered a series of activating mutations in the estrogen receptor (ESR1) in breast cancer patients. These activating mutations in ESR1 are a key mechanism in acquired endocrine resistance in breast cancer therapy. Overall, these discoveries made over the funding period contribute towards the understanding of the molecular and genetic etiology of breast cancer that will advance the development of targeted therapies.

**15. SUBJECT TERMS**
Gene fusions, sequencing, MAST,Notch

| | | | | | |
|---|---|---|---|---|---|
| **16. SECURITY CLASSIFICATION OF:** | | | **17. LIMITATION OF ABSTRACT** | **18. NUMBER OF PAGES** | **19a. NAME OF RESPONSIBLE PERSON**<br>USAMRMC |
| **a. REPORT**<br>U | **b. ABSTRACT**<br>U | **c. THIS PAGE**<br>U | UU | 130 | **19b. TELEPHONE NUMBER** *(include area code)* |

## Table of Contents

## INTRODUCTION:

In this project, ―**A Search for Gene Fusions/Translocations in Breast Cancer**" we have undertaken a systematic evaluation of breast cancer to map disease-specific, recurrent chromosomal or transcriptional chimeras in breast cancer that can be further characterized to develop novel biomarkers and therapeutic targets. Over the entire grant period, we have made high-impact discoveries and tremendous progress towards our goal of identifying molecular drivers of breast cancer that have prognostics/diagnostic value as well as potential as therapeutic targets.

During the early years of this grant, we reported the characterization of a subset of ER positive breast cancer patients characterized by the overexpression of AGTR1 who may be responsive to an available drug, losartan1 (Rhodes et al, 2009). We also provided a novel mechanistic framework for the overexpression of the polycomb group protein EZH2 in metastatic breast and prostate cancers, involving the genomic loss of its negative regulator, miR101 (Varambally et al, 2008). Additionally, we reported the development of a high throughput sequencing pipeline for a directed search for gene fusions in cancers using next generation transcriptome sequencing platforms (Maher et al, 2009). From those efforts, we identified numerous gene fusions (70 in over 40 cancer samples) that mapped to *loci* of genomic amplifications. We shortlisted several fusion candidates that involved kinase genes and other genes of interest related to oncogenesis for further study.

Subsequently we described the exciting discovery and characterization of two novel recurrent and actionable gene fusions in our breast cancer cohort involving MAST and Notch genes. Both MAST and Notch family gene fusions exerted significant phenotypic effects in breast epithelial cells (Robinson et al, 2011). We also reported the development of a novel bioinformatics tool designed to facilitate the discovery of gene fusions from next-generation sequencing data (Iyer et al, 2011); as well as a study that furthers our understanding of the role of microRNAs in cancer progression (Cao et al, 2011).

We characterized amplicon-associated gene fusions in breast cancer and found that most of them were ―passengers" rather than ―driver" fusions even when fusions involved known oncogenic kinases (Kalyana-Sundaram et al, June, 2012). We also developed a novel bioinformatics methodology to discover processed pseudogenes in cancer including those specifically expressed in breast cancer (Kalyana-Sundaram et al, August, 2012-2).

We analyzed transcriptome sequencing data from a compendium of 482 cancer and benign samples from 25 different tissue types to assess the complete landscape of a cancer‗s ―kinome" expression and determine which kinases are activated in specific tumor types. Here, we found frequent outlier kinase expression in breast cancer included therapeutic targets like ERBB2 and FGFR4 whereas MET, AKT2, and PLK2 were expressed in pancreatic cancer. The results of this study were published in Cancer Discovery (Kothari et al, 2013).

Recently, through our clinical sequencing study, we identified gene rearrangements of FGFR across multiple cancer types, including patients with breast cancer. All FGFR fusions had intact kinase domains and oligomerization capability. Importantly, two cell lines harboring FGFR fusions were sensitive to inhibitors *in vitro* and *in vivo* (Wu et al, 2013). In a separate study, we identified mutations in the estrogen receptor that is acquired after breast cancer patients take anti-estrogen therapies. These mutations uncover the mechanism by which patients often become resistant to hormone therapy (Robinson et al, 2013).

Overall, these discoveries contribute towards the understanding of the molecular and genetic etiology of breast cancer that will advance the development of targeted therapies.

## BODY:

***A detailed, itemized report of all major completed studies follows:***

### 1. Establishment of next generation transcriptome sequencing analysis

Breast cancer cell lines, immortalized normal mammary epithelial cell lines and primary cultures of normal mammary epithelial cells were obtained from ATCC and collaborators at University of California, San Diego. A total of 40+ cell lines were cultured and DNA, RNA and protein extracted. Breast cancer tissue samples, representing various clinic-pathological stages of breast cancer, were obtained from the University of Michigan Breast Cancer Program, and processed for RNA, DNA and protein in batches.

RNA isolated from experimental samples was assessed for quality and integrity through Bioanalyzer (Agilent). and 2 to 10 µg total RNA with RNA Integrity Number ≥8 was used to prepare transcriptome sequencing libraries. Briefly, total RNA was passed over oligo-dT bearing magnetic beads to purify mRNA, fragmented and converted into double stranded cDNA by reverse transcription followed by DNA polymerase reaction. The cDNA ends were modified by ligating short adaptor sequences (complementary to the oligos on the sequencing flowcell). The cDNA library was size-fractionated by agarose gel electrophoresis and a 300 base-pair region was cut out of the gel, purified, and PCR amplified using adaptor specific PCR primer. The purified PCR product was assessed for quality and concentration using the Bioanalyzer and

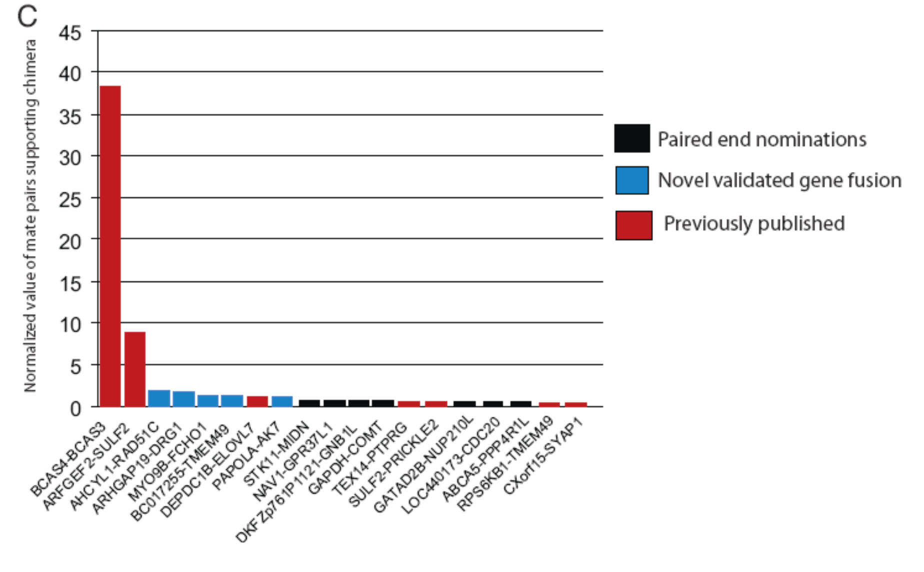


**Figure 1.** Discovery of gene fusions in MCF7 by Paired End Transcriptome Sequencing.

libraries with a clean, single peak (representing approximately 300bp) was applied on the flowcells for cluster generation. The experimental protocol for transcriptome sequencing was developed by Illumina scientists, and ***our group has served as the beta-test center for the fine-tuning and subsequent assembly of the kit for paired-end sequencing library preparation.***

Utilizing the pipeline described above, we conducted a proof of concept study to nominate gene fusions from paired end transcriptome sequence data (Maher et al, 2009). Here, we ―rediscovered" previously known gene fusions in the breast cancer cell line MCF7 including BCAS4-BCAS3 and ARGEF2-SUL2, as well as several novel gene fusions that were all nominated by sequence analysis and validated by fusion-specific real time PCR (**Figure 1**). This strategy was subsequently utilized in the discovery of gene fusions involving the MAST and NOTCH family members described below.

## 2. AGTR as a COPA candidate in breast cancer

In order to identify genes that display outlier expression in breast cancers, we employed our gene expression data compendium, Oncomine 3.0 (Rhodes et al, Jan-Feb, 2004, 2007) to perform Cancer Outlier Profile Analysis (COPA) previously used for the discovery of gene fusions in prostate cancer (Tomlins et al, 2005). Gene expression values obtained from microarray data-sets were median-centered, setting each gene‘s median expression value to zero and each gene expression value was divided by its median absolute deviation (MAD) to calculate COPA scores. Next, genes were rank-ordered by their COPA scores and outlier genes were defined as those that ranked in the top 100 COPA scores at the 75th, 90th or 95th percentile cutoffs. Genes showing outlier expression across multiple studies (meta-outlier genes) were scored as outliers in a significant fraction (p<1E-5) of datasets using MetaCopa analysis, described earlier (Rhodes et al, June, 2004).

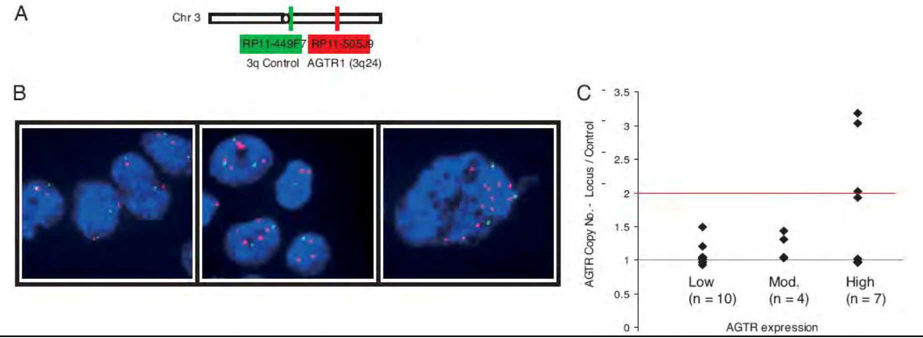


**Figure 2.** Copy number analysis of the AGTR1 locus**.** (A) A schematic of probes used for FISH analysis- Control (green) and AGTR1 (red). (B) Representative images from FISH analysis- left, representative negative case, middle and right, cases with copy number gains of AGTR1. (C) Association ofAGTR1overexpression with copy number gain. *Proc Natl Acad Sci U S A.* 2009 Jun 23;106(25):10284-9.

Meta-Copa analysis of breast cancer datasets on 31 breast cancer profiling studies comprising 3,157 microarray experiments led to the identification of a total of 159 significant meta outliers (P<1E-5). Among the top genes identified as outliers in a majority of datasets examined, the highest outlier in ERBB2-negative breast cancer samples was found to be AGTR1, the Angiotensin II Receptor Type I. Potential genomic rearrangement of AGTR1 locus was investigated as a likely mechanism for overexpression.

We performed FISH on tissue microarrays containing 311 cases of invasive breast cancer to test the AGTR1 locus for gene rearrangement or DNA copy number aberrations and observed an amplification of the AGTR1 locus rather than rearrangement associated with AGTR1 overexpression in 7 of 112 cases (6.25%) (**Figure 2**). This observation was confirmed by qRT-PCR analysis. Further analysis revealed that although copy number gain was always associated with overexpression, increased expression also occurred without copy number gain.

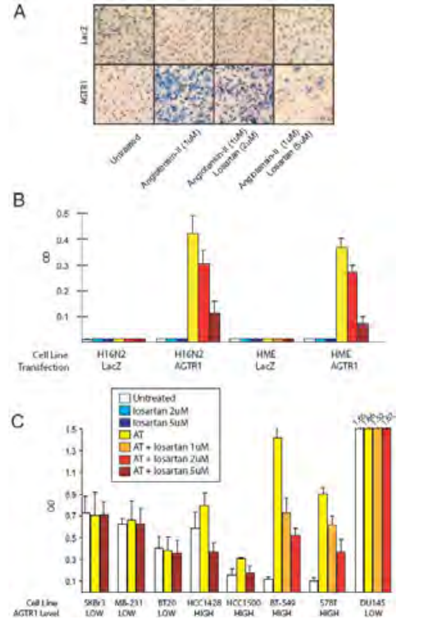


**Figure 3.** AGTR1 overexpression and effect on cell invasion. (A) Matrigel invasion assays of Human Mammary Epithelial Cells (HMEC) or immortalized normal mammary epithelial cells, H16N2 overexpressing AGTR1 or LacZ. Cells cultured with and without agonist, angiotensin (AT) or antagonist, losartan. Similar results were observed for HME cells. (B) Colorimetric readout of invasion assays with LacZ- or AGTR1-expressing H16N2 or HMEC cells treated with AT or losartan. (C) Colorimetric readout of invasion assays from a panel of 7 breast cancer cell lines and a prostate cancer cell line, DU145, after treatment with AT and/or losartan. *Proc Natl Acad Sci U S A.* 2009 Jun 23;106(25):10284-9.

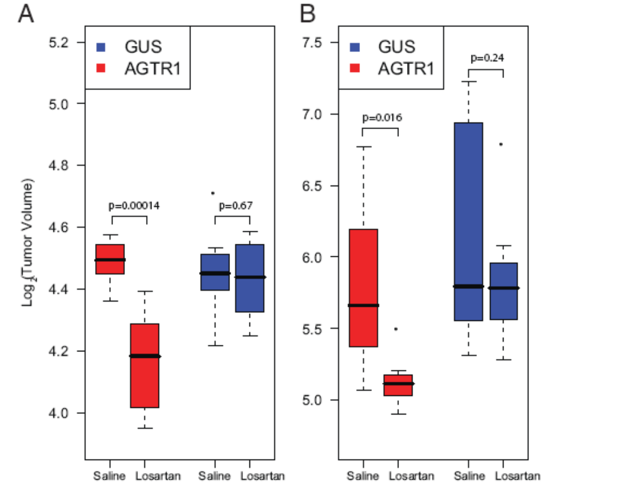


**Figure 4**. Effect of losartan treatment on AGTR1 expressing MCF7 cell xenografts. (A) Xenograft tumor size at 2 weeks. (B) Xenograft tumor size at 8 weeks. *Proc Natl Acad Sci U S A.* 2009 Jun 23;106(25):10284-9.

Ectopic overexpression of AGTR1 in primary mammary epithelial cells HMEC and H16N2 combined with angiotensin II stimulation led to a highly invasive phenotype that was attenuated by the AGTR1 antagonist losartan (**Figure 3**).

Similar to the observations of *in vitro* cell culture experiments, the AGTR inhibitor losartan exerted an inhibitory effect on AGTR1-positive breast cancer xenografts, reducing tumor growth by 30% (**Figure 4**).

Both, *in vitro* and *in vivo* studies indicate that a subpopulation of ER-positive, ERBB2-negative breast cancers that overexpress AGTR1 may benefit from targeted therapy with AGTR1 antagonists such as losartan.

## 3. Loss of microRNA101 leads to overexpression of EZH2

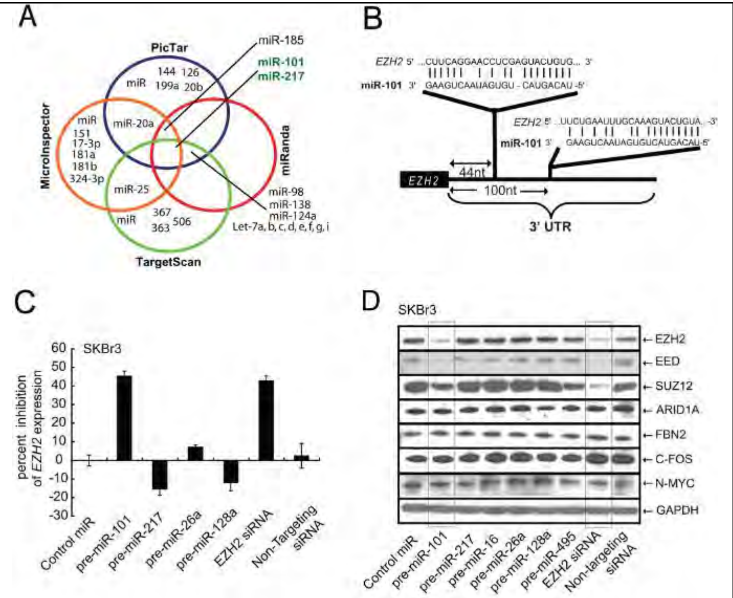


**Figure 5.** miR-101 inhibits EZH2 transcript and protein expression in breast cancer cell line SKBR3. (**A**) Venn diagram displaying miRNAs computationally predicted to target EZH2 using different target prediction programs. (**B**) Schematic of two predicted miR-101 binding sites in the EZH2 3'UTR. (**C**) miR-101 downregulates EZH2 transcript expression. qRT-PCR of EZH2 in SKBr3 cells transfected with precursor miR-101. (**D**), miR-101 downregulates Polycomb Group Complex 2 protein expression. miR-101 downregulates EZH2 protein as well as Polycomb members SUZ12 and EED in SKBr3 cells. *Science*. 2008 Dec 12;322(5908):1695-9

Enhancer of zeste homolog 2 (EZH2) is a mammalian histone methyltransferase that is overexpressed in aggressive solid tumors, including breast cancer (Kleer et al, 2003) and regulates the survival and metastasis of cancer cells through epigenetic silencing of target genes. We investigated the potential role of microRNAs in the regulation of expression of EZH2 following an integrative bioinformatic analysis of miRNA target prediction databases, and identified mir101 as a likely regulator of EZH2. Functional characterization of the association between EZH2 and mir101 expression led to the discovery of genomic loss of mir101 that led to increased expression of EZH2 in a cohort of aggressive prostate and breast cancers (Varambally,et al, 2008) (**Figure 5**).

To investigate the role of mir101 in breast cancer, the EZH2 overexpressing breast cancer cell line SKBR3 was used as a model system in various experiments. An inverse correlation between mir101 and EZH2 (and other polycomb group 2 genes) expression levels was observed (**Figure 6**).

To study the role mir101 in regulation of gene expression, we performed chromatin immunoprecipitation (ChIP) assays to evaluate promoter occupancy of the H3K27 histone mark, in SKBr3 cells and EZH2 siRNA–treated cells. We found considerable reduction in the trimethyl H3K27 histone mark at the promoter of known PRC2 target genes in (**Figure 6A**), and this resulted in increased gene expression of the target genes (**Figure 6B**). Gene-expression array analysis of SKBr3 cells transfected with either miR-101 or EZH2 siRNA duplexes showed significant overlap in gene expression. SKBr3 cells treated with precursor miR-101 or siRNA targeting EZH2 reduced proliferation, but ectopically overexpressing EZH2 lacking its 3'UTR rescued the proliferation levels, further confirming the regulation of EZH2 by mir101. Use of miR-101 antagonists (antagomiRs to miR101) induced an invasive phenotype in benign immortalized H16N2 breast epithelial cells (**Figure 7**).

The genomic loss of miR-101 in cancer leads to overexpression of EZH2 and concomitant dysregulation of epigenetic pathways, a key molecular event in cancer progression.

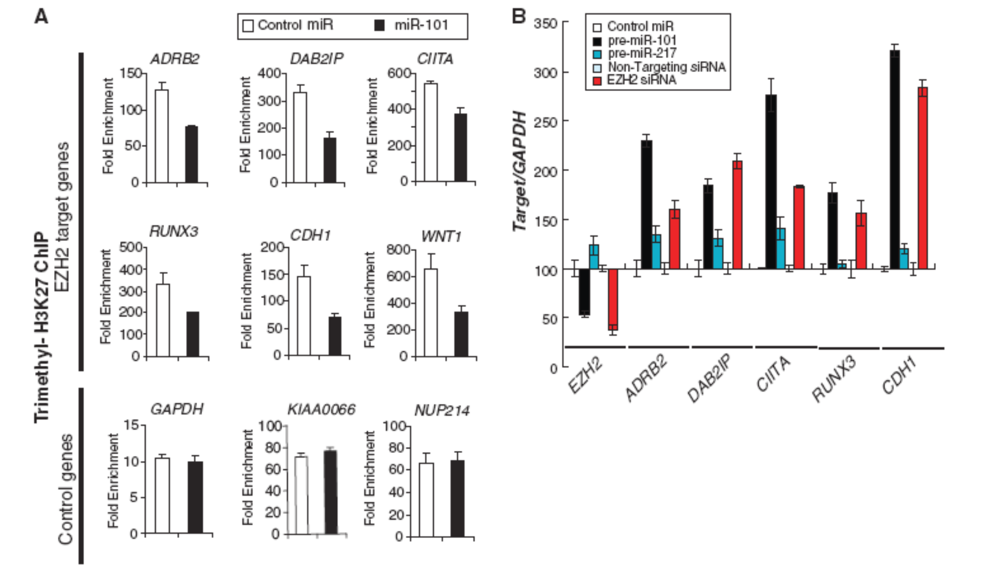


**Figure 6.** miR-101 regulation of the cancer epigenome through EZH2 and H3K27 tri-methylation. (**A**) Chromatin immunoprecipitation (ChIP) assay of the trimethyl H3K27 histone mark when miR-101 is overexpressed. Known PRC2 repression targets were examined in SKBr3 cells. ChIP was performed to test H3K27 trimethylation at the promoters of ADRB2, DAB2IP, CIITA, RUNX3, CDH1 and WNT1. GAPDH, KIAA0066 and NUP214 gene promoters served as controls. (**B**) qRT-PCR of EZH2 target genes was performed using SKBr3 cells transfected with miR-101. The EZH2 transcript and its known targets including ADRB2, DAB2IP, CIITA, RUNX3 and E-cadherin (CDH1) were measured. *Science*. 2008 Dec 12;322(5908):1695-9

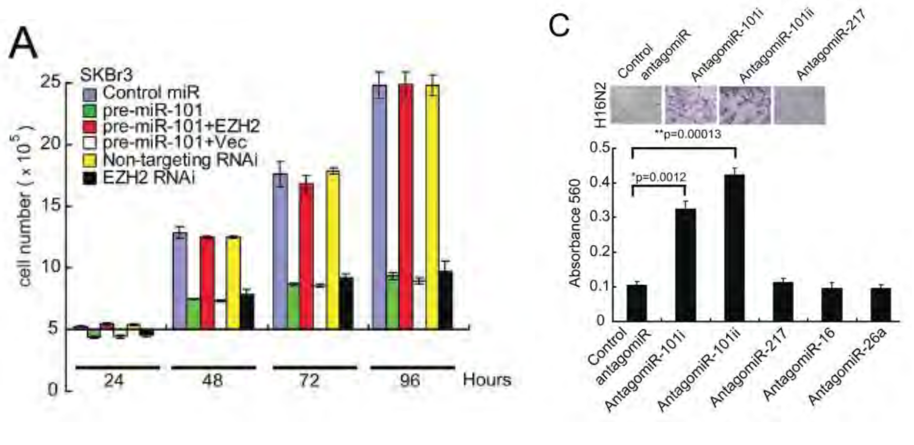


**Figure 7.** The role of miR-101 in regulating cell proliferation, invasion and tumor growth. (**C**) AntagomiRs to miR-101 induce invasion in benign immortalized H16N2 breast epithelial cells. *Science*. 2008 Dec 12;322(5908):1695-9

## 4. Recurrent rearrangements of the MAST kinase and Notch gene families in breast cancer

We employed paired-end transcriptome sequencing to explore the landscape of gene fusions on a panel of 89 breast cancer cell lines and tumors. We observed that individual breast cancers harbor an array of expressed gene fusions, most of which are likely ‗passenger‘ events. In this background of amplification- and rearrangement- induced complexity, we identified two classes of rare but recurrent gene rearrangements in breast cancer; we identified five instances of fusions of microtubule associated serine threonine kinase (MAST) family kinases (**Figure 8**) and eight instances of fusions of genes in the Notch family genes (**Figure 10**). Our analysis suggests that the MAST fusions were present in ~ 3-5% of breast cancers. Knockdown of MAST2 showed significant inhibitory effects on growth in MAST2-postive MDA-MB-468 cells but not in the fusion-negative cell line BT-483 or in benign TERT-HME1 breast cells (**Figure 9a**). We then cloned and overexpressed all five MAST1 and MAST2 fusions in TERT-HME1 cells; cells overexpressing the MAST fusion genes displayed increased cell proliferation (**Figure 9b**). MDA-MB-468 cells treated with MAST2 silencing shRNA had a reduction in growth as assessed by colony formation assay (**Figure 9c**), and in the mouse xenograft model, MDA-MB-468 cells transiently transfected with MAST2 shRNA did not establish palpable tumors over a time course of 4 weeks after transfection (**Figure 9d**). Our studies show that the MAST gene fusions are a key driver of tumorigenesis in a sub-set of breast cancer.

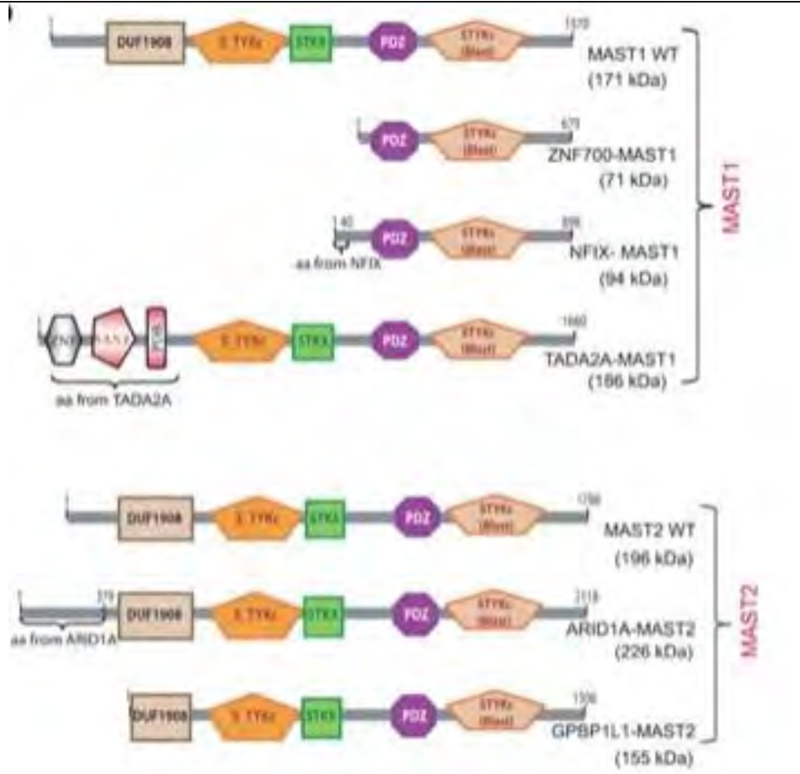


**Figure 8.** Schematic representation of functional domains retained in the putative chimeric proteins involving MAST1 and MAST2. *Nat Med.* 2011 Nov 20;17(12):1646-51.

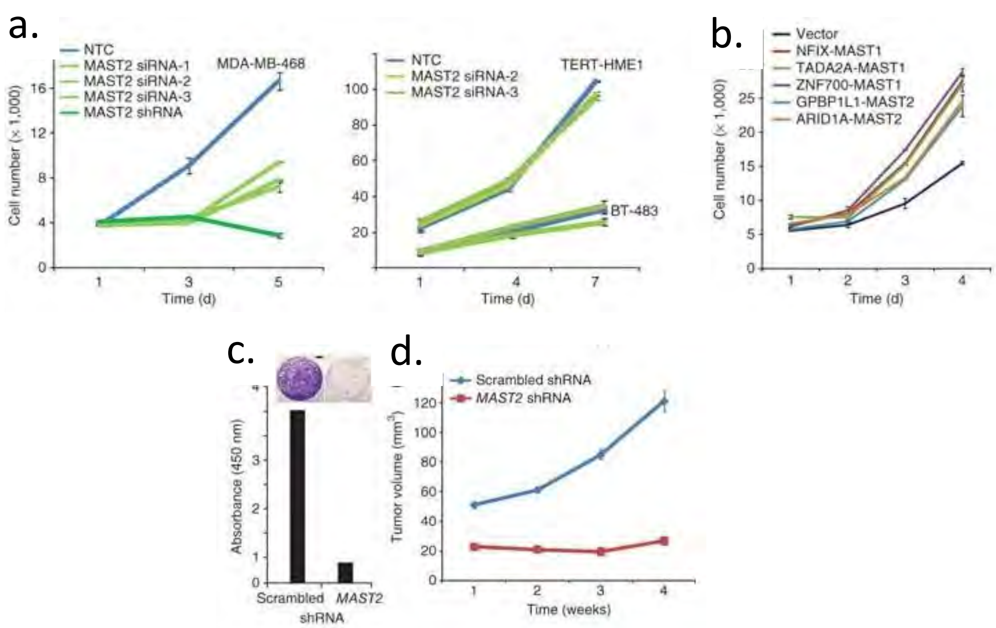


**Figure 9.** (a) Cell count proliferation assay following MAST2 knockdown in MDA-MB-468 cells (left) and fusion-negative TERT-HME1 or BT-483 cells (right). (b) Cell count proliferation assay following overexpression of MAST chimera in TERT-HME1 cells. (c) A colony formation assay of MDA-MB-468 cells treated with MAST2-specific shRNA or control scrambled shRNA. The inset shows crystal violet staining of cells treated with either scrambled or MAST2 shRNA. (d) Tumor volumes of MDA-MD-468 xenografts with MAST2 knockdown. Error bars represent s.d. *Nat Med.* 2011 Nov 20;17(12):1646-51.

We identified eight Notch gene fusions (NOTCH 1 or NOTCH 2, **Figure 10**), all of which occurred in ER$^-$ carcinoma (one was found in a triple-negative carcinoma). All the fusion transcripts retained the exons that encode the Notch intracellular domain (NICD) that is responsible for inducing the transcriptional program following Notch activation. We co-transfected three fusion alleles along with a Notch reporter plasmid and all three induced Notch-responsive transcription that was equivalent to native NICD (**Figure 11b**). Furthermore, the

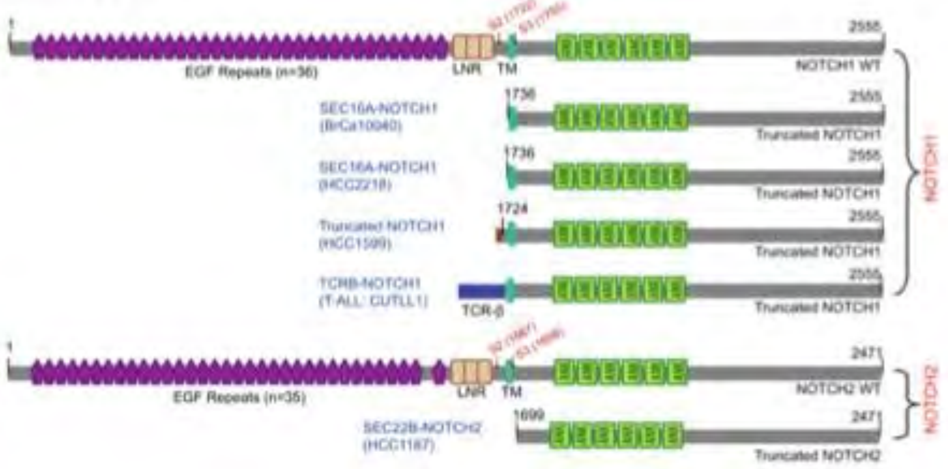


**Figure 10.** Schematic presentation of the predicted protein structures of the aberrant Notch genes. *Nat Med.* 2011 Nov 20;17(12):1646-51.

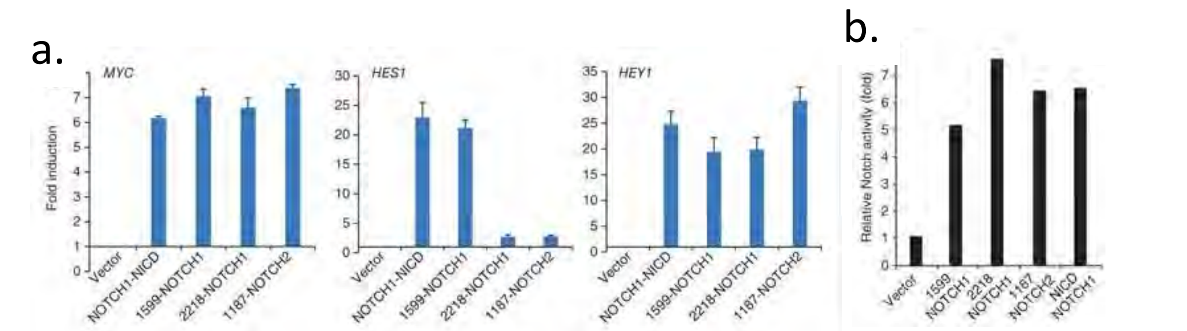


**Figure 11**. (a) mRNA expression levels of three Notch target genes inTERT-HME1 cells stably expressing Notch fusions, measured by quantitative RT-PCR. (b) Notch reporter luciferase levels in 293T cells, assayed following transient Notch expression. Error bars represent s.d. *Nat Med.* 2011 Nov 20;17(12):1646-51.

fusion alleles markedly induced expression of the Notch target genes MYC, HES1 and HEY1 (**Figure 11a**).

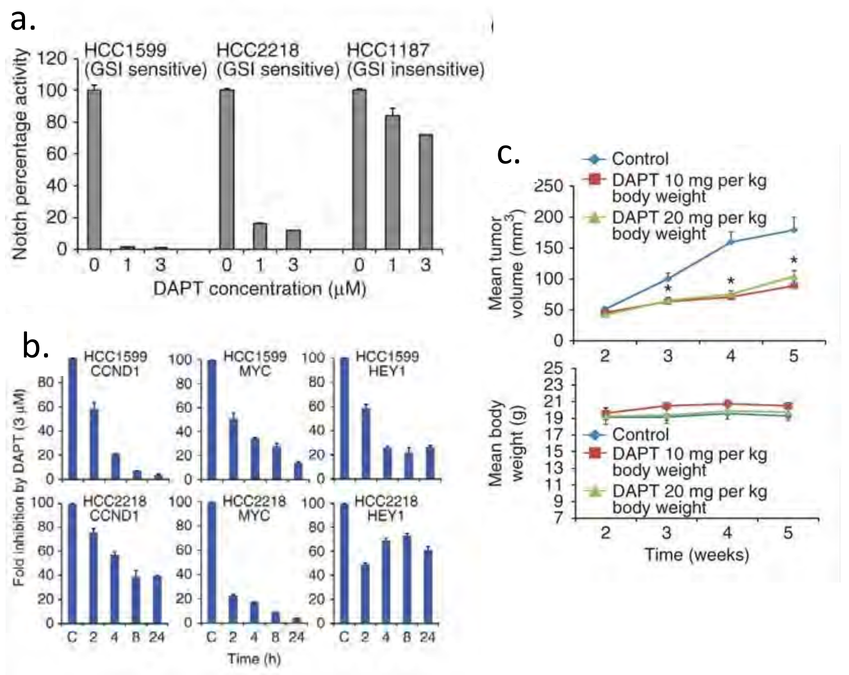


**Figure 12**. (a) Luciferase assay of the Notch signaling pathway following DAPT treatment. Breast cancer cells were co-infected with a Notch reporter construct, lenti-RBPJ firefly luciferase, and the internal control lenti-Renilla luciferase. Twenty-four hours after treatment with DAPT, luciferase activities were measured. (d) Expression of Notch target genes after treatment with DAPT, as measured by qRT-PCR. (e) Xenograft tumor volume and body mass after treatment with the γ-secretase inhibitor DAPT. Mice xenografted with HCC1599 cells were treated daily after tumors formed, and the size of the tumors was monitored. *Nat Med.* 2011 Nov 20;17(12):1646-51

The Notch fusions represent two functional classes with respect to dependence on the activity of γ-secretase. Fusions in BrCa10040, HCC2218 and HCC1599 cells are dependent on S3 cleavage for activity and are sensitive to γ-secretase inhibitors (GSIs). The fusion class in HCC1187 cells is independent of S3 cleavage. We established stable Notch reporter lines from each of the three Notch fusion index lines and treated them with the γ-secretase inhibitor N-[(3,5-difluorophenyl)acetyl]-L-al anyl-2-phenyl]glycine-1,1-dimethylethyl ester (DAPT). We saw a reduction of Notch reporter activity after treatment with DAPT in the HCC1599 and HCC2218 cells but not in HCC1187 cells (**Figure 12a**). Treatment with DAPT also repressed the expression of the Notch targets CCND1, MYC and HEY1 (**Figure 12b**). Finally, treatment with

DAPT significantly reduced tumor volume in a xenograft tumor model of HCC1599 cells but did not affect body weight (**Figure 12c**).

The discovery of functionally recurrent MAST and Notch fusions in a subset of breast carcinomas is a promising path for future research and treatment in breast cancer and illustrates the power of next-generation sequencing as a tool in the development of personalized medicine.

## 5. Next Generation Sequencing Analysis: ChimeraScan

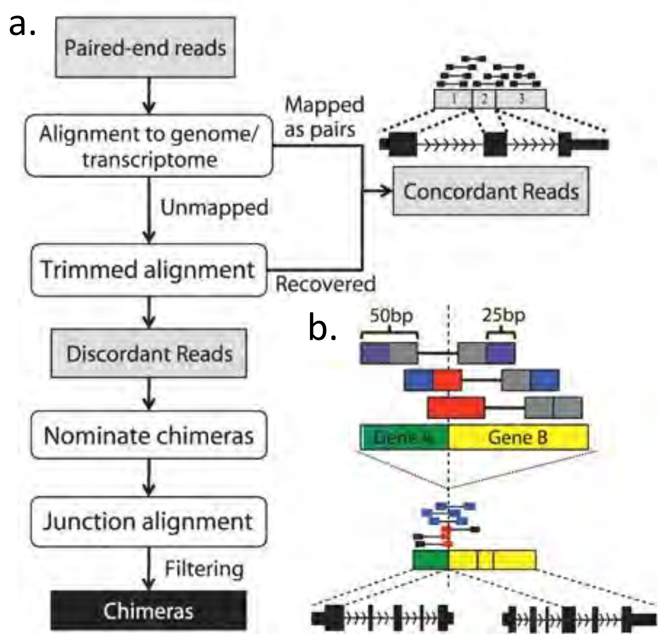


**Figure 13.** ChimeraScan flowchart. (a) Paired-end reads failing an initial alignment step are segmented and realigned to detect discordant reads. Discordant reads that pass filter criteria are realigned across putative chimeric junctions. (b) Chimera with encompassing (blue) and spanning (red) segments detected during realignment. *Bioinformatics.* 2011 Oct 15;27(20):2903-4

We previously used high-throughput paired-end transcriptome sequencing (RNA-Seq) to detect aberrant, chimeric RNAs and uncovered recurrent classes of clinically relevant gene fusions such as those found in breast cancer described above. This discovery was facilitated by the development of an open-source software package, ChimeraScan, for the discovery of chimeric transcription between two independent transcripts in high-throughput transcriptome sequencing data (schematic shown in **Figure 13**). ChimeraScan includes features such as the ability to process long (>75 bp) paired-end reads, processing of ambiguously mapping reads, detection of reads spanning a fusion junction, integration with the popular Bowtie aligner, supports the standardized SAM format and generation of HTML reports for easy investigation of results. Overall, we believe that the ChimeraScan will facilitate the discovery of additional gene fusions that may serve as clinically relevant targets in cancer.

## MicroRNAs mediate coordinate PRC2 and PRC1 function

Polycomb group (PcG) proteins form polycomb repressive complexes (PRC), PRC1 and PRC2. They play a critical role in normal development and when dysregulated, contribute to carcinogenesis. Earlier, we had identified EZH2, the methyltransferase subunit of the PRC2 complex, as a biomarker that can be used to molecularly stratify breast and prostate cancers to aid in the identification of aggressive disease (Kleer et al, 2003). We further studied the role and mechanism of EZH2 in cancer and discovered that microRNA miR-101 can attenuate EZH2 expression, and importantly miR-101 is significantly down-regulated in metastatic cancers (Varambally et al, 2008). Here, employing *in vitro* and *in vivo* cancer models and human tumor studies, we demonstrated for the first time that key microRNAs link PRC2 and PRC1 forming an integral regulatory axis of the epigenetic silencing machinery.

Using miRNA target analysis (www.targetscan.org), we identified 14 miRNAs as top candidates with the following properties: (1) upregulated by EZH2 knockdown in DU145 cancer cells which express high levels of PRC2; (2) higher in benign cell lines compared with DU145 cells, and (3) predicted to bind to the 3′ untranslated region (UTR) of target PRC1 components based on TargetScan (**Figure 14**).

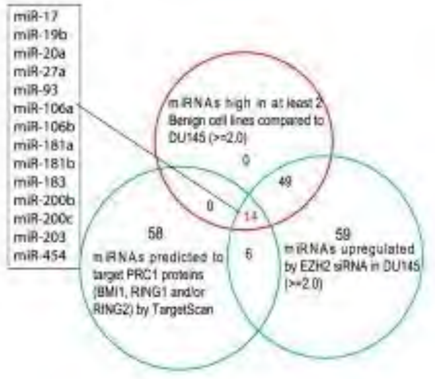

miR-17
miR-19b
miR-20a
miR-27a
miR-93
miR-106a
miR-106b
miR-181a
miR-181b
miR-183
miR-200b
miR-200c
miR-203
miR-454
miRNAs high in at least 2 Benign cell lines compared to DU145 (>=2.0)
0
49
0
14
58
59
6
miRNAs predicted to target PRC1 proteins (BMI1, RING1 and/or RING2) by TargetScan
miRNAs upregulated by EZH2 siRNA in DU145 (>=2.0)

**Figure 14.** A Venn diagram depicting 14 miRNAs that were upregulated by EZH2 knockdown, had high endogenous levels in normal cells, and were predicted to target PRC1 proteins. *Cancer Cell.* 2011 Aug 16;20(2):187-99.

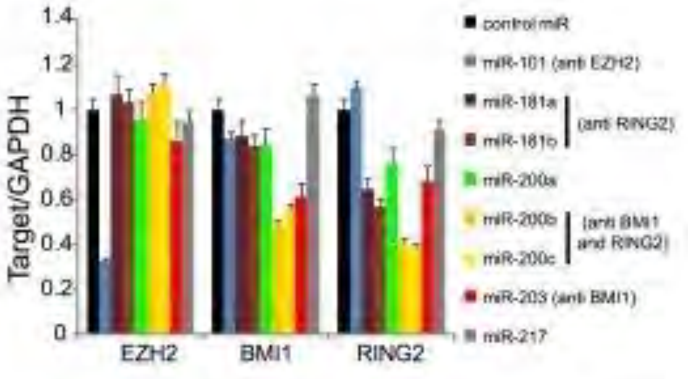


**Figure 15.** Overexpression of indicated miRs in BT-549 cells and expression of PRC components, EZH2, BMI1 and RING2, transcript levels were by qPCR. *Cancer Cell.* 2011 Aug 16;20(2):187-99.

We next overexpressed each of the EZH2-regulated miRNAs in BT-549 breast cancer cell line and found that miR-181a,b and miR-200b,c decreased RING2 transcript levels and miR-200b,c and miR-203 decreased BMI1 transcript levels (**Figure 15**), RING2 and BMI1 are both PRC1 proteins. We hypothesized that the EZH2-regulated microRNAs act as tumor suppressors. Consistent with this notion, overexpression of either miR-181a, miR-181b, miR-200a, miR-200b, miR-200c, or miR-203 markedly attenuated BT-549 cell proliferation to levels similar to that of cells transfected with EZH2 silencing siRNA, or cells overexpressing miR-101 (**Figure 16a**). Likewise, overexpression of either miR-181a, miR-181b, miR-200a, miR-200b, miR-200c, or miR-203 inhibited the *in vitro* invasive potential of BT-549 and

DU145 (prostate cancer) cells through modified Boyden chambers coated with Matrigel (**Figure 16b**). Thus this study furthers our insight on the molecular mechanisms by which EZH2 mediates its oncogenic effects and the microRNA pathway may serve as potential therapeutic targets.

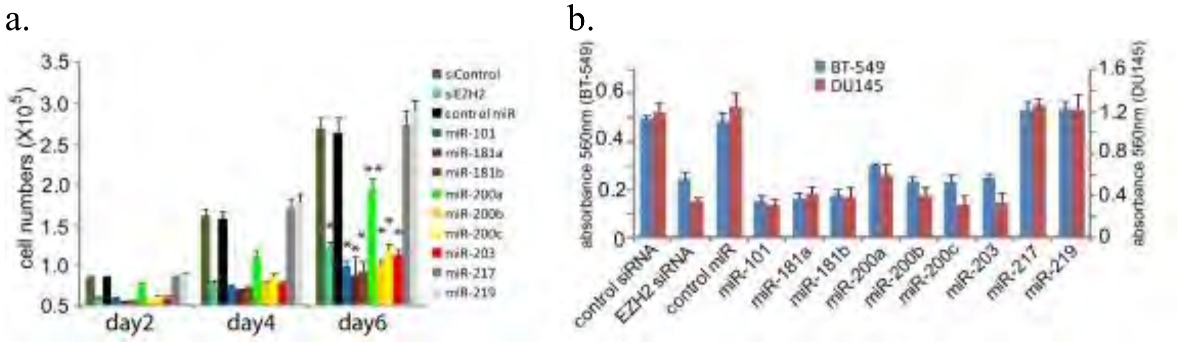


**Figure 16.** (a) Overexpression of PRC2-regulated miRNAs, but not control miR-217 or miR-219, inhibited BT-549 cell proliferation. EZH2 siRNA and miR-101 overexpression were positive controls and miR-217 and miR-219 overexpression were negative controls. (b) Overexpression of PRC2-regulated miRNAs decreased BT-549 and DU145 cell invasion *in vitro*. *Cancer Cell.* 2011 Aug 16;20(2):187-99.

## 6. Characterization of amplicon-associated gene fusions in breast cancer:

Application of high-throughput transcriptome sequencing has spurred highly sensitive detection and discovery of gene fusions in cancer, but distinguishing potentially oncogenic fusions from random, ―passenger" aberrations has proven challenging. We examined a distinctive group of gene fusions that involve genes present in the loci of chromosomal amplifications—a class of oncogenic aberrations that are widely prevalent in breast cancers. Integrative analysis of a panel of 14 breast cancer cell lines comparing gene fusions discovered by high-throughput transcriptome sequencing and genome-wide copy number aberrations assessed by array comparative genomic hybridization led to the identification of 77 gene fusions, of which more than 60% were localized to amplicons including 17q12, 17q23, 20q13, chr8q, among others. Many of these fusions appeared to be recurrent or involved highly expressed oncogenic drivers, frequently fused with multiple different partners, but sometimes displaying loss of functional domains.

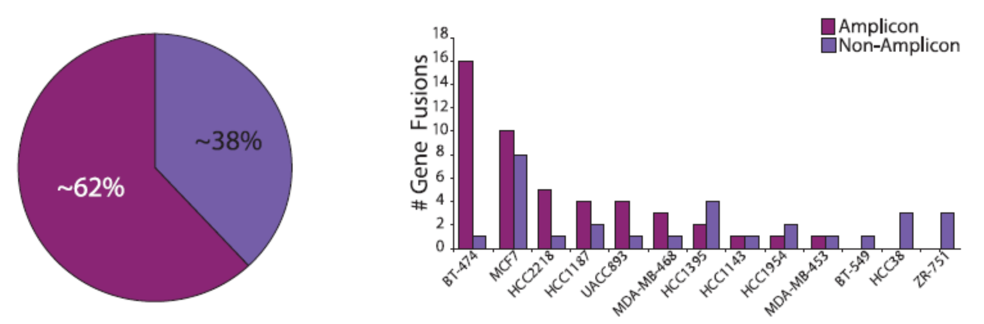


**Figure 17**. Distribution of gene fusions across breast cancer cell lines. Pie chart representation of the relative proportion of gene fusions associated with loci of genomic amplifications compared to unamplified loci (left) and bar graph representation of the relative distribution of gene fusions across different breast cancer cell lines (right). *Neoplasia*. 2012 Aug;14(8):702-8

Here we carried out a systematic analysis of the association between gene fusions and genomic amplification by integrating RNA-Seq data with array comparative genomic hybridization (aCGH)–based whole genome copy number profiling from a panel of breast cancer cell lines. We examined a set of ―amplicon-associated gene fusions” that refer to all the fusions where one or both gene partners are localized to a site of chromosomal amplification. We found that as many as 62% of the total fusions were associated with regions of amplifications (**Figure 17**).

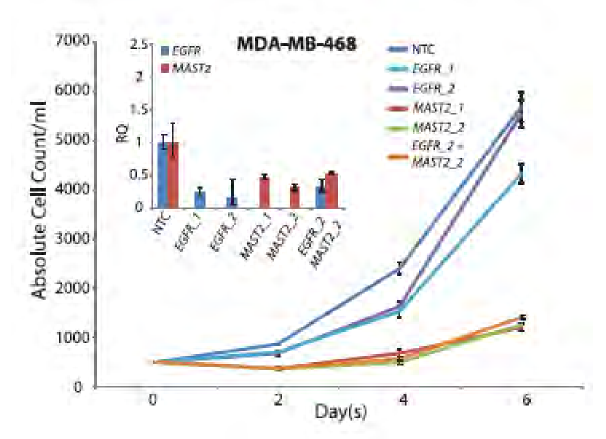


**Figure 18**. Proliferation assay showing absolute cell count (y axis) over a time course (x axis) after knockdown with EGFR and/or MAST2 siRNAs in MDA-MB-468. QPCR assessment of knockdown efficiencies relative to nontargeted control (NTC; inset). *Neoplasia*. 2012 Aug;14(8):702-8

We next assessed the functional relevance of two amplicon-associated fusion genes involving oncogenic kinases, EGFR and RPS6KB1, in the context of prioritizing fusion candidates important in tumorigenesis. In our transcriptome sequencing compendium of 89 breast cancer cell lines and tissues, the highest expression of EGFR is observed inMDA-MB-468, potentially resulting from a focal amplification at chr7p12. In addition, we detected an EGFR fusion transcript (EGFR-POLD1) in this cell line, encoding the N-terminal portion of EGFR, completely devoid of the tyrosine kinase domain. Considering that the MDA-MB-468 harbors both MAST2 and EGFR fusions, we wanted to assess its relative ―dependence” on both the kinases. Surprisingly, a profound reduction in cell proliferation was observed on siRNA knockdown of MAST2, whereas EGFR knockdown showed little effect (**Figure 18**). Next, testing the possibility of EGFR amplicon potentially cooperating with MAST2, we found that the effect of combined knockdown of EGFR and MAST2 was comparable with that of MAST2 knockdown alone (**Figure 18**), further suggesting that EGFR amplification does not signify a driver aberration.

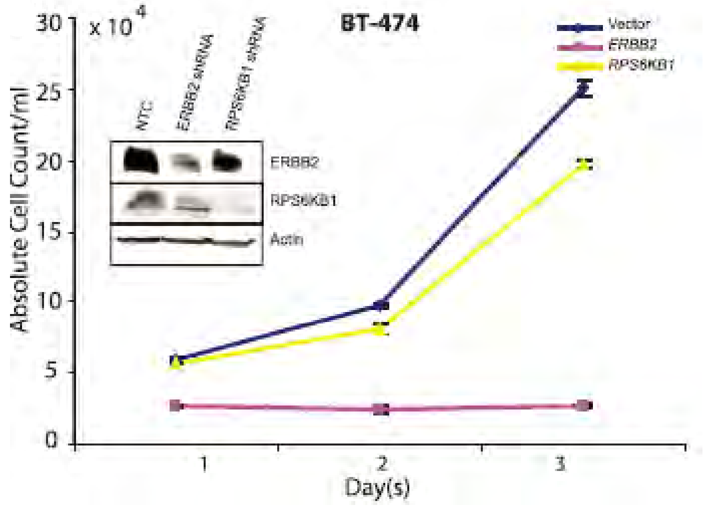


**Figure 19.** Proliferation assay showing absolute cell count (y axis) over a time course (x axis) after knockdown with EGFR and/or MAST2 siRNAs in MDA-MB-468. QPCR assessment of knockdown efficiencies relative to nontargeted control (NTC; inset). *Neoplasia*. 2012 Aug;14(8):702-8

Next, considering that BT-474 is an ERBB2-positive cell line, we tested potential dependence of these cells on the RPS6KB1 protein. Surprisingly, similar to our observations with EGFR knockdown in MDA-MB-468 cells, here we observed only a small effect on cell proliferation after shRNA knockdown of RPS6KB1, in dramatic contrast to the effect of ERBB2 knockdown

(**Figure 19**). Notably, the shRNA knockdown of RPS6KB1 led to a significant depletion of the full-length protein yet it did not affect cell proliferation compared with ERBB2 protein depletion (Figure 3, inset). Therefore, BT-474 cells do not display a dependence on RPS6KB1 protein, and considering that the RPS6KB1 fusion product is completely devoid of all functional domains of RPS6KB1, including the kinase domain, this fusion also likely represents a passenger event.

Overall, our study suggests that amplicon-associated gene fusions in breast cancer primarily represent a by-product of chromosomal amplifications that constitutes a subset of passenger aberrations and should be factored accordingly during prioritization of gene fusion candidates.

## 7. Next Generation Sequencing Analysis

Pseudogenes are a class of non-coding RNA transcripts that are dysfunctional relatives of known functional genes that have lost their protein coding ability and often not expressed. Aberrant expression of several functional non-coding RNA in cancer has been previously described, however genome-wide expression of pseudogenes had not been reported for any cancer type. We developed a pseudogene expression pipeline to analyze a large compendium of paired-end next generation sequencing (RNASeq) data generated from 293 samples, comprising 13 different epithelial cancers. Our integrative approach provided evidence of expression for 2,082 distinct pseudogenes that displayed lineage-specific, cancer-specific, as well as ubiquitous expression patterns.

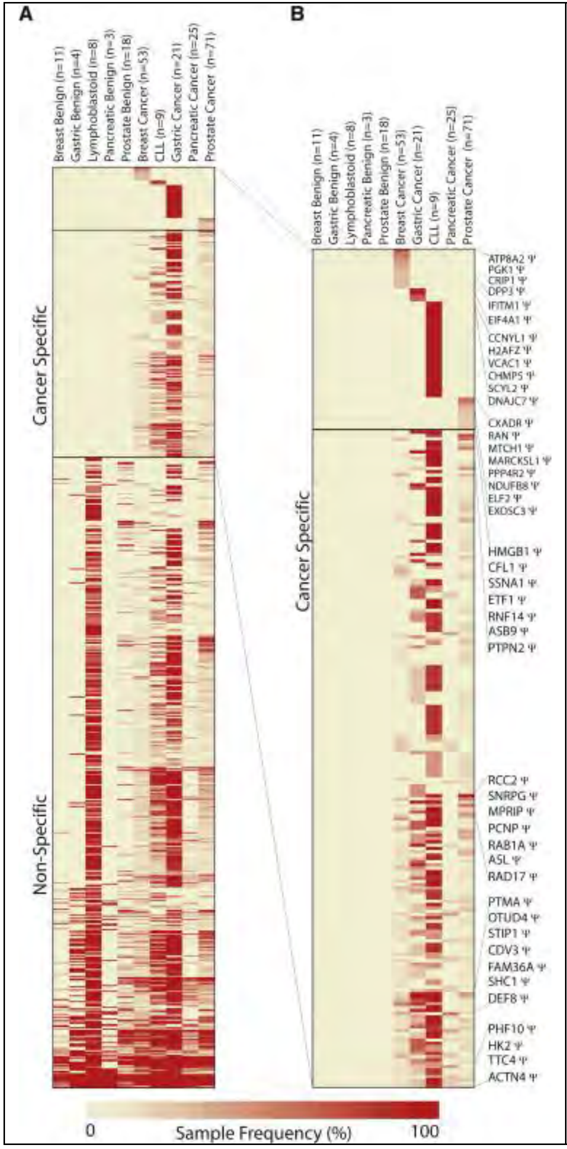


**Figure 20**. Cancer-Specific Pseudogene Expression Profiles(A) Heatmap of pseudogene expression sorted according to cancer-specific expression patterns displays pseudogene transcripts specific to individual cancers (top), common across multiple cancers (tissue-enriched; middle), and nonspecific (bottom).(B) Zoomed-in version of the top panel displaying individual cancer-specific expressed pseudogenes. The columns represent different tissues with the number of samples in parentheses. The rows represent individual clusters mapping to specific pseudogenes. The color intensity represents the frequency (%) of samples in a tissue type showing expression of a given pseudogenes (according to the scale indicated at the bottom). *Cell*. 2012 Jun 22;149(7):1622-34.

Though a majority of the pseudogenes examined were found in both cancer and benign samples, we observed 218 pseudogenes expressed only in cancer samples, of which 178 were observed in multiple cancers and 40 were found to have highly specific expression in a single cancer type only (**Figure 20**).

Among the pseudogene candidates in breast cancer, we identified an unprocessed pseudogene cognate to ATP8A2, a LIM domain-containing protein speculated to be associated with stress response and proliferative activity. ATP8A2-Ψ expression found to be restricted to breast samples, the highest levels seen in a subset of breast cancer tissues and cell lines (**Figure 18**). By contrast, ATP8A2-WT expression was highly variable across different tissue types and showed no correlation with ATP8A2-Ψ expression. To investigate a potential role of ATP8A2-Ψ expression in breast cancer, first we carried out siRNA-based knockdown of both the wild-type and pseudogene RNA in two independent breast cancer cell lines that expressed both the transcripts. Knockdown of ATP8A2-Ψ with two independent siRNAs was found to specifically inhibit the proliferation of overexpressing cell lines Cama-1 and HCC1806 (**Figure 21A**), but not the cell lines with no detectable levels of ATP8A2-Ψ, for example, the benign breast epithelial cell line H16N2 (**Figure 21A, right**). Knockdown of ATP8A2-Ψ (but not ATP8A2-WT) also resulted in reduced cell migration and invasion seen in *in vitro* Boyden Chamber assays (**Figure 21B**) as well as in *in vivo* intravasation and metastasis in chicken chorioallantoic mem-brane xenograft assay (**Figure 21C**). In contrast, knockdown of wild-type ATP8A2 had no effect on the proliferation of any of the cell lines tested, suggesting an un-expected growth reg-ulatory role for ATP8A2-Ψ.

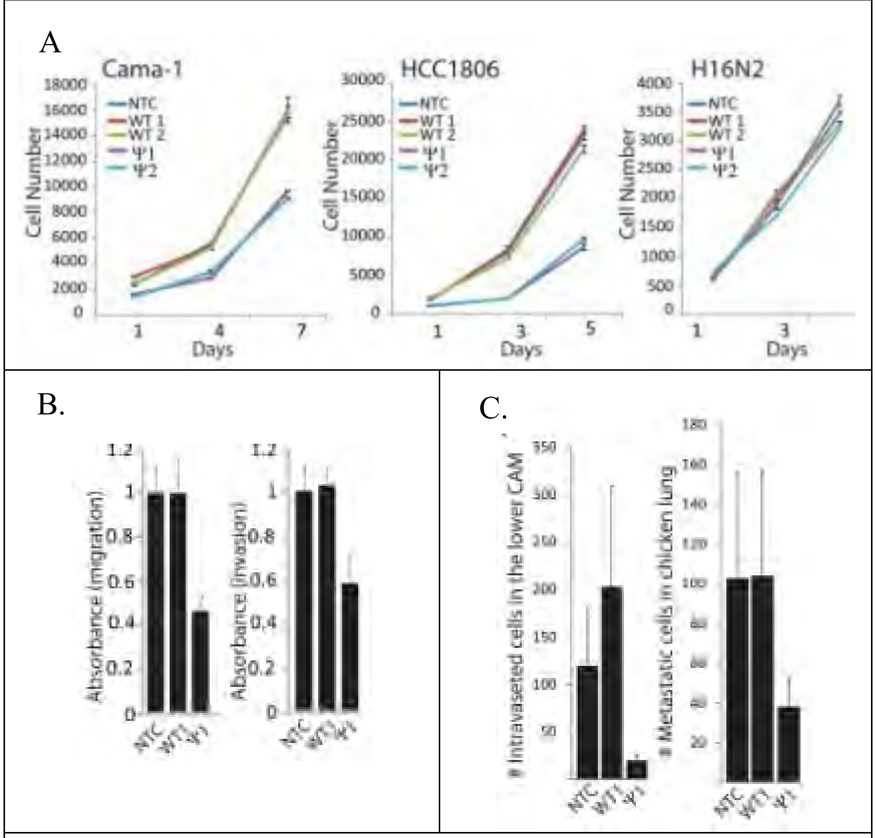


**Figure 21.** (A) Cell proliferation assays following siRNA knockdowns of ATP8A2-WT and -Ψ as indicated. NTC, nontargeting control; WT, siRNA against wild-type ATP8A2; Ψ, siRNA against ATP8A2-Ψ.(B) Boyden chamber assay showing cell migration (left) and invasion through matrigel (right). (C) chicken chorioallantoic membrane assay of HCC-1806 cells treated with nontargeting control siRNA, ATP8A2-WT, or ATPA2-Ψ siRNA showing relative number of cells intravasated in the lower CAM (left) and metastatic cells in chicken lung (right).Error bars represent means ± SE of the mean. *Cell.* 2012 Jun 22;149(7):1622-34.

This study is the first large-scale analysis of pseudogene expres-sion in human cancer using transcriptome sequencing data.

## 9. RNA-Seq identifies targetable kinases in cancer:

We analyzed transcriptome sequencing data from a compendium of 482 cancer and benign samples from 25 different tissue types to assess the complete landscape of a cancer‘s ―kinome” expression and determine which kinases are activated in specific tumor types. Protein kinases represent the most effective class of therapeutic targets in cancer; therefore, determination of kinase aberrations is a major focus of cancer genomic studies. Here, we found frequent outlier kinase expression in breast cancer included therapeutic targets like ERBB2 and FGFR4 whereas MET, AKT2, and PLK2 were expressed in pancreatic cancer. Outlier expression of polo-like kinases was observed in a subset of KRAS-dependent pancreatic cancer cell lines and conferred increased sensitivity to the pan-PLK inhibitor BI-6727. These results suggest that outlier kinases are effective therapeutic targets and can be readily identified through RNA sequencing of tumors.

## 10. Identification of targetable FGFR gene fusions in diverse cancers

Earlier, we established the Michigan Oncology Sequencing Center (MI-ONCOSEQ) clinical sequencing program that prospectively enrolls patients with advanced cancers for comprehensive mutational analysis with the goal of identifying informative and/or actionable mutations. In four index MI-ONCOSEQ cases, we identified gene rearrangements of FGFR2, including patients with cholangiocarcinoma, breast cancer, and prostate cancer. We then extended the screening of FGFR rearrangements across multiple tumor cohorts and identified additional FGFR fusions with intact kinase domains in lung squamous cell cancer, bladder cancer, thyroid cancer, oral cancer, glioblastoma, and head and neck squamous cell cancer (**Figure 22**). Four FGFR gene fusions were found specifically in breast cancer, three involving FGFR2 and one with FGFR1 (**Figure 23**).

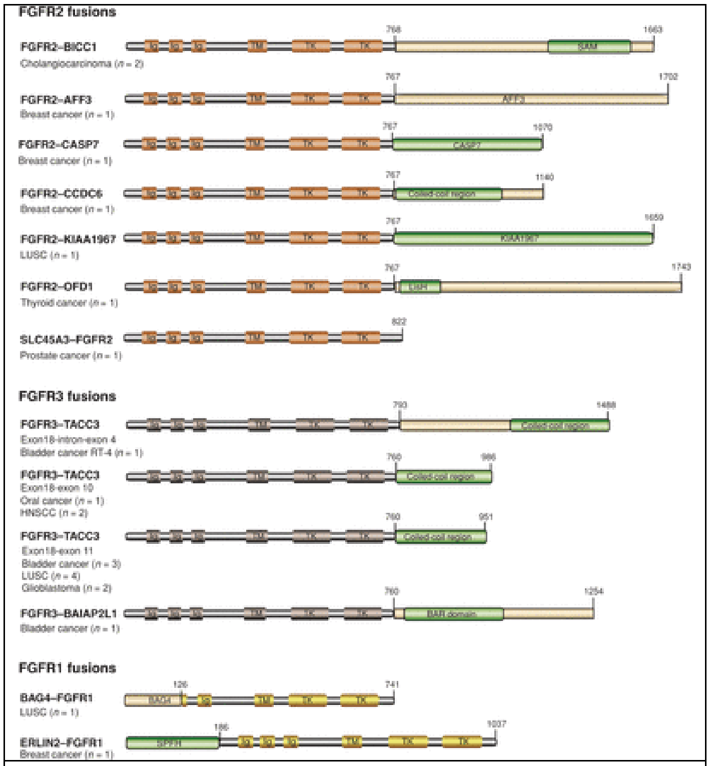


**Figure 22.** Schematic representations of the predicted FGFR fusions identified by transcriptome sequencing of human cancers. Data used include RNA sequencing results from the 4 index patients, our internal tumor cohort, and the TCGA compendium. Out of 4 FGFR receptor family members, FGFR1, FGFR2, and FGFR3 are involved in gene fusions with various partners located on several chromosomes. Eleven distinct fusion partners of FGFRs were identified. *Cancer Discov.* 2013 Jun;3(6):636-47

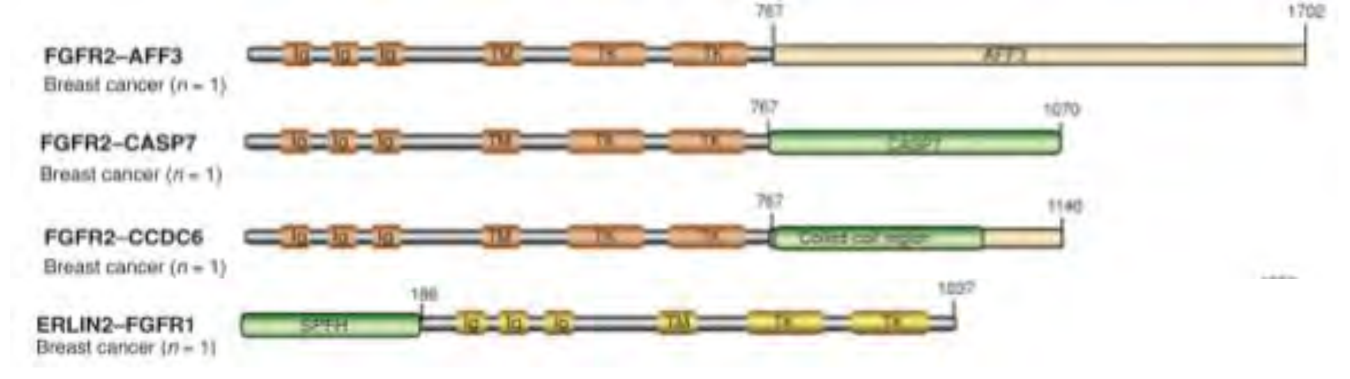


**Figure 23**. FGFR gene fusions found specifically in breast cancer. *Cancer Discov.* 2013 Jun;3(6):636-47.

All FGFR fusion partners tested exhibited oligomerization capability, suggesting a shared mode of kinase activation. Overexpression of FGFR fusion proteins induced cell proliferation in TERT-HME cells, including FGFR2-CCDC6 discovered in breast cancer (**Figure 24**).

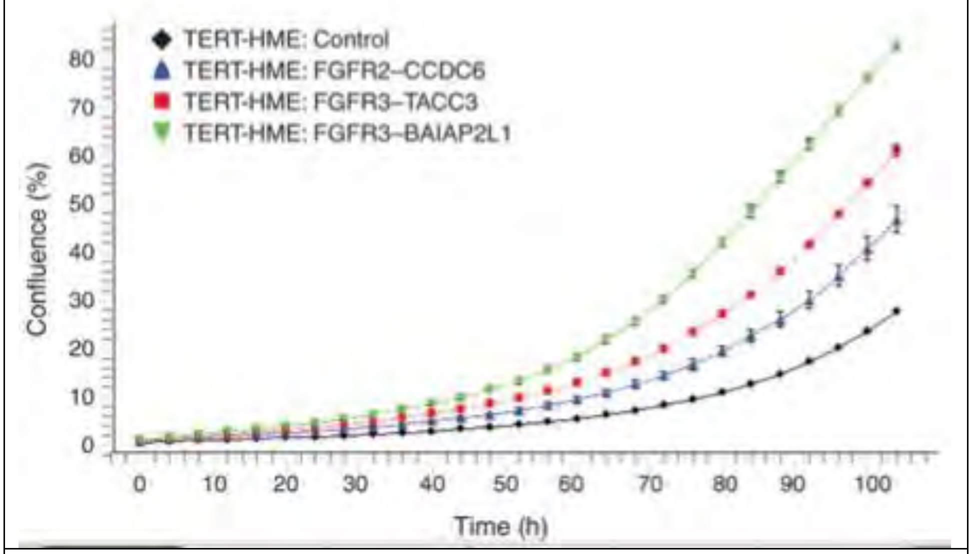


**Figure 24**. Overexpression of FGFR fusions induces cell proliferation in TERT-HME cells. Cell proliferation assays were conducted by Incucyte live-cell imaging. Data shown are cell confluence versus time at 3-hour intervals. Each data point is the mean of quadruplicates. *Cancer Discov.* 2013 Jun;3(6):636-47.

Two bladder cancer cell lines that harbor FGFR3 fusion proteins exhibited enhanced susceptibility to pharmacologic inhibition *in vitro* and *in vivo*. Because of the combinatorial possibilities of FGFR family fusion to a variety of oligomerization partners, clinical sequencing approaches that incorporate transcriptome analysis for gene fusions have the potential to identify rare, targetable FGFR fusions across diverse cancer types.

## 11. Gene sequencing reveals mutations in estrogen receptor

Through a prospective clinical sequencing program (MI-ONCOSEQ) for advanced cancers, we enrolled 9 ER-positive metastatic breast cancer patients. The samples are then subjected to integrative sequencing which includes whole exome and transcriptome analysis that allows a mutational landscape of coding genes including point mutations, indels, amplifications, deletions, gene fusions/translocations, and outlier gene expression.

The most remarkable observation in the mutational landscape of these treated ER positive patients was the finding of nonsynonymous mutations in the ligand binding domain (LBD) of ESR1 (n=4). The four index patients MO_1031, MO_1051, MO_1069, and MO_1129 had LBD mutations in amino acids L536Q, Y537S, D538G, and Y537S, respectively. All had been treated with anti-estrogens and estrogen deprivation therapies. A survey of The Cancer Genome Atlas (TCGA) identified 4 endometrial cancers with similar mutations of ESR1. The 5 novel LBD mutations of ESR1 identified here (L536Q, Y537S, Y537C, Y537N, and D538G, **Figure 25**) were shown to be constitutively active and continue to be responsive to anti-estrogen therapies *in vitro*. Taken together, these studies suggest that activating mutations of ESR1 are an important mechanism of acquired endocrine resistance in breast cancer therapy.

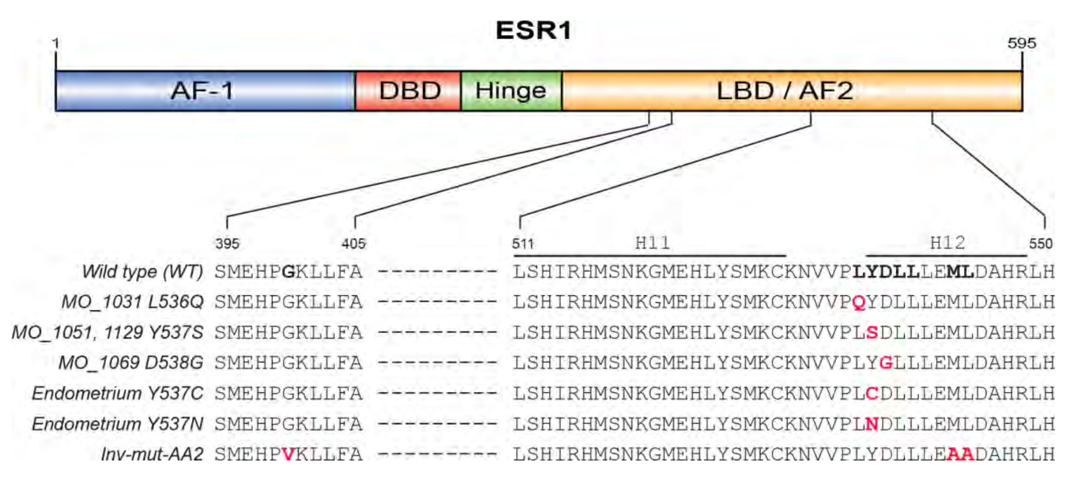

ESR1
1
595
AF-1
DBD
Hinge
LBD / AF2
395
405
511
H11
H12
550
Wild type (WT) SMEHPGKLLFA --------- LSHIRHMSNKGMEHLYSMKCKNVVPLYDLLLEMLDAHRLH
MO_1031 L536Q SMEHPGKLLFA --------- LSHIRHMSNKGMEHLYSMKCKNVVPQYDLLLEMLDAHRLH
MO_1051, 1129 Y537S SMEHPGKLLFA --------- LSHIRHMSNKGMEHLYSMKCKNVVPLSDLLLEMLDAHRLH
MO_1069 D538G SMEHPGKLLFA --------- LSHIRHMSNKGMEHLYSMKCKNVVPLYGLLLEMLDAHRLH
Endometrium Y537C SMEHPGKLLFA --------- LSHIRHMSNKGMEHLYSMKCKNVVPLCDLLLEMLDAHRLH
Endometrium Y537N SMEHPGKLLFA --------- LSHIRHMSNKGMEHLYSMKCKNVVPLNDLLLEMLDAHRLH
Inv-mut-AA2 SMEHPVKLLFA --------- LSHIRHMSNKGMEHLYSMKCKNVVPLYDLLLEAADAHRLH

**Figure 25**: The structural domains of ESR1 are illustrated on top, including the transcription activation function-1 domain (AF-1), the DNA-binding domain (DBD), the hinge domain and the ligand-binding domain (LBD/AF-2). Altered residues identified in mutants are marked in red, and reference residues are shown in bold in the wild-type sequence. Endometrium p.Tyr537Cys and p.Tyr537Asn are two alterations discovered in endometrial cancer samples from the TCGA study. Inv-mut-AA2 represents a ligand activity inversion mutant of ESR1 that confers inverted responses to anti-estrogen and estrogen. H11, helix 11; H12, helix 12. *Nat Genet.* 2013 Dec;45(12):1446-51.

## KEY RESEARCH ACCOMPLISHMENTS:

- We provided a robust and high throughput pipeline for a directed search for gene fusions in cancers using next generation transcriptome sequencing platforms. The comprehensive coverage afforded by this approach would help unravel the chimeric landscape of breast cancer transcriptome- the primary aim of our current project.

- We report the characterization of a subset of ER positive breast cancer patients. This group is characterized by the overexpression of AGTR1, and this subset may be responsive to an available drug, losartan. Our study is expected to lead to follow-up clinical trials.
- We succeeded in providing a novel mechanistic framework for the overexpression of the polycomb group protein EZH2 in metastatic breast and prostate cancers, involving the genomic loss of its negative regulator, mir101.
- We analyzed and screened a shortlist of potentially functional and recurrent gene fusions from a total of 89 breast cancer cell lines and tumors. We discovered two rare but recurrent gene fusions involving MAST and Notch genes. Moreover, these fusions could potentially be targeted by their respective inhibitors.
- We have developed a tool, ChimeraScan that facilitates the analysis of our transcriptome sequencing data and allows us to identify novel rare and common gene fusions in cancer.
- We have further extended our earlier microRNA studies and have identified several miRNAs that are regulated by PRC2 protein, EZH2. In addition these miRNAs in turn regulate PRC1 proteins. This is the first study to demonstrate a molecular link between PRC2 and PRC1 network that is mediated by miRNAs.
- We performed an integrated analysis combining RNASeq and aCGH to examine amplicon-associated gene fusions across 14 breast cancer cell lines. We found that many of these fusions, even when they involve known oncogenes, are often ―passenger” events that do not display oncogenic potential.
- We used a novel bioinformatics approach to analyze next generation sequencing data to discover novel expressed pseudogenes. Although many of the pseudogenes are ubiquitously expressed, we found a sub-set of them are expressed in a lineage and cancer-specific manner, including the breast cancer-specific pseudogene, ATP8A2Ψ.
- Using transcriptome sequencing, we identified targetable kinases in cancer.
- Through our clinical sequencing study, we identified FGFR gene rearrangements in various cancers including breast cancer; the fusion proteins express intact kinase domains that are potentially targetable.
- We identified mutations in the estrogen receptor that is acquired after breast cancer patients take anti-estrogen therapies. These mutations uncover the mechanism by which patients often become resistant to hormone therapy.

## REPORTABLE OUTCOMES (papers included in Appendix):

AGTR1 overexpression defines a subset of breast cancer and confers sensitivity to losartan, an AGTR1 antagonist. Rhodes DR, Ateeq B, Cao Q, Tomlins SA, Mehra R, Laxman B, Kalyana-Sundaram S, Lonigro RJ, Helgeson BE, Bhojani MS, Rehemtulla A, Kleer CG, Hayes DF, Lucas PC, Varambally S, Chinnaiyan AM. Proc Natl Acad Sci U S A. 2009 Jun 23;106(25):10284-9. Epub 2009 Jun 1. PMID: 19487683.

Genomic loss of microRNA-101 leads to overexpression of histone methyltransferase EZH2 in cancer. Varambally S, Cao Q, Mani RS, Shankar S, Wang X, Ateeq B, Laxman B, Cao X, Jing X, Ramnarayanan K, Brenner JC, Yu J, Kim JH, Han B, Tan P, Kumar-Sinha C, Lonigro RJ, Palanisamy N, Maher CA, Chinnaiyan AM. Science. 2008 Dec 12;322(5908):1695-9. Epub 2008 Nov 13. PMID: 19008416.

Chimeric transcript discovery by paired-end transcriptome sequencing. Maher CA, Palanisamy N, Brenner JC, Cao X, Kalyana-Sundaram S, Luo S, Khrebtukova I, Barrette TR, Grasso C, Yu J, Lonigro RJ, Schroth G, Kumar-Sinha C, Chinnaiyan AM. Proc Natl Acad Sci U S A. 2009 Jul 10. [Epub ahead of print]. PMID: 19592507.

Kim JH, Dhanasekaran SM, Prensner JR, Cao X, Robinson D, Kalyana-Sundaram S, Huang C, Shankar S, Jing X, Iyer M, Hu M, Sam L, Grasso C, Maher CA, Palanisamy N, Mehra R, Kominsky HD, Siddiqui J, Yu J, Qin ZS, Chinnaiyan AM. Deep sequencing reveals distinct patterns of DNA methylation in prostate cancer. Genome Res. 2011 Jul;21(7):1028-41. PubMed PMID: 21724842; PubMed Central PMCID: PMC3129246.

Iyer MK, Chinnaiyan AM, Maher CA. ChimeraScan: a tool for identifying chimeric transcription in sequencing data. Bioinformatics. 2011 Oct 15;27(20):2903-4. Epub 2011 Aug 11. PubMed PMID: 21840877; PubMed Central PMCID: PMC3187648.

Robinson D.R., Kalyana-Sundaram S., Wu Y.-I., Shankar S., Cao X., Ateeq B., Asangani I.A., Iyer M., Maher C.A., Grasso C.S., Lonigro R.J., Quist M., Siddiqui J., Mehra R., Jing X., Giordano T.J., Sabel M.S., Kleer C.G., Palanisamy N., Natrajan R., Lambros M.B., Reis-Filho J.S., Kumar-Sinha C., and Chinnaiyan A.M. Functionally Recurrent Rearrangements of the MAST Kinase and Notch Gene Families in Breast Cancer. Nat Med. 2011 Nov 20;17(12):1646-51. PMID: 22101766; PMCID: PMC3233654.

Kalyana-Sundaram S, Kumar-Sinha C, Shankar S, Robinson DR, Wu YM, Cao X, Asangani IA, Kothari V, Prensner JR, Lonigro RJ, Iyer MK, Barrette T, Shanmugam A, Dhanasekaran SM, Palanisamy N, Chinnaiyan AM. Expressed pseudogenes in the transcriptional landscape of human cancers. Cell. 2012 Jun 22;149(7):1622-34.
PubMed PMID: 22726445.

Kalyana-Sundaram S, Shankar S, Deroo S, Iyer MK, Palanisamy N, Chinnaiyan AM, Kumar-Sinha C. Gene fusions associated with recurrent amplicons represent a class of passenger aberrations in breast cancer. Neoplasia. 2012 Aug;14(8):702-8. PubMed PMID: 22952423; PubMed Central PMCID: PMC3431177.

Kothari V, Wei I, Shankar S, Kalyana-Sundaram S, Wang L, Ma LW, Vats P, Grasso CS, Robinson DR, Wu YM, Cao X, Simeone DM, Chinnaiyan AM, Kumar-Sinha C. Outlier kinase expression by RNA sequencing as targets for precision therapy. Cancer Discov. 2013 Mar;3(3):280-93. doi: 10.1158/2159-8290.CD-12-0336. Epub 2013 Feb 5. PubMed PMID: 23384775; PubMed Central PMCID: PMC3597439.

Wu YM, Su F, Kalyana-Sundaram S, Khazanov N, Ateeq B, Cao X, Lonigro RJ, Vats P, Wang R, Lin SF, Cheng AJ, Kunju LP, Siddiqui J, Tomlins SA, Wyngaard P, Sadis S, Roychowdhury S, Hussain MH, Feng FY, Zalupski MM, Talpaz M, Pienta KJ, Rhodes DR, Robinson DR, Chinnaiyan AM. Identification of targetable FGFR gene fusions in diverse cancers. Cancer Discov. 2013 Jun;3(6):636-47. doi: 10.1158/2159-8290.CD-13-0050. Epub 2013 Apr 4. PubMed PMID: 23558953; PubMed Central PMCID: PMC3694764.

Robinson DR, Wu YM, Vats P, Su F, Lonigro RJ, Cao X, Kalyana-Sundaram S, Wang R, Ning Y, Hodges L, Gursky A, Siddiqui J, Tomlins SA, Roychowdhury S, Pienta KJ, Kim SY, Roberts JS, Rae JM, Van Poznak CH, Hayes DF, Chugh R, Kunju LP, Talpaz M, Schott AF, Chinnaiyan AM. Activating ESR1 mutations in hormone-resistant metastatic breast cancer. Nat Genet. 2013 Nov 3. doi: 10.1038/ng.2823. [Epub ahead of print] PubMed PMID: 24185510.

## CONCLUSION:

We initiated a search for recurrent gene fusions in breast cancer, in the wake of our discovery and characterization of recurrent gene fusions in prostate cancer. While a majority prostate cancers harbor androgen-regulated Ets family gene fusions (predominantly TMPRSS2-ERG), we hypothesized that breast cancers might harbor estrogen regulated oncogenic gene fusions. Based on our first year‘s work, we have observed that breast cancers harbor multiple gene fusions in most of the samples examined, individual fusions likely do not recur as frequently as they do in prostate cancers. In this respect, breast cancer gene fusions appear closer to the scenario in lung cancer, where multiple gene fusions have been observed in much smaller cohorts of samples.

**“So what?”: Gene fusions represent exquisitely specific cancer biomarkers as well as therapeutic targets,** and while most of the previous gene fusion discoveries have been serendipitous, the development of ultra-high throughput sequencing technologies has enabled us to actively seek out genomic and transcriptomic aberrations. Indeed, our group has successfully applied these techniques to discover gene fusions in cancers at an unprecedented depth of coverage. We identified two rare but recurrent gene fusions in breast cancer cell lines and tissues involving the MAST and Notch genes. The most exciting aspect of our findings is that both MAST and Notch fusions are potentially ―actionable” and patients positive for those gene fusions may benefit from MAST and Notch inhibitors respectively. Our results indicate that breast cancers (like many other solid tumors) are heterogeneous in nature consisting of many rare molecular sub-types. Our ultimate goal is to identify ALL the specific ―actionable” driving gene fusions in individual breast cancer patients. Therefore, the tools that we have developed to identify novel gene fusions along with the functional analysis lays the framework for developing personalized breast cancer therapies based on driving fusion type.

In addition to the discovery of the MAST and NOTCH gene fusions in breast cancer, we make a number of other significant discoveries that advance our understanding of the molecular mechanisms that drive cancer progression and metastasis. Our recent findings of mutations in the estrogen receptor in breast cancer patients that develop after anti-estrogen therapies is a major step in understanding resistance mechanisms.

Overall, these discoveries made over the funding period contribute towards the understanding of the molecular and genetic etiology of breast cancer that will advance the development of targeted therapies.

## REFERENCES:

1. Kleer CG, Cao Q, Varambally S, Shen R, Ota I, Tomlins SA, Ghosh D, Sewalt RG, Otte AP, Hayes DF, Sabel MS, Livant D, Weiss SJ, Rubin MA, Chinnaiyan AM. EZH2 is a marker of aggressive breast cancer and promotes neoplastic transformation of breast epithelial cells. Proc Natl Acad Sci U S A. 2003 Sep 30;100(20):11606-11. Epub 2003 Sep 19. PubMed PMID: 14500907; PubMed Central PMCID: PMC208805.
2. Rhodes DR, Yu J, Shanker K, Deshpande N, Varambally R, Ghosh D, Barrette T, Pandey A, Chinnaiyan AM. ONCOMINE: a cancer microarray database and integrated data-mining platform. Neoplasia. 2004 Jan-Feb;6(1):1-6. PubMed PMID: 15068665; PubMed Central PMCID: PMC1635162.
3. Rhodes DR, Yu J, Shanker K, Deshpande N, Varambally R, Ghosh D, Barrette T, Pandey A, Chinnaiyan AM. Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. Proc Natl Acad Sci U S A. 2004 Jun 22;101(25):9309-14. Epub 2004 Jun 7. PubMed PMID: 15184677; PubMed Central PMCID: PMC438973.
4. 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. Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. Science. 2005 Oct 28;310(5748):644-8. PubMed PMID: 16254181.
5. Rhodes DR, Kalyana-Sundaram S, Mahavisno V, Varambally R, Yu J, Briggs BB, Barrette TR, Anstet MJ, Kincead-Beal C, Kulkarni P, Varambally S, Ghosh D, Chinnaiyan AM. Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles. Neoplasia. 2007 Feb;9(2):166-80. PubMed PMID: 17356713; PubMed Central PMCID: PMC1813932.
6. Varambally S, Cao Q, Mani RS, Shankar S, Wang X, Ateeq B, Laxman B, Cao X, Jing X, Ramnarayanan K, Brenner JC, Yu J, Kim JH, Han B, Tan P, Kumar-Sinha C, Lonigro RJ, Palanisamy N, Maher CA, Chinnaiyan AM: Genomic loss of microRNA-101 leads to overexpression of histone methyltransferase EZH2 in cancer, Science 2008, 322:1695-1699.
7. Rhodes DR, Ateeq B, Cao Q, Tomlins SA, Mehra R, Laxman B, Kalyana-Sundaram S, Lonigro RJ, Helgeson BE, Bhojani MS, Rehemtulla A, Kleer CG, Hayes DF, Lucas PC, Varambally S, Chinnaiyan AM: AGTR1 overexpression defines a subset of breast cancer and confers sensitivity to losartan, an AGTR1 antagonist, Proc Natl Acad Sci U S A 2009, 106:10284-10289.
8. Ateeq B, Tomlins SA, Chinnaiyan AM. AGTR1 as a therapeutic target in ER-positive and ERBB2-negative breast cancer cases. Cell Cycle. 2009 Dec;8(23):3794-5. PubMed PMID: 19934656; PubMed Central PMCID: PMC2940713.
9. Maher CA, Palanisamy N, Brenner JC, Cao X, Kalyana-Sundaram S, Luo S, Khrebtukova I, Barrette TR, Grasso C, Yu J, Lonigro RJ, Schroth G, Kumar-Sinha C, Chinnaiyan AM. Chimeric transcript discovery by paired-end transcriptome sequencing. Proc Natl Acad Sci U S A. 2009 Jul 28;106(30):12353-8. doi: 10.1073/pnas.0904720106. Epub 2009 Jul 10. PubMed PMID: 19592507; PubMed Central PMCID: PMC2708976.
10. Maher CA, Kumar-Sinha C, Cao X, Kalyana-Sundaram S, Han B, Jing X, Sam L, Barrette T, Palanisamy N, Chinnaiyan AM. Transcriptome sequencing to detect gene

fusions in cancer. Nature. 2009 Mar 5;458(7234):97-101. doi: 10.1038/nature07638. Epub 2009 Jan 11. PubMed PMID: 19136943; PubMed Central PMCID: PMC2725402.

11. Palanisamy N, Ateeq B, Kalyana-Sundaram S, Pflueger D, Ramnarayanan K, Shankar S, Han B, Cao Q, Cao X, Suleman K, Kumar-Sinha C, Dhanasekaran SM, Chen YB, Esgueva R, Banerjee S, LaFargue CJ, Siddiqui J, Demichelis F, Moeller P, Bismar TA, Kuefer R, Fullen DR, Johnson TM, Greenson JK, Giordano TJ, Tan P, Tomlins SA, Varambally S, Rubin MA, Maher CA, Chinnaiyan AM. Rearrangements of the RAF kinase pathway in prostate cancer, gastric cancer and melanoma. Nat Med. 2010 Jul;16(7):793-8. doi: 10.1038/nm.2166. Epub 2010 Jun 6. PubMed PMID: 20526349; PubMed Central PMCID: PMC2903732.
12. Cao Q, Mani RS, Ateeq B, Dhanasekaran SM, Asangani IA, Prensner JR, Kim JH, Brenner JC, Jing X, Cao X, Wang R, Li Y, Dahiya A, Wang L, Pandhi M, Lonigro RJ, Wu YM, Tomlins SA, Palanisamy N, Qin Z, Yu J, Maher CA, Varambally S, Chinnaiyan AM. Coordinated regulation of polycomb group complexes through microRNAs in cancer. Cancer Cell. 2011 Aug 16;20(2):187-99. PubMed PMID: 21840484; PubMed Central PMCID: PMC3157014.
13. Iyer MK, Chinnaiyan AM, Maher CA. ChimeraScan: a tool for identifying chimeric transcription in sequencing data. Bioinformatics. 2011 Oct 15;27(20):2903-4. Epub 2011 Aug 11. PubMed PMID: 21840877; PubMed Central PMCID: PMC3187648.
14. Wu YM, Su F, Kalyana-Sundaram S, Khazanov N, Ateeq B, Cao X, Lonigro RJ, Vats P, Wang R, Lin SF, Cheng AJ, Kunju LP, Siddiqui J, Tomlins SA, Wyngaard P, Sadis S, Roychowdhury S, Hussain MH, Feng FY, Zalupski MM, Talpaz M, Pienta KJ, Rhodes DR, Robinson DR, Chinnaiyan AM. Identification of targetable FGFR gene fusions in diverse cancers. Cancer Discov. 2013 Jun;3(6):636-47. doi: 10.1158/2159-8290.CD-13-0050. Epub 2013 Apr 4. PubMed PMID: 23558953; PubMed Central PMCID: PMC3694764.
15. Robinson D.R., Kalyana-Sundaram S., Wu Y.-I., Shankar S., Cao X., Ateeq B., Asangani I.A., Iyer M., Maher C.A., Grasso C.S., Lonigro R.J., Quist M., Siddiqui J., Mehra R., Jing X., Giordano T.J., Sabel M.S., Kleer C.G., Palanisamy N., Natrajan R., Lambros M.B., Reis-Filho J.S., Kumar-Sinha C., and Chinnaiyan A.M. Functionally Recurrent Rearrangements of the MAST Kinase and Notch Gene Families in Breast Cancer. Nat Med. 2011 Nov 20;17(12):1646-51. PMID: 22101766; PMCID: PMC3233654.

# AGTR1 overexpression defines a subset of breast cancer and confers sensitivity to losartan, an AGTR1 antagonist

Daniel R. Rhodes[a,b,1], Bushra Ateeq[a,b,1], Qi Cao[a,b,1], Scott A. Tomlins[a,b,1], Rohit Mehra[a,b], Bharathi Laxman[a,b], Shanker Kalyana-Sundaram[a,b], Robert J. Lonigro[a,c], Beth E. Helgeson[a,b], Mahaveer S. Bhojani[c,d], Alnawaz Rehemtulla[c,d], Celina G. Kleer[b,c], Daniel F. Hayes[c,e], Peter C. Lucas[b,c], Sooryanarayana Varambally[a,b,c], and Arul M. Chinnaiyan[a,b,c,f,g,2]

[a]Michigan Center for Translational Pathology, [f]Howard Hughes Medical Institute, and Departments of [g]Urology, and [b]Pathology, University of Michigan Medical School, 1301 Catherine Street, Ann Arbor, MI 48109-5602; [c]University of Michigan Comprehensive Cancer Center, 1500 East Medical Center Drive, Ann Arbor, MI 48109-5940; [d]Department of Radiation Oncology, University of Michigan Comprehensive Cancer Center, 1500 East Medical Center Drive, 2G332 UH, Ann Arbor, MI 48109-5054; and [e]Department of Internal Medicine, University of Michigan Comprehensive Cancer Center, 1500 East Medical Center Drive, 6312 CCC, Ann Arbor, MI 48109-5942

Edited by Owen N. Witte, David Geffen School of Medicine at the University of California, Los Angeles, CA, and approved April 10, 2009 (received for review January 12, 2009)

**Breast cancer patients have benefited from the use of targeted therapies directed at specific molecular alterations. To identify additional opportunities for targeted therapy, we searched for genes with marked overexpression in subsets of tumors across a panel of breast cancer profiling studies comprising 3,200 microarray experiments. In addition to prioritizing ERBB2, we found AGTR1, the angiotensin II receptor type I, to be markedly overexpressed in 10–20% of breast cancer cases across multiple independent patient cohorts. Validation experiments confirmed that AGTR1 is highly overexpressed, in several cases more than 100-fold. AGTR1 overexpression was restricted to estrogen receptor-positive tumors and was mutually exclusive with ERBB2 overexpression across all samples. Ectopic overexpression of AGTR1 in primary mammary epithelial cells, combined with angiotensin II stimulation, led to a highly invasive phenotype that was attenuated by the AGTR1 antagonist losartan. Similarly, losartan reduced tumor growth by 30% in AGTR1-positive breast cancer xenografts. Taken together, these observations indicate that marked AGTR1 overexpression defines a subpopulation of ER-positive, ERBB2-negative breast cancer that may benefit from targeted therapy with AGTR1 antagonists, such as losartan.**

A central aim in cancer research is to identify genetic alterations involved in the pathogenesis of cancer, thereby providing an opportunity to develop therapies that directly target the alterations. In breast cancer research, this strategy has been realized with the study of ERBB2, which is amplified and overexpressed in 25–30% of breast tumors (1, 2), directly contributing to tumorigenesis (3, 4). Targeting this genetic lesion with trastuzumab, a humanized monoclonal antibody directed against ERBB2, has significant clinical benefit in breast cancer management (5–7). Cancer genes are activated or inactivated by a variety of mechanisms, including those that alter the activity of proteins (e.g., activating Ras mutation, BCR-ABL fusion protein) and those that change expression levels of proteins (e.g., ERBB2 gene amplification, Ig-Myc DNA translocation, or p53 homozygous deletion). It is likely that only a fraction of such "driver" alterations have been identified to date, and furthermore, many of the identified alterations are not thought to be "druggable" by conventional means.

DNA microarrays have been widely applied to the study of gene expression in cancer. Although microarrays are not capable of directly detecting alterations affecting the activity of proteins, they are theoretically well suited to detect alterations that change the expression of genes and proteins, although it can be difficult to identify driver alterations directly related to tumorigenesis among hundreds or thousands of differentially expressed genes. As a strategy for using microarray data to identify genes directly related to cancer pathogenesis that may thus serve as therapeutic targets, we hypothesized that genes that show the most profound changes in gene expression (10-fold to more than 100-fold increase relative to baseline), termed "pathogenic overexpression," even if in only a small subset of cases, may play a direct role in cancer progression and may serve as optimal therapeutic targets for the subpopulations with overexpression. Because cancer is heterogeneous, distribution statistics that compare average expression values between classes of samples (e.g., cancer vs. normal) will often fail to identify these profound changes in expression, especially if the alterations occur in subsets of cases (e.g., Her2/neu amplification and overexpression in 25% of breast cancer). We previously developed a simple analytical method, termed "Cancer Outlier Profile Analysis" (COPA), to identify such gene expression profiles, nominating ERG and ETV1 as novel cancer genes in prostate cancer, which were shown to be activated by gene fusions with the androgen-regulated gene TMPRSS2 (8). Here, we extend the COPA approach to include a meta-analysis strategy, combining the search for profound changes in expression with multistudy validation. We focus our analysis on breast cancer because this disease has been most extensively analyzed by gene expression profiling. Interestingly, the majority of such analyses have focused on disease classification and prediction of patient outcome, rather than target discovery. We present a large-scale analysis spanning 31 gene expression profiling studies comprising nearly 3,200 microarray experiments. In addition to objectively identifying the prototypical breast cancer target, ERBB2, our analysis also nominates a number of previously unidentified genes which, based on their profound overexpression in subsets of tumors across independent cohorts, may play a role in tumorigenesis and may serve as therapeutic targets in their respective subpopulations.

## Results

We hypothesized that genes directly involved in breast tumorigenesis may be activated via pathological overexpression in specific subsets of tumors. Thus, we developed a methodology to

Author contributions: D.R.R., B.A., S.A.T., S.V., and A.M.C. designed research; D.R.R., B.A., Q.C., S.A.T., R.M., B.L., B.E.H., and S.V. performed research; M.S.B., A.R., C.G.K., and S.V. contributed new reagents/analytic tools; D.R.R., B.A., S.A.T., R.M., S.K.-S., R.J.L., D.F.H., P.C.L., S.V., and A.M.C. analyzed data; and D.R.R., S.A.T., and A.M.C. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

[1]D.R.R., B.A., Q.C., and S.A.T. contributed equally to this work.

[2]To whom correspondence should be addressed. E-mail: arul@umich.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0900351106/DCSupplemental.

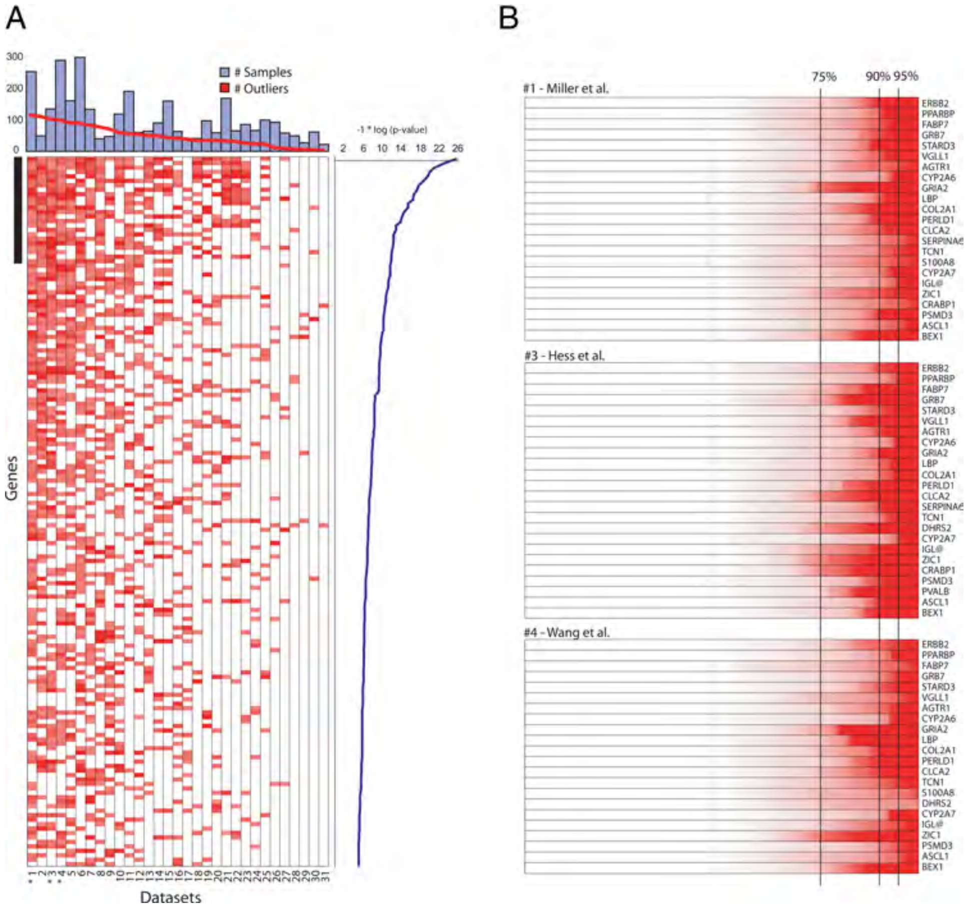


**Fig. 1.** MetaCOPA analysis of breast cancer gene expression data. (*A*) MetaCOPA map. Each column in the map represents a breast cancer gene expression dataset. The numbers at the base of the map correspond to dataset details (Table S1). Each row indicates a gene. A red cell indicates that the gene was deemed to have an outlier expression profile in the respective dataset because it scored in the top 1% of COPA values at 1 of 3 percentile cutoffs. The line graph along the *y* axis indicates the *P* value for a gene based on the number of datasets in which the gene was deemed an outlier. A total of 158 genes were called outliers in a significant fraction of datasets ($P <$ 1E-5). The bar graph indicates the number of samples in the respective datasets and the contribution of the dataset to the meta-analysis. The black bar on the left of the map indicates the top 25 meta-outliers, which are detailed in *B* for 3 datasets marked with an asterisk. (*B*) Heatmaps of COPA-normalized values for top-scoring meta-outliers across 3 highly contributory datasets: Miller et al. (26), Hess et al. (27), and Wang et al. (28). Genes are ranked by their MetaCOPA *P* values. For each gene, samples are ordered from left to right by their COPA-normalized expression values. Highest intensity of red indicates a COPA-normalized value of 6 or greater. White indicates a value of zero or less.

identify genes that display substantial changes in expression in subpopulations of tumors across independent cancer microarray datasets. The methodology, MetaCOPA, combines MetaAnalysis and COPA, 2 approaches that we have applied previously but separately to identify cancer genes (8, 9) (Fig. S1). We analyzed 31 breast cancer profiling datasets, comprising 3,157 microarrays (Table S1). We defined per dataset "outliers" as genes with the most dramatic overexpression in a subset of tumors, and"meta-outliers" as genes that were identified in a statistically significant fraction of datasets. We identified 159 significant meta-outliers ($P <$ 1E-5) (Fig. 1*A* and Table S2), of which ≈20 genes were identified as outliers in the majority of datasets examined (Fig. 1*B* and Table S3).

Notably, considering all human genes represented in the analysis, ERBB2 was the most significant meta-outlier, identified in 21 of 29 independent datasets (72%; $P$ = 3.6E-26), indicating that this established therapeutic target shows the most substantial and consistent overexpression in a fraction of breast tumors (Fig. S2*A*). Although ERBB2 did not have a no.1 ranked outlier expression profile in any individual dataset, it did score highest in the meta-analysis. Several other top-scoring meta-outliers localize within 1 Mb of ERBB2 on chromosome 17q. As expected from the past observation that ERBB2 and genomic neighbors are coamplified and coexpressed in breast cancer (10, 11), we observed a clear coexpression pattern of the 17q meta-outliers (Fig. S2*B*).

The next most consistently scoring outlier, excluding ERBB2 and genomic neighbors, was AGTR1, the gene encoding angiotensin II receptor type I, which is the target of the antihypertensive drug losartan (12) and has previously been linked to cancer (12–17) and cancer-related signaling pathways (18, 19). AGTR1 was called an outlier in 15 of 22 datasets (68%; $P$ = 2.0E-18). The microarray data clearly indicated that AGTR1 is highly overexpressed in a subset of tumors relative to normal tissue (Fig. 2*A*) and that high overexpression occurs exclusively in a subset of estrogen receptor-positive ($ER^+$) tumors (Fig. 2*C*). Furthermore, a coexpression analysis of AGTR1 and ERBB2 revealed a mutually exclusive relationship, with breast tumors overexpressing ERBB2 or AGTR1, but never both (Fig. 2 *B* and *D*). Additional evidence for the marked overexpression of AGTR1 in 10–20% of breast tumors, specifically $ER^+$, $ERBB2^-$ breast tumors, is presented in *SI Materials and Methods* (Figs. S3 and S4). AGTR1 overexpression was not significantly associated with 5-year recurrence-free survival in $ER^+$, $ERBB2^-$ breast cancer across 2 independent datasets (Fig. S5). We validated and quantified AGTR1 overexpression by quantitative RT-PCR in formalin-fixed, paraffin-embedded tissue from normal breast, primary breast cancer, and metastatic breast cancer. Consistent with the microarray data, we found AGTR1 to be more than 20-fold overexpressed in 7 of 45 tumors (15.5%) and more than 100-fold overexpressed in 2 primary tumors and 1 metastatic tumor (Fig. 2*E*).

Given the remarkable overexpression of AGTR1 in tumor subsets, we investigated potential mechanisms by which AGTR1 becomes overexpressed. First, using Oncomine, we examined AGTR1 coexpression data from 5 independent datasets, and in each case we found no more than one additional gene correlated with AGTR1 ($R >$ 0.5), providing preliminary evidence that AGTR1 is not regulated as part of a larger transcriptional program. Second, we examined AGTR1 overexpression in the context of genes that neighbor AGTR1 on chromosome 3q. Unlike ERBB2, AGTR1 did not display any correlated expression with genomic neighbors (Fig. S6).

Next, we performed FISH on tissue microarrays to test the AGTR1 locus for gene rearrangement or DNA copy number aberration. Using a split probe strategy (8), we found that 5′ and 3′

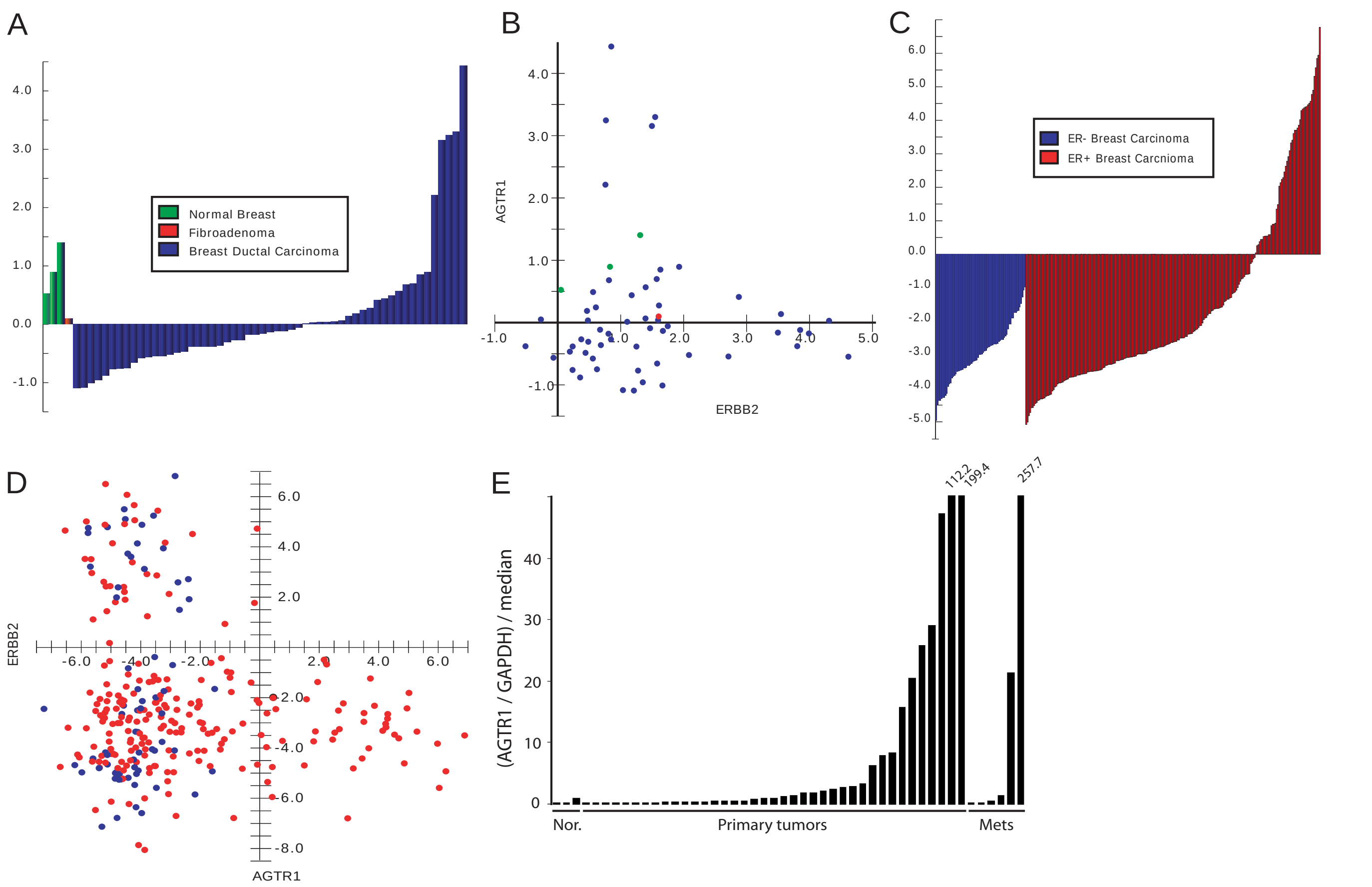


**Fig. 2.** AGTR1 outlier expression in breast cancer. (*A*) AGTR1 expression profile in the Perou et al. (29) cDNA microarray dataset ($n = 55$). (*B*) In the same dataset, AGTR1 expression vs. ERBB2 expression. (*C*) AGTR1 expression profile in the van de Vijver et al. (30) oligonucleotide dataset, segregated by ER status ($n = 295$). (*D*) AGTR1 expression vs. ERBB2 expression in the same dataset. (*E*) AGTR1 expression by quantitative RT-PCR in formalin-fixed, paraffin-embedded tissue. Expression of AGTR1 was assessed in 3 normal breast tissue specimens, 36 primary breast tumor specimens, and 9 metastatic breast cancer specimens. Expression levels were normalized to GAPDH expression and then scaled by the median AGTR/GADPH ratio.

AGTR1 probes never demonstrated consistent split signals, and thus concluded that rearrangement of the AGTR1 locus is not involved in AGTR1 overexpression. AGTR1 copy number was also evaluated in 112 breast carcinoma cases. Definitive copy number gain [locus/control (L/C) $> 1.5$] was observed in 7 of 112 cases (6.25%), of which 6 were invasive ductal carcinoma and 1 was ductal carcinoma in situ (Fig. 3 *A* and *B*). To study the association between DNA copy number and overexpression, we identified available cases for qRT-PCR analysis, including 14 cases with no gain (L/C $<$ 1.2), 3 cases with questionable gain ($1.2 <$ L/C $< 1.5$), and 4 cases with definitive DNA copy number gain (L/C $> 1.5$). We observed a significant concordance between high AGTR1 expression and definitive copy number gain ($P = 0.006$; Fig. 3*C*). All 4 cases tested with definitive copy number gain also had high AGTR1 expression; however, high expression was also observed in 3 of 17 cases without definitive copy number gain. Thus, in this small sample set, copy number gain was always associated with overexpression, but overexpression also occurred without copy number gain.

To study the function of AGTR1 overexpression in breast epithelial cells, we generated an adenovirus construct expressing AGTR1. Human mammary epithelial cells (H16N2 and HME) were infected with AGTR1-expressing virus or control LacZ-expressing virus and cultured in serum-free media (Fig. S7). We assayed AGTR1-overexpressing cells and control cells for cell proliferation and invasion both in serum-free media and upon stimulation with angiotensin II (AT), the ligand of AGTR1. Overexpression of AGTR1 alone or in combination with AT did not affect cell proliferation. However, in both cell lines, we did observe that overexpression of AGTR1 with AT stimulation did significantly promote cell invasion in a reconstituted basement membrane invasion chamber assay (Fig. 4 *A* and *B*). The control experiment, in which the LacZ gene was transfected, did not exhibit increased invasion with AT stimulation. Importantly, AGTR1 and AT-mediated invasion was attenuated in a dose-dependent manner with inclusion of the AGTR1 blocker, losartan. Losartan had no effect on the LacZ-transfected cells or the AGTR1-transfected cells not stimulated with AT (Fig. 4*B*). To confirm that losartan inhibition of invasion is specific to AGTR1 transfection, we also infected H16N2 and HME cells with EZH2-expressing adenovirus, a gene known to induce invasion and, as expected, found that EZH2-mediated invasion was not attenuated by losartan treatment (Fig. S8). Thus, in 2 benign breast epithelial cell lines, AGTR1 overexpression in the presence of AT led to a markedly invasive tumorigenic phenotype, which is specifically reversed by treatment with losartan. We also tested the AGTR1-overexpressing mammary epithelial cells for activation of the MAPK and PI3K pathways, as measured by ERK phosphorylation and AKT phosphorylation, respectively. We found that AGTR1 overexpression combined with AT stimulation did increase ERK phosphorylation but not AKT phosphorylation. Losartan treatment (10 $\mu$M) inhibited the AT-stimulated increase in ERK phosphorylation (Fig. S9).

Next, we identified and tested a panel of breast cancer cell lines with endogenous AGTR1 overexpression. By using Oncomine (20), we identified 4 breast cancer cell lines with validated AGTR1

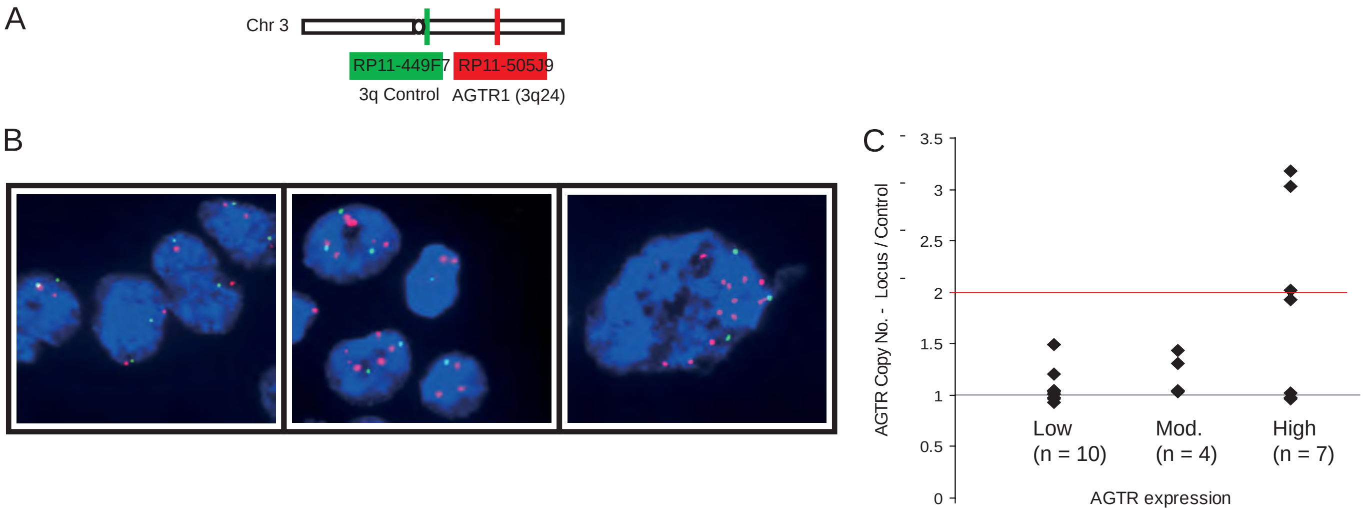


**Fig. 3.** Copy number analysis of the AGTR1 locus. (*A*) A schematic of probes used for FISH analysis. (*B*) Representative image from FISH analysis. *Left* is taken from a representative negative case. *Middle* and *Right* are images from a representative case with definitive copy number gain of AGTR1. Red signal is the AGTR1 locus probe, and green signal is the probe near the chromosome 3 centromere. (*C*) Association of AGTR1 overexpression with copy number gain. Three expression bins were defined based on AGTR1/GAPDH ratios: low (<1.0), moderate (1.0–2.0), and high (>2.0).

overexpression and 3 breast cancer cell lines with little or no expression of AGTR1 (Fig. S10). As an additional negative control, we also included the highly invasive prostate cancer cell line DU145, which has low expression of AGTR1. By using the reconstituted basement membrane invasion chamber assay, we tested the cell line panel with and without 1 $\mu$M AT and losartan. In each of the 4 AGTR1-overexpressing cell lines, we observed an increase in invasion upon stimulation with 1 $\mu$M AT, which was reversible by addition of losartan, whereas none of the 3 breast cancer cell lines with low AGTR1 expression, nor DU145, showed an increase in invasion upon 1 $\mu$M AT stimulation (Fig. 4*C*). Thus, we confirmed that our ectopic AGTR1 overexpression results can be generalized to breast cancer cells with endogenous overexpression but not those with low expression, and that losartan-mediated decrease in invasion is specific to invasion related to AT stimulation and AGTR1 overexpression.

Next, we stably transfected AGTR1 into MCF7 human breast cancer cells and performed mouse xenograft studies. We implanted MCF7-AGTR1 cells or MCF7-GUS control cells into the mammary fat pad of nude mice and treated animals with 90 mg/kg losartan per day or vehicle control. We studied the impact of losartan on tumor growth at 2 weeks and 8 weeks. Ten mice were studied in each group: MCF7-AGTR1 plus saline, MCF7-AGTR1 plus losartan, MCF7-GUS plus saline, and MCF7-GUS plus losartan. MCF7-AGTR1 tumors did not display increased growth at 2 weeks or 8 weeks relative to MCF7-GUS control tumors. Losartan treatment did, however, significantly reduce early and late tumor growth in MCF7-AGTR1-implanted mice but had no effect on tumor growth in MCF7-GUS control-implanted mice. At 2 weeks after implantation, the median tumor size of MCF7-AGTR1 tumors treated with losartan was 20% smaller than MCF7-AGTR1 tumors treated with vehicle control ($P$ = 1.4E-4; Fig. 5*A*). On the contrary, there was no significant change in tumor size at 2 weeks in MCF7-GUS tumors treated with losartan relative to vehicle control ($P$ = 0.67). Similarly, at 8 weeks, median tumor size of MCF7-AGTR1 tumors treated with losartan was 31% smaller than those treated with control ($P$ = 0.016; Fig. 5*B*). Again, no significant change in median tumor size of MCF7-GUS tumors was observed upon losartan treatment ($P$ = 0.24). In summary, although AGTR1 transfection into MCF7 breast cancer cells did not increase tumor size, it did significantly sensitize tumors to growth inhibition with losartan treatment.

## Discussion

In summary, we performed a large-scale meta-analysis of outlier expression profiles across several large cohorts of breast tumors. Our analysis prioritized genes with marked overexpression in subsets of tumors. This approach correctly prioritized the prototypical breast cancer oncogene and drug target ERBB2. In addition, several new genes were identified, demonstrating consistent and dramatic overexpression in tumor subsets. We suspect that our analysis has uncovered a new crop of potentially important breast cancer genes.

AGTR1, the angiotensin II receptor, was found to be one of the most highly overexpressed genes in 10–20% of breast cancers across independent breast cancer microarray studies. This has potential clinical importance because AGTR1 is antagonized by commonly prescribed antihypertensive agents (12), such as losartan, which have been shown to have antitumorigenic effects in model systems (12–17). Interestingly, AGTR1 always displayed high overexpression in ER-positive, ERBB2-negative tumors, potentially providing insights into the selective pressures governing AGTR1 activation in breast cancer. Contrary to expectation, ER in fact down-regulates the AGTR1 transcript via cytosolic mRNA-binding proteins (21). Thus, we hypothesize that the paradoxical marked overexpression of AGTR1 in a subset of $ER^+$ breast tumors may be the result of a genetic aberration that put the AGTR1 transcript under the positive control of the ER. Based on the mutually exclusive expression pattern with ERBB2 and the reported overlapping downstream pathways affected by AGTR1 and ERBB2, we suspect that AGTR1 activation and ERBB2 activation may represent alternative but functionally related events in tumorigenesis. Our AGTR1 transfection experiments in HME cells confirmed that ERK phosphorylation, a MAPK pathway readout, increases upon angiotensin stimulation.

We applied computational and experimental strategies to uncover mechanisms for AGTR1 overexpression. Coexpression analysis revealed that AGTR1 is not likely to be part of a larger transcriptional program, because other genes were not found to be highly coexpressed with AGTR1. FISH analysis demonstrated that chromosomal rearrangements do not occur at the AGTR1 locus, making gene fusions an unlikely cause of overexpression. DNA copy number analysis did identify a small fraction (6.5%) of breast tumors with increased copy number at the AGTR1 locus, and copy number gain occurred only in cases with overexpression. However,

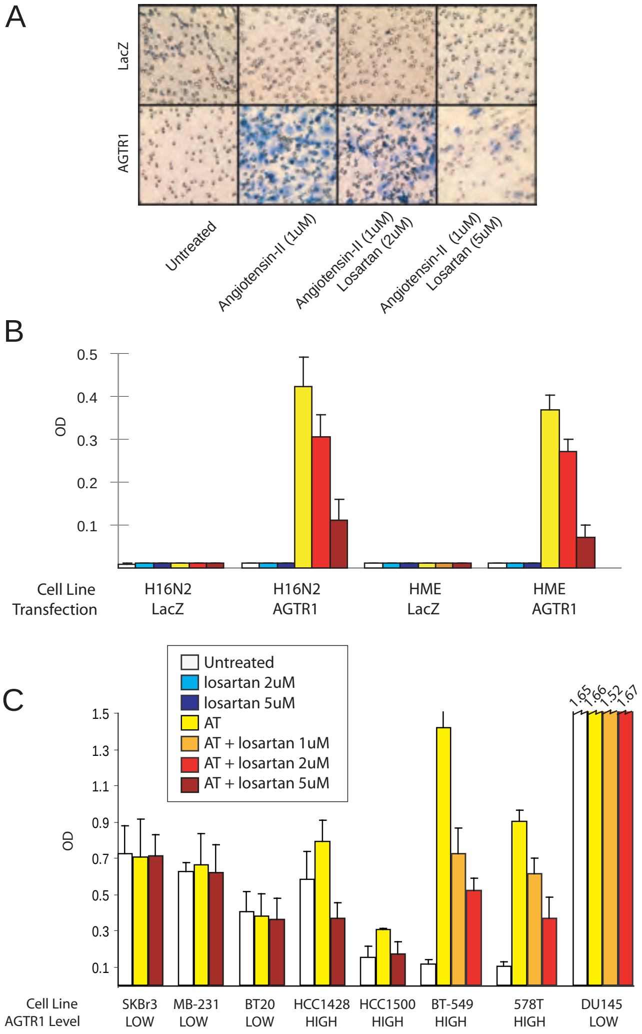


**Fig. 4.** AGTR1 overexpression and analysis of angiotensin II (AT) and losartan effects on cell invasion. (*A*) Matrigel invasion assays of H16N2 cells infected with adenovirus expressing AGTR1 or LacZ. Cells were cultured in serum-free media and were pretreated with and without AT and losartan. Similar results were observed for HME cells. (*B*) Colorimetry readout of invasion assays from transfection experiments. LacZ- or AGTR1-expressing adenovirus was infected into H16N2 and HME immortalized mammary epithelial cells, and cells were treated with or without 1 $\mu$M AT and losartan. Because of absent baseline invasion, the optical density (OD) measurements were background subtracted, and values below 0.01 were set to 0.01. (*C*) Colorimetry readout of invasion assays from a panel of cancer cell lines. Seven breast cancer cell lines and a prostate cancer cell line, DU145, were examined for invasion after treatment with or without 1 $\mu$M AT and losartan. AGTR1 expression levels are indicated and were obtained from published microarray data and qRT-PCR analysis (Fig. S7). The quantification of invasion was done as described in *B*.

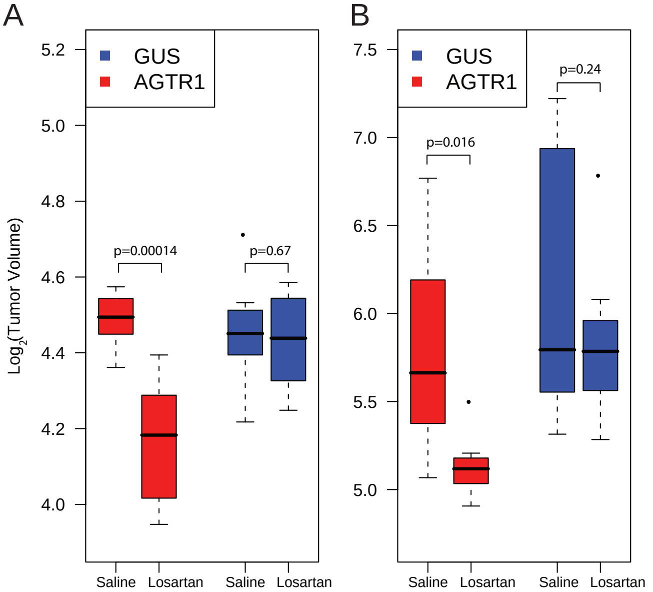


**Fig. 5.** Effect of losartan treatment on AGTR1- or GUS-overexpressing MCF7 cell xenografts. Female BALB/C nu/nu mice were implanted with 2.5 × $10^6$ stable MCF7 cells overexpressing AGTR1 or GUS resuspended in 100 $\mu$L of saline with 20% Matrigel into the mammary fat pad of anesthetized mice. Mice from both groups: MCF7-AGTR1 or MCF7-GUS ($n = 10$ for each group) were treated every day with losartan (90 mg/kg body weight) or vehicle control. All animals were monitored at weekly intervals for tumor growth, and tumor sizes were recorded using the formula $(\pi/6)\,(L \times W^2)$, where $L$ = length of tumor and $W$ = width. Box plots of $\log_2$ tumor volumes are shown. *P* values from 2-sided Student's *t* tests indicate statistical significance. (*A*) Xenograft tumor size at 2 weeks. (*B*) Xenograft tumor size at 8 weeks.

some overexpressing cases did not have copy number gain, and the level of copy number gain observed in positive cases was not proportional to the degree of overexpression observed. Thus, we suspect that copy number gain contributes to overexpression in some cases but is not likely to be the predominant mechanism. Future studies to investigate the mechanism of AGTR1 overexpression should include high-resolution array comparative genomic hybridization and sequencing of the AGTR1 locus.

Regardless of the mechanism, AGTR1 undergoes profound deregulation in a subset of breast cancers, and our in vitro and in vivo studies demonstrate a functional role for AGTR1 overexpression in breast cancer and, more importantly, the potential for targeting AGTR1$^+$ breast tumors with an available therapy. Past work has shown that in breast cancer cell lines, angiotensin II stimulation evokes an invasive phenotype, which is inhibited by losartan treatment (22). Furthermore, it was demonstrated that the increase in invasion is coincident with decreased expression of integrins, possibly via protein kinase C signaling. Although these observations were made in transformed breast cancer cells naturally expressing AGTR1, our work shows that activated AGTR1 pathway, by way of artificial AGTR1 overexpression, in normal breast epithelial cells is sufficient to activate an invasive phenotype, suggesting that this pathway may be especially important in breast tumors with high overexpression. Furthermore, we studied a panel of cell lines with either high or low levels of AGTR1 and showed a clear correlation between AT-mediated invasion and level of AGTR1 expression.

Our in vivo data provide further evidence that losartan may be a viable therapy for women with AGTR1-overexpressing breast tumors. Breast cancer xenografts overexpressing AGTR1 were differentially sensitive to losartan treatment, demonstrating a 30% reduction in growth at 8 weeks, whereas control xenografts had no reductin in tumor size. It is interesting that MCF7-AGTR1 xenografts did not display increased growth relative to MCF7 control xenografts, but they did display a significantly increased losartan effect. This suggests that AGTR1 does not provide an additive growth signal to MCF7 cells, which do harbor an activating PI3K mutation. We suspect that the stable transfection of AGTR1 reprogrammed MCF7 cells to be at least partially dependent on AGTR1 as a growth or survival signal; hence, the differential response to losartan. We anticipate that de novo AGTR1-positive primary tumors may be even more dependent on the AGTR1 signal, and thus more sensitive to inhibition.

Interestingly, past studies have linked polymorphisms in the angiotensin pathway with breast cancer incidence (23, 24), documenting a significant increase in breast cancer incidence in women with the D/D angiotensin-converting enzyme (ACE) allele, which is associated with increased circulating ACE levels, and thus increased levels of angiotensin II, the ligand for AGTR1. Other studies have examined the relationship between antihypertensive therapy (AHT), which often involves modulation of the angiotensin axis, and breast cancer incidence. The largest of such studies did not observe a significant relationship (25); however, the study examined a variety of AHT modalities and was likely not powered to detect a small change incidence that might be expected from a response only in the AGTR1$^+$ subpopulation.

In summary, this study provides a rationale for a clinical trial that includes losartan in the treatment of breast cancer patients with tumors positive for AGTR1. We demonstrated that AGTR1 transcript levels and DNA copy number can be effectively measured from formalin-fixed, paraffin-embedded tissue specimens, thus enabling the identification of the appropriate patient population.

## Materials and Methods

**MetaCOPA Analysis.** COPA analysis was performed on 31 breast cancer gene expression datasets in Oncomine (www.oncomine.org) as described previously (8). Genes scoring in the top 1% of COPA scores at any of the 3 percentile cutoffs (75th, 90th, and 95th) were deemed outliers in their respective datasets. Meta-outliers were defined as genes deemed outliers in a significant fraction ($P <$ 1E-5) of datasets as assessed by the binomial distribution. Analysis details are provided in *SI Materials and Methods*.

**Quantitative PCR (QPCR).** QPCR was performed by using SYBR Green dye on an Applied Biosystems 7300 Real Time PCR system (Applied Biosystems) essentially as described previously (8). Details and primer sequences are available in *SI Materials and Methods*.

**AGTR1 Transfection.** The benign human mammary epithelial cells HME and H16N2 were transfected with AGTR1-expressing adenovirus and assayed for cell invasion with or without losartan and angiotensin II treatment. Details are available in *SI Materials and Methods*.

**Cell Invasion Assay.** Breast cell lines BT-549, Hs578T, HME, H16N2, HCC1528, HCC1500 and prostate carcinoma line DU145 were assayed for cell invasion with or without losartan and angiotensin II treatment using Matrigel invasion chambers. Details are available in *SI Materials and Methods*.

**AGTR1 Amplification Assessment.** A breast cancer tissue microarray containing 311 cases of invasive breast cancer was tested for AGTR1 locus amplification by flourscence in situ hybridization. Details are available in *SI Materials and Methods*.

**Mammary Fat Pad Xenograft Model.** Balb/C nu/nu mice were implanted with MCF7 cells stably overexpressing AGTR1 or Gus and then treated daily with losartan vehicle control. Details are available in *SI Materials and Methods*.

**ACKNOWLEDGMENTS.** We thank V. Mahavisno, R. Varambally, T. Barrette, and D. Gibbs for Oncomine support, and Diane Roulston and Lisa Smith for assistance with FISH. We thank Merck USA for providing losartan. This work is supported by the Department of Defense Era of Hope Scholar Award (to A.M.C., C.G.K, and D.F.H.), the Early Detection Research Network Biomarker Developmental Lab Grant UO1 CA111275-01 (to A.M.C.), Department of Defense Grants PC040517 (to R.M.) and PC020322 (to A.M.C.), and the Cancer Center Bioinformatics Core Support Grant 5P30 CA46592. S.A.T. is supported by a Rackham Predoctoral Fellowship. A.M.C. is supported by a Clinical Translational Research Award from the Burroughs Welcome Foundation and a Doris Duke Charitable Foundation Distinguished Clinical Scientist Award.

1. King CR, Kraus MH, Aaronson SA (1985) Amplification of a novel v-erbB-related gene in a human mammary carcinoma. *Science* 229:974–976.
2. Slamon DJ, et al. (1987) Human breast cancer: Correlation of relapse and survival with amplification of the HER-2/neu oncogene. *Science* 235:177–182.
3. Di Fiore PP, et al. (1987) erbB-2 is a potent oncogene when overexpressed in NIH/3T3 cells. *Science* 237:178–182.
4. Hudziak RM, et al. (1989) p185HER2 monoclonal antibody has antiproliferative effects in vitro and sensitizes human breast tumor cells to tumor necrosis factor. *Mol Cell Biol* 9:1165–1172.
5. Piccart-Gebhart MJ, et al. (2005) Trastuzumab after adjuvant chemotherapy in HER2-positive breast cancer. *N Engl J Med* 353:1659–1672.
6. Romond EH, et al. (2005) Trastuzumab plus adjuvant chemotherapy for operable HER2-positive breast cancer. *N Engl J Med* 353:1673–1684.
7. Slamon DJ, et al. (2001) Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2. *N Engl J Med* 344:783–792.
8. Tomlins SA, et al. (2005) Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science* 310:644–648.
9. Rhodes DR, et al. (2004) Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. *Proc Natl Acad Sci USA* 101:9309–9314.
10. Bertucci F, et al. (2004) Identification and validation of an ERBB2 gene expression signature in breast cancers. *Oncogene* 23:2564–2575.
11. Kauraniemi P, Barlund M, Monni O, Kallioniemi A (2001) New amplified and highly expressed genes discovered in the ERBB2 amplicon in breast cancer by cDNA microarrays. *Cancer Res* 61:8235–8240.
12. Timmermans PB (1999) Angiotensin II receptor antagonists: An emerging new class of cardiovascular therapeutics. *Hypertens Res* 22:147–153.
13. Miyajima A, et al. (2002) Angiotensin II type I antagonist prevents pulmonary metastasis of murine renal cancer by inhibiting tumor angiogenesis. *Cancer Res* 62:4176–4179.
14. Fujimoto Y, Sasaki T, Tsuchida A, Chayama K (2001) Angiotensin II type 1 receptor expression in human pancreatic cancer and growth inhibition by angiotensin II type 1 receptor antagonist. *FEBS Lett* 495:197–200.
15. Rivera E, Arrieta O, Guevara P, Duarte-Rojo A, Sotelo J (2001) AT1 receptor is present in glioma cells; its blockage reduces the growth of rat glioma. *Br J Cancer* 85:1396–1399.
16. Uemura H, et al. (2003) Angiotensin II receptor blocker shows antiproliferative activity in prostate cancer cells: A possibility of tyrosine kinase inhibitor of growth factor. *Mol Cancer Ther* 2:1139–1147.
17. Suganuma T, et al. (2005) Functional expression of the angiotensin II type 1 receptor in human ovarian carcinoma cells and its blockade therapy resulting in suppression of tumor invasion, angiogenesis, and peritoneal dissemination. *Clin Cancer Res* 11:2686–2694.
18. Muscella A, Greco S, Elia MG, Storelli C, Marsigliante S (2003) PKC-zeta is required for angiotensin II-induced activation of ERK and synthesis of C-FOS in MCF-7 cells. *J Cell Physiol* 197:61–68.
19. Amaya K, et al. (2004) Angiotensin II activates MAP kinase and NF-kappaB through angiotensin II type I receptor in human pancreatic cancer cells. *Int J Oncol* 25:849–856.
20. Rhodes DR, et al. (2004) ONCOMINE: A cancer microarray database and integrated data-mining platform. *Neoplasia* 6:1–6.
21. Krishnamurthi K, et al. (1999) Estrogen regulates angiotensin AT1 receptor expression via cytosolic proteins that bind to the 5′ leader sequence of the receptor mRNA. *Endocrinology* 140:5435–5438.
22. Puddefoot JR, Udeozo UK, Barker S, Vinson GP (2006) The role of angiotensin II in the regulation of breast cancer cell adhesion and invasion. *Endocr Relat Cancer* 13:895–903.
23. Gonzalez-Zuloeta Ladd AM, et al. (2005) Angiotensin-converting enzyme gene insertion/deletion polymorphism and breast cancer risk. *Cancer Epidemiol Biomarkers Prev* 14:2143–2146.
24. Gonzalez-Zuloeta Ladd AM, et al. (2007) Differential roles of Angiotensinogen and Angiotensin Receptor type 1 polymorphisms in breast cancer risk. *Breast Cancer Res Treat* 101:299–304.
25. Fryzek JP, et al. (2006) A cohort study of antihypertensive medication use and breast cancer among Danish women. *Breast Cancer Res Treat* 3:3.
26. Miller LD, et al. (2005) An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. *Proc Natl Acad Sci USA* 102:13550–13555.
27. Hess KR, et al. (2006) Pharmacogenomic predictor of sensitivity to preoperative chemotherapy with paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide in breast cancer. *J Clin Oncol* 24:4236–4244.
28. Wang Y, et al. (2005) Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet* 365:671–679.
29. Perou CM, et al. (2000) Molecular portraits of human breast tumours. *Nature* 406:747–752.
30. van de Vijver MJ, et al. (2002) A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med* 347:1999–2009.

This model makes the prediction that the stalk is tilted (extending from the AAA+ ring toward the minus end, as in Fig. 4B) at the time when a productive power stroke occurs. The MTBD may preferentially bind to the microtubule at such an angle, as suggested by our cryo-EM map (Fig. 2B) and a recent reconstruction of a whole axonemal dynein in its pre–power stroke state (*19*). Alternatively, the MTBD may rebind to the microtubule at various angles, but respond to a power stroke differently depending on its angle of attachment. A force-producing conformational change would produce a productive, minus end–directed displacement of the cargo if the stalk were pointing toward the minus end (e.g., Fig. 4C), whereas the MTBD would release if the stalk were pointing in the opposite direction (*18*, *20*). Further work will be needed to define the orientation of the stalk at different stages of the motility cycle and to learn how dynein might be able to reverse its direction of motion, as has been reported for a mammalian dynein (*21*).

The model for dynein motility presented here (Fig. 4) differs from the swinging lever arm model developed for myosin and kinesin (*22*). The dynein stalk does not serve as a rigid lever, as proposed elsewhere (*23*), but rather acts as a tether that allows the detached MTBD to explore a range of potential microtubule-binding sites and transmit tension between the AAA+ ring and the MTBD. The large AAA+ ring and its associated linker domain undergo ATP-dependent conformational changes (*8*, *9*) that pull along the stalk axis. This is consistent with the known actions of other AAA+ proteins (*24*) and the previous proposal that dynein acts as a winch (*8*). And it is the MTBD—one of the smallest elements of the large dynein motor protein—that governs the directionality of the motor.

### References and Notes

1. I. R. Gibbons, *Cell Motil. Cytoskeleton* **32**, 136 (1995).
2. B. J. Howell *et al.*, *J. Cell Biol.* **155**, 1159 (2001).
3. K. T. Vaughan, *Biochim. Biophys. Acta* **1744**, 316 (2005).
4. R. B. Vallee *et al.*, *J. Neurobiol.* **58**, 189 (2004).
5. M. A. Zariwala *et al.*, *Annu. Rev. Physiol.* **69**, 423 (2007).
6. M. Hafezparast *et al.*, *Science* **300**, 808 (2003).
7. D. Zuccarello *et al.*, *Hum. Reprod.* **23**, 1957 (2008).
8. S. A. Burgess *et al.*, *Nature* **421**, 715 (2003).
9. T. Kon *et al.*, *Nat. Struct. Mol. Biol.* **12**, 513 (2005).
10. M. A. Gee *et al.*, *Nature* **390**, 636 (1997).
11. I. R. Gibbons *et al.*, *J. Biol. Chem.* **280**, 23960 (2005).
12. G. Woehlke *et al.*, *Cell* **90**, 207 (1997).
13. M. P. Koonce, I. Tikhonenko, *Mol. Biol. Cell* **11**, 523 (2000).
14. N. Mizuno *et al.*, *EMBO J.* **23**, 2459 (2004).
15. M. Kikkawa, N. Hirokawa, *EMBO J.* **25**, 4187 (2006).
16. C. V. Sindelar, K. H. Downing, *J. Cell Biol.* **177**, 377 (2007).
17. M. Hulko *et al.*, *Cell* **126**, 929 (2006).
18. S. L. Reck-Peterson *et al.*, *Cell* **126**, 335 (2006).
19. T. Oda *et al.*, *J. Cell Biol.* **177**, 243 (2007).
20. A. Gennerich *et al.*, *Cell* **131**, 952 (2007).
21. J. L. Ross *et al.*, *Nat. Cell Biol.* **8**, 562 (2006).
22. R. D. Vale, R. A. Milligan, *Science* **288**, 88 (2000).
23. N. Mizuno *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **104**, 20832 (2007).
24. P. A. Tucker, L. Sallai, *Curr. Opin. Struct. Biol.* **17**, 641 (2007).

25. We thank J. Welburn and E. Nogales for advice and equipment; M. Kikkawa for sending his group's electron density map; and N. Bradshaw, J. Kardon, K. Aathavan, A. Dosé, A. Gennerich, A. Roll-Mecak, and S. Reck-Peterson for critically reading the manuscript. Supported by the Jane Coffin Childs Foundation (A.P.C.); NIH grants GM52468-14 (R.A.M.), P01-AR42895 (R.D.V.) and GM30401-29 (I.R.G.)]; the Agouron Institute (A.P.C.); the Leukemia and Lymphoma Society (A.P.C.); and the Howard Hughes Medical Institute. Crystallography data were collected at beamline 8.3.1 of the Advanced Light Source at Lawrence Berkeley National Laboratory. The atomic coordinates and structure factors have been deposited in the Protein Data Bank (code 3ERR). Cryo-EM data were collected in part at the National Resource for Automated Molecular Microscopy at the Scripps Research Institute (NIH P41 RR-17573). Maps were deposited at the EMDataBank (EMD-1581).

**Supporting Online Material**
www.sciencemag.org/cgi/content/full/322/5908/1691/DC1
Materials and Methods
Tables S1 to S3
Figs. S1 to S7
Movies S1 to S4
References

8 August 2008; accepted 6 November 2008
10.1126/science.1164424

# Genomic Loss of microRNA-101 Leads to Overexpression of Histone Methyltransferase EZH2 in Cancer

**Sooryanarayana Varambally,[1,2,3]* Qi Cao,[1,2]* Ram-Shankar Mani,[1,2] Sunita Shankar,[1,2] Xiaosong Wang,[1,2] Bushra Ateeq,[1,2] Bharathi Laxman,[1,2] Xuhong Cao,[1,4] Xiaojun Jing,[1,2] Kalpana Ramnarayanan,[5] J. Chad Brenner,[1,2,6] Jindan Yu,[1,2] Jung H. Kim,[1,3] Bo Han,[1,2] Patrick Tan,[5,7] Chandan Kumar-Sinha,[1,2] Robert J. Lonigro,[1,3] Nallasivam Palanisamy,[1,2,5] Christopher A. Maher,[1,2] Arul M. Chinnaiyan[1,2,3,4,6,8]†**

Enhancer of zeste homolog 2 (EZH2) is a mammalian histone methyltransferase that contributes to the epigenetic silencing of target genes and regulates the survival and metastasis of cancer cells. EZH2 is overexpressed in aggressive solid tumors by mechanisms that remain unclear. Here we show that the expression and function of EZH2 in cancer cell lines are inhibited by microRNA-101 (miR-101). Analysis of human prostate tumors revealed that miR-101 expression decreases during cancer progression, paralleling an increase in EZH2 expression. One or both of the two genomic loci encoding miR-101 were somatically lost in 37.5% of clinically localized prostate cancer cells (6 of 16) and 66.7% of metastatic disease cells (22 of 33). We propose that the genomic loss of miR-101 in cancer leads to overexpression of EZH2 and concomitant dysregulation of epigenetic pathways, resulting in cancer progression.

Polycomb group proteins, including enhancer of zeste homolog 2 (EZH2), play a master regulatory role in controlling important cellular process such as maintaining stem cell pluripotency (*1*–*3*), cell proliferation (*4*, *5*), early embryogenesis (*6*), and X chromosome inactivation (*7*). EZH2 functions in a multiprotein complex called polycomb repressive complex 2 (PRC2), which includes SUZ12 (suppressor of zeste 12) and EED (embryonic ectoderm development) (*8*, *9*). The primary activity of the EZH2 protein complex is to trimethylate histone H3 lysine 27 (H3K27) at target gene promoters, leading to epigenetic silencing (*10*, *11*). Mounting evidence suggests that EZH2 has properties consistent with those of an oncogene because overexpression promotes cell proliferation, colony formation, and increased invasion of benign cells in vitro (*4*, *5*, *12*) and induces xenograft tumor growth in vivo (*13*). Likewise, knockdown of EZH2 in cancer cells results in growth arrest (*4*, *13*) as well as diminished tumor growth (*10*) and metastasis in vivo (*14*).

EZH2 was initially found to be elevated in a subset of aggressive clinically localized prostate cancers and almost all metastatic prostate cancers (*4*). Subsequently, EZH2 has also been found to be aberrantly overexpressed in breast cancer (*12*), melanoma (*15*), bladder cancer (*16*), gastric cancer (*17*), and other cancers (*5*). Thus, although EZH2 is broadly overexpressed in aggressive solid tumors and has properties of an oncogene, the genetic mechanism of EZH2 elevation in cancer is unclear.

[1]Michigan Center for Translational Pathology, University of Michigan Medical School, Ann Arbor, MI 48109, USA. [2]Department of Pathology, University of Michigan Medical School, Ann Arbor, MI 48109, USA. [3]Comprehensive Cancer Center, University of Michigan Medical School, Ann Arbor, MI 48109, USA. [4]Howard Hughes Medical Institute, University of Michigan Medical School, Ann Arbor, MI 48109, USA. [5]Genome Institute of Singapore, 60 Biopolis Street, 02-01, Genome, Singapore 138672, Singapore. [6]Department of Cellular and Molecular Biology, University of Michigan Medical School, Ann Arbor, MI 48109, USA. [7]National Cancer Centre, 11 Hospital Drive, Singapore 169610, Singapore. [8]Department of Urology, University of Michigan Medical School, Ann Arbor, MI 48109, USA.

*These authors contributed equally to this work.
†To whom correspondence should be addressed. E-mail: arul@umich.edu

Because microRNAs (miRNAs) have gained considerable attention as regulators of gene expression (*18*) and play important roles in cellular differentiation and embryonic stem cell development (*19*), we postulated that they may play a role in modulating EZH2 expression. To test whether miRNAs play a role in controlling EZH2 expression, we computationally nominated those that might contribute to *EZH2* regulation. Because intersecting the results of multiple prediction algorithms can increase specificity at the cost of lower sensitivity (*20*), we integrated the results of the prediction software programs PicTar (*21*), TargetScan (*22*), miRanda (*23*), and miRInspector (*24*). Overall, only 29 miRNAs were found by any program to target *EZH2*, whereas only microRNA-101 (miR-101) and miR-217 were found by all four programs to be predicted to regulate *EZH2* (Fig. 1A and table S1) (*25*). Furthermore, PicTar, miRanda, and TargetScan predicted two miR-101–binding sites within the *EZH2* 3′ untranslated region (3′UTR) (Fig. 1B), whereas PicTar and TargetScan predicted two miR-217 binding sites within the *EZH2* 3′UTR. Of the 34 miRNAs predicted to regulate EZH2 by at least one program (table S1), only miR-101 had a strong negative association with prostate cancer progression from benign to localized disease to metastasis.

To examine whether miR-101 regulates the 3′UTR of *EZH2*, we generated luciferase reporters encoding the normal, antisense, and mutated versions of the EZH2 3′UTR. Overexpression of miR-101, but not miR-217 or control miRNA, decreased the activity of the luciferase reporter encoding the 3′UTR of EZH2 (fig. S1). Similarly, the antisense and mutant EZH2 3′UTR activities were not inhibited by miR-101. To explore whether the 3′UTR binding by miR-101 results in down-regulation of the *EZH2* transcript, we transfected SKBr3 breast cancer cells (which express high levels of endogenous EZH2) with precursors of miR-101, miR-217, and a control miRNA, as well as several other unrelated miRNAs. Quantitative reverse transcription polymerase chain reaction (RT-PCR) demonstrated a reduction in *EZH2* transcript levels by miR-101 (Fig. 1C) but not miR-217 or other control miRs.

To determine whether miR-101 represses EZH2 protein expression, we performed immunoblot analysis using an EZH2-specific antibody as well as antibodies to other PRC2 members, including EED and SUZ12 (Fig. 1D). In addition to miR-101, we included other miRNAs that were predicted to regulate EZH2, including miR-217 and miR-26a. Control miR-495 was predicted by TargetScan to target the PRC1 component BMI-1. Only miR-101 and EZH2 small interfering RNA (siRNA) attenuated EZH2 protein expression. miR-101 overexpression also leads to repression of EZH2's tight binding partners in the PRC2 complex: EED and, to a lesser extent, SUZ12. These proteins are thought to form a coregulated functional complex, and altering the expression of one affects the expression of the others (*5*, *26*, *27*). In this particular case, upon further inspection of the 3′UTRs of the PRC2 components, miR-101 binding sites were found in *EED* (fig. S2) but not in *SUZ12*. Because miRNAs are known to regulate multiple target genes, and in some cases hundreds of genes (*18*), we used the prediction algorithm TargetScan to nominate targets of miR-101. In addition to EZH2 and EED, we tested four predicted targets of miR-101 (table S2) that have been implicated in cancer pathways, including n-Myc, c-Fos, AT-rich interactive domain 1A (also called SWI-like and ARID1A), and fibrillin 2 (FBN2). None of these putative miR-101 targets were affected by overexpression of miR-101 (Fig. 1D). To support the findings from our miR-101–overexpression experiments, we employed antagomiR technology (*28*) to specifically inhibit miR-101 expression in benign immortalized breast epithelial cells (fig. S3). Two independent antagomiRs to miR-101 (i and ii) induced expression of EZH2 protein in benign breast epithelial cells.

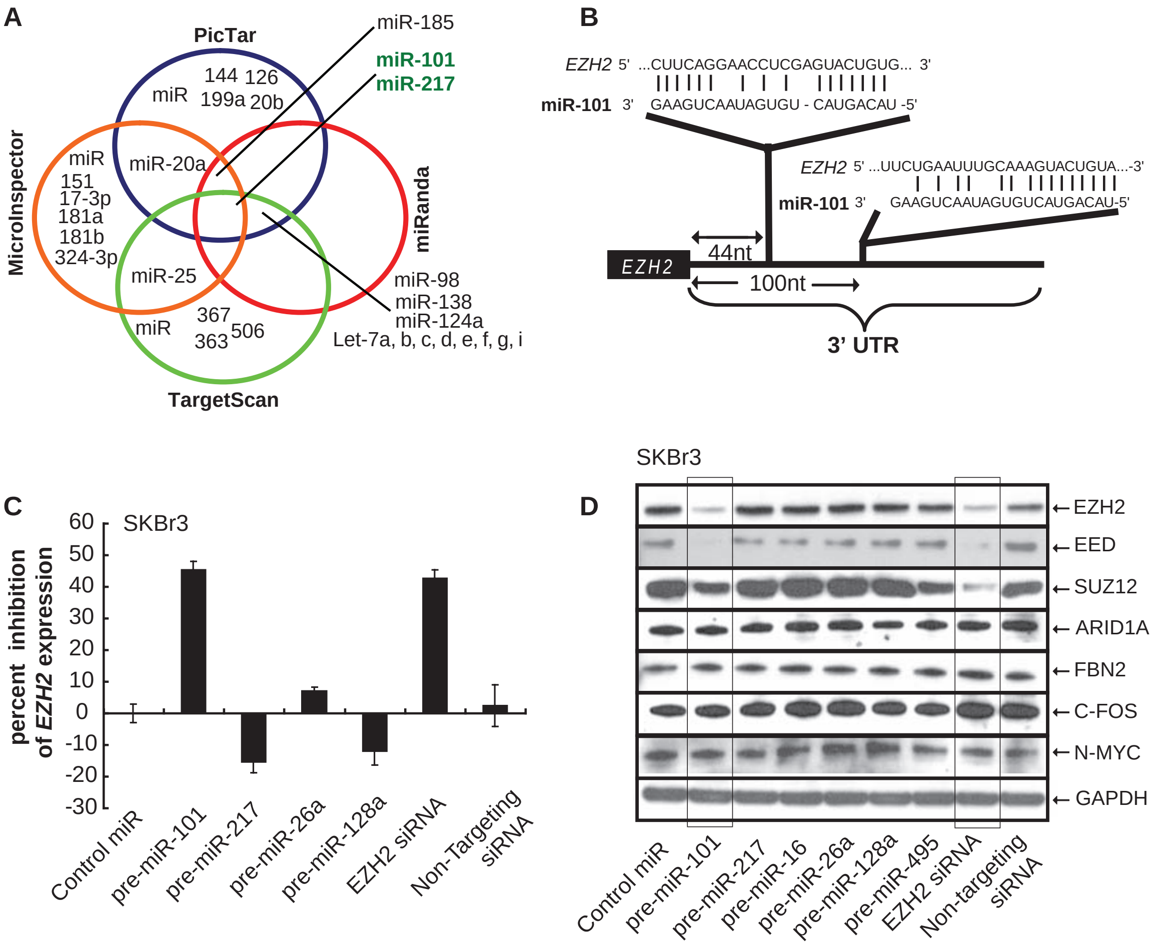


**Fig. 1.** miR-101 regulates EZH2 transcript and protein expression. (**A**) Venn diagram displaying miRNAs computationally predicted to target EZH2 by PicTar (blue), miRanda (red), TargetScan (green), and MicroInspector (orange). (**B**) Schematic of two predicted miR-101–binding sites in the *EZH2* 3′UTR. (**C**) miR-101 regulates *EZH2* transcript expression. Quantitative RT-PCR of *EZH2* in SKBr3 cells transfected with precursor miR-101 is shown. Control miR and other precursor miRNAs (miR-26a, miR-128a, and miR-217) were also used for transfection. (**D**) miR-101 regulates PRC2 protein expression. miR-101 down-regulates EZH2 protein as well as PRC2 members SUZ12 and EED in SKBr3 cells. Control miRs and EZH2-specific siRNA were also used for transfection. The experiment was performed three independent times and a representative result is displayed. GAPDH, glyceraldehyde-3-phosphate dehydrogenase.

To determine whether miR-101 affects EZH2 and PRC2 function, we evaluated cellular proliferation, a property known to be regulated by EZH2 (*4*, *5*). miR-101 overexpression in SKBr3 and DU145 cells markedly attenuated cell proliferation (Fig. 2A and fig. S4). Overexpression of EZH2 (without an endogenous 3′UTR) rescued the inhibition of cell growth by miR-101, which suggests target specificity.

We previously showed that upon overexpression, EZH2 can induce cell invasion in matrigel-coated basement membrane invasion assays (*12*). Here we show that miR-101 overexpression markedly inhibits the in vitro invasive potential of DU145 prostate-cancer cells (Fig. 2B) and SKBr3 breast cancer cells (fig. S5). Similarly, stable expression of miR-101 in DU145 cells showed a reduction in *EZH2* expression and reduced invasion (fig. S6, A and B). Overexpression of EZH2 rescued the inhibition that was mediated by miR-101. Another in vitro readout for tumorigenic potential, increased cell migration, was also inhibited by miR-101 (fig. S7). Because overexpression of miR-101 attenuates cancer invasion, inhibition of miR-101 should enhance this neoplastic phenotype. Two independent antagomiRs targeting miR-101 (i and ii) induced an invasive phenotype when transfected into benign immortalized breast epithelial cell lines H16N2 or HME (Fig. 2C and fig. S8).

To assess whether miR-101 inhibits anchorage-independent growth, we used a soft-agar assay. DU145 prostate cancer cells stably overexpressing miR-101 exhibited markedly reduced colony formation relative to the parental cells or vector controls (fig. S9). Furthermore, in vivo, DU145 cells expressing miR-101 grew significantly slower than the vector control xenografts ($P$ = 0.0001) (Fig. 2D), demonstrating that miR-101 has properties consistent with that of a tumor suppressor in these particular assays.

Because EZH2 and PRC2 regulate gene expression by trimethylating H3K27, we hypothesized that miR-101 overexpression would result in decreased overall H3K27 trimethylation in cancer cells. SKBr3 breast cancer and DU145 prostate cancer cells transfected with miR-101 or EZH2 siRNA for 7 days displayed a global decrease in trimethyl H3K27 levels (fig. S10A). The effect of miR-101 on H3K27 methylation was negated by overexpression of EZH2 (fig. S10B).

To test the level of promoter occupancy of the H3K27 histone mark, we performed chromatin immunoprecipitation (ChIP) assays in cancer cells overexpressing miR-101. We found significant reduction in the trimethyl H3K27 histone mark at the promoter of known PRC2 target genes

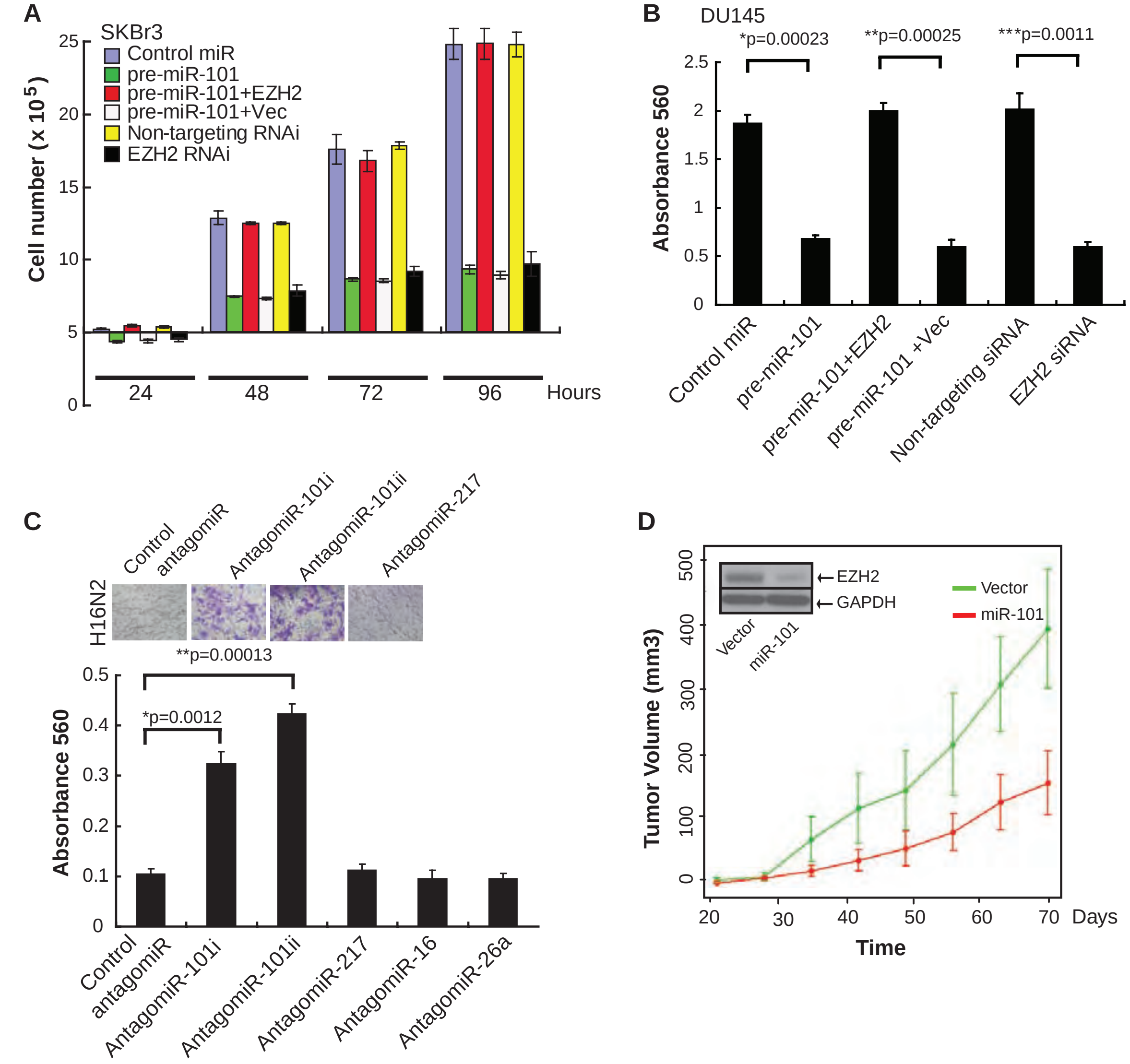


**Fig. 2.** The role of miR-101 in regulating cell proliferation, invasion, and tumor growth. (**A**) miR-101 overexpression reduces cell proliferation. A cell growth assay of SKBr3 cells treated with either precursor miR-101 or siRNA targeting EZH2 is shown. Cell growth relative to the control miRNA and control siRNA duplex was measured. Rescue experiments were performed by overexpressing EZH2 (minus its endogenous 3′UTR) in miR-101–treated cells. (**B**) miR-101 expression decreases cell invasion of DU145 prostate carcinoma cells. We transfected cells with miR-101, EZH2-specific siRNA, control miR, and nontargeting siRNA. miR-101 was also overexpressed in those cells that overexpressed EZH2 by andenoviral infection. All cells were subjected to a matrigel invasion assay. (**C**) AntagomiRs to miR-101 induce the invasiveness of benign immortalized H16N2 breast epithelial cells. Representative fields of invaded and stained cells are shown in the inset. *P* values were calculated between control antagomiR, antagomiR-101i, and antagomiR-101ii. (**D**) Overexpression of miR-101 attenuates prostate tumor growth. Overexpression of miR-101 reduces DU145 tumor growth in a mouse xenograft model. Plot of mean tumor-volume trajectories over time for the mice inoculated with (red) miR-101– and (green) vector-expressing DU145 cells. Error bars represent the SE of the mean at each time point. The inset displays the decrease of EZH2 protein levels in miR-101–expressing cell lines.

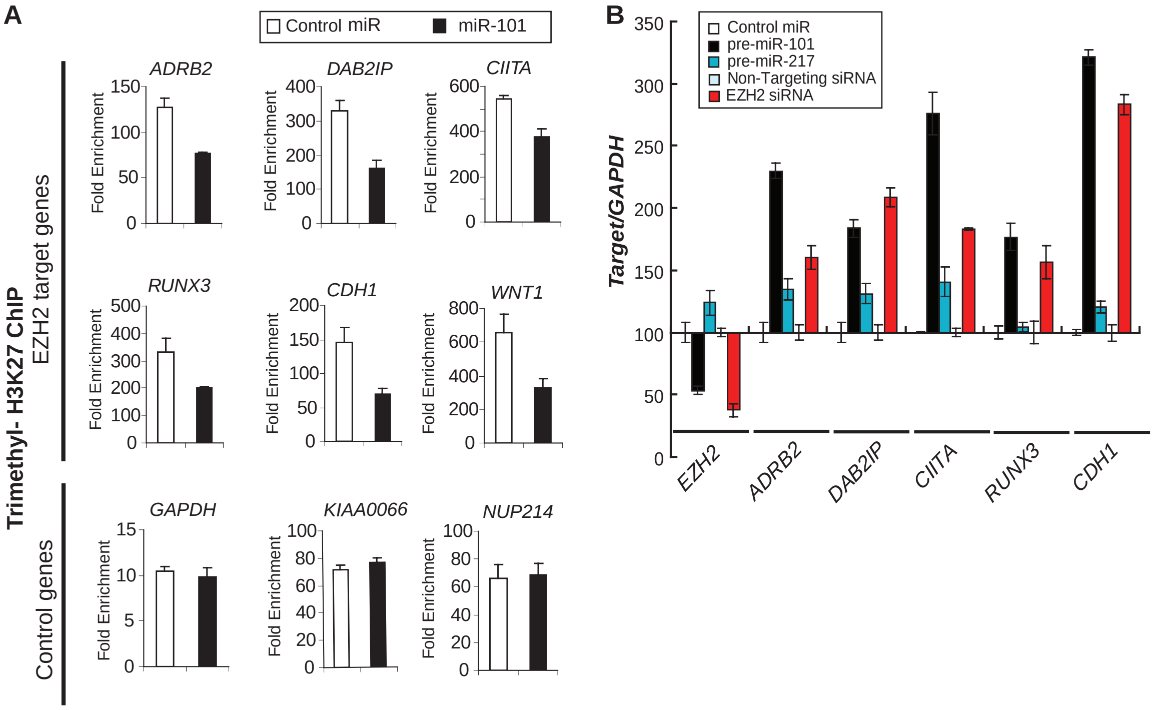


**Fig. 3.** miR-101 regulation of the cancer epigenome through EZH2 and H3K27 trimethylation. (**A**) ChIP assay of the trimethyl H3K27 histone mark when miR-101 is overexpressed. Known PRC2 repression targets were examined in SKBr3 cells. ChIP was performed to test H3K27 trimethylation at the promoters of *ADRB2*, *DAB2IP*, *CIITA*, *RUNX3*, *CDH1*, and *WNT1*. *GAPDH*, *KIAA0066*, and *NUP214* gene promoters served as controls. (**B**) Quantitative RT-PCR of EZH2 target genes was performed with SKBr3 cells transfected with miR-101. The EZH2 transcript and its known targets, including *ADRB2*, *DAB2IP*, *CIITA*, *RUNX3*, and E-cadherin (*CDH1*) were measured.

such as *ADRB2*, *DAB2IP*, *CIITA*, and *WNT1* in miR-101–overexpressing SKBr3 cells and EZH2 siRNA–treated cells (Fig. 3A and fig. S11). To determine whether the reduced promoter occupancy by H3K27 results in concomitant reduction of gene expression, we performed quantitative RT-PCR on the PRC2 targets tested by ChIP assay. As expected, there was a significant increase in target gene expression in both miR-101– and EZH2 siRNA–treated cells (Fig. 3B). To further explore miR-101 regulation of EZH2 and its possible similarity with EZH2-specific RNA interference (RNAi), we examined whether miR-101 overexpression and *EZH2* knockdown generated similar gene expression profiles. To assess this, we conducted gene-expression array analysis of SKBr3 cells transfected with either miR-101 or EZH2 siRNA duplexes. Genes that were overexpressed at the twofold threshold were significantly overlapping in both the miR-101– and EZH2 siRNA–transfected cells ($P = 6.08 \times 10^{-17}$) (fig. S12). Similarly, those genes that were repressed also had significant overlap ($P = 3.24 \times 10^{-27}$).

We next investigated whether miR-101 expression inversely correlates with EZH2 levels in human tumors. A meta-analysis of a majority of the publicly available miRNA expression data sets suggested that miR-101 is significantly underexpressed in prostate, breast, ovarian, lung, and colon cancers (table S3). Because EZH2 was initially found to be overexpressed in a subset of aggressive clinically localized prostate cancers and almost universally elevated in metastatic disease (*4*), we examined miR-101 in a similar context of prostate cancer progression by doing quantitative PCR analysis (Fig. 4A and fig. S13). As expected, metastatic prostate cancers expressed significantly higher levels of EZH2 as compared with those of clinically localized disease or benign adjacent prostate tissue ($P <$ 0.0001). Consistent with a functional connection between miR-101 and EZH2, miR-101 expression was significantly decreased in metastatic prostate cancer relative to that in clinically localized disease or benign adjacent prostate tissue ($P < 0.0001$). miR-217, which like miR-101 was predicted to regulate EZH2, did not exhibit significant differences between metastatic disease and clinically localized prostate cancer or benign prostate tissue ($P = 0.35$ and 0.13, respectively).

To investigate the mechanism for miR-101 transcript loss in prostate cancer progression, we performed quantitative genomic PCR for miR-101. miR-101 has two genomic loci that are on chromosome 1 (miR-101–1) and chromosome 9 (miR-101–2) (fig. S14, A and B). Based on genomic PCR, 2 of 16 clinically localized prostate cancer s and 17 of 33 metastatic prostate cancers exhibited loss of the miR-101–1 locus (Fig. 4B). Similarly, 4 of 16 clinically localized prostate cancers and 8 of 33 metastatic prostate cancers displayed loss of miR-101–2 (Fig. 4B). Figure 4C displays a heat-map representation of matched samples in which miR-101 transcript, *EZH2* transcript, miR-101–1 genomic loci, and miR-101–2 genomic loci were monitored. *EZH2* transcript levels were inversely associated with miR-101 transcript levels across prostate cancer progression to metastasis ($P <$ 0.0001). *EZH2* tended to be uniformly elevated in samples in which the miR-101–1 or miR-101–2 genomic loci exhibited a loss in copy number ($P = 0.004$, permutation test).

To formally demonstrate that genomic loss of miR-101 loci was somatic in nature, we identified nine metastatic prostate cancers that exhibited loss of miR-101–1 and obtained DNA from matched normal tissue. As expected, eight of nine cases exhibited a marked decrease in relative levels of miR-101–1 copy number in the cancer as compared with that in matched normal tissue (Fig. 4D). We also explored miR-101 genomic loss in other tumor types. Using a number of experimental platforms, we demonstrated focal loss (~20 kB) of miR-101–1 in a subset of breast, gastric, and prostate cancers (figs. S15 to S17). Furthermore, we explored public-domain high-density array comparative genomic hybridization and single-nucleotide polymorphism array data sets and observed a genomic loss of one or both miR-101 loci in a subset of glioblastoma multiforme, lung adenocarcinoma, and acute lymphocytic leukemia (fig. S18) (*25*).

miR-101, by virtue of its regulation of EZH2, may have profound control over the epigenetic pathways that are active not only in cancer cells but in normal pluripotent embryonic stem cells. Overexpression of miR-101 may configure the histone code of cancer cells to that associated with a more benign cellular phenotype. Because the loss of miR-101 and concomitant elevation of EZH2 are most pronounced in metastatic cancer, we postulate that miR-101 loss may represent a progressive molecular lesion in the development of more aggressive disease. Approaches to reintroduce miR-101 into tumors may have therapeutic benefit by reverting the epigenetic program of tumor cells to a more normal state.

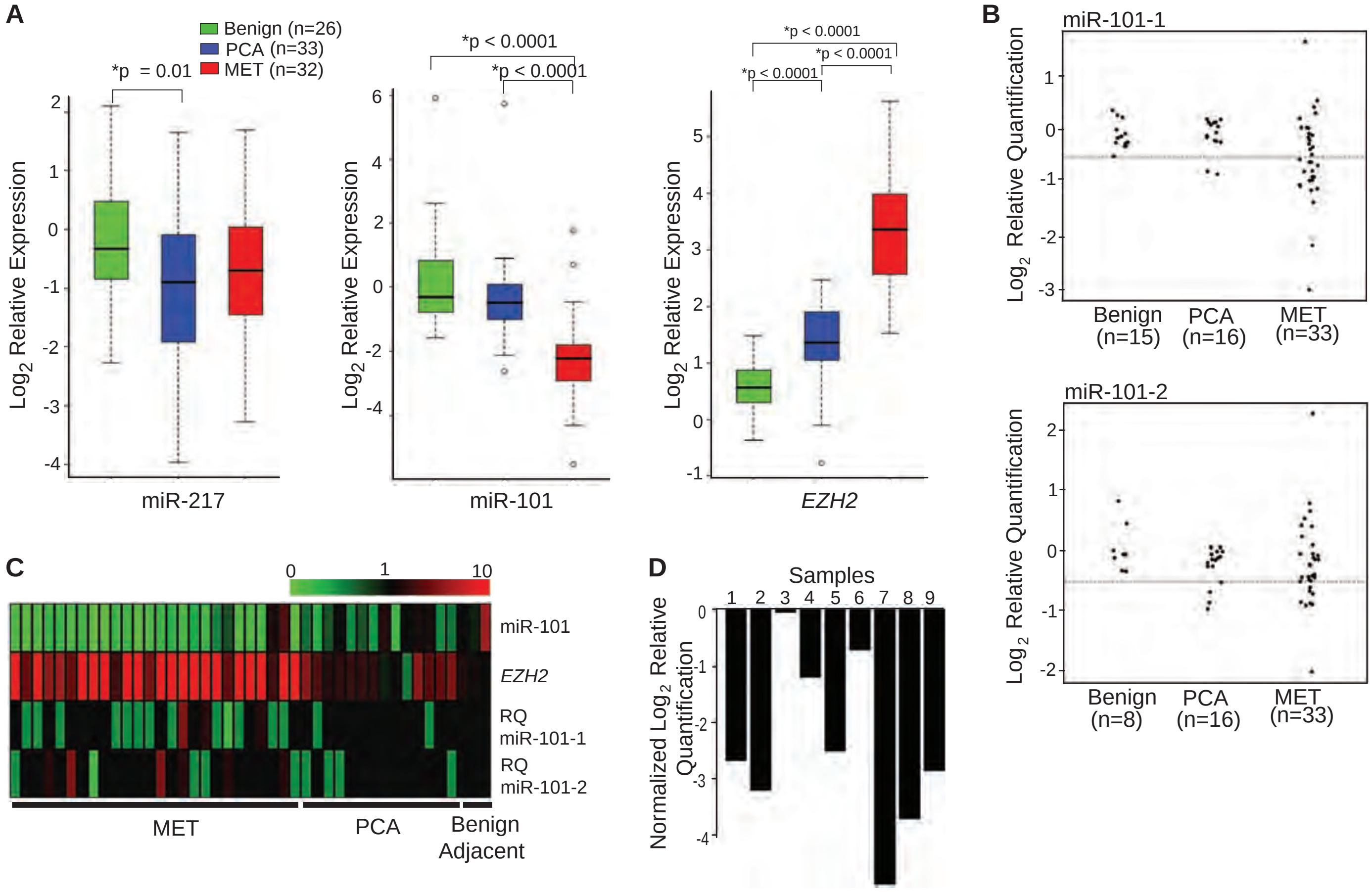


**Fig. 4.** Genomic loss of the miR-101 locus may explain overexpression of EZH2 in solid tumors. (**A**) miR-101 transcript levels are inversely correlated with EZH2 expression in prostate cancer progression. We performed quantitative PCR for miR-101 and miR-217 by using total RNA from benign adjacent prostate, prostate cancer (PCA), and metastatic (MET) prostate cancer tissue. *EZH2* expression was analyzed from the same RNA samples. (**B**) Genomic PCR of miR-101–1 and miR-101–2 in prostate cancer progression. Vertical axes represent log (base 2) relative quantification values; dashed lines are shown at the deletion threshold of $\log_2(0.7) \approx -0.51$. For clarity, points have been horizontally displaced within each sample class. (**C**) Heat-map representation of matched normal, tumor, and metastatic samples (from right to left) in which miR-101 transcript, *EZH2* transcript, and both miR-101–1 and miR-101–2 relative copy number were assessed. miR-101 and *EZH2* expression is represented by a color scale highlighting down-regulation (green), no alteration (black), and up-regulation (red) of transcripts. miR-101–1 and miR-101–2 relative quantitation (RQ) of copy number are represented as homozygous loss (<0.3; bright green), single-copy loss (<0.7; light green), no copy number change (≥0.7 and ≤1.3; black), single-copy gain (>1.3; light red), and double-copy gain (>1.7; bright red). (**D**) Evidence that the miR-101–1 locus is somatically lost in tumors samples relative to matched normal samples. Nine metastatic prostate cancers were chosen that had copy number loss in the miR-101–1 locus, and matched normal tissue was analyzed for comparison. Bar heights represent differences in $\log_2(RQ)$ values between metastatic and matched normal tissues.

### References and Notes

1. L. A. Boyer *et al.*, *Nature* **441**, 349 (2006).
2. T. I. Lee *et al.*, *Cell* **125**, 301 (2006).
3. F. Sher *et al.*, *Stem Cells* **26**, 2875 (2008).
4. S. Varambally *et al.*, *Nature* **419**, 624 (2002).
5. A. P. Bracken *et al.*, *EMBO J.* **22**, 5323 (2003).
6. S. Erhardt *et al.*, *Development* **130**, 4235 (2003).
7. K. Plath *et al.*, *Science* **300**, 131 (2003).
8. R. Cao *et al.*, *Science* **298**, 1039 (2002).
9. A. Kuzmichev, K. Nishioka, H. Erdjument-Bromage, P. Tempst, D. Reinberg, *Genes Dev.* **16**, 2893 (2002).
10. J. Yu *et al.*, *Cancer Cell* **12**, 419 (2007).
11. Q. Cao *et al.*, *Oncogene* 10:1038/onc.2008.333 (2008).
12. C. G. Kleer *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 11606 (2003).
13. P. A. Croonquist, B. Van Ness, *Oncogene* **24**, 6269 (2005).
14. F. Takeshita *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **102**, 12177 (2005).
15. I. M. Bachmann *et al.*, *J. Clin. Oncol.* **24**, 268 (2006).
16. S. Weikert *et al.*, *Int. J. Mol. Med.* **16**, 349 (2005).
17. Y. Matsukawa *et al.*, *Cancer Sci.* **97**, 484 (2006).
18. B. P. Lewis, C. B. Burge, D. P. Bartel, *Cell* **120**, 15 (2005).
19. A. Marson *et al.*, *Cell* **134**, 521 (2008).
20. P. Sethupathy, M. Megraw, A. G. Hatzigeorgiou, *Nat. Methods* **3**, 881 (2006).
21. A. Krek *et al.*, *Nat. Genet.* **37**, 495 (2005).
22. B. P. Lewis, I. H. Shih, M. W. Jones-Rhoades, D. P. Bartel, C. B. Burge, *Cell* **115**, 787 (2003).
23. B. John *et al.*, *PLoS Biol.* **2**, e363 (2004).
24. V. Rusinov, V. Baev, I. N. Minkov, M. Tabler, *Nucleic Acids Res.* **33**, W696 (2005).
25. Materials and methods are available as supporting material on *Science* Online.
26. D. Pasini, A. P. Bracken, M. R. Jensen, E. Lazzerini Denchi, K. Helin, *EMBO J.* **23**, 4061 (2004).
27. W. Fiskus *et al.*, *Mol. Cancer Ther.* **5**, 3096 (2006).
28. J. Krutzfeldt *et al.*, *Nature* **438**, 685 (2005).

29. We thank S. M. Dhanasekaran, S. Tomlins, and J. Yu for helpful discussions; J. Siddiqui and M. Pandhi for technical assistance; and J. Granger and K. Giles for critically reading the manuscript. A.M.C. is supported by a Burroughs Welcome Foundation Award in Clinical Translational Research. This work was supported in part by NIH (Prostate Specialized Program of Reasearch Excellence grant P50CA69568 to A.M.C. and Early Detection Research Network grant UO1 111275 to A.M.C.) and the U.S. Department of Defense (Era of Hope Scholar grant BC075023 to A.M.C., PC051081 to A.M.C. and S.V., and BC083217 to J.C.B.). C.A.M. was supported by an NIH postdoctoral training grant and currently derives support from the American Association of Cancer Research Amgen Fellowship in Clinical/Translational Research and the Canary Foundation and American Cancer Society Early Detection Postdoctoral Fellowship. The microarray data used in this study have been deposited in the National Center for Biotechnology Information Gene Expression Omnibus with the accession number GSE13286.

### Supporting Online Material

www.sciencemag.org/cgi/content/full/1165395/DC1
Materials and Methods
Figs. S1 to S18
Tables S1 to S10
References

3 September 2008; accepted 3 November 2008
Published online 13 November 2008;
10.1126/science.1165395
Include this information when citing this paper.

# Chimeric transcript discovery by paired-end transcriptome sequencing

Christopher A. Maher[a,b], Nallasivam Palanisamy[a,b], John C. Brenner[a,b], Xuhong Cao[a,c], Shanker Kalyana-Sundaram[a,b], Shujun Luo[d], Irina Khrebtukova[d], Terrence R. Barrette[a,b], Catherine Grasso[a,b], Jindan Yu[a,b], Robert J. Lonigro[a,b], Gary Schroth[d], Chandan Kumar-Sinha[a,b], and Arul M. Chinnaiyan[a,b,c,e,f,1]

[a]Michigan Center for Translational Pathology, Ann Arbor, MI 48109; Departments of [b]Pathology and [e]Urology, University of Michigan, Ann Arbor, MI 48109; [c]Howard Hughes Medical Institute and [f]Comprehensive Cancer Center, University of Michigan Medical School, Ann Arbor, MI 48109; and [d]Illumina Inc., 25861 Industrial Boulevard, Hayward, CA 94545

Communicated by David Ginsburg, University of Michigan Medical School, Ann Arbor, MI, May 4, 2009 (received for review March 16, 2009)

**Recurrent gene fusions are a prevalent class of mutations arising from the juxtaposition of 2 distinct regions, which can generate novel functional transcripts that could serve as valuable therapeutic targets in cancer. Therefore, we aim to establish a sensitive, high-throughput methodology to comprehensively catalog functional gene fusions in cancer by evaluating a paired-end transcriptome sequencing strategy. Not only did a paired-end approach provide a greater dynamic range in comparison with single read based approaches, but it clearly distinguished the high-level "driving" gene fusions, such as *BCR-ABL1* and *TMPRSS2-ERG*, from potential lower level "passenger" gene fusions. Also, the comprehensiveness of a paired-end approach enabled the discovery of 12 previously undescribed gene fusions in 4 commonly used cell lines that eluded previous approaches. Using the paired-end transcriptome sequencing approach, we observed read-through mRNA chimeras, tissue-type restricted chimeras, converging transcripts, diverging transcripts, and overlapping mRNA transcripts. Last, we successfully used paired-end transcriptome sequencing to detect previously undescribed ETS gene fusions in prostate tumors. Together, this study establishes a highly specific and sensitive approach for accurately and comprehensively cataloguing chimeras within a sample using paired-end transcriptome sequencing.**

bioinformatics | gene fusions | prostate cancer | breast cancer | RNA-Seq

One of the most common classes of genetic alterations is gene fusions, resulting from chromosomal rearrangements (1). Intriguingly, >80% of all known gene fusions are attributed to leukemias, lymphomas, and bone and soft tissue sarcomas that account for only 10% of all human cancers. In contrast, common epithelial cancers, which account for 80% of cancer-related deaths, can only be attributed to 10% of known recurrent gene fusions (2–4). However, the recent discovery of a recurrent gene fusion, *TMPRSS2-ERG*, in a majority of prostate cancers (5, 6), and *EML4-ALK* in non-small-cell lung cancer (NSCLC) (7), has expanded the realm of gene fusions as an oncogenic mechanism in common solid cancers. Also, the restricted expression of gene fusions to cancer cells makes them desirable therapeutic targets. One successful example is imatinib mesylate, or Gleevec, that targets *BCR-ABL1* in chronic myeloid leukemia (CML) (8–10). Therefore, the identification of novel gene fusions in a broad range of cancers is of enormous therapeutic significance.

The lack of known gene fusions in epithelial cancers has been attributed to their clonal heterogeneity and to the technical limitations of cytogenetic analysis, spectral karyotyping, FISH, and microarray-based comparative genomic hybridization (aCGH). Not surprisingly, *TMPRSS2-ERG* was discovered by circumventing these limitations through bioinformatics analysis of gene expression data to nominate genes with marked overexpression, or outliers, a signature of a fusion event (6). Building on this success, more recent strategies have adopted unbiased high-throughput approaches, with increased resolution, for genome-wide detection of chromosomal rearrangements in cancer involving BAC end sequencing (11), fosmid paired-end sequences (12), serial analysis of gene expression (SAGE)-like sequencing (13), and next-generation DNA sequencing (14). Despite unveiling many novel genomic rearrangements, solid tumors accumulate multiple nonspecific aberrations throughout tumor progression; thus, making causal and driver aberrations indistinguishable from secondary and insignificant mutations, respectively.

The deep unbiased view of a cancer cell enabled by massively parallel transcriptome sequencing has greatly facilitated gene fusion discovery. As shown in our previous work, integrating long and short read transcriptome sequencing technologies was an effective approach for enriching "expressed" fusion transcripts (15). However, despite the success of this methodology, it required substantial overhead to leverage 2 sequencing platforms. Therefore, in this study, we adopted a single platform paired-end strategy to comprehensively elucidate novel chimeric events in cancer transcriptomes. Not only was using this single platform more economical, but it allowed us to more comprehensively map chimeric mRNA, hone in on driver gene fusion products due to its quantitative nature, and observe rare classes of transcripts that were overlapping, diverging, or converging.

## Results

**Chimera Discovery via Paired-End Transcriptome Sequencing.** Here, we employ transcriptome sequencing to restrict chimera nominations to "expressed sequences," thus, enriching for potentially functional mutations. To evaluate massively parallel paired-end transcriptome sequencing to identify novel gene fusions, we generated cDNA libraries from the prostate cancer cell line VCaP, CML cell line K562, universal human reference total RNA (UHR; Stratagene), and human brain reference (HBR) total RNA (Ambion). Using the Illumina Genome Analyzer II, we generated 16.9 million VCaP, 20.7 million K562, 25.5 million UHR, and 23.6 million HBR transcriptome mate pairs ($2 \times 50$ nt). The mate pairs were mapped against the transcriptome and categorized as (*i*) mapping to same gene, (*ii*) mapping to different genes (chimera candidates), (*iii*) nonmapping, (*iv*) mitochondrial, (*v*) quality control, or (*vi*) ribosomal (Table S1). Overall, the chimera candidates represent a minor fraction of the mate pairs, comprising ≈<1% of the reads for each sample.

We believe that a paired-end strategy offers multiple advantages over single read based approaches such as alleviating the reliance on sequencing the reads traversing the fusion junction, increased coverage provided by sequencing reads from the ends of a tran-

Author contributions: C.A.M. and A.M.C. designed research; C.A.M., N.P., J.C.B., X.C., S.L., I.K., T.R.B., R.J.L., G.S., C.K.-S., and A.M.C. performed research; C.A.M., S.L., I.K., R.J.L., and G.S. contributed new reagents/analytic tools; C.A.M., N.P., J.C.B., S.K.-S., C.G., J.Y., R.J.L., G.S., C.K.-S., and A.M.C. analyzed data; and C.A.M., N.P., X.C., C.K.-S., and A.M.C. wrote the paper.

The authors declare no conflict of interest.

[1]To whom correspondence should be addressed. E-mail: arul@umich.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0904720106/DCSupplemental.

scribed fragment, and the ability to resolve ambiguous mappings (Fig. S1). Therefore, to nominate chimeras, we leveraged each of these aspects in our bioinformatics analysis. We focused on both mate pairs encompassing and/or spanning the fusion junction by analyzing 2 main categories of sequence reads: chimera candidates and nonmapping (Fig. S2*A*). The resulting chimera candidates from the nonmapping category that span the fusion boundary were merged with the chimeras found to encompass the fusion boundary revealing 119, 144, 205, and 294 chimeras in VCaP, K562, HBR, and UHR, respectively.

**Comparison of a Paired-End Strategy Against Existing Single Read Approaches.** To assess the merit of adopting a paired-end transcriptome approach, we compared the results against existing single read approaches. Although current RNA sequencing (RNA-Seq) studies have been using 36-nt single reads (16, 17), we increased the likelihood of spanning a fusion junction by generating 100-nt long single reads using the Illumina Genome Analyzer II. Also, we chose this length because it would facilitate a more comparable amount of sequencing time as required for sequencing both 50-nt mate pairs. In total, we generated 7.0, 59.4, and 53.0 million 100-nt transcriptome reads for VCaP, UHR, and HBR, respectively, for comparison against paired-end transcriptome reads from matched samples.

Because the UHR is a mixture of cancer cell lines, we expected to find numerous previously identified gene fusions. Therefore, we first assessed the depth of coverage of a paired-end approach against long single reads by directly comparing the normalized frequency of sequence reads supporting 4 previously identified gene fusions [*TMPRSS2-ERG* (5, 6), *BCR-ABL1* (18), *BCAS4-BCAS3* (19), and *ARFGEF2-SULF2* (20)]. As shown in Fig. 1*A*, we observed a marked enrichment of paired-end reads compared with long single reads for each of these well characterized gene fusions.

We observed that *TMPRSS2-ERG* had a >10-fold enrichment between paired-end and single read approaches. The schematic representation in Fig. 1*B* indicates the distribution of reads confirming the *TMPRSS2-ERG* gene fusion from both paired-end and single read sequencing. As expected, the longer reads improve the number of reads spanning known gene fusions. For example, had we sequenced a single 36-mer (shown in red text), 11 of the 17 chimeras, shown in the bottom portion of the long single reads, would not have spanned the gene fusion boundary, but instead, would have terminated before the junction and, therefore, only aligned to *TMPRSS2*. However, despite the improved results only 17 chimeric reads were generated from 7.0 million long single read sequences. In contrast, paired-end sequencing resulted in 552 reads supporting the *TMPRSS2-ERG* gene fusion from ≈17 million sequences.

Because we are using sequence based evidence to nominate a chimera, we hypothesized that the approach providing the maximum nucleotide coverage is more likely to capture a fusion junction. We calculated an *in silico* insert size for each sample using mate pairs aligning to the same gene, and found the mean insert size of ≈200 nt. Then, we compared the total coverage from single reads (coverage is equivalent to the total number of pass filter reads against the read length) with the paired-end approach (coverage is equivalent to the sum of the insert size with the length of each read) (Fig. S2*B*). Overall, we observed an average coverage of 848.7 and 757.3 MB using single read technology, compared with 2,553.3 and 2,363 MB from paired-end in UHR and HBR, respectively. This increase in ≈3-fold coverage in the paired-end samples compared with the long read approach, per lane, could explain the increased dynamic range we observed using a paired-end strategy.

Next we wanted to identify chimeras common to both strategies. The long read approach nominated 1,375 and 1,228 chimeras, whereas with a paired-end strategy, we only nominated 225 and 144 chimeras in UHR and HBR, respectively. As shown in the Venn diagram (Fig. 1*C*), there were 32 and 31 candidates common to both technologies for UHR and HBR, respectively. Within the common UHR chimeric candidates, we observed previously identified gene fusions *BCAS4-BCAS3*, *BCR-ABL1*, *ARFGEF2-SULF2*, and *RPS6KB1-TMEM49* (13). The remaining chimeras, nominated by both approaches, represent a high fidelity set. Therefore, to further assess whether a paired-end strategy has an increased dynamic range, we compared the ratio of normalized mate pair reads against single reads for the remaining chimeras common to both technologies. We observed that 93.5 and 93.9% of UHR and HBR candidates, respectively, had a higher ratio of normalized mate pair reads to single reads (Table S2), confirming the increased dynamic range offered by a paired-end strategy. We hypothesize that the greater number of nominated candidates specific to the long read approach represents an enrichment of false positives, as observed when using the 454 long read technology (15, 21).

**Paired-End Approach Reveals Novel Gene Fusions.** We were interested in determining whether the paired-end libraries could detect novel gene fusions. Among the top chimeras nominated from VCaP, HBR, UHR, and K562, many were already known, including *TMPRSS2-ERG*, *BCAS4-BCAS3*, *BCR-ABL1*, *USP10-ZDHHC7*, and *ARFGEF2-SULF2.* Also ranking among these well known gene fusions in UHR was a fusion on chromosome 13 between *GAS6* and *RASA3* (Fig. S3*A* and Table S2). The fact that *GAS6-RASA3* ranked higher than *BCR-ABL1* suggests that it may be a driving fusion in one of the cancer cell lines in the RNA pool.

Another observation was that there were 2 candidates among the top 10 found in both UHR and K562. This observation was intriguing, because hematological malignancies are not considered to have multiple gene fusion events. In addition to *BCR-ABL1*, we were able to detect a previously undescribed interchromosomal gene fusion between exon 23 of *NUP214* located at chromosome 9q34.13 with exon 2 of *XKR3* located at chromosome 22q11.1. Both of these genes reside on chromosome 22 and 9 in close proximity to *BCR* and *ABL1*, respectively (Fig. S3*B*). We confirmed the presence of *NUP214-XKR3* in K562 cells using qRT-PCR, but were unable to detect it across an additional 5 CML cell lines tested (SUP-B15, MEG-01, KU812, GDM-1, and Kasumi-4) (Fig. S3*C*). These results suggest that *NUP214-XKR3* is a "private" fusion that originated from additional complex rearrangements after the translocation that generated *BCR-ABL1* and a focal amplification of both gene regions.

Although we were able to detect *BCR-ABL1* and *NUP214-XKR3* in both UHR and K562, there was a marked reduction in the mate pairs supporting these fusions in UHR. Although a diluted signal is expected, because UHR is pooled samples, it provides evidence that pooling samples can serve as a useful approach for nominating top expressing chimeras, and potentially enrich for "driver" chimeras.

**Previously Undescribed Prostate Gene Fusions.** Our previous work using integrative transcriptome sequencing to detect gene fusions in cancer revealed multiple gene fusions, demonstrating the complexity of the prostate transcriptomes of VCaP and LNCaP (15). Here, we exploit the comprehensiveness of a paired-end strategy on the same cell lines to reveal novel chimeras. In the circular plot shown in Fig. S4*A*, we displayed all experimentally validated paired-end chimeras in the larger red circle. We found that all of the previously discovered chimeras in VCaP and LNCaP comprised a subset of the paired-end candidates, as displayed in the inner black circle.

As expected, *TMPRSS2-ERG* was the top VCaP candidate. In addition to "rediscovering" the *USP10-ZDHHC7*, *HJURP-INPP4A*, and *EIF4E2-HJURP* gene fusions, a paired-end approach revealed several previously undescribed gene fusions in VCaP. One such example was an interchromosomal gene fusion between *ZDHHC7*, on chromosome 16, with *ABCB9*, residing on chromosome 12, that was validated by qRT-PCR (Fig. S3*D*). Interestingly, the 5′ partner, *ZDHHC7*, had previously been validated as a complex intrachro-

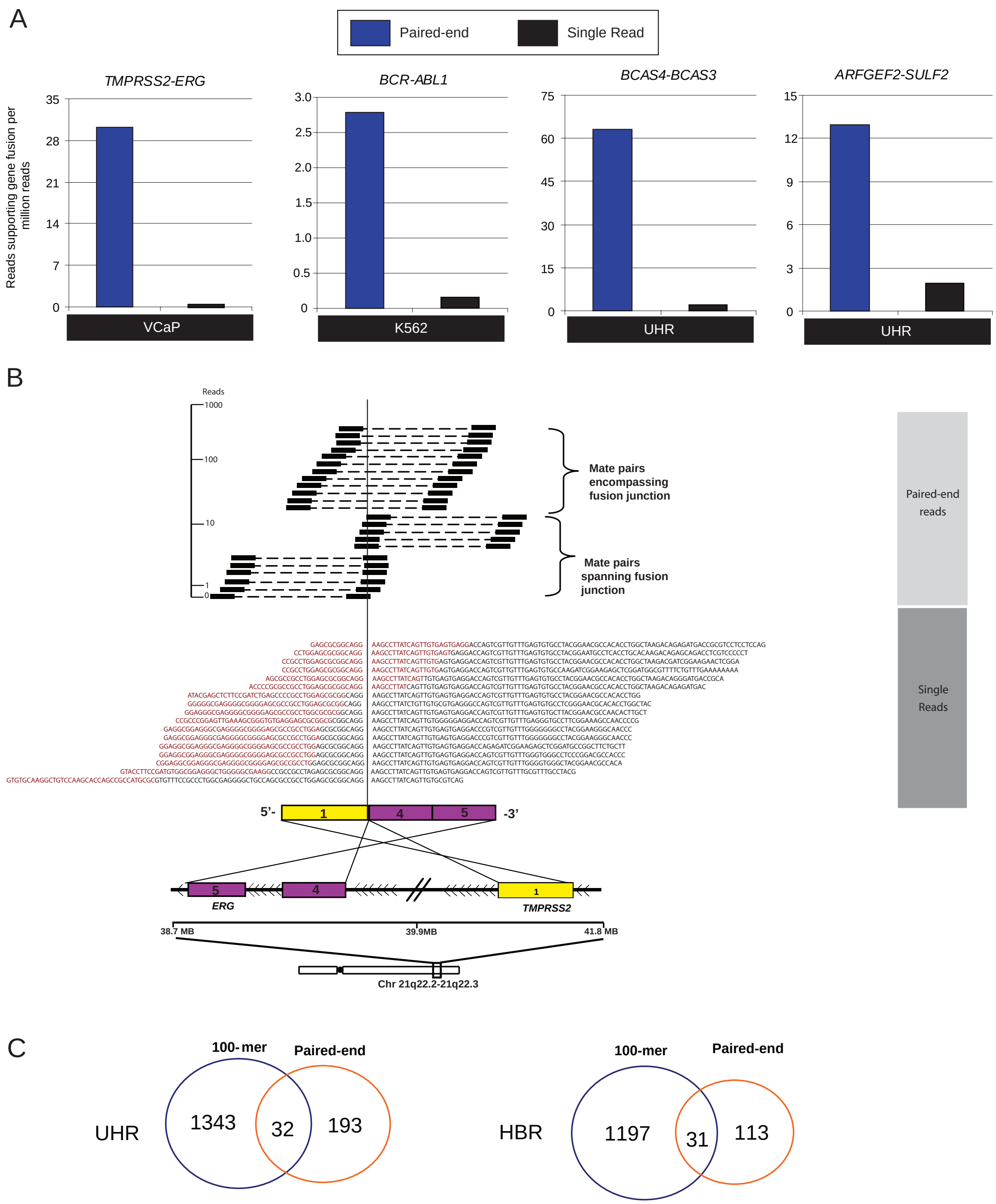


**Fig. 1.** Dynamic range and sensitivity of the paired-end transcriptome analysis relative to single read approaches. (*A*) Comparison of paired-end (blue) and long single transcriptome reads (black) supporting known gene fusions *TMPRSS2-ERG*, *BCR-ABL1*, *BCAS4-BCAS3*, and *ARFGEF2-SULF2*. (*B*) Schematic representation of *TMPRSS2-ERG* in VCaP, comparing mate pairs with long single transcriptome reads. (*Upper*) Frequency of mate pairs, shown in log scale, are divided based on whether they encompass or span the fusion boundary; (*Lower*) 100-mer single transcriptome reads spanning *TMPRSS2-ERG* fusion boundary. First 36 nt are highlighted in red. (*C*) Venn diagram of chimera nominations from both a paired-end (orange) and long single read (blue) strategy for UHR and HBR.

mosomal gene fusion with *USP10* (15). Both fusions have mate pairs aligning to the same exon of *ZDHHC7* (15), suggesting that their breakpoints are in adjacent introns (Fig. S3*D*).

Another previously undescribed VCaP interchromosomal gene fusion that we discovered was between exon 2 of *TIA1*, residing on chromosome 2, with exon 3 of *DIRC2*, or disrupted in renal carcinoma 2, located on chromosome 3. *TIA1-DIRC2* was validated by qRT-PCR and FISH (Fig. S5). In total, we confirmed an additional 4 VCaP and 2 LNCaP chimeras (Fig. S6). Overall, these fusions demonstrate that paired-end transcriptome sequencing can nominate candidates that have eluded previous techniques, including other massively parallel transcriptome sequencing approaches.

**Distinguishing Causal Gene Fusions from Secondary Mutations.** We were next interested in determining whether the dynamic range provided by paired-end sequencing can distinguish known high-

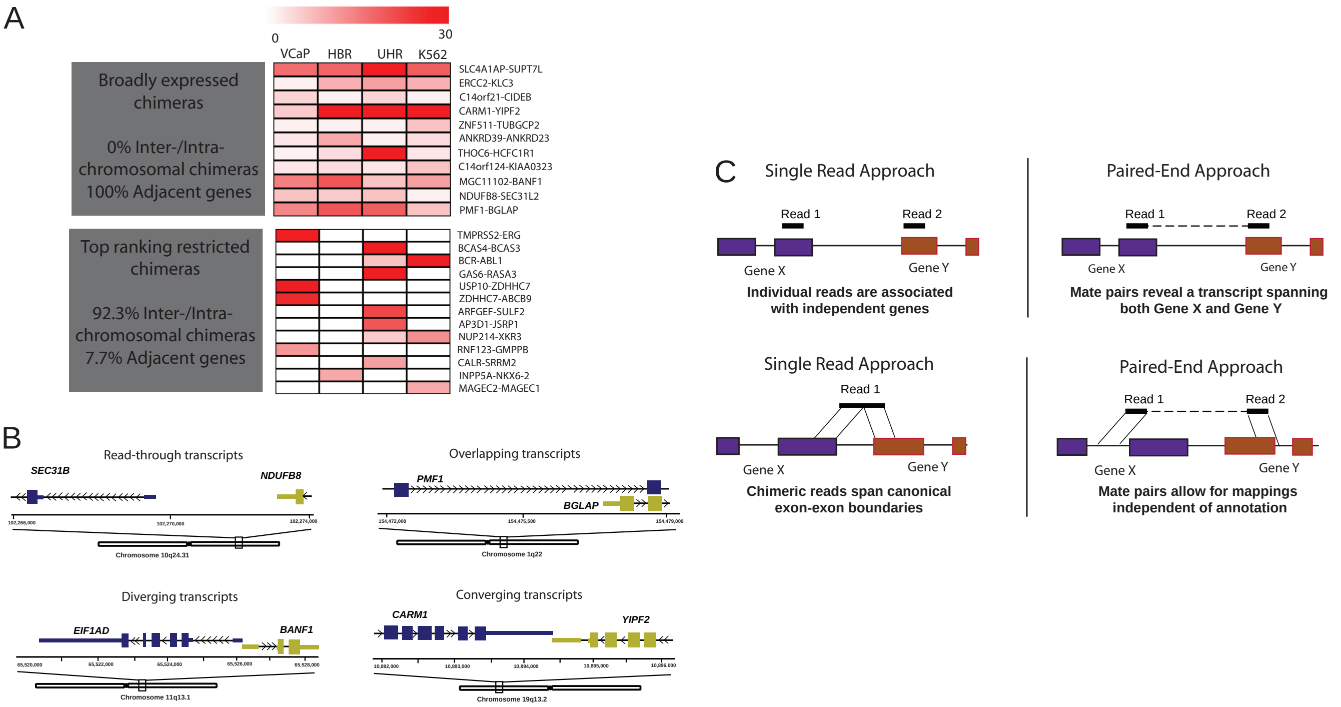


**Fig. 2.** RNA based chimeras. (*A*) Heatmaps showing the normalized number of reads supporting each read-through chimera across samples ranging from 0 (white) to 30 (red). (*Upper*) The heatmap highlights broadly expressed chimeras in UHR, HBR, VCaP, and K562. (*Lower*) The heatmap highlights the expression of the top ranking restricted gene fusions that are enriched with interchromosomal and intrachromosomal rearrangements. (*B*) Illustrative examples classifying RNA-based chimeras into (*i*) read-throughs, (*ii*) converging transcripts, (*iii*) diverging transcripts, and (*iv*) overlapping transcripts. (*C Upper*) Paired-end approach links reads from independent genes as belonging to the same transcriptional unit (*Right*), whereas a single read approach would assign these reads to independent genes (*Left*). (*Lower*) The single read approach requires that a chimera span the fusion junction (*Left*), whereas a paired-end approach can link mate pairs independent of gene annotation (*Right*).

level "driving" gene fusions, such as known recurrent gene fusions *BCR-ABL1* and *TMPRSS2-ERG*, from lower level "passenger" fusions. Therefore, we plotted the normalized mate pair coverage at the fusion boundary for all experimentally validated gene fusions for the 2 cell lines that we sequenced harboring recurrent gene fusions, VCaP and K562. As shown in Fig. S4*B*, we observed that both driver fusions, *TMPRSS2-ERG* and *BCR-ABL1*, show the highest expression among the validated chimeras in VCaP and K562, respectively. This observation suggests a paired-end nomination strategy for selecting putative driver gene fusions among private nonspecific gene fusions that lack detectable levels of expression across a panel of samples (15).

**Previously Undescribed Breast Cancer Gene Fusions.** Our ability to detect previously undescribed prostate gene fusions in VCaP and LNCaP demonstrated the comprehensiveness of paired-end transcriptome sequencing compared with an integrated approach, using short and long transcriptome reads. Therefore, we extended our paired-end analysis by using breast cancer cell line MCF-7, which has been mined for fusions using numerous approaches such as expressed sequence tags (ESTs) (22), array CGH (23), single nucleotide polymorphism arrays (24), gene expression arrays (25), end sequence profiling (20, 26), and paired-end diTag (PET) (13).

A histogram (Fig. S4*C*) of the top ranking MCF-7 candidates highlights *BCAS4-BCAS3* and *ARFGEF-SULF2* as the top 2 ranking candidates, whereas other previously reported candidates, such as *SULF2-PRICKLE*, *DEPDC1B-ELOVL7*, *RPS6KB1-TMEM49*, and *CXorf15-SYAP1*, were interspersed among a comprehensive list of previously undescribed putative chimeras. To confirm that these previously undescribed nominations were not false positives, we experimentally validated 2 interchromosomal and 3 intrachromosomal candidates using qRT-PCR (Fig. S6). Overall, not only was a paired-end approach able to detect gene fusions that have eluded numerous existing technologies, it has revealed 5 previously undescribed mutations in breast cancer.

**RNA-Based Chimeras.** Although many of the inter and intrachromosomal rearrangements that we nominated were found within a single sample, we observed many chimeric events shared across samples. We identified 11 chimeric events common to UHR, VCaP, K562, and HBR (Table S3). Via heatmap representation (Fig. 2*A*) of the normalized frequency of mate pairs supporting each chimeric event, we can observe these events are broadly transcribed in contrast to the top restricted chimeric events. Also, we found that 100% of the broadly expressed chimeras resided adjacent to one another on the genome, whereas only 7.7% of the restricted candidates were neighboring genes. This discrepancy can be explained by the enrichment of inter and intrachromosomal rearrangements in the restricted set.

Unlike, previously characterized restricted read-throughs, such as *SLC45A3-ELK4* (15), which are found adjacent to one another, but in the same orientation, we found that the majority of the broadly expressed chimera candidates resided adjacent to one another in different orientations. Therefore, we have categorized these events as (*i*) read-throughs, adjacent genes in the same orientation, (*ii*) diverging genes, adjacent genes in opposite orientation whose 5′ ends are in close proximity, (*iii*) convergent genes, adjacent genes in opposite orientation whose 3′ ends are in close proximity, and (*iv*) overlapping genes, adjacent genes who share common exons (Fig. 2*B*). Based on this classification, we found 1 read-through, 2 convergent genes, 6 divergent genes, and 2 overlapping genes. Also, we found that ≈81.8% of these chimeras had at least 1 supporting EST, providing independent confirmation of the event (Table S3). In contrast to paired-end, single read approaches would likely miss these instances as each mate would have aligned to their respective genes based on the current annotations (Fig. 2*C*). Also, these instances may represent extensions of a transcriptional unit, which would not be detectable by a single read approach that identifies chimeric reads that span exon boundaries of independent genes. Overall, we believe that many of these broadly expressed RNA chimeras represent instances where mate pairs are revealing previously undescribed annotation for a transcriptional unit.

**Previously Undescribed ETS Gene Fusions in Clinically Localized Prostate Cancer.** Given the high prevalence of gene fusions involving ETS oncogenic transcription factor family members in prostate tumors, we applied paired-end transcriptome sequencing for gene fusion discovery in prostate tumors lacking previously reported ETS fusions. For 2 prostate tumors, aT52 and aT64, we generated 6.2 and 7.4 million transcriptome mate pairs, respectively. In aT64, we found that *HERPUD1*, residing on chromosome 16, juxtaposed in front of exon 4 of *ERG* (Fig. 3*A*), which was validated by qRT-PCR (Fig. S6) and FISH (Fig. 3*B*), thus identifying a third 5′ fusion partner for *ERG*, after *TMPRSS2* (6) and *SLC45A3* (27), and presumably, *HERPUD1* also mediates the overexpression of ERG in a subset of prostate cancer patients. Also, just as *TMPRSS2* and *SLC45A3* have been shown to be androgen regulated by qRT-PCR (5), we found *HERPUD1* expression, via RNA-Seq, to be responsive to androgen treatment (Fig. S7). Also, ChIP-Seq analysis revealed androgen binding at the 5′ end of *HERPUD1* (Fig. S7).

Also, in the second prostate tumor sample (aT52), we discovered an interchromosomal gene fusion between the 5′ end of a prostate cDNA clone, *AX747630* (*FLJ35294*), residing on chromosome 17, with exon 4 of *ETV1*, located on chromosome 7 (Fig. 3*C*), which was validated via qRT-PCR (Fig. S6) and FISH (Fig. 3*D*). Interestingly, this fusion has previously been reported in an independent sample found by a fluorescence in situ hybridization screen (27); thus, demonstrating that it is recurrent in a subset of prostate cancer patients. As previously reported, gene expression via RNA-Seq confirmed that *AX747630* is an androgen-inducible gene (Fig. S7). Also, ChIP-Seq revealed androgen occupancy at the 5′ end of *AX747630* (Fig. S7).

## Discussion

This study demonstrates the effectiveness of paired-end massively parallel transcriptome sequencing for fusion gene discovery. By using a paired-end approach, we were able to rediscover known gene fusions, comprehensively discover previously undescribed gene fusions, and hone in on causal gene fusions. The ability to detect 12 previously undescribed gene fusions in 4 commonly used cell lines that eluded any previous efforts conveys the superior sensitivity of a paired-end RNA-Seq strategy compared with existing approaches. Also, it suggests that we may be able to unveil previously undescribed chimeric events in previously characterized samples believed to be devoid of any known driver gene fusions as exemplified by the discovery of previously undescribed ETS gene fusions in 2 clinically localized prostate tumor samples that lacked known driver gene fusions.

By analyzing the transcriptome at unprecedented depth, we have revealed numerous gene fusions, demonstrating the prevalence of a relatively under-represented class of mutations. However, one of the major goals remains to discover recurrent gene fusions and to distinguish them from secondary, nonspecific chimeras. Although quantifying expression levels is not proof of whether a gene fusion is a driver or passenger, because a low-level gene fusion could still be causative, it still of major significance that a paired-end strategy clearly distinguished known high-level driving gene fusions, such as *BCR-ABL1* and *TMPRSS2-ERG*, from potential lower level passenger chimeras. Overall, these fusions serve as a model for employing a paired-end nomination strategy for prioritizing leads likely to be high-level driving gene fusions, which would subsequently undergo further functional and experimental evaluation.

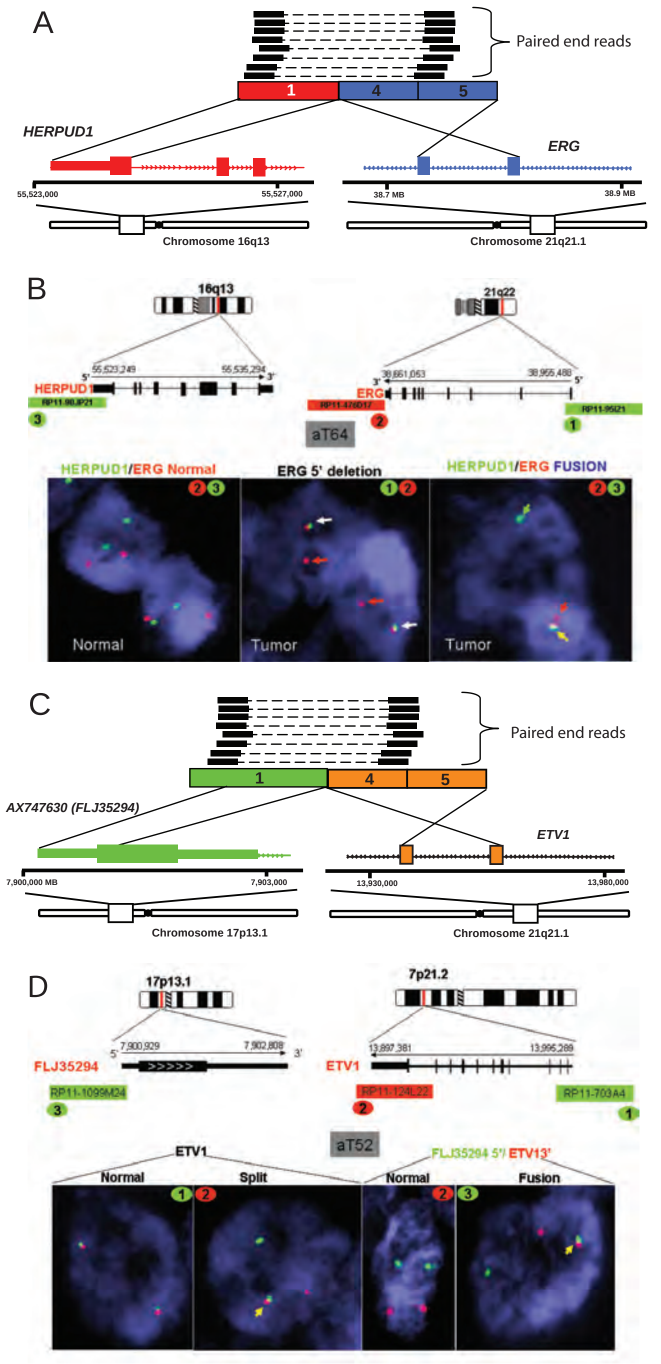


**Fig. 3.** Discovery of previously undescribed ETS gene fusions in localized prostate cancer. (*A*) Schematic representation of the interchromosomal gene fusion between exon 1 of *HERPUD1* (red), residing on chromosome 16, with exon 4 of *ERG* (blue), located on chromosome 21. (*B*) Schematic representation showing genomic organization of *HERPUD1* and *ERG* genes. Horizontal red and green bars indicate the location of BAC clones. (*Lower*) FISH analysis using BAC clones showing *HERPUD1* and *ERG* in a normal tissue (*Left*), deletion of the ERG 5′ region in tumor (*Center*), and *HERPUD1-ERG* fusion in a tumor sample (*Right*). (*C*) Schematic representation of the interchromosomal gene fusion between *FLJ35294* (green), residing on chromosome 17, with exon 4 of *ETV1* (orange) located on chromosome 21. (*D Upper*) Schematic representation of the genomic organization of *FLJ35294* and *ETV1* genes. (*Lower*) FISH analysis using BAC clones showing split of *ETV1* in tumor sample (*Left*) and the colocalization of *FLJ35294* and *ETV1* in a tumor sample (*Right*).

One of the major advantages of using a transcriptome approach is that it enables us to identify rearrangements that are not detectable at the DNA level. For example, conventional cytogenetic methods would miss gene fusions produced by paracentric inversions, or sub microscopic events, such as *GAS6-RASA3*. Also, transcriptome sequencing can unveil RNA chimeras, lacking DNA aberrations, as demonstrated by the discovery of a recurrent, prostate specific, read-through of *SLC45A3* with *ELK4* in prostate cancers. Further classification of RNA based events using paired-end sequencing revealed numerous broadly expressed chimeras between adjacent genes. Although these events were not necessarily read-throughs events, because they typically had different orientations, we believe they represent extensions of transcriptional units beyond their annotated boundaries. Unlike single read based approaches, which require chimeras to span exon boundaries of independent genes, we were able to detect these events using paired-end sequencing, which could have significant impact for improving how we annotate transcriptional units.

Overall, we have demonstrated the advantages of employing a paired-end transcriptome strategy for chimera discovery, established a methodology for mining chimeras, and extensively catalogued chimeras in a prostate and hematological cancer models. We believe that the sensitivity of this approach will be of broad impact and significance for revealing novel causative gene fusions in various cancers while revealing additional private gene fusions that may contribute to tumorigenesis or cooperate with driver gene fusions.

## Methods

**Paired-End Gene Fusion Discovery Pipeline.** Mate pair transcriptome reads were mapped to the human genome (hg18) and Refseq transcripts, allowing up to 2 mismatches, using Efficient Alignment of Nucleotide Databases (ELAND) pair within the Illumina Genome Analyzer Pipeline software. Illumina export output files were parsed to categorize passing filter mate pairs as (*i*) mapping to the same transcript, (*ii*) ribosomal, (*iii*) mitochondrial, (*iv*) quality control, (*v*) chimera candidates, and (*vi*) nonmapping. Chimera candidates and nonmapping categories were used for gene fusion discovery. For the chimera candidates category, the following criteria were used: (*i*) mate pairs must be of high mapping quality (best unique match across genome), (*ii*) best unique mate pairs do not have a more logical alternative combination (i.e., best mate pairs suggest an interchromosomal rearrangement, whereas the second best mapping for a mate reveals the pair have a alignment within the expected insert size), (*iii*) the sum of the distances between the most 5′ and 3′ mate on both partners of the gene fusion must be <500 nt, and (*iv*) mate pairs supporting a chimera must be nonredundant.

In addition to mining mate pairs encompassing a fusion boundary, the nonmapping category was mined for mate pairs that had 1 read mapping to a gene, whereas its corresponding read fails to align, because it spans the fusion boundary. First, the annotated transcript that the ''mapping'' mate pair aligned against was extracted, because this transcript represents one of the potential partners involved in the gene fusion. The ''nonmapping'' mate pair was then aligned against all of the exon boundaries of the known gene partner to identify a perfect partial alignment. A partial alignment confirms that the nonmapping mate pair maps to our expected gene partner while revealing the portion of the nonmapping mate pair, or overhang, aligning to the unknown partner. The overhang is then aligned against the exon boundaries of all known transcripts to identify the fusion partner. This process is done using a Perl script that extracts all possible University of California Santa Cruz (UCSC) and Refseq exon boundaries looking for a single perfect best hit.

Mate pairs spanning the fusion boundary are merged with mate pairs encompassing the fusion boundary. At least 2 independent mate pairs are required to support a chimera nomination, which can be achieved by (*i*) 2 or more nonredundant mate pairs spanning the fusion boundary, (*ii*) 2 or more nonredundant mate pairs encompassing a fusion boundary, or (*iii*) 1 or more mate pairs encompassing a fusion boundary and 1 or more mate pairs spanning the fusion boundary. All chimera nominations were normalized based on the cumulative number of mate pairs encompassing or spanning the fusion junction per million mate pairs passing filter.

**RNA Chimera Analysis.** Chimeras found from UHR, HBR, VCaP, and K562 were grouped based on whether they showed expression in all samples, ''broadly expressed,'' or a single sample, ''restricted expression.'' Because UHR is comprised of K562, chimeras found in only these 2 samples were also considered as restricted. Heatmap visualization was conducted by using TIGR's MultiExperiment Viewer (TMeV) version 4.0 (www.tm4.org).

**Additional Details.** Additional details can be found in *SI Text*.

**ACKNOWLEDGMENTS.** We thank Lu Zhang, Eric Vermaas, Victor Quijano, and Juying Yan for assistance with sequencing, Shawn Baker and Steffen Durinck for helpful discussions, Rohit Mehra and Javed Siddiqui for collecting tissue samples, and Bo Han and Kalpana Ramnarayanan for technical assistance. C.A.M. was supported by a National Institutes of Health (NIH) Ruth L. Kirschstein postdoctoral training grant, and currently derives support from the American Association of Cancer Research Amgen Fellowship in Clinical/Translational Research and the Canary Foundation and American Cancer Society Early Detection Postdoctoral Fellowship. J.Y. was supported by NIH Grant 1K99CA129565-01A1 and Department of Defense (DOD) Grant PC080665. A.M.C. was supported in part by the NIH (Prostate SPORE P50CA69568, R01 R01CA132874), the DOD (BC075023, W81XWH-08-0110), the Early Detection Research Network (U01 CA111275), a Burroughs Welcome Foundation Award in Clinical Translational Research, a Doris Duke Charitable Foundation Distinguished Clinical Investigator Award, and the Howard Hughes Medical Institute. This work was also supported by National Center for Integrative Biomedical Informatics Grant U54 DA021519.

1. Futreal PA, et al. (2004) A census of human cancer genes. *Nat Rev* 4:177–183.
2. Kumar-Sinha C, Tomlins SA, Chinnaiyan AM (2008) Recurrent gene fusions in prostate cancer. *Nat Rev* 8:497–511.
3. Mitelman F, Johansson B, Mertens F (2004) Fusion genes and rearranged genes as a linear function of chromosome aberrations in cancer. *Nat Genet* 36:331–334.
4. Mitelman F, Mertens F, Johansson B (2005) Prevalence estimates of recurrent balanced cytogenetic aberrations and gene fusions in unselected patients with neoplastic disorders. *Gene Chromosome Canc* 43:350–366.
5. Tomlins SA, et al. (2007) Distinct classes of chromosomal rearrangements create oncogenic ETS gene fusions in prostate cancer. *Nature* 448:595–599.
6. Tomlins SA, et al. (2005) Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science* 310:644–648.
7. Soda M, et al. (2007) Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer. *Nature* 448:561–566.
8. Druker BJ, et al. (2006) Five-year follow-up of patients receiving imatinib for chronic myeloid leukemia. *New Engl J Med* 355:2408–2417.
9. Druker BJ, et al. (1996) Effects of a selective inhibitor of the Abl tyrosine kinase on the growth of Bcr-Abl positive cells. *Nat Med* 2:561–566.
10. Kantarjian H, et al. (2002) Hematologic and cytogenetic responses to imatinib mesylate in chronic myelogenous leukemia. *New Engl J Med* 346:645–652.
11. Volik S, et al. (2003) End-sequence profiling: Sequence-based analysis of aberrant genomes. *Proc Natl Acad Sci USA* 100:7696–7701.
12. Tuzun E, et al. (2005) Fine-scale structural variation of the human genome. *Nat Genet* 37:727–732.
13. Ruan Y, et al. (2007) Fusion transcripts and transcribed retrotransposed loci discovered through comprehensive transcriptome analysis using Paired-End diTags (PETs). *Genome Res* 17:828–838.
14. Campbell PJ, et al. (2008) Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat Genet* 40:722–729.
15. Maher CA, et al. (2009) Transcriptome sequencing to detect gene fusions in cancer. *Nature* 458:97–101.
16. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y (2008) RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res* 18:1509–1517.
17. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* 5:621–628.
18. Shtivelman E, Lifshitz B, Gale RP, Canaani E (1985) Fused transcript of abl and bcr genes in chronic myelogenous leukaemia. *Nature* 315:550–554.
19. Barlund M, et al. (2002) Cloning of BCAS3 (17q23) and BCAS4 (20q13) genes that undergo amplification, overexpression, and fusion in breast cancer. *Gene Chromosome Canc* 35:311–317.
20. Hampton OA, et al. (2009) A sequence-level map of chromosomal breakpoints in the MCF-7 breast cancer cell line yields insights into the evolution of a cancer genome. *Genome Res* 19:167–177.
21. Zhao Q, et al. (2009) Transcriptome-guided characterization of genomic rearrangements in a breast cancer cell line. *Proc Natl Acad Sci USA* 106:1886–1891.
22. Hahn Y, et al. (2004) Finding fusion genes resulting from chromosome rearrangement by analyzing the expressed sequence databases. *Proc Natl Acad Sci USA* 101:13257–13261.
23. Shadeo A, Lam WL (2006) Comprehensive copy number profiles of breast cancer cell model genomes. *Breast Cancer Res* 8:R9.
24. Huang J, et al. (2004) Whole genome DNA copy number changes identified by high density oligonucleotide arrays. *Hum Genom* 1:287–299.
25. Neve RM, et al. (2006) A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. *Cancer Cell* 10:515–527.
26. Volik S, et al. (2006) Decoding the fine-scale structure of a breast cancer genome and transcriptome. *Genome Res* 16:394–404.
27. Han B, et al. (2008) A fluorescence in situ hybridization screen for E26 transformation-specific aberrations: Identification of DDX5-ETV4 fusion protein in prostate cancer. *Cancer Res* 68:7629–7637.

# Deep sequencing reveals distinct patterns of DNA methylation in prostate cancer

Jung H. Kim,[1,11] Saravana M. Dhanasekaran,[1,2,11] John R. Prensner,[1] Xuhong Cao,[1] Daniel Robinson,[1] Shanker Kalyana-Sundaram,[1,3] Christina Huang,[1] Sunita Shankar,[1] Xiaojun Jing,[1] Matthew Iyer,[1] Ming Hu,[4,12] Lee Sam,[1,2] Catherine Grasso,[1] Christopher A. Maher,[1,2,5] Nallasivam Palanisamy,[1] Rohit Mehra,[1] Hal D. Kominsky,[1] Javed Siddiqui,[1] Jindan Yu,[6] Zhaohui S. Qin,[7] and Arul M. Chinnaiyan[1,2,5,8,9,10,13]

*[1] Michigan Center for Translational Pathology, University of Michigan Medical School, Ann Arbor, Michigan 48109, USA; [2] Department of Pathology, University of Michigan Medical School, Ann Arbor, Michigan 48109, USA; [3] Department of Environmental Biotechnology, Bharathidasan University, Tiruchirappalli 620 024, India; [4] Department of Biostatistics, University of Michigan Medical School, Ann Arbor, Michigan 48109, USA; [5] Center for Computational Medicine and Bioinformatics, University of Michigan Medical School, Ann Arbor, Michigan 48109, USA; [6] Robert H. Lurie Comprehensive Cancer Center, Northwestern University Feinberg School of Medicine, Chicago, Illinois 60611, USA; [7] Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University, Atlanta, Georgia 30322, USA; [8] Department of Urology, University of Michigan Medical School, Ann Arbor, Michigan 48109, USA; [9] Comprehensive Cancer Center, University of Michigan Medical School, Ann Arbor, Michigan 48109, USA; [10] Howard Hughes Medical Institute, University of Michigan Medical School, Ann Arbor, Michigan 48109, USA*

Beginning with precursor lesions, aberrant DNA methylation marks the entire spectrum of prostate cancer progression. We mapped the global DNA methylation patterns in select prostate tissues and cell lines using MethylPlex–next-generation sequencing (M-NGS). Hidden Markov model–based next-generation sequence analysis identified ~68,000 methylated regions per sample. While global CpG island (CGI) methylation was not differential between benign adjacent and cancer samples, overall promoter CGI methylation significantly increased from ~12.6% in benign samples to 19.3% and 21.8% in localized and metastatic cancer tissues, respectively ($P$-value < $2 \times 10^{-16}$). We found distinct patterns of promoter methylation around transcription start sites, where methylation occurred not only on the CGIs, but also on flanking regions and CGI sparse promoters. Among the 6691 methylated promoters in prostate tissues, 2481 differentially methylated regions (DMRs) are cancer-specific, including numerous novel DMRs. A novel cancer-specific DMR in the *WFDC2* promoter showed frequent methylation in cancer (17/22 tissues, 6/6 cell lines), but not in the benign tissues (0/10) and normal PrEC cells. Integration of LNCaP DNA methylation and H3K4me3 data suggested an epigenetic mechanism for alternate transcription start site utilization, and these modifications segregated into distinct regions when present on the same promoter. Finally, we observed differences in repeat element methylation, particularly LINE-1, between *ERG* gene fusion-positive and -negative cancers, and we confirmed this observation using pyrosequencing on a tissue panel. This comprehensive methylome map will further our understanding of epigenetic regulation in prostate cancer progression.

[Supplemental material is available for this article. The next-generation sequencing and microarray data from this study have been submitted to the NCBI Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/) under accession nos. GSE29155 and GSE27619, respectively.]

CpG residues, the targets of DNA methylation, have an asymmetric distribution in mammalian genomes and are often found in small clusters termed CpG islands (CGIs) (Bird 2002). Approximately 60% of all human gene promoters overlap with CGIs (Illingworth and Bird 2009), and accumulation of promoter DNA methylation is associated with gene silencing (Jones and Baylin 2007). Previously, DNA methylation studies in prostate cancer have used methodologies of variable scale, focusing on either a few promoters (Li et al. 2005) or several thousand genomic regions with a CpG island array (Kron et al. 2009). Alternatively, functional approaches that monitored gene expression changes after treatment with the demethylating agent 5-aza-2′-deoxycytidine (5-Aza) have also been used (Yegnasubramanian et al. 2004, 2008). However, to date, only 115 genes are reported as methylation targets in prostate cancer, 85 of which are listed in the Pubmeth database (http://www.pubmeth.org) (Ongenaert et al. 2008).

The advent of next-generation sequencing (NGS) now presents a novel approach to assess genome-wide epigenetic changes without the limitations of probe-based microarray platforms. MethylC-seq, a bisulfite conversion approach, was previously used to analyze the methylome at single-base resolution for *Arabidopsis* (Cokus et al. 2008) and recently for human H1 embryonic stem cells and fetal lung fibroblasts (Harris et al. 2010). Meissner et al. (2008) produced methylation maps by reduced representation bisulfite sequencing of MspI-digested, genomic

**[11] These authors contributed equally to this work.**
**[12] Present address: Department of Statistics, Harvard University, Cambridge, MA 02138, USA.**
**[13] Corresponding author.**
**E-mail arul@umich.edu.**

Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.119347.110.

DNA from pluripotent and differentiated cells, and the same method was used by Gu et al. (2010) more recently on colon cancer samples. Alternatively, several groups enriched methylated fragments based on their affinities to anti-5′-methylcytosine antibody (Weber et al. 2005; Down et al. 2008; Feber et al. 2011) and methylated DNA binding protein Mbd2b (Rauch and Pfeifer 2009), susceptibility to methylation-sensitive restriction enzymes (Brunner et al. 2009), or capture technology (Weber et al. 2005; Hodges et al. 2009) before sequencing.

While the enrichment-based and bisulfite conversion methods identified largely comparable methylation events, variation was observed in CpG coverage, resolution, quantitative accuracy, and other measures (Bock et al. 2010; Harris et al. 2010). We used a novel MethylPlex technology described here to enrich methylated regions present in genomic DNA from LNCaP prostate cancers, normal prostate epithelial cells (PrEC), and clinical prostate specimens ($n$ = 17). Massively parallel sequencing of the enriched product identified differentially methylated regions (DMRs) and revealed novel insights regarding the genomic placement and functional consequences of DNA methylation in cancer.

## Results

### Characterization of DNA methylation by M-NGS in prostate cells

To perform a genome-wide analysis of DNA methylation in prostate cancer, we used MethylPlex–next generation sequencing (M-NGS) methodology, which enriches methylated DNA using restriction enzymes and requires minimal input genomic DNA (i.e., 50 ng). The ability of M-NGS to identify methylated genomic regions was first evaluated in a prostate cancer cell line, LNCaP, and normal PrEC cells. A schematic describing sequencing library generation is provided in Supplemental Figure 1. Briefly, MethylPlex libraries were constructed by digesting input genomic DNA isolated from samples with a cocktail of methylation-sensitive restriction enzymes, followed by ligation of adaptors containing universal primers sequences and PCR-based amplification. A second round of enzymatic treatment depleted non-GC-rich sequences, followed by an additional amplification step to ensure enrichment of highly methylated DNA fragments. The amplification adaptors were enzymatically removed prior to NGS library preparation (Supplemental Fig. 1). The MethylPlex libraries described above were constructed through the commercial service provided by Rubicon Genomics Inc.

For initial standardization, we used two different concentrations (1 and 5 μg) of each MethylPlex sample from LNCaP and PrEC cells as input DNA to obtain single-read sequencing on the Illumina Genome Analyzer II (for protocol details, see Methods). For each cell type (LNCaP and PrEC), a total of four sequencing libraries were prepared corresponding to 200- and 400-bp size selections of 1 μg and 5 μg of MethylPlex product. We obtained an average of 5 million mappable reads for each M-NGS sample (Supplemental Table 1). CG dinucleotides were enriched by the MethylPlex procedure up to threefold in mapped reads from M-NGS compared to previously obtained control ChIP-sequencing data, namely, pan-histone ChIP-seq (Supplemental Table 1; Yu et al. 2010).

To demonstrate experimental consistency, a comparative analysis of data from 1 and 5 μg of MethylPlex DNA exhibited high correlation both for reads mapping to chromosome 21 and for reads mapping to all CpG islands (Supplemental Fig. 2). Data from 400 bp–5 μg were most enriched for CG-rich sequences (Supplemental Table 1) and showed maximum overlap (~70%) with methylation identified by hybridizing the MethylPlex product to a CpG island array (Supplemental Fig. 3A; Supplemental Table 2). We therefore selected these data for further analysis.

A hidden Markov model (HMM)–based algorithm previously used for ChIP-seq data analysis (Qin et al. 2010) was used to locate peaks from mapped reads obtained in each sequencing run (Supplemental Table 1). We found a 70% overlap in methylated genomic regions between LNCaP (56,727 regions) and PrEC (61,615 regions) cells (Fig. 1A). Methylation located in intergenic and intronic regions of the genomes analyzed had a similar distribution (Fig. 1B); additionally, in LNCaP cells, we also used MeDIP-seq, a methodology that uses 5′-methylcytosine antibody to enrich methylated regions, and we identified approximately 68,000 methylated regions in this cell line, which was comparable to the M-NGS results. Moreover, there was an overall 62% concordance between all the genomic regions (data not shown) and >83% in CGIs identified by M-NGS and MeDIP-seq, thereby validating the two methodologies (Supplemental Fig. 3B).

The cancer-derived LNCaP cells displayed frequent methylation among the 56 previously reported methylated promoter regions in prostate cancer tissues (36/56 in LNCaP M-NGS and 40/56 in LNCaP MeDIP-seq) compared to PrEC cells (7/56 in PrEC M-NGS) (Supplemental Table 10). However, this difference was absent when we examined the promoters and gene body of known imprinted genes (24/29 in PrEC M-NGS, 23/29 in LNCaP M-NGS, and 26/29 in LNCaP MeDIP-seq) (Supplemental Table 10; Morison et al. 2005).

### Global differences in CGI methylation

Because hypermethylation in CpG-rich promoters is a common feature of tumorigenesis (Issa 2004), we compared the extent of CpG island methylation between LNCaP and PrEC cells. Of the 68,508 (72.74 Mb) CpG islands identified using Takai Jones criteria (Takai and Jones 2002) in the human genome, 6865 (7.6 Mb) and 5767 (6.1 Mb) CpG islands were methylated in LNCaP and PrEC, respectively. Globally, we observed a 1.7-fold increase in uniquely methylated CpG islands between LNCaP and PrEC, and this ratio increased to approximately sevenfold specifically in CpG islands associated within gene promoters but not among CGIs located elsewhere (Fig. 1C). In LNCaP cells, methylation in >88% of CpG islands located within promoters and 83% of CGIs in non-promoters detected by M-NGS were corroborated by the MeDIP-seq data (Supplemental Fig. 3B).

Aberrant promoter methylation is thought to contribute to tumorigenesis by repressing transcription of tumor-suppressor genes (Jones and Baylin 2007). We next looked for methylation on RefSeq gene promoters (±1500 bp flanking the transcription start site) and identified 3496 that were methylated in at least one sample (Supplemental Fig. 4). Visualization of these methylation marks in the context of promoter CGIs revealed the presence of several distinct methylation patterns on gene promoters (Supplemental Fig. 4). Broadly, the promoters fell into two groups based on the presence or absence of a CpG island within this specified region. Interestingly, although 35% of promoters ($n$ = 1232) lacked CpG islands, they exhibited methylation around the transcription start site (TSS) (Supplemental Fig. 4; Supplemental Table 3). The remaining 65% ($n$ = 2264) had CpG islands spanning the TSS, and three distinct methylation patterns were observed in this group: (1) Methylation was mostly confined (39.6%, $n$ = 1383) to the

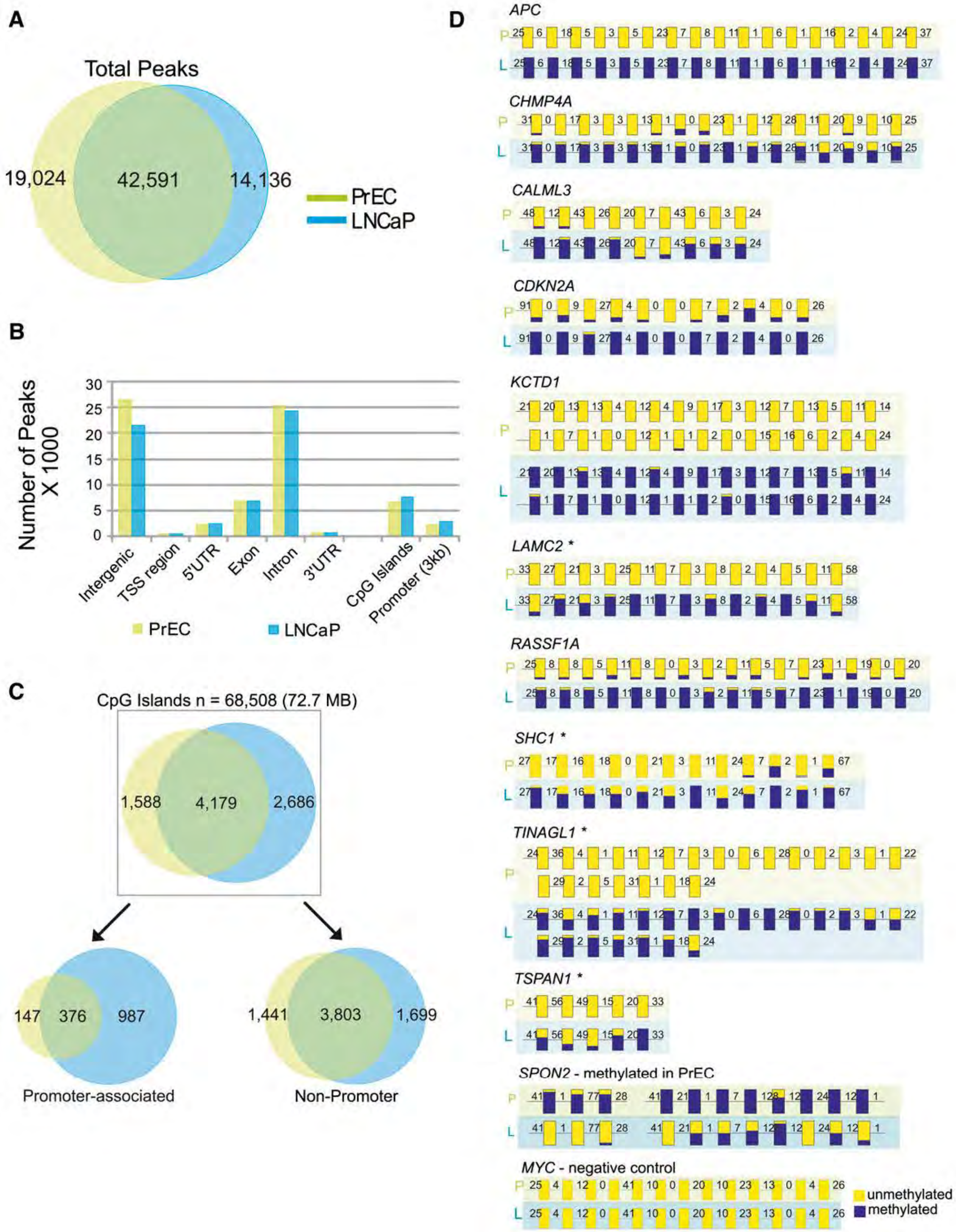


**Figure 1.** Characterization of genome-wide methylation patterns in prostate cells by M-NGS. (*A*) The Venn diagram represents a 70% overlap between the regions methylated in LNCaP (blue) and PrEC (green) cells. (*B*) In LNCaP (blue) and PrEC (green), the majority of DNA methylation occurred in intergenic and intronic regions, and the genomic distribution of methylation peaks was similar. (*C*) Promoter-associated CpG islands displayed a sevenfold difference in methylation between LNCaP (blue) and PrEC (green) cells. (*D*) DNA methylation in *APC*, *CHMP4A*, *CALML3*, *CDKN2A*, *KCTD1*, *LAMC2*, *RASSF1*, *SHC1*, *TINAGL1*, and *TSPAN1* gene promoters in LNCaP (L) cells. *SPON2* in PrEC (P) cells and a negative control region in *MYC* were validated by bisulfite sequencing. The methylation status of each CG residue from 10 clones sequenced on both strands was analyzed using the BIQ Analyzer (Bock et al. 2005) program, where the height of the blue bar indicates the percent methylation at a given position, yellow indicates no methylation, and the numbers indicate the distance between analyzed CG dinucleotides. *CpG islands were absent in these promoters. Additional details are provided in Supplemental Figures 5–7 and Supplemental Table 5. Validation of additional candidates, including *NAP1L5*, *C9orf125*, *AOX1*, *AMT*, *NTN4*, and *PPP1R3C*, are presented in Supplemental Figure 8.

island, and interestingly with much higher frequency (greater than sixfold difference) in LNCaP ($n = 952$) compared to PrEC ($n = 147$) cells (Supplemental Fig. 4); (2) methylation was positioned 5′ to the CpG island (11.8%, $n = 412$); and (3) methylation was positioned 3′ to the CpG island (13.4%, $n = 469$). In total, methylation flanking 5′ or 3′ of promoter CpG islands accounted for 25.2% of all methylation observed ($n = 881$). To explore the role of these methylation patterns in prostate cancer pathogenesis, we identified 812 out of 1171 unique gene promoters to be methylated only in LNCaP (Supplemental Table 4) and were considered for further analysis. The remaining 359 promoters were methylated in both LNCaP and PrEC cells.

## Validation of DMRs

We next selected 18 regions based on M-NGS data and validated their methylation status using a standard bisulfite sequencing technique in LNCaP and PrEC cells. This included 15 DMRs in LNCaP (*RASSF1*, *KCTD1*, *CHMP4A*, *APC*, *CDKN2A*, *SHC1*, *LAMC2*, *TSPAN1*, *CALML3*, *AOX1*, *AMT*, *C9orf125*, and *TINAGL1*), one gene in PrEC cells (*SPON2*), one region methylated in both LNCaP and PrEC cells (*NAP1L5*), and a control *MYC* promoter region that was unmethylated in both cell types. The UCSC Genome Browser view of methylation in the two samples by M-NGS and methylation in LNCaP by MeDIP-seq, along with gene schematic, primer sequences, and bisulfite sequence amplicon locations, are presented in Supplemental Figures 5–7 and Supplemental Table 5. Notably, the results for all 18 regions confirmed the data generated by M-NGS (Fig. 1D; Supplemental Fig. 8).

In addition, we observed overexpression of a significant number of LNCaP methylated genes following 5-Aza treatment of cells in a functional validation strategy using gene expression arrays. A total of 973 out of 1171 methylated genes in LNCaP were present in gene expression array data. Significance Analysis of Microarray (SAM) results showed up-regulation of 246 out of 973 methylated genes at a 5% false discovery rate (Supplemental Fig. 9; Supplemental Table 6), supporting epigenetic regulation of these genes.

To identify molecular concepts enriched in our DMRs, we analyzed our data set using the Molecular Concept Map (MCM) analysis derived from the Oncomine database (Rhodes et al. 2007a; Tomlins et al. 2007b). MCM analysis of 789 out of 813 genes methylated only in LNCaP that mapped to the Oncomine database (Supplemental Fig. 10; Supplemental Tables 4, 7) revealed preferential enrichment with underexpressed gene signatures from localized and metastatic PCa samples (lowest $P$-value < $1.90 \times 10^{-14}$) from several independent studies. Furthermore, the signatures, "genes previously known to be methylated in prostate cancer" ($P$-value < $1.40 \times 10^{-6}$) (Ongenaert et al. 2008), and "Gene Ontology-tumor suppressor genes" ($P$-value < 0.009) were significantly enriched (Supplemental Fig. 10A; Supplemental Table 7). In contrast, PrEC cells did not share this enrichment, and MCM analysis of PrEC-only methylated regions revealed only concepts pertaining to histone modifications and that were common to both PrEC and LNCaP MCM analysis (Supplemental Fig. 10B). Finally, integration with RNA-seq data revealed an association between gene repression and promoter methylation, globally by Gene Set Enrichment Analysis (GSEA) (Supplemental Fig. 11) and upon specific evaluation of select genes (Supplemental Fig. 12). For example, *TIG1*, *GSTP1*, *CALML3*, *TASCTD2*, and *KCTD1* were methylated and repressed specifically in LNCaP, compared to *SPON2* and *GAGE* genes, which were methylated and repressed only in PrEC cells. *HIC1* showed basal transcript expression and was methylated in both cell types.

## Characterization of DNA methylation in prostate cancer tissues

Having established the robustness of M-NGS to identify highly methylated regions in cell line models, we next characterized 17 prostate tissues (six benign adjacent, two normal, five localized prostate cancer, and four metastatic prostate cancer specimens) (Supplemental Table 8). A genome-wide assessment of both benign adjacent and cancer tissues showed a similar number of methylation events within intergenic and intronic regions (Fig. 2A). Of the total 68,508 CGIs present genome-wide, 18.5%, 19.7%, and 20.2% of all CGIs were methylated in benign, localized, and metastatic cancer samples, respectively (Fig. 2B). Importantly, a significant increase in promoter-associated CGI methylation (Pearson's $\chi^2$ test, $P$-value < $2 \times 10^{-16}$) paralleled prostate cancer progression (benign 12.6%, localized PCa 19.3%, and metastatic PCa 21.8%),

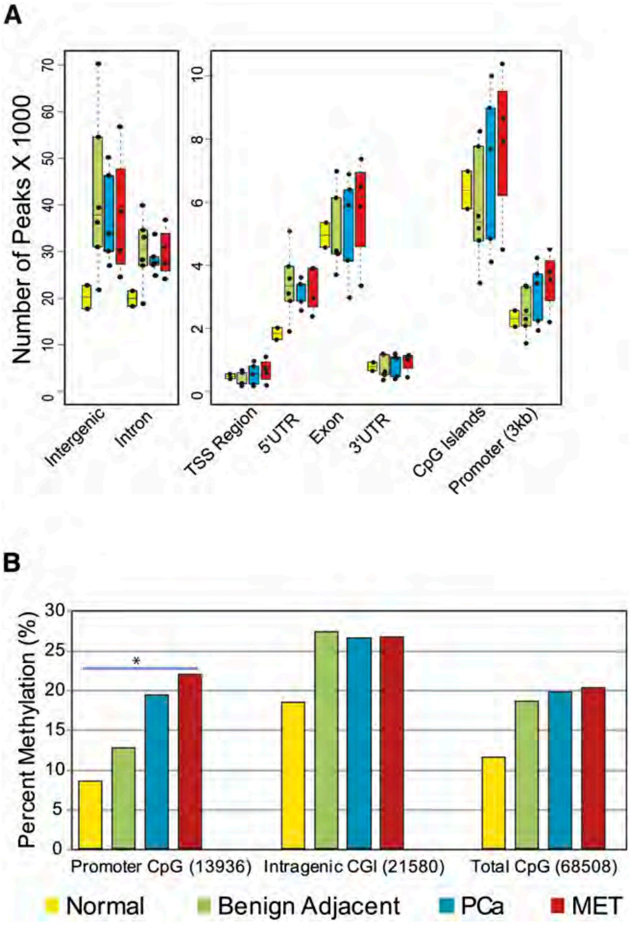


**Figure 2.** DNA methylation pattern in prostate tissues. (*A*) Genome-wide distribution of DNA methylation in various prostate sample groups analyzed. The majority of methylation peaks are confined to intergenic and intronic regions similar to cell lines. (Yellow) Normal prostate; (green) benign adjacent; (blue) localized PCa; (red) metastatic PCa. (*B*) A gradual increase in percent methylation, with cancer progression among promoter CGIs compared with CGIs located in other genomic regions, was observed. (*) Pearson's $\chi^2$ test, $P$-value < $2 \times 10^{-16}$.

whereas methylation of intragenic CGIs remained essentially unchanged (~26.5%) among the three groups (Fig. 2B).

Next, we identified 6619 promoter methylation events (within ±1500 bp flanking the transcriptional start site) present in either normal, benign adjacent, localized, or metastatic prostate cancer samples (Fig. 3; Supplemental Table 9). Of 6619 total methylation events, 2737 were found in all samples, and 1401 of the remaining 3882 were absent in normal prostate samples and PrEC cells but present in benign adjacent prostates. This left 2481 cancer-specific methylation events that may warrant further characterization (Fig. 3). Nearly all of the 56 previously reported prostate cancer methylated regions from pubmeth.org and a recent study (Kron et al. 2009) showed increased methylation in cancer tissues (Supplemental Table 10).

To identify DMRs with functional significance, we next examined promoter methylation events associated with transcriptional changes. Promoters methylated in cancer were significantly associated with gene repression regardless of whether that promoter contained ($p < 0.001$) or lacked ($p < 0.001$) a CpG island by GSEA, while genes that displayed coding exon methylation tended to be overexpressed ($p < 0.024$) (Fig. 4). Oncomine meta-analysis with a data set of 13 different prostate cancer genes' expression further supported methylated candidates' association with gene repression (Supplemental Table 11). Several previously characterized methylation targets (*GSTM2*, *GSTM1*, *S100A6*, *PYCARD*, and *RARRES1*) were present among this list, thereby validating the approach.

We next used MethylProfiler PCR (Jaspers et al. 2010) as an independent evaluation of the methylation status of a novel target region in *WFDC2* (WAP four-disulphide core domain protein 2, previously called HE4), the recently reported prostate methylation target *TACSTD2* (Ibragimova et al. 2010), and the well-characterized *GSTP1*, all identified in this M-NGS study. *WFDC2*, which ranked 25th in Oncomine meta-analysis, was methylated in 100% (6/6) of transformed prostate cell lines and 77% (17/22) of cancer tissues but not in benign tissues or PrEC (Fig. 5A). In addition, *WFDC2* methylation in select samples was independently confirmed by bisulfite sequencing (Supplemental Fig. 13). In comparison, the *TACSTD2* promoter was less frequently methylated, with 21% (5/23) of cancer tissues and 9% (1/11) of benign tissues showing hypermethylation, and prostate cell lines similarly exhibited variable levels of methylation (Fig. 5B). In contrast, the well-characterized *GSTP1* promoter showed frequent methylation in cancer tissues (86%) and in all transformed cell lines (100%), similar to *WFDC2* (Fig. 5C).

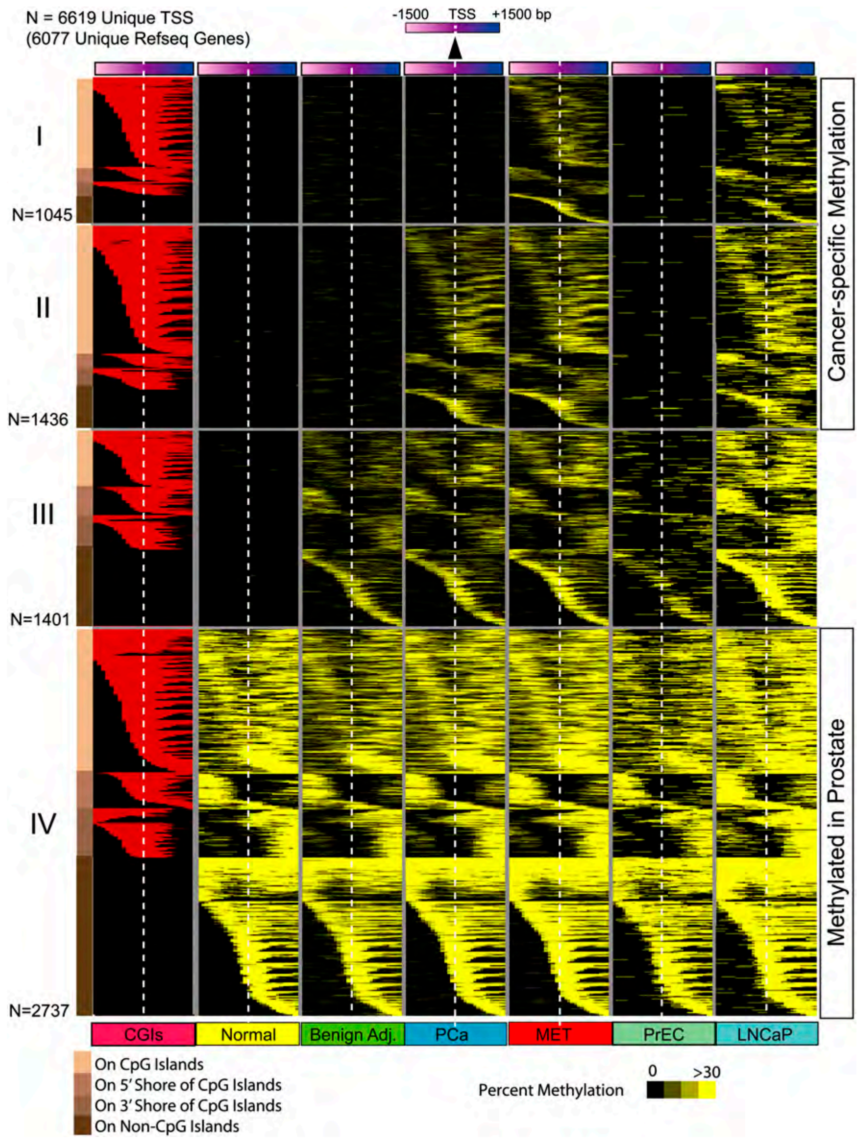


**Figure 3.** Promoter DNA methylation during prostate cancer progression. A total of 6619 gene promoters from 6077 unique RefSeq genes harbored DNA methylation (yellow) among the various sample groups analyzed (normal, benign adjacent, PCa, or MET). Promoter methylation percentage in sample groups is represented by varying shades of yellow. Each row represents a unique promoter region at 100-bp window size, covering ±1500 bp flanking the transcription start site, indicated by the white dotted line. The location of a CpG island (red) in methylated gene promoters is shown in the first column. Promoters in group IV ($n = 2737$) are methylated in all sample groups analyzed, promoters in group III ($n = 1401$) are methylated in all sample groups except normal tissues, while promoters in groups II ($n = 1436$) and I ($n = 1045$) are methylated specifically in cancer samples. Promoters are ordered by the location of methylation on a CpG island, adjacent to the island (shores) or on promoters that lacked CpG islands as represented with different shades of brown on the *left* for groups I to IV. Methylation patterns in prostate cells PrEC and LNCaP are presented alongside for comparison.

## Regulation of transcript variant expression by DNA methylation

We also observed that a subset of genes displayed selective promoter methylation in a transcript isoform-specific manner, suggesting a mechanism for regulating transcript variant expression in cancer. A well-known example, *RASSF1*, frequently

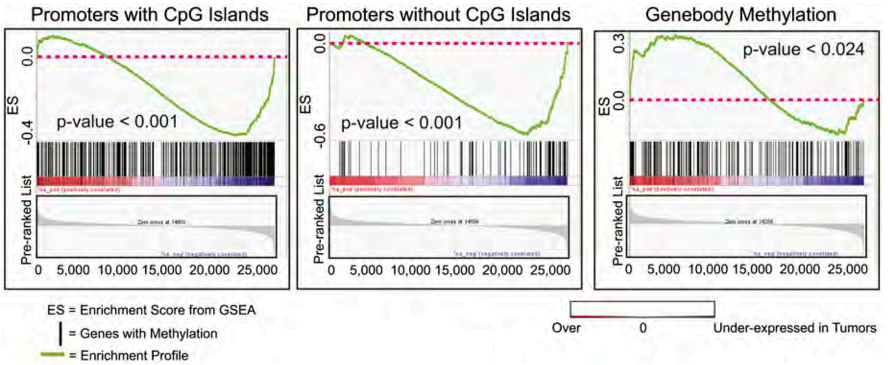


**Figure 4.** Promoter methylation and gene repression. Promoter methylation is associated with gene repression. Gene Set Enrichment Analysis (GSEA) of promoters methylated in prostate cancer was performed on microarray expression data from corresponding samples. Significant correlation was observed between gene repression and promoter methylation among both promoters with ($P$-value < 0.001) or without ($P$-value < 0.001) CpG islands. Overexpressed transcripts were enriched among genes with gene body methylation ($P$-value < 0.024).

inactivated by epigenetic alteration in human cancers (Dammann et al. 2005), is composed of three distinct variants. In LNCaP, we observed DNA methylation–mediated silencing of the longer transcript of *RASSF1*, variant 1, while the smaller isoform, variant 3, that codes for an N-terminal variant protein expressed in multiple cancer cell lines and tissues including PCa, retains high expression (Fig. 6A,B; Dammann et al. 2000; Kuzmin et al. 2002). Active transcription of variant 3 in LNCaP cells is supported by histone 3 lysine 4 trimethylation (H3K4me3) as observed in previously obtained ChIP-seq data (Yu et al. 2010), and 5′ rapid amplification of cDNA ends (5′-RACE) showed the presence of shorter transcripts but not variant 1 in LNCaP (Fig. 6A). Isoform-specific methylation of variant 1 was confirmed by preferential reexpression of this transcript upon 5-Aza treatment of LNCaP cells (Fig. 6B). Interestingly, we found segregation of epigenetic marks into distinct genomic regions in promoters containing CpG islands when we superimposed the promoter methylation and H3K4me3 ChIP-seq data from LNCaP cells (Fig. 7; Supplemental Table 12; Yu et al. 2010). While integration of other epigenetic marks is necessary for a full analysis, these data further suggest that multiple epigenetic modifications may co-occur in distinct patterns to regulate transcript expression in cancer.

Since our M-NGS methodology accurately detected DNA methylation events of *RASSF1*, we queried our data for differential methylation of transcript variants compared to H3K4me3 marks and identified 34 genes in LNCaP that exhibit isoform-specific promoter methylation (Supplemental Table 13). We validated two genes from this list, namely, *NDRG2* and *APC* (Fig. 6D; Supplemental Fig. 14A). In both of these candidates, the transcript variants (variants 1–4 in *NDRG2* and variants 2 and 3 in *APC*) showing DNA methylation were confirmed to be underexpressed in LNCaP cells compared to PrEC cells by qRT-PCR and 5′-RACE (Fig. 6E; Supplemental Fig. 11A). Furthermore, these variants were preferentially reexpressed upon 5-Aza treatment of LNCaP cells. To determine whether patient tissues demonstrated similar isoform-specific expression patterns, we tested *NDRG2* isoforms in two normals, three adjacent normals, five localized PCas, and two metastatic samples by qRT-PCR. Similar to LNCaP cells, variants 1–4 were significantly underexpressed compared to variants 5–8 in localized PCa ($P$-value = 0.034) and adjacent benign prostate ($P$-value = 0.012), but not in normal (non–prostate cancer) tissues (Supplemental Fig. 14B). In addition, previously obtained RNA-seq data from LNCaP cells supported the above observation for *RASSF1* and *NDRG2* genes (Fig. 6C,F).

### Methylation differences between ETS-positive and ETS-negative tissues

Transcription factor occupancy is suggested to have a protective role in limiting the spread of DNA methylation into affected CpG islands (Gebhard et al. 2010). In prostate cancer, gene fusions involving ETS transcription factors (most commonly *ERG* and *ETV1*) occur in ~40%–50% of patients and serve as the most frequent genetic aberration in this disease (Kumar-Sinha et al. 2008). DNA methylation differences between patients harboring or lacking an ETS gene fusion might therefore provide insights into the transcriptional program of *ERG* in prostate cancer. We compared the five ERG fusion-positive (ETS-positive) patients and four fusion-negative (ETS-negative) patients in our cohort and observed more than 40 Mb of DMRs specifically associated with ETS-positive samples. Interestingly, the majority of DMRs in ETS-negative samples were also shared with benign samples (Fig. 8A). ETS-positive samples also contained higher repeat-element methylation compared to ETS-negative samples (Fig. 8B). In particular, assessment of global LINE-1 methylation by an independent pyrosequencing analysis on a prostate tissue cohort ($n$ = 20) revealed a significant decrease in LINE-1 element methylation ($P$-value < 0.0001) in ETS-negative compared to ETS-positive samples (Fig. 8C). These data suggest that previous studies documenting global hypomethylation of LINE-1 elements in prostate cancer may miss subtleties present in different molecular subtypes of this disease.

## Discussion

In this study, we characterized genome-wide methylation patterns in prostate tissues and cell lines using a novel M-NGS methodology. Compared to the bisulfite-based MethylC-seq and enrichment-based MeDIP-seq and MBD-seq, which require microgram quantities of genomic DNA, M-NGS and the reduced representation bisulfite sequencing (Gu et al. 2010) need only nanogram quantities of input DNA and are promising options to characterize clinical samples with limited material availability. Using MeDIP-seq, bisulfite sequencing, and 5-Aza treatments as validation, we demonstrate the accuracy and utility of M-NGS to detect genome-wide methylated regions. A recent study using MeDIP-seq reported methylation in 16% (~4428/27,679) of all CGIs in the human brain (Maunakea et al. 2010), which is comparable to our MeDIP-seq data (20%) and M-NGS data (up to 20%). The high overlap (>83%) in the methylated CGIs identified by MeDIP-seq and M-NGS in LNCaP cells suggests a comparable performance of these two methodologies. However, a comparative analysis similar to those by Bock et al. (2010) and Harris et al. (2010) may further characterize the advantages and limitations of M-NGS compared to other existing technologies.

This study reveals important DMRs and methylation patterns in both intragenic and intergenic regions in prostate cancer. While globally the total number of genomic regions methylated in all samples was comparable, several thousand DMRs appear to be specific to either the benign or cancer samples. Consistent with prior studies, we found an increasing number of promoter CGIs to accumulate DNA methylation and that this phenomenon correlated with target gene repression (Perry et al. 2010).

We noted patterns of promoter methylation spanning the CGI, but also in the 5′ and 3′ regions flanking the CGI. For these latter categories, Irizarry et al. (2009) have used microarrays to demonstrate ~70% methylation in regions up to 2 kb away from CpG islands, which were termed "shores" in colon cancer samples. Methylation also occurred on promoters that lacked CGIs, which may also have functional significance. A previous study by Eckhardt et al. (2006) determined that repression of the Oncostatin (*OSM*) gene occurs by promoter methylation despite the absence of a CGI in the *OSM* promoter region. Hence the promoter DMRs identified here (including promoter CGI/shores methylation and methylation in promoters that lack CGI) will likely regulate the cancer transcriptome.

Using Gene Set Enrichment Analysis (GSEA) of M-NGS and an expression array data set for promoters methylated in cancer, we revealed enrichment for gene repression regardless of whether the promoter contained a CGI. A similar association between promoter methylation and gene repression was found in multiple public expression data sets using Oncomine meta-analysis. This analysis nominated a novel methylation target, *WFDC2*, previously shown to be repressed in prostate cancer, starting with the prostatic intraepithelial neoplasia (PIN) stage (Ashida et al. 2004). Methylprofiler PCR analysis showed *WFDC2* promoter methylation specifically in >77% cancer but not in benign specimens, indicating that *WFDC2* repression is mediated by this epigenetic modification (Fig. 5A). Interestingly, in ovarian cancer, *WFDC2* is up-regulated and serves as a biomarker, suggesting that this gene may have different functions in different malignancies (Schummer et al. 1999; Bouchard et al. 2006). In comparison, Ibragimova et al. (2010) recently reported that 17% of prostate cancers contain *TACSTD2* gene promoter methylation, while we observed methylation of 26% of cancer specimens in our panel (Fig. 5B). Given the probe limitations

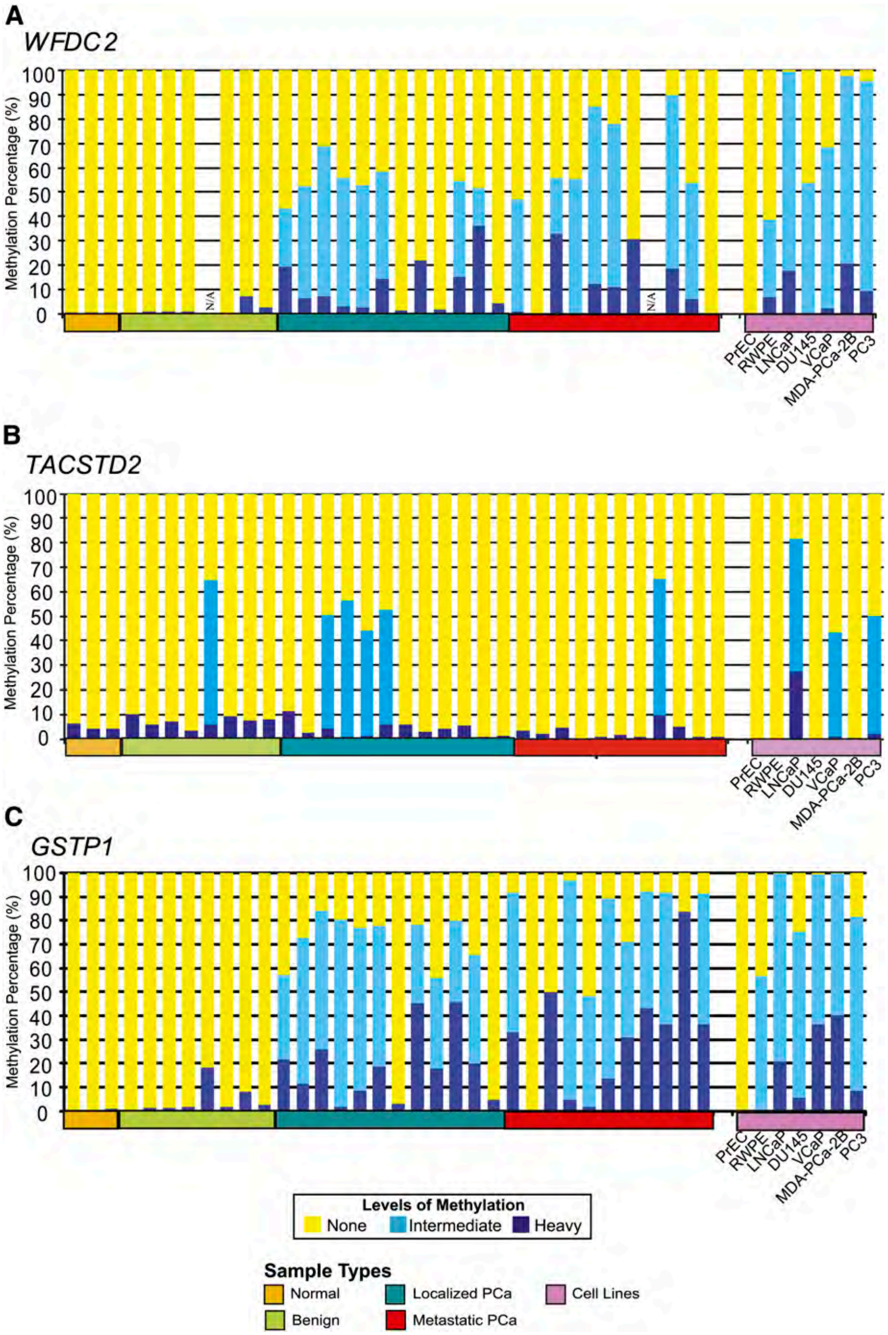


**Figure 5.** *WFDC2*, *TACSTD2* and *GSTP1* methylation in prostate tissue panel. MethylProfiler qPCR was used to determine DNA methylation of the *WFDC2* (*A*), *TACSTD2* and *GSTP1* (*B*) genes. 17/22 prostate cancer tissues and 6/6 transformed prostate cell lines showed methylation of the *WFDC2* promoter, whereas there was no detectable methylation in normal (0/3), benign adjacent tissues (0/7), or the normal PrEC cells. In each sample, the height of the yellow bars indicates no methylation; light blue bars indicate moderate methylation levels; and dark blue bars indicate heavy levels of DNA methylation. Select samples were independently validated by bisulfite sequencing of the corresponding region (Supplemental Fig. 10). (*B*) Methylation of the *TACSTD2* promoter in prostate tissues and cell lines was assessed by MethylProfiler qPCR. Twenty-one percent cancer tissues (5/23) and prostate cancer cell lines, VCaP, LNCaP, and PC3, were methylated. (*C*) Methylprofiler qPCR analysis of *GSTP1*. 20/22 prostate cancer tissues, 1/7 benign adjacent tissues, and 6/6 transformed prostate cell lines showed methylation of the *GSTP1* promoter, whereas there was no detectable methylation in normal tissues (0/3) or the normal PrEC cells.

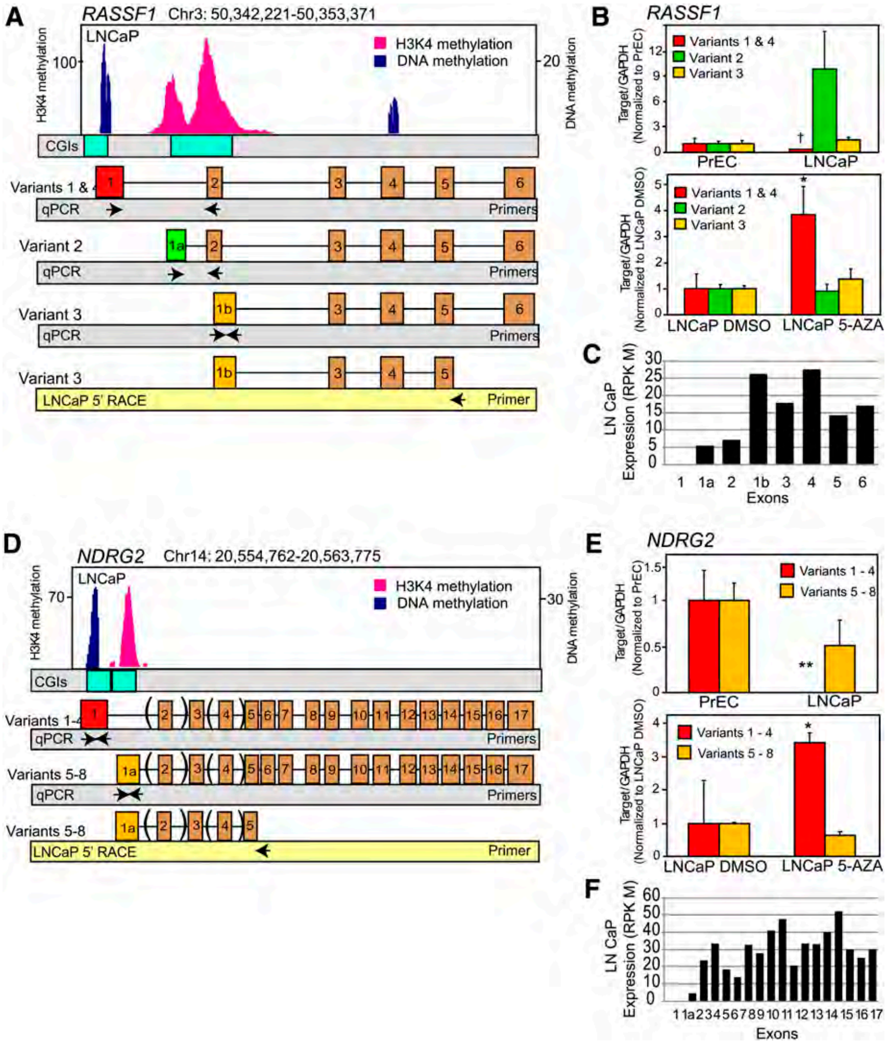


**Figure 6.** Regulation of alternate transcription start site utilization by DNA methylation. (*A*,*D*) Cancer-specific DNA methylation enables switching of alternative transcriptional start sites (TSS) leading to transcript isoform regulation. Next-generation sequencing for DNA methylation and histone 3 lysine 4 trimethylation (H3K4me3) in LNCaP cells reveals genome-wide patterning that couples CpG methylation with H3K4 marks to repress or activate specific transcript variants. Independent epigenetic modifications mark specific alternative TSS. In *RASSF1* (*A*) and *NDRG2* (*D*), CpG methylation occurs at the TSS of the longer variants, with H3K4me3 marks positioned on the TSS of the shorter variants. (*B*,*E*) Preferential silencing and 5-Aza-induced re-expression of CpG-methylated variants in LNCaP cells. Variants exhibiting CpG methylation on their TSSs show preferential silencing compared to variants with H3K4me3 marks in LNCaP cells. These variants show preferential reexpression upon treatment of LNCaP cells with 6 μM 5-Aza for 48 h. qRT-PCR data are normalized to variant expression levels in PrEC prostate primary epithelial cells or DMSO-treated LNCaP cells in the respective panels. (*B*,*D*) 5′-RACE results validated *RASSF1* variant-3 and *NDRG2* variants 5–8 expression in LNCaP cells. (*C*,*F*) Exon expression values from LNCaP RNA-seq data support the corresponding variant transcription of *RASSF1* and *NDRG2* genes.

of the microarray platforms upon which our analyses are based, we expect that an integrative analysis with NGS transcriptome data will expand our understanding of the role of DNA methylation in cancer further. In addition to *WFDC2*, several other novel DMRs (*MAGI2*, *MEIS2*, *NTN4*, *GPRC5B*, *C9orf125*, *FGFR2*, *AOX1*, *VAMP5*, *C14orf159*, *PPP1R3C*, *S100A16*, and *AMT* genes) ranked among the top 30 in the meta-analysis (Supplemental Table 11) and merit further examination. Of particular interest, a recent prostate cancer genome sequencing study revealed inactivating deletions in the PTEN-interacting protein MAGI2 (Berger et al. 2011). We observe a DMR in the MAGI2 promoter, thereby suggesting additional regulatory mechanisms to inactivate this gene and indicating that this gene may shave a role in suppressing prostate cancer progression.

While gene silencing mediated by DNA methylation in prostate cancer has been well described (Nelson et al. 2009), a growing body of evidence now supports a role for epigenetic modification in alternate transcription start site utilization. Regulation of alternate transcription by DNA methylation for the *PIP5KIA* gene in colon cancer (Irizarry et al. 2009) and *PARP12* in human B-cells (Rauch et al. 2009) was identified using microarray experiments. More recently, tissue-specific DNA methylation regulating intragenic promoter activity in the *SHANK3* locus was demonstrated using MeDIP-seq (Maunakea et al. 2010). Our integrative analysis of DNA methylation and H3K4me3 data nominated candidates for alternate transcription start site utilization as demonstrated in *RASSF1*, *NDRG2*, and *APC* genes. Our analysis further demonstrates that, when present on the same promoter, H3K4me3 modifications and DNA methylation have mutually exclusive boundaries. A similar pattern was observed in mouse neural stem cells, where the *GBX2* locus harbors proximal promoter regions containing H3K4me3 marks that are flanked by DNMT3a-bound CpG rich regions containing DNA methylation marks (Wu et al. 2010). Interestingly, H3K4me3, previously considered an active histone mark, is now known to occupy promoters of transcriptionally inactive genes, albeit at threefold lower levels compared to active promoters (Bernstein et al. 2006; Guenther et al. 2007). More recently, binding of CFP1 protein to CpG-rich regions and a 98% overlap between H3K4me3-modified regions and CFP1-binding sites were shown (Thomson et al. 2010). Thus, while the regulation of transcription by DNA methylation and H3K4me3 is well explored, the role for adjacent H3K4me3 and DNA methylation marks in some promoters needs further investigation.

Finally, molecular classification based on ETS gene fusions has enabled subtype-specific analyses of prostate cancer showing distinct copy number aberrations and gene expression patterns in this subtype (Kim et al. 2007; Tomlins et al. 2007b). Here, we define DNA methylation patterns unique to ETS-positive and ETS-negative samples. Specifically, we observed a decrease in repeat-element methylation in ETS-negative compared to ETS-positive samples. Previous work has shown pronounced reduction in both global 5′ methyl cytosine content and LINE-1 hypomethylation in metastatic samples compared to localized PCa (Yegnasubramanian et al. 2008). However, all previous reported assessments of LINE-1

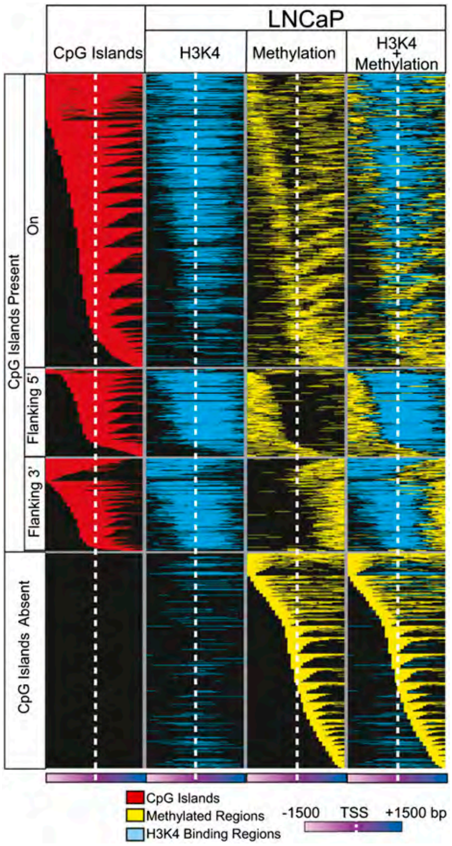


**Figure 7.** Mutually exclusive patterns of promoter DNA methylation and histone H3K4me3 marks in LNCaP cells. Integration of M-NGS DNA methylation data with H3K4me3 ChIP-seq data indicates that DNA methylation and H3K4me3 may be present on the same gene promoter but remain nonoverlapping, adjacent modifications in these promoters. Each row represents a unique promoter region, ±1500 bp flanking the transcription start site (white dotted line) at 100-bp window size. The CpG island location is indicated in red in the first column. The second column represents histone H3K4me3 marks (blue), and the third column (yellow) depicts DNA methylation observed in the corresponding location in LNCaP. Superimposed data are displayed in the fourth column.

or repeat elements have been global measurements that did not distinguish between specific genomic locations contributing to these measurements. In contrast, next-generation sequencing, including M-NGS, is able to resolve individual repeat elements at certain locations in the genome using uniquely mapped reads. Recent identification of methylation changes in repeat elements at specific genomic locations using MeDIP-seq in malignant nerve sheath tumors demonstrated this potential (Feber et al. 2011). On this basis, we find that ETS-negative prostate cancers show decreased levels of LINE-1 methylation when compared to ETS-positive cancers. Subsequent pyrosequencing validation in a prostate tissue cohort confirmed this difference in LINE-1 methylation, thereby corroborating our M-NGS results. While the mechanism of LINE1 hypomethylation in ETS-negative prostate cancers is unknown, it is interesting to note that previous studies identified *TDRD1* (Tudor domain containing protein 1) as a gene that is overexpressed in ERG-positive prostate cancers (Jhavar et al. 2008). Recent data by Reuter et al. (2009) showed derepression of L1 transposons accompanied by a loss of DNA methylation at their 5′ regulatory region in *TDRD1* knockout mice, suggesting that *TDRD1* may have a role in ETS-specific repeat-element methylation.

Aberrant DNA methylation in prostate cancer is believed to occur in two waves, where epigenetic alteration of some genes, such as GSTP1, can be detected in early disease stages, whereas other genes, such as ESR1, are frequently subject to aberrant DNA methylation in metastatic disease and are considered late events (Nelson et al. 2009). Drug therapies aiming to reverse epigenetic changes, especially those found in castration-resistant prostate cancers, are currently being investigated (Perry et al. 2010). By analyzing both localized and metastatic prostate cancer tissues by M-NGS, we have now identified several hundred differentially methylated regions (DMRs), and

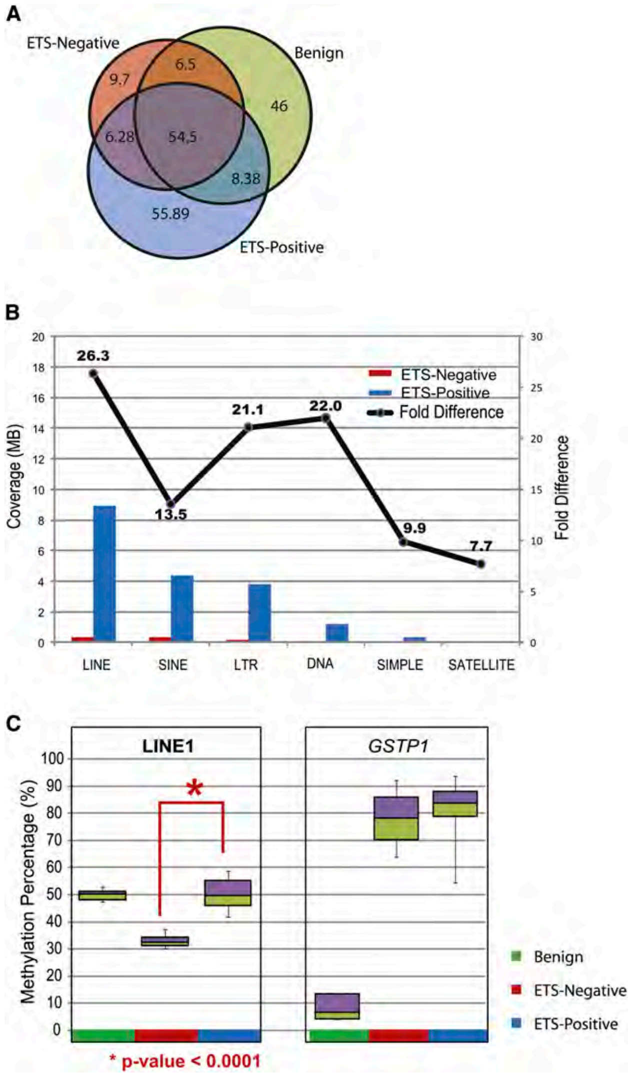


**Figure 8.** Differentially methylated regions between ETS-positive and ETS-negative samples. (*A*) Venn diagram displays the methylation overlap observed between ETS-positive (blue), ETS-negative (red), and benign (green) prostate tissue samples. The inset numbers represent the coverage in each section. (*B*) The coverage for various repeat elements was higher in ETS-positive compared to ETS-negative samples, indicating higher methylation in the former. The fold difference for methylation in each class of repeat element is indicated by the line plot *above*. (*C*) Percent methylation was assessed independently by pyro-sequencing assays for LINE-1 elements and *GSTP1* gene promoter methylation in prostate tissue panel (benign $n$ = 5, ETS-positive cancers $n$ = 10, and ETS-negative cancers $n$ = 4). LINE-1 methylation was significantly lower ($P$-value < 0.0001) in ETS-negative samples compared to ETS-positive tissues, while the *GSTP1* gene promoter was highly methylated in both cancer subgroups and not in benign.

because our sample cohort spans the stages of disease progression, we can identify the specific epigenetic alterations associated with early- and late-phase disease (Yegnasubramanian et al. 2004). Ultimately, this information may be used to elucidate epigenetic diagnostic and prognosis markers; a primary example of this is GSTP1, a gene frequently methylated in prostate cancer that may also be detected from clinical samples obtained in a noninvasive fashion (Nelson et al. 2009). The present study thus provides vital information on genomic locations of cancer-specific DMRs that may now facilitate high-throughput screening analyses for prostate cancer disease markers. Moreover, future genome-wide analyses of DNA methylation may improve with larger patient sample sets and with the incorporation of multiple NGS methodologies, such as MeDIP-seq and others, to completely chart an epigenetic landscape (Laird 2010).

In summary, we used a high-throughput M-NGS strategy to characterize the DNA methylome map of prostate cancer tissues and cells using a minimal amount of input DNA. We observe distinct patterns of DNA methylation around TSSs that frequently occur on promoters either containing or lacking a CpG island. This study has uncovered several hundred novel cancer-specific DMRs, similar to the region we characterized in *WFDC2*, and this information will be used in a future high-throughput screen. We also found additional evidence in prostate that selective regional DNA methylation regulates expression of specific transcript isoforms between normal and cancer cells. Finally, we identified genome-wide differences in DNA methylation between ETS-positive and ETS fusion-negative prostate cancer specimens, along with differences in repeat-element methylation. The comprehensive prostate methylome map generated here provides the precise genomic locations that undergo methylation changes, which will be a highly valuable public resource for investigations aimed at understanding epigenetic regulation of the prostate cancer genome.

## Methods

### Reagents, cell lines, and prostate tissue samples

Human primary prostate epithelial cells were purchased from Lonza, and the prostate cancer cell line LNCaP was obtained from ATCC. The PrEC and LNCaP cells were grown in PrEGM media (Lonza) and RPMI 1640 containing 10% FBS (Life Technologies), respectively. Human prostate tissue samples were obtained from the University of Michigan SPORE program (Supplemental Table 8). All samples were collected with informed consent of the patients and prior institutional review board approval. CpG island microarrays were purchased from Agilent Technologies. Genomic DNA was isolated from cultured cells and tissue using the DNeasy Blood and Tissue kit (QIAGEN) according to the manufacturer's instructions. 5-Aza-2′-deoxycytidine (5-Aza) was purchased from Sigma-Aldrich and used at 6 μM final concentration dissolved in DMSO.

### M-NGS library generation

MethylPlex library synthesis and GC enrichment were obtained through a commercial service at Rubicon Genomics, Inc. (Supplemental Fig. 1). Briefly, 50 ng of gDNA from tissues or cells was digested with methylation-sensitive restriction enzymes 1 and 2 (MSRE1 and MSRE2; Rubicon Genomics) in a 100-μL reaction volume for 12 h at 37°C followed by incubation for 2 h at 60°C. The samples were precipitated with two volumes of ethanol in the presence of sodium acetate (pH 5.2) and pellet paint (VWR). DNA pellets were washed with 70% ethanol, air dried, and suspended in 20 μL of TE buffer (pH 8.0). To prepare MethylPlex libraries, 10 μL of the samples from the previous step was denatured for 4 min at 95°C, cooled to 4°C, and mixed with 4 μL of library synthesis mix (Rubicon Genomics). The tubes were incubated for 2 min at 95°C and returned to 4°C before adding 1 μL of library synthesis enzyme (Rubicon Genomics). The reaction was carried in a thermocycler under the following conditions: 20 min at 16°C, 20 min at 24°C, 20 min at 37°C, and 10 min at 75°C, then returned to 4°C. Subsequently, 15 μL of the MethylPlex library was amplified in a Bio-Rad iCycler real-time PCR machine after mixing with 60 μL of library amplification mix (Rubicon Genomics), under the following cycle conditions, 2 min at 95°C (1 cycle), followed by 9 to 13 cycles of 20 sec at 96°C, 2 min at 65°C, and 1 min at 75°C. The amplified DNA was purified using the QIAquick PCR purification kit (QIAGEN), eluted in a 50-μL volume and subjected to GC enrichment following the manufacturer's protocol (Rubicon Genomics). The GC-enriched DNA was purified using the DNA Clean and Concentrator kit (Zymo Research) and eluted in 35 μL of Tris-EDTA buffer (pH 8.0). One and five micrograms of the purified products from each cell line were directly incorporated into the genomic DNA sequencing sample preparation kit procedure from Illumina at the end repair step, skipping the nebulization process. An adenine base was then added to the purified end repaired products using Klenow exo (3′-to-5′ exo minus) enzyme. The reaction product was purified, ligated to Illumina adaptors with DNA ligase, and resolved on an agarose gel. For LNCaP and PrEC libraries, gel pieces were excised at 200-bp and 400-bp positions, and the DNA was extracted using a gel extraction kit (QIAGEN). Subsequently for all tissue samples, a 350–450-bp gel cut was used. One microliter of this eluate was used as a template in a PCR amplification reaction with Phusion DNA polymerase (Finnzymes) to enrich for the adapter-modified DNA fragments. The PCR product was purified and analyzed by Bioanalyzer (Agilent Technologies) before using it for flow cell generation, where 10 nM library was used to prepare flowcells with approximately 30,000 clusters per lane. The raw sequencing image data were analyzed by the Illumina analysis pipeline and aligned to the unmasked human reference genome (NCBI v36, hg18) using the ELAND software (Illumina) to generate sequence reads of 25–32 bp. Additional information on sequencing runs for all cells and tissue sample runs can be found in Supplemental Tables 1 and 8. The M-NGS data have been deposited under accession number GSE27619 in the GEO database.

### Total RNA isolation and quantitative real-time PCR (QPCR)

Total RNA was isolated from cells using the RNeasy mini kit (QIAGEN) according to the manufacturer's instructions. A DNase I treatment step was included during the total RNA isolation procedure to remove genomic DNA from the samples. One microgram of total RNA was used in cDNA synthesis using Superscript III reverse transcriptase (Invitrogen). Quantitative real-time PCR (QPCR) was performed on prostate-cell-line cDNA samples using SYBR Green Mastermix (Applied Biosystems) on an Applied Biosystems 7900 Real Time PCR system as described (Tomlins et al. 2007b). All oligonucleotide primers were synthesized by Integrated DNA Technologies and are listed in Supplemental Table 5. *GAPDH* primer sequences were as described (Vandesompele et al. 2002). The amount of target transcript and *GAPDH* in each sample was normalized by standard ddCt methodology, and then to the reference PrEC or DMSO-treated LNCaP samples accordingly.

### CpG island annotation

The genomic coordinates for human CGIs were downloaded from ftp://ftp.ncbi.nih.gov/genomes/H_sapiens/ARCHIVE/BUILD.36.1/mapview/seq_cpg_islands.md.gz. Only islands annotated as strict CpGs were used in this study.

## RNA-seq library preparation

Poly(A) RNA from LNCaP and PrEC cells (200 ng) was isolated from total RNA using SeraMag Magnetic Oligo(dT) Beads (Thermo Fisher Scientific). RNA was fragmented for 5 min at 70°C in a fragmentation buffer (Ambion) and converted to first-strand cDNA using SuperscriptII (Life Technologies). Second-strand cDNA synthesis was performed with *Escherichia coli* DNA Pol I (Life Technologies). The double-stranded cDNA library was further processed following the Illumina Genomic DNA sample preparation protocol, which involved end repair using T4 DNA polymerase, Klenow DNA polymerase, and T4 Polynucleotide kinase followed by a single "A" base addition using Klenow 3′-to-5′ exo$^-$ polymerase. Illumina's adaptor oligo was ligated using T4 DNA ligase. The adaptor-ligated library was size-selected by separating on a 4% agarose gel and cutting out the library smear at 200 bp. The library was PCR-amplified by Phusion polymerase (Finnzymes) and purified by a PCR purification kit (QIAGEN). The library was quantified with a Bio-Analyzer (Agilent Technologies), and 10 nM each library was used to prepare flowcells with approximately 30,000 clusters per lane. The GEO accession number for the LNCaP and PrEC RNA-seq libraries is GSE29155.

## Statistical analysis

### *HMM analysis of M–NGS data*

Hidden Markov model (HMM)–based next-generation sequencing analysis is conducted in a two-step process that takes in raw reads and outputs refined boundaries of enriched chromosomal regions (Qin et al. 2010). The first step includes the formation of hypothetical DNA fragments (HDFs) from uniquely mapped reads, where the coverage of HDFs is determined by the specified DNA fragment size, and overlapped HDFs are merged to represent one consecutive genomic region. The second step is designed to refine the boundaries of the enriched region using HMM with a bin size of 25 bp (by default). Under the null hypothesis, raw reads are assumed to land on the genome following a Poisson distribution with the background rate of $r^0$, and enriched regions are expected to have more HDFs with statistical significance. The rate of the Poisson distributions in a given sample is assumed to be $r^1$, and the transition probabilities are estimated empirically, based on the inferred enriched regions defined in the first step. The output from HMM is selected based on the posterior probability of being in the enriched regions and then further filtered using maximum read counts. The threshold for maximum read counts is determined from a Bonferroni-corrected *P*-value of 0.001 calculated using a Poisson distribution with background rate $r^0$. The output is provided in BED format as well as Wiggle format for UCSC Genome Browser visualization. The output file annotation field contains information such as enriched genomic position and length, maximum height, GC content, repeated sequencing genomic position and length, mean and standard deviation of conservative scores for the enriched region, relationship with nearest genes including whether the enriched region is located within the gene or between genes, gene name, GB accession number, strand, and distance to the gene transcription start site.

## Calculating gene expression from RNA-seq data

Gene expression levels of passing filter reads from RNA-seq data that mapped by ELAND to exons (March 2006 assembly of UCSC KnownGene table) in LNCaP and PrEC cell lines are quantified as described (Maher et al. 2009).

## One-class SAM analysis

Significance analysis of microarray (SAM) (Tusher et al. 2001) (http://www-stat.stanford.edu/~tibs/SAM/) was performed on the gene expression data set obtained from 5-Aza and DMSO-treated LNCaP cells by selecting genes that were methylated in LNCaP. From 1171 methylated genes from LNCaP M-NGS (Supplemental Table 4), a total of 973 genes was mapped to Agilent expression profiling data. One-class SAM analysis was done using default settings, and significant genes were calculated with a false discovery rate (FDR) of 0.05.

## Gene Set Enrichment Analysis

Gene Set Enrichment Analysis (GSEA) is a computational method that assesses whether a defined set of genes shows statistically significant, concordant differences between any two given conditions. The fold change between the raw counts from RNA-seq NGS data on LNCaP and PrEC (representing 24,167 unique genes) was calculated, and genes were ranked by the order of expression in LNCaP. This list was uploaded as a pre-ranked gene list to GSEA v2.04 (Broad Institute, Cambridge, MA), and using respective gene lists of methylated targets in LNCaP and PrEC cell lines, GSEA was performed using a weighted enrichment statistic and default normalization mode. Similarly, the fold change between the average expression value from normal/benign ($n = 4$) and cancer samples ($n = 9$) profiled on the Agilent Human GE 44K microarray was calculated and preranked (representing 27,928 unique probes). This list was uploaded to GSEA, and enrichment analysis was performed using methylation target gene lists (the methylation present in promoters with CGIs and without CGIs, and in the gene body) in tumor samples.

## Oncomine meta-analysis

A complete description of meta-analysis performed in Oncomine is available (Rhodes et al. 2007a). In brief, a genelist of interest is uploaded to the Oncomine database, and the built-in meta-analysis tool rank-orders the genelist by the *P*-value, which is determined by a Student's *t*-test for comparisons made within each available data set (e.g., cancer vs. normal). The ranked genes were visualized with pink and green shades (top-ranked ones with darker shades, pink for overexpression, and green for repression) in heatmap format, with each row representing genes and each column representing the data set. The final order of the genes is determined by averaging ranks across the data sets.

## Molecular Concepts Map analysis

A complete description of the methods used to identify biological concept signatures in Molecular Concepts Map (MCM) is available (Rhodes et al. 2007b). In addition to more than 15,000 biological concepts from Oncomine, which include manual curation of the literature, target gene sets from genome-scale regulatory motif analyses, and reference gene sets from several gene and protein annotation databases, we have uploaded a gene list from differentially methylated regions identified from an independent Differential Methylation Hybridization profiling (concept named "DMH-Tissue Methylated in PCa") (data not shown), as well as known methylated genes in cancers provided from the Pubmeth database. In brief, MCM analysis uses a Fisher's exact test to find various significantly enriched concepts in an uploaded gene list and provides visual interaction networks.

## Repeat-element methylation analysis

The list of repeat elements predicted by the RepeatMasker (RepeatMasker Open-3.0; http://www.repeatmasker.org) program was downloaded from the UCSC Genome Browser. The

MethylPlex-NGS data from localized and metastatic prostate tissue samples were divided into two groups based on their ETS gene fusion status (ETS-positive and ETS-negative). The samples in each group were pooled together for HMM analysis, and the regions identified were mapped to repeat-element location.

### ChIP-sequencing

LNCaP cells' ChIP-seq data obtained using the H3K4me3 antibody (Abcam) and PanH3 (Abcam) were reported previously by Yu et al. (2010); the GEO accession number for the data set is GSE14097. ChIP samples were prepared for sequencing using the Genomic DNA sample prep kit (Illumina) following the manufacturer's protocols. To facilitate ChIP-seq data analysis, a hidden Markov model (HMM)–based enriched region identifying algorithm (described in the Methods section under "Statistical Analysis") was used.

### MeDIP-sequencing

Six micrograms of genomic DNA isolated from LNCaP cells was sonicated to an ~100–500-bp range and purified using the QIAGEN PCR purification kit. Using standard Illumina protocol/reagents, we end-repaired, A-tailed, and added adaptors to the fragmented DNA. The DNA was then heat-denatured for 10 min at 95°C and snap-cooled on ice. The DNA was incubated with 6 μg of anti-methyl cytosine antibody in IP buffer (10 mM sodium phosphate buffer containing 140 mM sodium chloride and 0.05% Triton X-100) overnight at 4°C in a shaker. The methylated fragments were collected by incubating with 100 μL of protein A beads (Invitrogen) for 2 h at 4°C. The beads were washed four times at 4°C in IP buffer and resuspended in 200 μL of TE buffer containing 0.25% SDS and 5 μg of proteinase K and incubated for 2 h at 55°C. The samples were purified using a DNA Clean and Concentrator-5 kit (Zymo Research), and the libraries were prepared following Illumina ChIP-seq protocol. The library was quantified with a Bio-Analyzer (Agilent Technologies), and 10 nM each library was used to prepare flowcells with approximately 30,000 clusters per lane.

### Methyl-Profiler

Methyl-Profiler (SABiosciences) is a restriction enzyme digestion–based novel technology for CGI methylation profiling, requiring <500 ng of input genomic DNA (Jaspers et al. 2010). The samples were first digested with methylation-sensitive (Ms) and/or methylation-dependent (Md) restriction enzymes along with mock digestion according to the manufacturer's instruction. PCR reactions were performed with an ABI StepOne qPCR machine (Applied Biosystems) with RT$^2$ SYBR Green/ROX qPCR Master Mix (SABiosciences) and primers targeting the region of interest. The PCR reactions were carried out with the following conditions: 10 min at 95°C, followed by 40 cycles of 15 min at 97°C and 1 min at 72°C, as described in the manufacturer's protocol. Using delta-Ct values, the relative amounts of methylation are calculated using an automated Excel-based data analysis template provided by the manufacturer. The mock-digested template is used for initial DNA input quantification, the Ms enzyme is used for hypermethylation quantification, and the Md enzyme is used for quantifying unmethylated DNA. A mixture of these two enzymes (Msd) is used to quantify the undigested amount of DNA. A methylation rate below 5% is considered not significant. While the calculated methylation percentage between 10% and 60% is considered intermediate, the values above 60% are taken as heavy methylation.

### Bisulfite sequencing

Bisulfite conversion was carried out using an EZ DNA methylation gold kit (Zymo Research) according to the manufacturer's instructions. Briefly, 500 ng of genomic DNA from either LNCaP or PrEC cells in a 20-μL volume was mixed with 130 μL of CT conversion reagent and was initially incubated for 10 min at 98°C followed by incubation for 2.5 h at 64°C. M-biding buffer (600 μL) was added to the above reaction and DNA was purified using a Zymo spin column. Sequential washes were performed with 100 μL of M-Wash buffer, 200 μL of M-sulphonation buffer, and 200 μL of M-wash buffer was carried out before eluting the DNA in 30 μL of M-elution buffer. Purified DNA (2 μL) was used as template for PCR reactions with primers (Integrated DNA Technologies) and synthesized according to bisulfite-converted DNA sequences for the regions of interest using the Methprimer software (Li and Dahiya 2002). The PCR product was gel-purified and cloned into the pCR4 TOPO TA sequencing vector (Life Technologies). Plasmid DNA isolated from 10 colonies from each sample was sequenced by conventional Sanger Sequencing (University of Michigan DNA Sequencing Core). The "BIQ Analyzer" (Bock et al. 2005) online tool was used to calculate the methylation percentage and to generate the bar graphs.

### Microarray profiling

#### *Expression profiling of 5-Aza–treated LNCaP cells*

For 5-Aza stimulation experiments, LNCaP cells cultured in RPMI 1640 were treated with vehicle, dimethyl sulfoxide (DMSO), or 6 μM 5-Aza for 4 or 6 d in duplicates. Total RNA was isolated with TRIzol (Life Technologies) and further purified using the RNAeasy Micro Kit (QIAGEN) according to the manufacturer's instructions. Expression profiling was performed using the Agilent 44K expression array. One microgram of total RNA was converted to cRNA and then labeled according to the manufacturer's protocol (Agilent). Hybridizations were performed for 16 h at 65°C. Scanned images from an Agilent microarray scanner were analyzed and extracted using Agilent Feature Extraction Software 9.1.3.1 with linear and lowess normalization performed for each array. A total of four hybridizations were performed including two 4-d and two 6-d 5-Aza-treated samples (Cy5) against control DMSO-treated samples (Cy3). The accession number for the gene expression data set in the GEO database is GSE27619.

### Expression profiling of prostate tissues

Prostate tissues characterized by M-NGS, normal/benign ($n$ = 4) and cancer ($n$ = 9), were profiled on an Agilent Human GE 44K microarray as described for LNCaP cells above. Total RNA from pooled normal prostate tissues obtained from a commercial source (Clontech Laboratories) was used as the reference. This microarray data set was used in GSEA analysis to study the association between DNA methylation and gene expression. The data set has been deposited in the GEO database.

### MethylPlex library Agilent CpG array hybridization

Two micrograms of the purified products from each PrEC and LNCaP MethylPlex DNA were labeled following the mammalian ChIP-on-chip protocol (Agilent Technologies) starting at the sample labeling stage, which uses a random primed, Klenow-based extension protocol. The samples were hybridized to an Agilent Human CpG 244K array (Cat# G4492A; Agilent Technologies), where LNCaP sample was coupled with Cy5 and PrEC to Cy3. The slides were washed according to the manufacturer's instructions. A dye-flip experiment was also performed. The scanned images were analyzed and extracted using Agilent Feature Extraction Software 9.1.3.1. Methylated regions identified by the array data were compared to M-NGS targets; their overlap is presented in Supplemental Figure 3, and the data are provided in Supplemental Table 2. This data set has been deposited in GEO under accession number GSE27619.

### 5′ rapid amplification of cDNA ends (5′-RACE)

5′-RACE was performed as previously described (Han et al. 2008). First-strand cDNA was amplified with gene-specific reverse primers *RASSF1*, *APC*, and *NDRG2* (Supplemental Table 5) and 5′ GeneRacer primers (Life Technologies) using Platinum Taq High Fidelity enzyme (Life Technologies) after the touchdown PCR protocol according to the manufacturer's instructions. PCR amplification products were cloned into a pCR4-TOPO TA vector (Life Technologies) and sequenced bidirectionally using vector primers as described (Tomlins et al. 2007a).

### Pyrosequencing

LINE-1 element methylation was estimated using the PyroMark Q24 LINE-1 methylation assay (QIAGEN) according to the manufacturer's instructions. Briefly, bisulfite-converted gDNA (described above), LINE-1 primers, and components of Hotstart Master Mix (QIAGEN) were used in a PCR reaction to amplify LINE regions from the sample. The amplification was obtained from 45 cycles of 20 sec at 95°C, 20 sec at 50°C, and 20 sec at 72°C, after an initial denaturation/enzyme activation for 15 min at 95°C, and final elongation of 5 min at 72°C. The PCR products were captured on Streptavidin Sepharose beads (GE Healthcare), denatured to produce single strands, washed, and annealed to sequencing primer, and the sequence was determined using the PyroMark Q24 system (QIAGEN). The mean methylation of three individual positions within the PCR product is considered in this assay.

## Competing interest statement

The University of Michigan has filed a patent on the findings discussed in the manuscript. A.M.C., S.M.D. and J.H.K. are named as co-inventors. A.M.C. has served on the scientific advisory board of Rubicon Genomics in the past.

## Acknowledgments

We thank Emmanuel Kamberov, Takao Kurihara, Tim Tesmer, Vladimir Marakarov, and John Langmore from Rubicon Genomics for early access to the MethylPlex technology. We thank Terrence Barrette, Mike Quist, Bharathi Laxman, Chandan Kumar, Jianjun Yu, Hao Wu, Stephanie Huang, and Yi-Mi Wu for help with data analysis; Roger Morey for help with Illumina sequencing; Bo Han for providing prostate tissue samples; and Jyoti Athanikar for editing the manuscript. C.A.M. was supported by the American Association of Cancer Research Amgen Fellowship in Clinical/Translational Research, Canary Foundation and American Cancer Society Early Detection Postdoctoral Fellowship, and a Prostate Cancer Foundation Young Investigator Award. J.Y. was supported by NIH Grant 1K99CA129565-01A1 and Department of Defense (DOD) Grant PC080665. Z.S.Q. and M.H. were supported by NIH Grant R01HG005119. J.R.P. was supported by NIH Cancer Biology Training Grant CA009676-18. J.R.P. is a Fellow of the University of Michigan Medical Scientist Training Program. A.M.C. was supported in part by the NIH (Prostate SPORE P50CA69568, R01 R01CA132874), the DOD (BC075023, W81XWH-08-0110), the Early Detection Research Network (U01 CA111275), a Burroughs Welcome Foundation Award in Clinical Translational Research, and a Doris Duke Charitable Foundation Distinguished Clinical Investigator Award. This work was also supported by National Center for Integrative Biomedical Informatics Grant U54 DA021519.

## References

Ashida S, Nakagawa H, Katagiri T, Furihata M, Iiizumi M, Anazawa Y, Tsunoda T, Takata R, Kasahara K, Miki T, et al. 2004. Molecular features of the transition from prostatic intraepithelial neoplasia (PIN) to prostate cancer: Genome-wide gene-expression profiles of prostate cancers and PINs. *Cancer Res* **64:** 5963–5972.

Berger MF, Lawrence MS, Demichelis F, Drier Y, Cibulskis K, Sivachenko AY, Sboner A, Esgueva R, Pflueger D, Sougnez C, et al. 2011. The genomic complexity of primary human prostate cancer. *Nature* **470:** 214–220.

Bernstein BE, Mikkelsen TS, Xie X, Kamal M, Huebert DJ, Cuff J, Fry B, Meissner A, Wernig M, Plath K, et al. 2006. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* **125:** 315–326.

Bird A. 2002. DNA methylation patterns and epigenetic memory. *Genes Dev* **16:** 6–21.

Bock C, Reither S, Mikeska T, Paulsen M, Walter J, Lengauer T. 2005. BiQ Analyzer: visualization and quality control for DNA methylation data from bisulfite sequencing. *Bioinformatics* **21:** 4067–4068.

Bock C, Tomazou EM, Brinkman AB, Muller F, Simmer F, Gu H, Jager N, Gnirke A, Stunnenberg HG, Meissner A. 2010. Quantitative comparison of genome-wide DNA methylation mapping technologies. *Nat Biotechnol* **28:** 1106–1114.

Bouchard D, Morisset D, Bourbonnais Y, Tremblay GM. 2006. Proteins with whey-acidic-protein motifs and cancer. *Lancet Oncol* **7:** 167–174.

Brunner AL, Johnson DS, Kim SW, Valouev A, Reddy TE, Neff NF, Anton E, Medina C, Nguyen L, Chiao E, et al. 2009. Distinct DNA methylation patterns characterize differentiated human embryonic stem cells and developing human fetal liver. *Genome Res* **19:** 1044–1056.

Cokus SJ, Feng S, Zhang X, Chen Z, Merriman B, Haudenschild CD, Pradhan S, Nelson SF, Pellegrini M, Jacobsen SE. 2008. Shotgun bisulphite sequencing of the *Arabidopsis* genome reveals DNA methylation patterning. *Nature* **452:** 215–219.

Dammann R, Li C, Yoon JH, Chin PL, Bates S, Pfeifer GP. 2000. Epigenetic inactivation of a RAS association domain family protein from the lung tumour suppressor locus 3p21.3. *Nat Genet* **25:** 315–319.

Dammann R, Schagdarsurengin U, Seidel C, Strunnikova M, Rastetter M, Baier K, Pfeifer GP. 2005. The tumor suppressor RASSF1A in human carcinogenesis: an update. *Histol Histopathol* **20:** 645–663.

Down TA, Rakyan VK, Turner DJ, Flicek P, Li H, Kulesha E, Graf S, Johnson N, Herrero J, Tomazou EM, et al. 2008. A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. *Nat Biotechnol* **26:** 779–785.

Eckhardt F, Lewin J, Cortese R, Rakyan VK, Attwood J, Burger M, Burton J, Cox TV, Davies R, Down TA, et al. 2006. DNA methylation profiling of human chromosomes 6, 20 and 22. *Nat Genet* **38:** 1378–1385.

Feber A, Wilson GA, Zhang L, Presneau N, Idowu B, Down TA, Rakyan VK, Noon LA, Lloyd AC, Stupka E, et al. 2011. Comparative methylome analysis of benign and malignant peripheral nerve sheath tumours. *Genome Res* **21:** 515–524.

Gebhard C, Benner C, Ehrich M, Schwarzfischer L, Schilling E, Klug M, Dietmaier W, Thiede C, Holler E, Andreesen R, et al. 2010. General transcription factor binding at CpG islands in normal cells correlates with resistance to de novo DNA methylation in cancer cells. *Cancer Res* **70:** 1398–1407.

Gu H, Bock C, Mikkelsen TS, Jager N, Smith ZD, Tomazou E, Gnirke A, Lander ES, Meissner A. 2010. Genome-scale DNA methylation mapping of clinical samples at single-nucleotide resolution. *Nat Methods* **7:** 133–136.

Guenther MG, Levine SS, Boyer LA, Jaenisch R, Young RA. 2007. A chromatin landmark and transcription initiation at most promoters in human cells. *Cell* **130:** 77–88.

Han B, Mehra R, Dhanasekaran SM, Yu J, Menon A, Lonigro RJ, Wang X, Gong Y, Wang L, Shankar S, et al. 2008. A fluorescence in situ hybridization screen for E26 transformation-specific aberrations: Identification of DDX5-ETV4 fusion protein in prostate cancer. *Cancer Res* **68:** 7629–7637.

Harris RA, Wang T, Coarfa C, Nagarajan RP, Hong C, Downey SL, Johnson BE, Fouse SD, Delaney A, Zhao Y, et al. 2010. Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications. *Nat Biotechnol* **28:** 1097–1105.

Hodges E, Smith A, Kendall J, Xuan Z, Ravi K, Rooks M, Zhang M, Ye K, Battacharjee A, Brizuela L, et al. 2009. High definition profiling of mammalian DNA methylation by array capture and single molecule bisulfite sequencing. *Genome Res* **19:** 1593–1605.

Ibragimova I, Ibanez de Caceres I, Hoffman AM, Potapova A, Dulaimi E, Al-Saleem T, Hudes GR, Ochs MF, Cairns P. 2010. Global reactivation of epigenetically silenced genes in prostate cancer. *Cancer Prev Res (Phila)* **3:** 1084–1092.

Illingworth RS, Bird AP. 2009. CpG islands–'a rough guide.' *FEBS Lett* **583:** 1713–1720.

Irizarry RA, Ladd-Acosta C, Wen B, Wu Z, Montano C, Onyango P, Cui H, Gabo K, Rongione M, Webster M, et al. 2009. The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nat Genet* **41:** 178–186.

Issa JP. 2004. CpG island methylator phenotype in cancer. *Nat Rev Cancer* **4:** 988–993.

Jaspers I, Horvath KM, Zhang W, Brighton LE, Carson JL, Noah TL. 2010. Reduced expression of IRF7 in nasal epithelial cells from smokers after infection with influenza. *Am J Respir Cell Mol Biol* **43:** 368–375.

Jhavar S, Reid A, Clark J, Kote-Jarai Z, Christmas T, Thompson A, Woodhouse C, Ogden C, Fisher C, Corbishley C, et al. 2008. Detection of TMPRSS2-ERG translocations in human prostate cancer by expression profiling using GeneChip Human Exon 1.0 ST arrays. *J Mol Diagn* **10:** 50–57.

Jones PA, Baylin SB. 2007. The epigenomics of cancer. *Cell* **128:** 683–692.

Kim JH, Dhanasekaran SM, Mehra R, Tomlins SA, Gu W, Yu J, Kumar-Sinha C, Cao X, Dash A, Wang L, et al. 2007. Integrative analysis of genomic aberrations associated with prostate cancer progression. *Cancer Res* **67:** 8229–8239.

Kron K, Pethe V, Briollais L, Sadikovic B, Ozcelik H, Sunderji A, Venkateswaran V, Pinthus J, Fleshner N, van der Kwast T, et al. 2009. Discovery of novel hypermethylated genes in prostate cancer using genomic CpG island microarrays. *PLoS ONE* **4:** e4830. doi: 10.1371/journal.pone.0004830.

Kumar-Sinha C, Tomlins SA, Chinnaiyan AM. 2008. Recurrent gene fusions in prostate cancer. *Nat Rev Cancer* **8:** 497–511.

Kuzmin I, Gillespie JW, Protopopov A, Geil L, Dreijerink K, Yang Y, Vocke CD, Duh FM, Zabarovsky E, Minna JD, et al. 2002. The RASSF1A tumor suppressor gene is inactivated in prostate tumors and suppresses growth of prostate carcinoma cells. *Cancer Res* **62:** 3498–3502.

Laird PW. 2010. Principles and challenges of genome-wide DNA methylation analysis. *Nat Rev Genet* **11:** 191–203.

Li LC, Dahiya R. 2002. MethPrimer: designing primers for methylation PCRs. *Bioinformatics* **18:** 1427–1431.

Li LC, Carroll PR, Dahiya R. 2005. Epigenetic changes in prostate cancer: Implication for diagnosis and treatment. *J Natl Cancer Inst* **97:** 103–115.

Maher CA, Kumar-Sinha C, Cao X, Kalyana-Sundaram S, Han B, Jing X, Sam L, Barrette T, Palanisamy N, Chinnaiyan AM. 2009. Transcriptome sequencing to detect gene fusions in cancer. *Nature* **458:** 97–101.

Maunakea AK, Nagarajan RP, Bilenky M, Ballinger TJ, D'Souza C, Fouse SD, Johnson BE, Hong C, Nielsen C, Zhao Y, et al. 2010. Conserved role of intragenic DNA methylation in regulating alternative promoters. *Nature* **466:** 253–257.

Meissner A, Mikkelsen TS, Gu H, Wernig M, Hanna J, Sivachenko A, Zhang X, Bernstein BE, Nusbaum C, Jaffe DB, et al. 2008. Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* **454:** 766–770.

Morison IM, Ramsay JP, Spencer HG. 2005. A census of mammalian imprinting. *Trends Genet* **21:** 457–465.

Nelson WG, De Marzo AM, Yegnasubramanian S. 2009. Epigenetic alterations in human prostate cancers. *Endocrinology* **150:** 3991–4002.

Ongenaert M, Van Neste L, De Meyer T, Menschaert G, Bekaert S, Van Criekinge W. 2008. PubMeth: a cancer methylation database combining text-mining and expert annotation. *Nucleic Acids Res* **36:** D842–D846.

Perry AS, Watson RW, Lawler M, Hollywood D. 2010. The epigenome as a therapeutic target in prostate cancer. *Nat Rev Urol* **7:** 668–680.

Qin ZS, Yu J, Shen J, Maher CA, Hu M, Kalyana-Sundaram S, Chinnaiyan AM. 2010. HPeak: an HMM-based algorithm for defining read-enriched regions in ChIP-Seq data. *BMC Bioinformatics* **11:** 369.

Rauch TA, Pfeifer GP. 2009. The MIRA method for DNA methylation analysis. *Methods Mol Biol* **507:** 65–75.

Rauch TA, Wu X, Zhong X, Riggs AD, Pfeifer GP. 2009. A human B cell methylome at 100-base pair resolution. *Proc Natl Acad Sci* **106:** 671–678.

Reuter M, Chuma S, Tanaka T, Franz T, Stark A, Pillai RS. 2009. Loss of the Mili-interacting Tudor domain-containing protein-1 activates transposons and alters the Mili-associated small RNA profile. *Nat Struct Mol Biol* **16:** 639–646.

Rhodes DR, Kalyana-Sundaram S, Mahavisno V, Varambally R, Yu J, Briggs BB, Barrette TR, Anstet MJ, Kincead-Beal C, Kulkarni P, et al. 2007a. Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles. *Neoplasia* **9:** 166–180.

Rhodes DR, Kalyana-Sundaram S, Tomlins SA, Mahavisno V, Kasper N, Varambally R, Barrette TR, Ghosh D, Varambally S, Chinnaiyan AM. 2007b. Molecular concepts analysis links tumors, pathways, mechanisms, and drugs. *Neoplasia* **9:** 443–454.

Schummer M, Ng WV, Bumgarner RE, Nelson PS, Schummer B, Bednarski DW, Hassell L, Baldwin RL, Karlan BY, Hood L. 1999. Comparative hybridization of an array of 21,500 ovarian cDNAs for the discovery of genes overexpressed in ovarian carcinomas. *Gene* **238:** 375–385.

Takai D, Jones PA. 2002. Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proc Natl Acad Sci* **99:** 3740–3745.

Thomson JP, Skene PJ, Selfridge J, Clouaire T, Guy J, Webb S, Kerr AR, Deaton A, Andrews R, James KD, et al. 2010. CpG islands influence chromatin structure via the CpG-binding protein Cfp1. *Nature* **464:** 1082–1086.

Tomlins SA, Laxman B, Dhanasekaran SM, Helgeson BE, Cao X, Morris DS, Menon A, Jing X, Cao Q, Han B, et al. 2007a. Distinct classes of chromosomal rearrangements create oncogenic ETS gene fusions in prostate cancer. *Nature* **448:** 595–599.

Tomlins SA, Mehra R, Rhodes DR, Cao X, Wang L, Dhanasekaran SM, Kalyana-Sundaram S, Wei JT, Rubin MA, Pienta KJ, et al. 2007b. Integrative molecular concept modeling of prostate cancer progression. *Nat Genet* **39:** 41–51.

Tusher VG, Tibshirani R, Chu G. 2001. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci* **98:** 5116–5121.

Vandesompele J, De Preter K, Pattyn F, Poppe B, Van Roy N, De Paepe A, Speleman F. 2002. Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biol* **3:** RESEARCH0034. doi: 10.1186/gb-2002-3-7-research0034.

Weber M, Davies JJ, Wittig D, Oakeley EJ, Haase M, Lam WL, Schubeler D. 2005. Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nat Genet* **37:** 853–862.

Wu H, Coskun V, Tao J, Xie W, Ge W, Yoshikawa K, Li E, Zhang Y, Sun YE. 2010. Dnmt3a-dependent nonpromoter DNA methylation facilitates transcription of neurogenic genes. *Science* **329:** 444–448.

Yegnasubramanian S, Kowalski J, Gonzalgo ML, Zahurak M, Piantadosi S, Walsh PC, Bova GS, De Marzo AM, Isaacs WB, Nelson WG. 2004. Hypermethylation of CpG islands in primary and metastatic human prostate cancer. *Cancer Res* **64:** 1975–1986.

Yegnasubramanian S, Haffner MC, Zhang Y, Gurel B, Cornish TC, Wu Z, Irizarry RA, Morgan J, Hicks J, DeWeese TL, et al. 2008. DNA hypomethylation arises later in prostate cancer progression than CpG island hypermethylation and contributes to metastatic tumor heterogeneity. *Cancer Res* **68:** 8954–8967.

Yu J, Mani RS, Cao Q, Brenner CJ, Cao X, Wang X, Wu L, Li J, Hu M, Gong Y, et al. 2010. An integrated network of androgen receptor, polycomb, and TMPRSS2-ERG gene fusions in prostate cancer progression. *Cancer Cell* **17:** 443–454.

*Received December 14, 2010; accepted in revised form April 28, 2011.*

BIOINFORMATICS APPLICATIONS NOTE

Vol. 27 no. 20 2011, pages 2903–2904
doi:10.1093/bioinformatics/btr467

Genome analysis

Advance Access publication August 11, 2011

# ChimeraScan: a tool for identifying chimeric transcription in sequencing data

Matthew K. Iyer[1,2], Arul M. Chinnaiyan[1,2,3,4,5] and Christopher A. Maher[1,2,3,*]

[1]Michigan Center for Translational Pathology, [2]Center for Computational Medicine and Biology, [3]Department of Pathology, [4]Howard Hughes Medical Institute and [5]Department of Urology, University of Michigan Medical School, Ann Arbor, MI 48109, USA

Associate Editor: Alfonso Valencia

## ABSTRACT

**Summary:** Next generation sequencing (NGS) technologies have enabled *de novo* gene fusion discovery that could reveal candidates with therapeutic significance in cancer. Here we present an open-source software package, ChimeraScan, for the discovery of chimeric transcription between two independent transcripts in high-throughput transcriptome sequencing data.

**Availability:** http://chimerascan.googlecode.com

**Contact:** cmaher@dom.wustl.edu

**Supplementary Information:** Supplementary data are available at *Bioinformatics* online.

Received on March 4, 2011; revised on July 26, 2011; accepted on August 3, 2011

## 1 INTRODUCTION

High-throughput transcriptome sequencing (RNA-Seq) facilitates detection of aberrant, chimeric RNAs (Maher *et al.*, 2009a; Maher *et al.*, 2009b). Methods for chimera detection have already uncovered recurrent classes of clinically relevant gene fusions in prostate (Palanisamy *et al.*, 2010) and lymphoid cancers (Steidl *et al.* 2011). Therefore, the continued development of accurate and efficient software tools for chimera discovery is of major clinical significance. To this end, we have developed a chimera discovery methodology, or ChimeraScan, and offer it as open-source software package for the community to utilize for their own sequencing efforts. ChimeraScan includes features such as the ability to process long (>75 bp) paired-end reads, processing of ambiguously mapping reads, detection of reads spanning a fusion junction, integration with the popular Bowtie aligner (Langmead *et al.*, 2009), supports the standardized SAM format and generation of HTML reports for easy investigation of results. Overall, we believe that the ChimeraScan will facilitate the discovery of additional gene fusions that may serve as clinically relevant targets in cancer.

## 2 METHODS

*Initial paired-end alignment*: ChimeraScan uses Bowtie to align paired-end reads to a combined genome-transcriptome reference. An indexing program creates the combined index from genomic sequences (FASTA format) and transcript features (UCSC GenePred format). Paired alignments within the fragment size range (default: 0–1000) are referred to as concordantly mapping reads (Fig. 1A). ChimeraScan uses these alignments to estimate the insert size distribution of the library, which will later be used to filter out likely false positive chimeras.

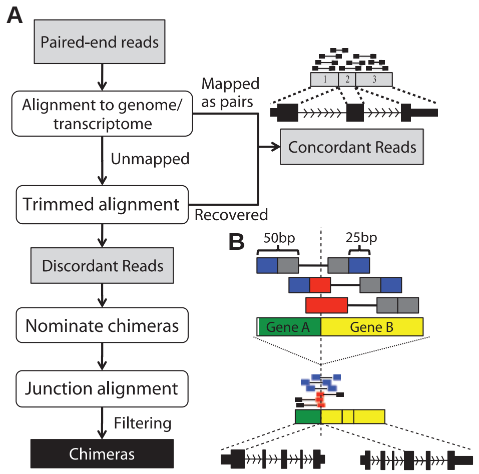


**Fig. 1.** ChimeraScan flowchart. (**A**) Paired-end reads failing an initial alignment step are segmented and realigned to detect discordant reads. Discordant reads that pass filter criteria are realigned across putative chimeric junctions. (**B**) Chimera with encompassing (blue) and spanning (red) segments detected during realignment.

*Trimmed paired-end alignment*: read pairs that could not be aligned concordantly are trimmed into smaller segments (default = 25 bp) and realigned. Trimming increases the chance that neither read alignment spans a chimeric junction, thereby improving sensitivity for nominating chimeras.

*Nomination of chimera candidates*: the trimmed alignments are scanned for evidence of discordant read pairs, or reads that align to distinct references or distant genomic locations (as determined by the fragment size range) of the same reference. Reads aligning to overlapping transcripts are not considered discordant. ChimeraScan clusters the discordant reads and produces a list of putative 5′–3′ transcript pairs that serve as chimera candidates.

*Detection of reads spanning the chimeric junction*: ChimeraScan builds a new reference index from the set of putative chimeric junction sequences, and realigns candidate junction-spanning reads to this index. Candidate spanning reads are either (i) discordant reads with trimmed alignments bordering a junction or (ii) unmapped reads whose mates align to a predicted chimera (Fig. 1B). A read that spans a junction by more than a minimum ‘anchor’ length is denoted as a ‘spanning’ read. We compute the required ‘anchor’ length separately for each chimera by insisting that the number of bases overlapping its junction be greater than number of homologous bases between the 5′ and 3′ genes at the breakpoint plus the number of mismatches allowed.

*To whom correspondence should be addressed.

*Filtering false-positive chimeras*: after spanning reads are incorporated, ChimeraScan filters chimeras with few supporting reads (default is <3 reads) and chimeras with fragment sizes far outside the range of the distribution (default is >99% of all fragment sizes). When isoforms of the same gene support a fusion ChimeraScan only retains the isoform(s) with highest coverage.

*Reporting chimeras*: ChimeraScan produces a tabular text file describing each chimera, and optionally generates a user-friendly HTML page with links to detailed descriptions of the chimeric genes.

## 3 RESULTS

To evaluate the results from ChimeraScan, we applied it to three well-characterized cancer cell lines known to harbor multiple chimeric transcripts: VCaP (prostate cancer, 2 × 53 bp) (Tomlins *et al.*, 2005), LNCaP (prostate cancer, 2 × 34 bp) and MCF7 (breast cancer, 2 × 35 bp) (Hampton *et al.*, 2009; Volik *et al.*, 2006). Sequence data are deposited in GenBank under the accession number GSE29098. We aligned to human genome (VR-hg19) and UCSC known transcripts (December 2010), allowing for up to two mismatches and no >100 alignments per read. The trimmed alignment step was performed with 25 bp segments.

As our initial benchmark, we confirmed that ChimeraScan was able to recapitulate experimentally validated candidates, our 'gold standard' (Supplementary Table 1) (Maher *et al.*, 2009b). ChimeraScan was able to detect 9/10, 4/4 and 12/13 chimeras from VCaP, LNCaP and MCF-7, respectively.

In addition to recapitulating previously reported results, we have identified novel candidates that demonstrate ChimeraScan's ability to identify and prioritize high-quality chimeras. Overall, ChimeraScan nominated 335 novel chimeras (78 in VCaP, 105 in LNCaP and 152 in MCF7) from the three cell lines (Supplementary Table 2–4). Interestingly, we detected an interchromosomal rearrangement *TBL1XR1-RGS17* detected in the MCF-7 cell line. While not originally reported within NGS data (Maher *et al.*, 2009b), *TBL1XR1-RGS17* was previously detected by a paired-end diTag approach and experimentally confirmed (Ruan *et al.*, 2007). Another novel candidate was the intrachromosomal rearrangement, *NDUFAF2-MAST4*, in VCaP that is supported by just two encompassing reads and one spanning reads. The ability to identify a high-quality spanning read that uniquely confirms the fusion junction (Supplementary Table 2), thereby increasing our confidence in *NDUFAF2-MAST4*, demonstrates the sensitivity of ChimeraScan.

We next compared ChimeraScan with publicly available tools deFuse (McPherson *et al.*, 2011), shortFuse (Kinsella *et al.*, 2011) and MapSplice (Wang *et al.*, 2010) using the 10 experimentally validated VCaP chimeras (Supplementary Table 5). While deFuse nominated the fewest chimeras, it only detected 60% of the true positives. In comparison, ChimeraScan detected 90% of the true positives from 78 predicted chimeras. Of the remaining programs, MapSplice nominated 400 chimeras while detecting 60% of the true positives and ShortFuse nominated 245 chimeras while confirming 70% of the true positives. Overall, these results suggest that ChimeraScan is among the more stringent programs while enriching for true positives.

## 4 CONCLUSION

Here, we present an optimized publicly available chimera discovery methodology for identifying novel therapeutically targetable gene fusions in human cancers. Our results suggest that ChimeraScan produces a stringent list of predictions that are enriched with true positives. Furthermore, due to its trimmed alignment steps we believe ChimeraScan will be scalable when longer reads are available to provide increased coverage of fusion junctions. Overall, we feel that with the existing features ChimeraScan is a user-friendly tool that will enable other research groups to make discoveries within their own RNA-Seq data collections.

*Funding*: Department of Defense Breast Cancer Predoctoral Grant (to M.K.I.); Prostate Cancer Foundation Young Investigator Award and National Institutes of Health Pathway to Independence (K99 CA149182-01) Award (to C.A.M.); National Institutes of Health, Department of Defense and Early Detection Research Network (to A.M.C.).

*Conflict of Interest*: none declared.

## REFERENCES

Hampton,O.A. *et al.* (2009) A sequence-level map of chromosomal breakpoints in the MCF-7 breast cancer cell line yields insights into the evolution of a cancer genome. *Genome Res.*, **19**, 167–177.

Kinsella,M. *et al.* (2011) Sensitive gene fusion detection using ambiguously mapping RNA-Seq read pairs. *Bioinformatics*, **27**, 1068–1075.

Langmead,B. *et al.* (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.

Maher,C.A. *et al.* (2009a) Transcriptome sequencing to detect gene fusions in cancer. *Nature*, **458**, 97–101.

Maher,C.A. *et al.* (2009b) Chimeric transcript discovery by paired-end transcriptome sequencing, *Proc. Natl Acad. Sci. USA*, **106**, 12353–12358.

McPherson,A. *et al.* (2011) deFuse: an algorithm for gene fusion discovery in tumor RNA-Seq data. *PLoS Comput. Biol.*, **7**, e1001138.

Palanisamy,N. *et al.* (2010) Rearrangements of the RAF kinase pathway in prostate cancer, gastric cancer and melanoma. *Nat. Med.*, **16**, 793–798.

Ruan,Y. *et al.* (2007) Fusion transcripts and transcribed retrotransposed loci discovered through comprehensive transcriptome analysis using Paired-End diTags (PETs). *Genome Res.*, **17**, 828–838.

Steidl,C. *et al.* (2011) MHC class II transactivator CIITA is a recurrent gene fusion partner in lymphoid cancers. *Nature*, **471**, 377–381.

Tomlins,S.A. *et al.* (2005) Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science*, **310**, 644–648.

Volik,S. *et al.* (2006) Decoding the fine-scale structure of a breast cancer genome and transcriptome. *Genome Res.*, **16**, 394–404.

Wang,K. *et al.* (2010) MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Res.*, **38**, e178.

nature medicine

# Functionally recurrent rearrangements of the MAST kinase and Notch gene families in breast cancer

Dan R Robinson[1,2,10], Shanker Kalyana-Sundaram[1–3,10], Yi-Mi Wu[1,2], Sunita Shankar[1,2], Xuhong Cao[1,2,4], Bushra Ateeq[1,2], Irfan A Asangani[1,2], Matthew Iyer[1,5], Christopher A Maher[1,2,5], Catherine S Grasso[1,2], Robert J Lonigro[1,2], Michael Quist[1,2], Javed Siddiqui[1,2], Rohit Mehra[1,2], Xiaojun Jing[1,2], Thomas J Giordano[2,6], Michael S Sabel[6,7], Celina G Kleer[2,6], Nallasivam Palanisamy[1,2], Rachael Natrajan[8], Maryou B Lambros[8], Jorge S Reis-Filho[8], Chandan Kumar-Sinha[1,2] & Arul M Chinnaiyan[1,2,4–6,9]

**Breast cancer is a heterogeneous disease that has a wide range of molecular aberrations and clinical outcomes. Here we used paired-end transcriptome sequencing to explore the landscape of gene fusions in a panel of breast cancer cell lines and tissues. We observed that individual breast cancers have a variety of expressed gene fusions. We identified two classes of recurrent gene rearrangements involving genes encoding microtubule-associated serine-threonine kinase (MAST) and members of the Notch family. Both MAST and Notch-family gene fusions have substantial phenotypic effects in breast epithelial cells. Breast cancer cell lines harboring Notch gene rearrangements are uniquely sensitive to inhibition of Notch signaling, and overexpression of *MAST1* or *MAST2* gene fusions has a proliferative effect both *in vitro* and *in vivo*. These findings show that recurrent gene rearrangements have key roles in subsets of carcinomas and suggest that transcriptome sequencing could identify individuals with rare, targetable gene fusions.**

Recurrent gene fusions and translocations have long been associated with hematologic malignancies and rare soft-tissue tumors as being 'driving' genetic lesions[1–3]. Over the last few years, it has become apparent that these genetic rearrangements are also present in common solid tumors, including a large subset of prostate cancers[4,5] and smaller subsets of lung cancer, among other types of tumors[6]. Secretory breast cancer, a rare subtype of breast cancer, is characterized by recurrent gene fusions of *ETV6* and *NTRK3* (ref. 7). Although multiple breast cancer genomes have been sequenced[8,9], and complex somatic rearrangements have been observed[10], the driving recurrent gene fusions have not been identified.

We used paired-end transcriptome sequencing on a panel of 89 breast cancer cell lines and tumors (**Supplementary Fig. 1**) and then applied our previously developed chimera discovery pipeline[11,12]. This panel represented a spectrum of breast carcinoma and included 42 estrogen receptor (ER)-positive, 21 v-erb-b2 erythroblastic leukemia viral oncogene homolog 2, neuro/glioblastoma derived oncogene homolog (ERBB2)-positive and 27 triple negative (ER$^-$, progesterone receptor–negative (PR$^-$) and ERBB2$^-$) samples (**Supplementary Table 1**). Investigation of fusion transcripts led to the identification of 384 expressed gene fusions at an average of nearly five fusions per breast cancer sample, with a slightly higher number of gene fusions in the cell lines compared to the primary tumors (**Supplementary Fig. 1b** and **Supplementary Table 2**). Notably, we found that only *SEC16A-NOTCH1* was recurrent in our compendium, even though several fusion genes appeared in combination with different fusion partners. Overall, we found 24 genes to be recurrent fusion partners (**Supplementary Table 2**). To focus on potentially tumorigenic driver fusions, we prioritized the gene fusions based on the known cancer-associated functions of component genes. Although there were many singleton fusions in our compendium that met these criteria, we identified five instances of fusions of MAST family kinases and eight instances of fusions of genes in the Notch family (**Fig. 1** and **Supplementary Fig. 2**).

The genes encoding members of the MAST kinase family are characterized by the presence of a serine-threonine kinase domain, a second 3′ MAST domain with some similarity to kinase domains and a PDZ domain[13]. Little is known about the biological role of MAST kinases, and somatic alterations have not previously been described in cancer. Initially, we identified three independent instances of MAST gene fusions using transcriptome analyses: fusions of *ARID1A* and *MAST2*, *ZNF700* and *MAST1*, and *NFIX* and *MAST1* (**Fig. 1a**). We devised a targeted sequencing approach to screen additional samples for MAST gene fusions. We generated and captured a transcriptome library of 74 pooled breast carcinoma RNAs with baits encompassing

[1]Michigan Center for Translational Pathology, University of Michigan, Ann Arbor, Michigan, USA. [2]Department of Pathology, University of Michigan, Ann Arbor, Michigan, USA. [3]Department of Environmental Biotechnology, Bharathidasan University, Tiruchirappalli, India. [4]Howard Hughes Medical Institute, University of Michigan, Ann Arbor, Michigan, USA. [5]Center for Computational Medicine and Biology, University of Michigan, Ann Arbor, Michigan, USA. [6]Comprehensive Cancer Center, University of Michigan, Ann Arbor, Michigan, USA. [7]Department of Surgery, University of Michigan, Ann Arbor, Michigan, USA. [8]The Breakthrough Breast Cancer Research Centre, Institute of Cancer Research, London, UK. [9]Department of Urology, University of Michigan, Ann Arbor, Michigan, USA. [10]These authors contributed equally to this work. Correspondence should be addressed to A.M.C. (arul@umich.edu) or C.K.-S. (chakumar@med.umich.edu).

Received 19 April; accepted 24 October; published online 20 November 2011; doi:10.1038/nm.2580

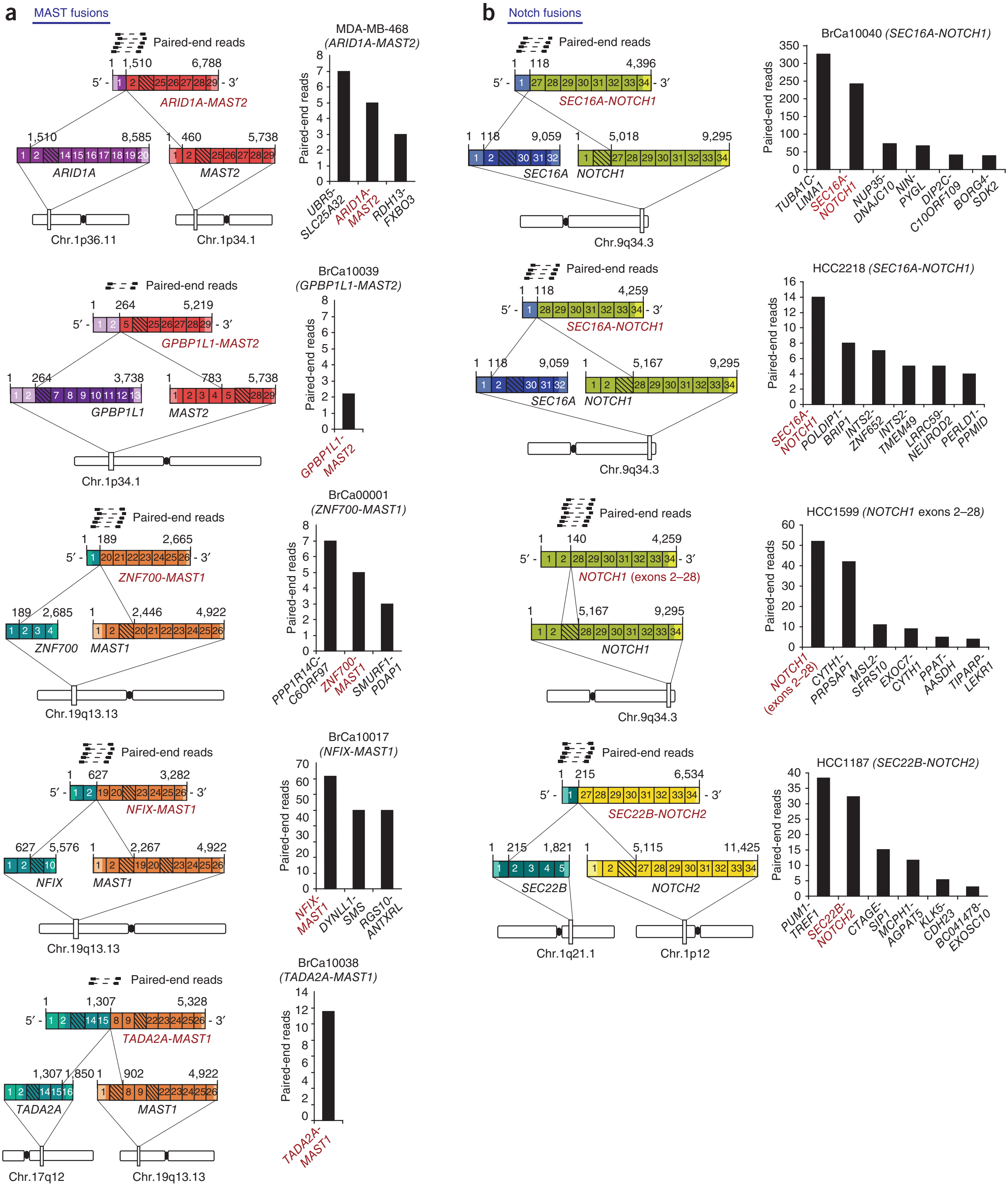


**Figure 1** Discovery of the MAST kinase and Notch gene fusions in breast cancer identified by paired-end transcriptome sequencing. (**a**) MAST family gene fusions. (**b**) Notch-family gene fusions. Fusion junctions with respective exon numbers (and nt positions) comprising the chimeric transcripts are shown. Bar plots of the top ranked gene fusions by number of paired-end reads supporting each nominated fusion in the index samples (shown on the right), with MAST or Notch fusion genes shown in red.

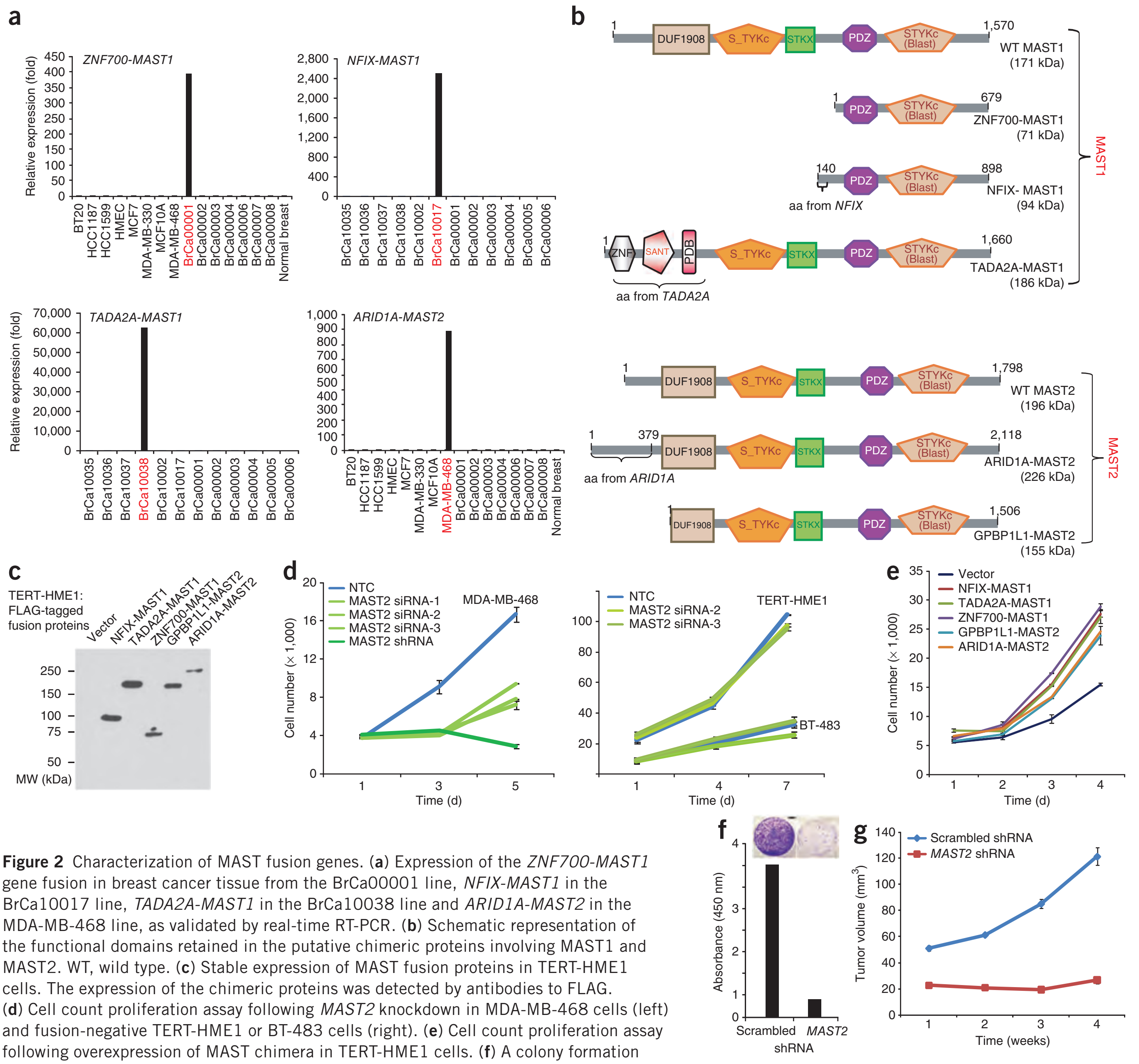
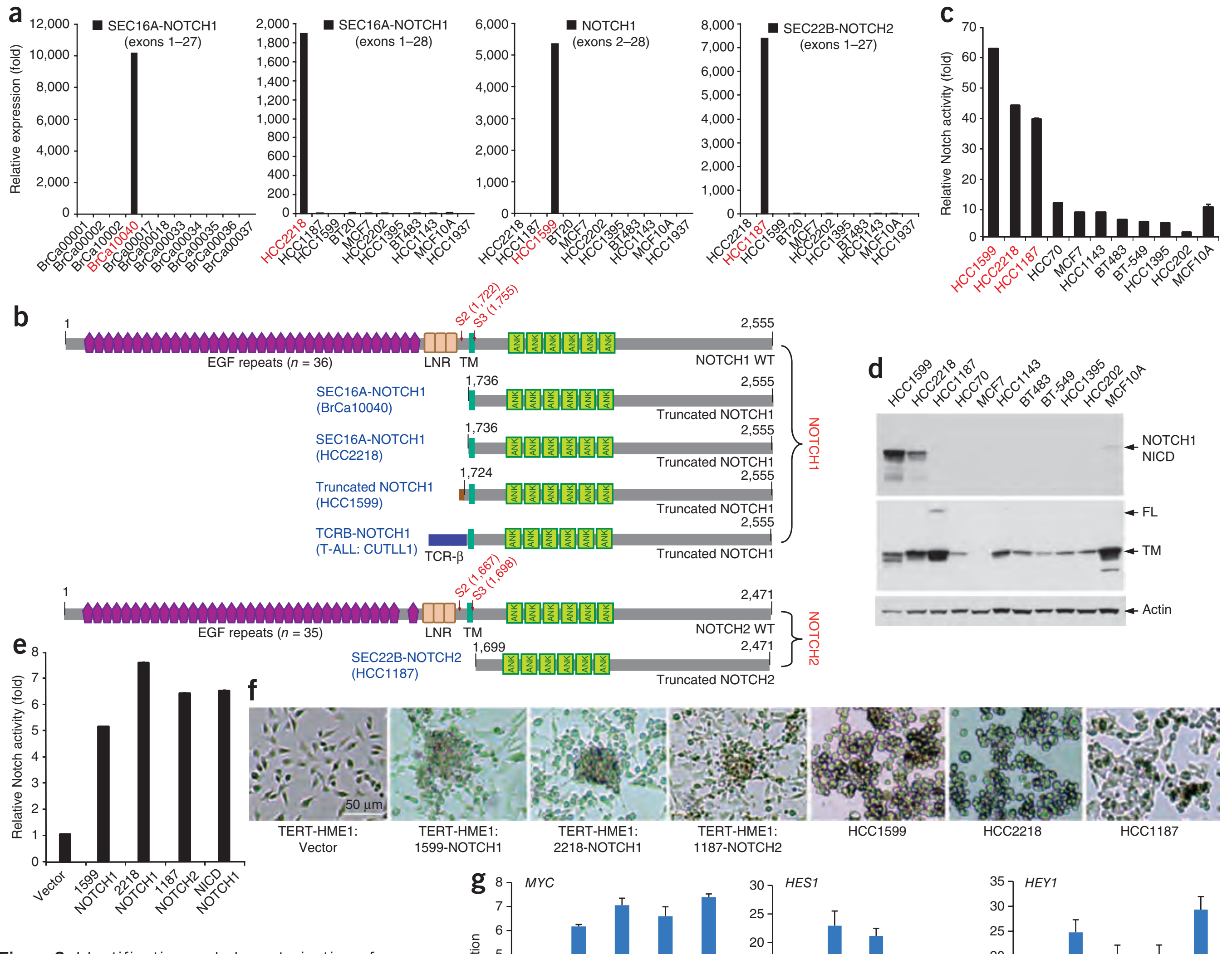


**Figure 2** Characterization of MAST fusion genes. (**a**) Expression of the *ZNF700-MAST1* gene fusion in breast cancer tissue from the BrCa00001 line, *NFIX-MAST1* in the BrCa10017 line, *TADA2A-MAST1* in the BrCa10038 line and *ARID1A-MAST2* in the MDA-MB-468 line, as validated by real-time RT-PCR. (**b**) Schematic representation of the functional domains retained in the putative chimeric proteins involving MAST1 and MAST2. WT, wild type. (**c**) Stable expression of MAST fusion proteins in TERT-HME1 cells. The expression of the chimeric proteins was detected by antibodies to FLAG. (**d**) Cell count proliferation assay following *MAST2* knockdown in MDA-MB-468 cells (left) and fusion-negative TERT-HME1 or BT-483 cells (right). (**e**) Cell count proliferation assay following overexpression of MAST chimera in TERT-HME1 cells. (**f**) A colony formation assay of MDA-MB-468 cells treated with *MAST2*-specific shRNA or control scrambled shRNA. The inset shows crystal violet staining of cells treated with either scrambled or *MAST2* shRNA. Scale bar, 5 mm. (**g**) Tumor volumes of MDA-MD-468 xenografts with *MAST2* knockdown. MW, molecular weight; aa, amino acids. Error bars represent s.d.

*MAST1* and *MAST2*. After sequencing, we discovered two new MAST gene fusions: *TADA2A-MAST1* and *GPBP1L1-MAST2* (**Fig. 1a**). The samples with MAST gene fusions are distinct from those with Notch family gene fusions (**Fig. 1b**).

We investigated the function of the MAST fusions (**Fig. 2**) and confirmed the fusions using fusion-specific PCR (**Fig. 2a**). All five MAST fusions encoded contiguous open reading frames (ORFs), some of which retained the canonical serine-threonine kinase domain and all of which retained the PDZ domain and the 3′ kinase-like domain (**Fig. 2b**). Therefore, in total, we discovered five new gene fusions encoding MAST1 and MAST2 in a cohort of approximately 100 breast cancer samples and more than 40 cell lines, suggesting that the newly identified MAST gene fusions are present in a subset of 3–5% of breast cancers.

The *ZNF700-MAST1* fusion transcript encodes a truncated MAST1 protein that retains the 3′ kinase-like and PDZ domains. We cloned the ORF of the *ZNF700-MAST1* fusion gene to test its phenotypic effects and used a full-length *MAST2* expression construct to mimic the function of *ARID1A-MAST2* overexpression. To assess the potential oncogenic functions of genes encoding MAST, we ectopically overexpressed epitope-tagged truncated MAST1 and full-length MAST2 in the benign breast cell line TERT-HME1 (**Supplementary Fig. 3a**–**h**). We then cloned and expressed all five *MAST1* and *MAST2* fusions. Consistent with the earlier observations, TERT-HME1 cells overexpressing the five MAST fusions (**Fig. 2c**) had greater cell proliferation (**Fig. 2e**). Overall, these results suggest that ectopic expression of the MAST fusions results in growth and a proliferative advantage in benign breast epithelial cells.



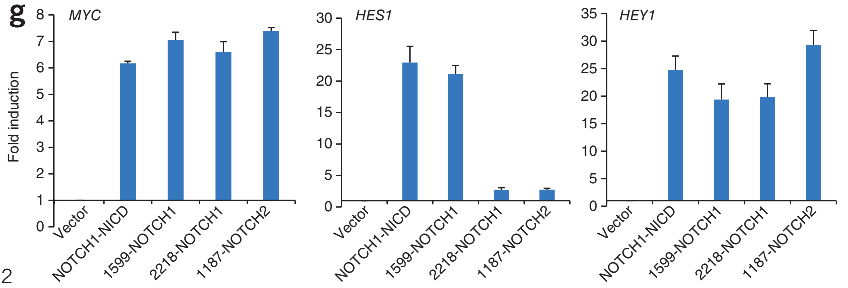


**Figure 3** Identification and characterization of Notch gene aberrations in breast carcinomas. (**a**) Detection of new Notch transcripts by quantitative RT-PCR (qRT-PCR). (**b**) Schematic presentation of the predicted protein structures of the aberrant Notch genes. (**c**) Notch reporter activities in the index lines that have Notch gene fusions. The fold activations of the Notch pathway were calculated using the cell line HCC202 as the reference. All the data was normalized to *Renilla* luciferase activity. (**d**) Western blot analysis of NOTCH1-NICD expression, detected with an antibody specifically recognizing the active NOTCH1-NICD protein after cleavage by γ-secretase. FL, full length NOTCH1; TM, transmembrane NOTCH1. (**e**) Notch reporter luciferase levels in 293T cells, assayed following transient Notch expression. (**f**) Bright-field images of control vector and TERT-HME1 cells expressing the Notch fusion alleles. Breast cancer lines positive for Notch fusion are shown for comparison. (**g**) mRNA expression levels of three Notch target genes inTERT-HME1 cells stably expressing Notch fusions, measured by quantitative RT-PCR. Error bars represent s.d.

Because the endogenous *ARID1A-MAST2* fusion is present in the breast cancer cell line MDA-MB-468, we used multiple independent siRNAs specific to *MAST2* or the *ARID1A-MAST2* fusion to achieve knockdown of the ARID1A-MAST2 fusion protein (**Supplementary Fig. 3i–s**). Knockdown of MAST2 showed significant inhibitory effects on growth in MDA-MB-468 cells but not in the fusion-negative cell line BT-483 or in benign TERT-HME1 breast cells (**Fig. 2d**). To further characterize the effects of the *ARID1A-MAST2* fusion in MDA-MB-468 cells, we used shRNA targeting *MAST2*, which showed efficient knockdown of *ARID1A-MAST2* fusion transcript and protein (**Supplementary Fig. 3k,l**). MDA-MB-468 cells treated with *MAST2* shRNA had a reduction in growth, as shown in a colony formation assay (**Fig. 2f**), and showed increased apoptosis and S-phase arrest (**Supplementary Fig. 3m,n**). In the mouse xenograft model, MDA-MB-468 cells transiently transfected with *MAST2* shRNA did not establish palpable tumors over a time course of 4 weeks after transfection (**Fig. 2g**). Our knockdown studies showed that the *ARID1A-MAST2* fusion is a key driver fusion in MDA-MB-468 cells.

In addition to MAST fusions, we found a total of eight rearrangements involving either *NOTCH1* or *NOTCH2* (**Fig. 1b** and **Supplementary Fig. 2**). We found all of these rearrangements in ER$^-$ breast carcinomas ($P = 0.008$) and all but one rearrangement in triple-negative breast carcinomas. We focused on one ER$^-$ tumor and three ER$^-$ breast cancer cell lines with 3′ *NOTCH1* or *NOTCH2*

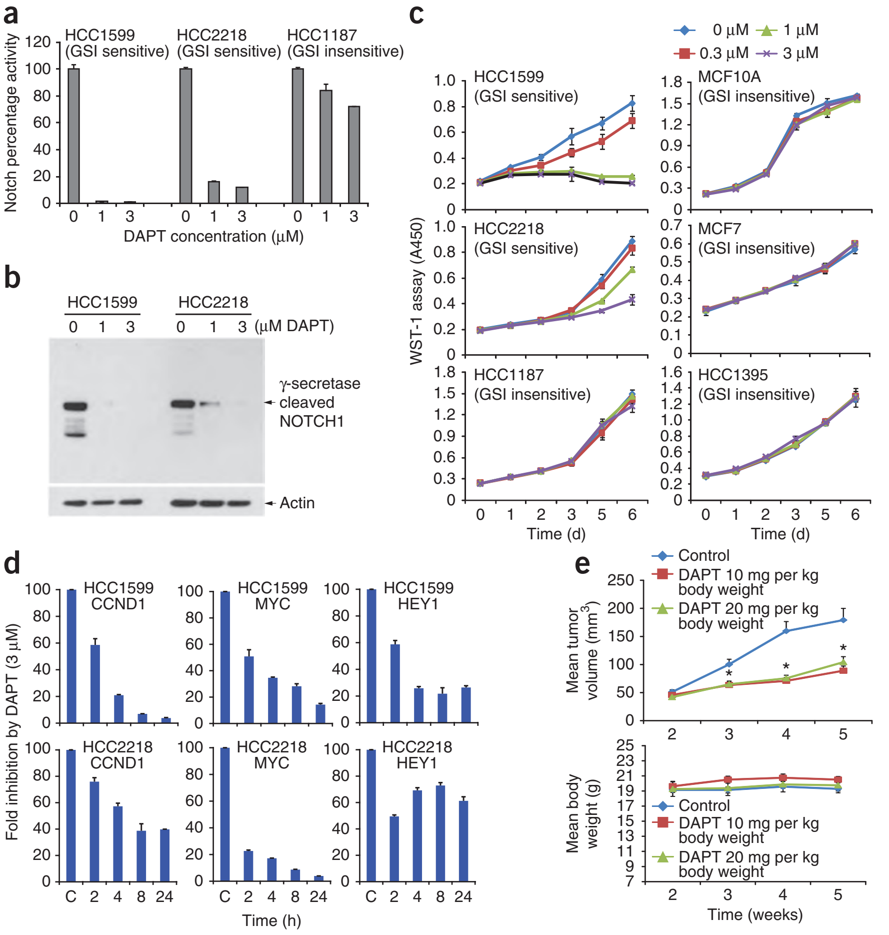


**Figure 4** γ-secretase inhibitor DAPT effects in fusion positive and negative breast carcinoma cell lines. (**a**) Luciferase assay of the Notch signaling pathway following DAPT treatment. Breast cancer cells were co-infected with a Notch reporter construct, lenti-RBPJ (recombination signal binding protein for immunoglobulin κJ) firefly luciferase, and the internal control lenti-*Renilla* luciferase. Twenty-four hours after treatment with DAPT, luciferase activities were measured. (**b**) NICD levels after treatment with DAPT detected using an antibody specific to active NOTCH1-NICD after cleavage by γ-secretase. (**c**) WST-1 cell proliferation assays of six breast cell lines after DAPT treatment. (**d**) Expression of Notch target genes after treatment with DAPT, as measured by qRT-PCR. (**e**) Xenograft tumor volume and body mass after treatment with the γ-secretase inhibitor DAPT. Mice xenografted with HCC1599 cells were treated daily after tumors formed, and the size of the tumors was monitored. *$P < 0.005$.

fusion transcripts in our functional studies. The Notch fusion transcripts were abundantly expressed and were specific to the samples with DNA rearrangements (**Fig. 3a**). All the fusion transcripts retained the exons that encode the Notch intracellular domain (NICD), which is responsible for inducing the transcriptional program following Notch activation (**Fig. 3b**). We characterized the DNA breakpoints associated with Notch fusions by mate-pair genomic library sequencing or by long-range genomic PCR (**Supplementary Fig. 4a,b**).

We categorized the predicted ORFs for the *NOTCH1* and *NOTCH2* fusion transcripts into two classes (**Fig. 3b**). For both the *SEC16A-NOTCH1* fusions and the intragenic *NOTCH1* fusion in the HCC1599 cell line, the predicted ORFs initiated after the S2 cleavage site but before the S3 γ-secretase cleavage site, similar to that seen in the *TCRB-NOTCH1* fusion in the adult lymphocytic leukemia T cell line CUTLL1 (ref. 14). In contrast, we predicted the *SEC22B-NOTCH2* fusion ORF to initiate just after the γ-secretase S3 cleavage site. The resulting protein would be nearly identical to NICD, and we predict that it would be highly active and independent of cleavage by γ-secretase (**Fig. 3b**).

We saw substantially higher Notch responsive transcriptional activity in the three cell lines with Notch fusions compared to the other breast cell lines using a Notch luciferase reporter (**Fig. 3c**). Therefore, each of the three Notch fusions is capable of activating the expression of Notch-responsive genes. Using an antibody specific to the γ-secretase cleaved active form of the NOTCH1 NICD, both HCC1599 and HCC2218 showed high concentrations of NICD, consistent with the fusion protein acting as a substrate for activation by γ-secretase (**Fig. 3d**). The HCC1187 cell line, which has a *NOTCH2* fusion gene, contains little NOTCH1 NICD. Most breast cancer lines express wild-type *NOTCH1* (**Fig. 3d**, middle); however, only the two cell lines with *NOTCH1* fusion alleles showed high concentrations of activated NICD. Each of the three fusion alleles, which we co-transfected with a Notch reporter plasmid, induced Notch-responsive transcription that was equivalent to NICD (**Fig. 3e**).

The three breast cell lines containing the Notch fusions showed decreased cell-matrix adhesion and grew in suspension or as weakly adherent clusters, which was in contrast to the majority of breast carcinoma cell lines. When we transduced *NOTCH1* and *NOTCH2* fusion alleles to create stable pools of TERT-HME1 cells, we observed notable morphological changes (**Fig. 3f**). TERT-HME1 cells had adherent epithelial properties, whereas cells expressing Notch fusion lost adherence and propagated as weakly attached clusters, similar to the index lines with Notch fusions and consistent with the previously reported effects of NICD expression in MCF10A cells[15]. Furthermore, the fusion alleles markedly induced expression of the Notch target genes *MYC*, *HES1* and *HEY1* (**Fig. 3g**).

The Notch fusions represent two functional classes with respect to dependence on the activity of γ-secretase. Fusions in BrCa10040, HCC2218 and HCC1599 cells are dependent on S3 cleavage for activity and are sensitive to γ-secretase inhibitors (GSIs). The fusion class in HCC1187 cells is independent of S3 cleavage. We established stable Notch reporter lines from each of the three Notch fusion index lines and treated them with the γ-secretase inhibitor *N*-[(3,5-difluorophenyl)acetyl]-L-al anyl-2-phenyl]glycine-1,1-dimethylethyl ester (DAPT)[16]. We saw a reduction of Notch reporter activity after treatment with DAPT in the HCC1599 and HCC2218 fusion alleles (**Fig. 4a**). However, Notch reporter activity was only slightly diminished by treatment with DAPT in HCC1187 cells, which express a γ-secretase–independent Notch fusion allele that is capable of activating Notch reporter activity. DAPT treatment also substantially reduced NICD protein concentrations in both of the γ-secretase inhibitor–sensitive cell lines (**Fig. 4b**). Furthermore, the index cell lines showed dependence on Notch signaling for proliferation and survival

(**Fig. 4c**). The HCC1599 and HCC2218 cell lines showed marked reductions in proliferation after treatment with DAPT. The HCC1187 cell line, which expresses GSI-independent *NOTCH2* fusion, had no reduction in proliferation after DAPT treatment, which is also the case in breast cell lines not expressing Notch fusion alleles.

Treatment with DAPT repressed the expression of the Notch targets *MYC* and *CCND1* (**Fig. 4d**), two genes that have a key role in mouse mammary tumorigenesis induced by Notch[17,18], which further supports the idea GSIs could be useful in treating cancers that have activated Notch alleles. Consistent with this, treatment with DAPT significantly reduced tumor volume in a xenograft tumor model of HCC1599 cells (**Fig. 4e**).

Since the discovery of the *TMPRSS2-ERG* gene fusion in approximately 50% of prostate cancers, emerging evidence has suggested that recurrent gene fusions have a more substantial role in common solid tumors than was previously known. The MAST and Notch aberrations in breast cancer are new classes of rare but functionally recurrent gene fusions with therapeutic implications (similar to the anaplastic lymphoma receptor tyrosine kinase (ALK) fusions in lung cancer). MAST kinase and Notch gene rearrangements were mutually exclusive aberrations in the samples we tested, and, together, may be present in up to 5–7% of breast cancers. The discovery of functionally recurrent MAST and Notch fusions in a subset of breast carcinomas is a promising path for future research and treatment in breast cancer and illustrates the power of next-generation sequencing as a tool in the development of personalized medicine.

## METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturemedicine/.

*Note: Supplementary information is available on the Nature Medicine website.*

### ACKNOWLEDGMENTS

We thank R. Morey for high-throughput sequencing support; T. Barrette for hardware and database management; R. Wang, N. Consul, C. Malla, L. Ma, J. Milton, L. Cai and M. Mei for technical help; and K. Suleman and W. Yan for help with cytogenetic analysis. D. Appledorn from Essen Bioscience performed the IncuCyte analyses. The aims of this project were defined by the Department of Defense Breast Cancer Research Program (W81XWH-08-0110) to A.M.C. This project was supported in part by an American Association for Cancer Research Stand Up to Cancer (SU2C) award to A.M.C. and J.S.R.-F., grant R01 CA125577 to C.G.K. The National Functional Genomics Center (W81XWH-11-1-0520), which is supported by the Department of Defense (A.M.C.) and, in part, by the US National Institutes of Health through the University of Michigan's Cancer Center Support grant 5 P30 CA46592. A.M.C. is supported by the US National Cancer Institute's Early Detection Research Network (U01 CA111275), the Doris Duke Charitable Foundation Clinical Scientist Award and the Burroughs Welcome Foundation Award in Clinical Translational Research. R.N., M.B.L. and J.S.R.-F. are funded in part by Breakthrough Breast Cancer. A.M.C. is an American Cancer Society Research professor and Taubman Scholar. C.K.S. and A.M.C. share senior authorship.

### AUTHOR CONTRIBUTIONS

D.R.R., C.K.-S. and A.M.C. conceived of the experiments. D.R.R., C.K.-S., Y.-M.W. and X.C. performed transcriptome sequencing. D.R.R., Y.-M.W. and X.C. performed target capture screening and sequencing. S.K.-S., C.A.M. and M.I. performed the bioinformatics analysis of high-throughput sequencing data and the nomination of gene fusions. C.S.G., R.J.L. and M.Q. performed bioinformatic analysis of high-throughput sequencing data for the gene expression profiling. C.K.-S., D.R.R. and Y.-M.W. performed the gene fusion validations. S.S. performed the *in vitro* experiments of MAST. I.A.A. performed the chorioallantoic membrane assays. B.A. performed the xenograft experiments. D.R.R. and Y.-M.W. performed the *in vitro* experiments of Notch. X.J. performed the microarray experiments. J.S., M.S.S., C.G.K., T.J.G., N.P., R.N., M.B.L. and J.S.R.-F. provided breast cancer tissue samples and the associated clinical annotation. N.P. performed fluorescence *in situ* hybridization experiments, and R.M. evaluated the fluorescence *in situ* hybridization results. D.R.R., C.K.-S. and A.M.C. wrote the manuscript, which was reviewed by all authors.

### COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/naturemedicine/.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Delattre, O. *et al.* Gene fusion with an ETS DNA-binding domain caused by chromosome translocation in human tumours. *Nature* **359**, 162–165 (1992).
2. Nowell, P.C. & Hungerford, D.A. Chromosome studies on normal and leukemic human leukocytes. *J. Natl. Cancer Inst.* **25**, 85–109 (1960).
3. Rowley, J.D. The critical role of chromosome translocations in human leukemias. *Annu. Rev. Genet.* **32**, 495–519 (1998).
4. Kumar-Sinha, C., Tomlins, S.A. & Chinnaiyan, A.M. Recurrent gene fusions in prostate cancer. *Nat. Rev. Cancer* **8**, 497–511 (2008).
5. Tomlins, S.A. *et al.* Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science* **310**, 644–648 (2005).
6. Prensner, J.R. & Chinnaiyan, A.M. Oncogenic gene fusions in epithelial carcinomas. *Curr. Opin. Genet. Dev.* **19**, 82–91 (2009).
7. Tognon, C. *et al.* Expression of the ETV6-NTRK3 gene fusion as a primary event in human secretory breast carcinoma. *Cancer Cell* **2**, 367–376 (2002).
8. Ding, L. *et al.* Genome remodelling in a basal-like breast cancer metastasis and xenograft. *Nature* **464**, 999–1005 (2010).
9. Shah, S.P. *et al.* Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. *Nature* **461**, 809–813 (2009).
10. Stephens, P.J. *et al.* Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature* **462**, 1005–1010 (2009).
11. Maher, C.A. *et al.* Transcriptome sequencing to detect gene fusions in cancer. *Nature* **458**, 97–101 (2009).
12. Maher, C.A. *et al.* Chimeric transcript discovery by paired-end transcriptome sequencing. *Proc. Natl. Acad. Sci. USA* **106**, 12353–12358 (2009).
13. Garland, P., Quraishe, S., French, P. & O'Connor, V. Expression of the MAST family of serine/threonine kinases. *Brain Res.* **1195**, 12–19 (2008).
14. Palomero, T. *et al.* CUTLL1, a novel human T-cell lymphoma cell line with t(7;9) rearrangement, aberrant NOTCH1 activation and high sensitivity to gamma-secretase inhibitors. *Leukemia* **20**, 1279–1287 (2006).
15. Mazzone, M. *et al.* Dose-dependent induction of distinct phenotypic responses to Notch pathway activation in mammary epithelial cells. *Proc. Natl. Acad. Sci. USA* **107**, 5012–5017 (2010).
16. Dovey, H.F. *et al.* Functional gamma-secretase inhibitors reduce beta-amyloid peptide levels in brain. *J. Neurochem.* **76**, 173–181 (2001).
17. Klinakis, A. *et al.* Myc is a Notch1 transcriptional target and a requisite for Notch1-induced mammary tumorigenesis in mice. *Proc. Natl. Acad. Sci. USA* **103**, 9262–9267 (2006).
18. Ling, H., Sylvestre, J.R. & Jolicoeur, P. Notch1-induced mammary tumor development is cyclin D1-dependent and correlates with expansion of pre-malignant multipotent duct-limited progenitors. *Oncogene* **29**, 4543–4554 (2010).

## ONLINE METHODS

**Cell lines and specimen collection.** Breast cancer cell lines were purchased from the American Type Culture Collection. The tissue was collected under approval of the University of Michigan Institutional Review Board IRBMED under approved protocol HUM00041989, and breast cancer samples were obtained with informed consent at the University of Michigan and the Breakthrough Breast Cancer Research Centre, Institute of Cancer Research (London, UK).

**Paired-end transcriptome sequencing.** Total RNA was extracted from healthy and cancer breast cell lines and breast tumor tissues, and the quality of the RNA was assessed with the Agilent Bioanalyzer. Transcriptome libraries from the mRNA fractions were generated following the RNA-Seq protocol (Illumina). Each sample was sequenced in a single lane with the Illumina Genome Analyzer II (with a 40- to 80-nt read length) or with the Illumina HiSeq 2000 (with a 100-nt read length). Paired-end transcriptome reads passing our filters were mapped to the human reference genome (hg18) and to UCSC genes using Illumina Efficient Alignment of Nucleotide Databases (ELAND) software. Sequence alignments were then processed to nominate gene fusions using a previously described method[11,12].

**qRT-PCR and long-range PCR.** qRT-PCR assays using SYBR Green Master Mix (Applied Biosystems) were carried out with the StepOne Real-Time PCR System (Applied Biosystems). Relative mRNA levels of each chimera were normalized to the expression of *GAPDH*. To detect the genomic fusion junction in HCC1187 cells, primers were designed that flanked the predicted fusion position, and PCR reactions were performed to amplify the fusion fragments. Oligonucleotide primer sequences are listed in **Supplementary Table 3**.

**Immunoblot detection of the MAST2 fusion protein and NOTCH1.** An immunoblot analysis of MAST2 was performed using an antibody to MAST2 obtained from Novus Biologicals. Antibody to human β-actin (Sigma-Aldrich) was used as a loading control. For the detection of NOTCH1, cells were lysed in radioimmunoprecipitation assay buffer containing protease inhibitor cocktail (Pierce). Proteins were separated by SDS-PAGE, transferred to nitrocellulose membranes and probed with antibodies recognizing total NOTCH1 (Cell Signaling), γ-secretase–cleaved NOTCH1 (NICD; Cell Signaling) or β-actin (Santa Cruz).

**Constructs used for overexpression studies.** The *ZNF700-MAST1* fusion ORFs from the BrCa00001 cell line were cloned into a Gateway pcDNA-DEST40 mammalian expression vector (Invitrogen) using LR Clonase II. A plasmid with a C-terminus V5 tag was generated and tested for protein expression after transfection into HEK293 cells. A full-length expression construct of *MAST2* with a DDK tag was obtained from OriGene.

**Establishment of stable pools of TERT-HME1 cells.** The five MAST fusion alleles were cloned with an N-terminal Flag epitope tag into the lentiviral vector pCDH510-B (SABiosciences). The lentivirus was produced by cotransfecting each of the MAST plasmids using the ViraPower packaging mix (Invitrogen) into 293T cells using FuGENE HD transfection reagent (Roche). Thirty-six hours after transfection, the viral supernatants were collected, centrifuged and then filtered through a 0.45-μm Steriflip filter unit (Millipore). TERT-HME1 cells were infected at a multiplicity of infection of 20 with polybrene at 8 μg ml$^{-1}$. Forty-eight hours after infection, the cells were split and placed into puromycin-selective medium. Stable pools of TERT-HME1 cells expressing the NOTCH fusion alleles as well as a control NOTCH1 intracellular domain were generated using the same procedures.

**Knockdown assay.** For siRNA knockdown experiments, multiple independent MAST2 siRNAs from Thermo were used (J-004633-06, J-004633-07 and J-004633-08). All siRNA transfections were performed using Oligofectamine reagent (Life Sciences). Similar experiments were performed with multiple custom siRNA sequences targeting the ARID1A-MAST2 fusion (Thermo). Lentiviral particles expressing the MAST2 shRNA (Sigma, TRCN0000001733) were transduced using polybrene according to the manufacturer's instructions.

**Colony formation assay.** MDA-MB-468 cells transduced with scrambled or *MAST2* shRNA lentivirus particles were plated and selected using puromycin. After 7–8 d, the plates were stained with crystal violet to visualize the number of colonies formed. For quantification of the differential staining, the plates were treated with 10% acetic acid, and absorbance was read at a wavelength of 750 nm.

**Mouse xenograft models.** Four-week-old female severe compromised immunodeficiency C.B17 mice were procured from a breeding colony at University of Michigan that is maintained by K. Pienta. Mice were anesthetized using a cocktail of xylazine (80 mg per kg of body weight intraperitoneally (i.p.)) and ketamine (10 mg per kg of body weight i.p.) for chemical restraint. Breast cancer cells with *MAST2* shRNA or scrambled shRNA knockdown ($n$ = 4 million) or the HCC1599 breast cancer cell line positive for the *NOTCH1* fusion allele ($n$ = 5 million) were resuspended in 100 μl of 1× PBS with 20% Matrigel (BD Biosciences) and implanted into the right and left abdominal inguinal mammary fat pads of the mice. Ten mice were included in each group. Two weeks after tumor implantation, HCC1599 xenografted mice were treated daily with the γ-secretase inhibitor DAPT, which was dissolved in 5% ethanol and corn oil (i.p.). All procedures involving mice were approved by the University Committee on Use and Care of Animals of the University of Michigan.

**Additional methods.** Detailed methodology is described in the **Supplementary Methods**.

# Expressed Pseudogenes in the Transcriptional Landscape of Human Cancers

Shanker Kalyana-Sundaram,[1,2,6,7] Chandan Kumar-Sinha,[1,2,7] Sunita Shankar,[1,2] Dan R. Robinson,[1,2] Yi-Mi Wu,[1,2] Xuhong Cao,[1,3] Irfan A. Asangani,[1,2] Vishal Kothari,[1] John R. Prensner,[1,2] Robert J. Lonigro,[1,2] Matthew K. Iyer,[1] Terrence Barrette,[1,2] Achiraman Shanmugam,[6] Saravana M. Dhanasekaran,[1,2] Nallasivam Palanisamy,[1,2] and Arul M. Chinnaiyan[1,2,3,4,5,*]

[1]Michigan Center for Translational Pathology
[2]Department of Pathology
[3]Howard Hughes Medical Institute
[4]Department of Urology
[5]Comprehensive Cancer Center
University of Michigan, Ann Arbor, MI 48109, USA
[6]Department of Environmental Biotechnology, Bharathidasan University, Tiruchirappalli 620 024, India
[7]These authors contributed equally to this work
*Correspondence: arul@umich.edu

DOI 10.1016/j.cell.2012.04.041

## SUMMARY

Pseudogene transcripts can provide a novel tier of gene regulation through generation of endogenous siRNAs or miRNA-binding sites. Characterization of pseudogene expression, however, has remained confined to anecdotal observations due to analytical challenges posed by the extremely close sequence similarity with their counterpart coding genes. Here, we describe a systematic analysis of pseudogene "transcription" from an RNA-Seq resource of 293 samples, representing 13 cancer and normal tissue types, and observe a surprisingly prevalent, genome-wide expression of pseudogenes that could be categorized as ubiquitously expressed or lineage and/or cancer specific. Further, we explore disease subtype specificity and functions of selected expressed pseudogenes. Taken together, we provide evidence that transcribed pseudogenes are a significant contributor to the transcriptional landscape of cells and are positioned to play significant roles in cellular differentiation and cancer progression, especially in light of the recently described ceRNA networks. Our work provides a transcriptome resource that enables high-throughput analyses of pseudogene expression.

## INTRODUCTION

Pseudogenes are ancestral copies of protein-coding genes that arise from genomic duplication or retrotransposition of mRNA sequences into the genome followed by accumulation of deleterious mutations due to loss of selection pressure, degenerating eventually into so-called genetic fossils (Sasidharan and Gerstein, 2008). Pseudogenes pervade the genome, representing virtually every coding gene, and due to their extremely close sequence similarity with their cognate genes, complicate whole-genome sequencing and gene expression analyses. A growing body of evidence strongly suggests their potential roles in regulating cognate wild-type gene expression/function by serving as a source of endogenous siRNA (Tam et al., 2008; Watanabe et al., 2008), antisense transcripts (Zhou et al., 1992), competitive inhibitors of translation of wild-type transcripts (Kandouz et al., 2004), and perhaps dominant-negative peptides (Katoh and Katoh, 2003). Pseudogene transcription has also been shown to regulate cognate wild-type gene expression by sequestering miRNAs (Poliseno et al., 2010). The recently described competing endogenous RNA (ceRNA) networks comprising sets of coordinately expressed genes with shared miRNA response elements (MREs) provide an additional dimension of (post-) transcriptional regulation in which the role of pseudogenes might overlap with those of protein-coding genes (Salmena et al., 2011; Sumazin et al., 2011).

Previous genome-wide studies of pseudogenes focused on the identification of their chromosomal coordinates and annotations based on diverse computational approaches (Karro et al., 2007; Zhang and Gerstein, 2004), including PseudoPipe (Zhang et al., 2006), HAVANA (Solovyev et al., 2006), PseudoFinder (Lu and Haussler, 2006, ASHG, conference), and Retrofinder (Zheng and Gerstein, 2006). These individual pipelines were subsequently consolidated into an integrated consensus platform, ENCyclopedia Of DNA Elements (ENCODE), which now serves as the definitive database of manually curated and annotated pseudogenes as well as pseudogene transcripts (Zheng et al., 2007). By contrast, genome-wide analyses of pseudogene expression have been somewhat arbitrary, mainly relying upon

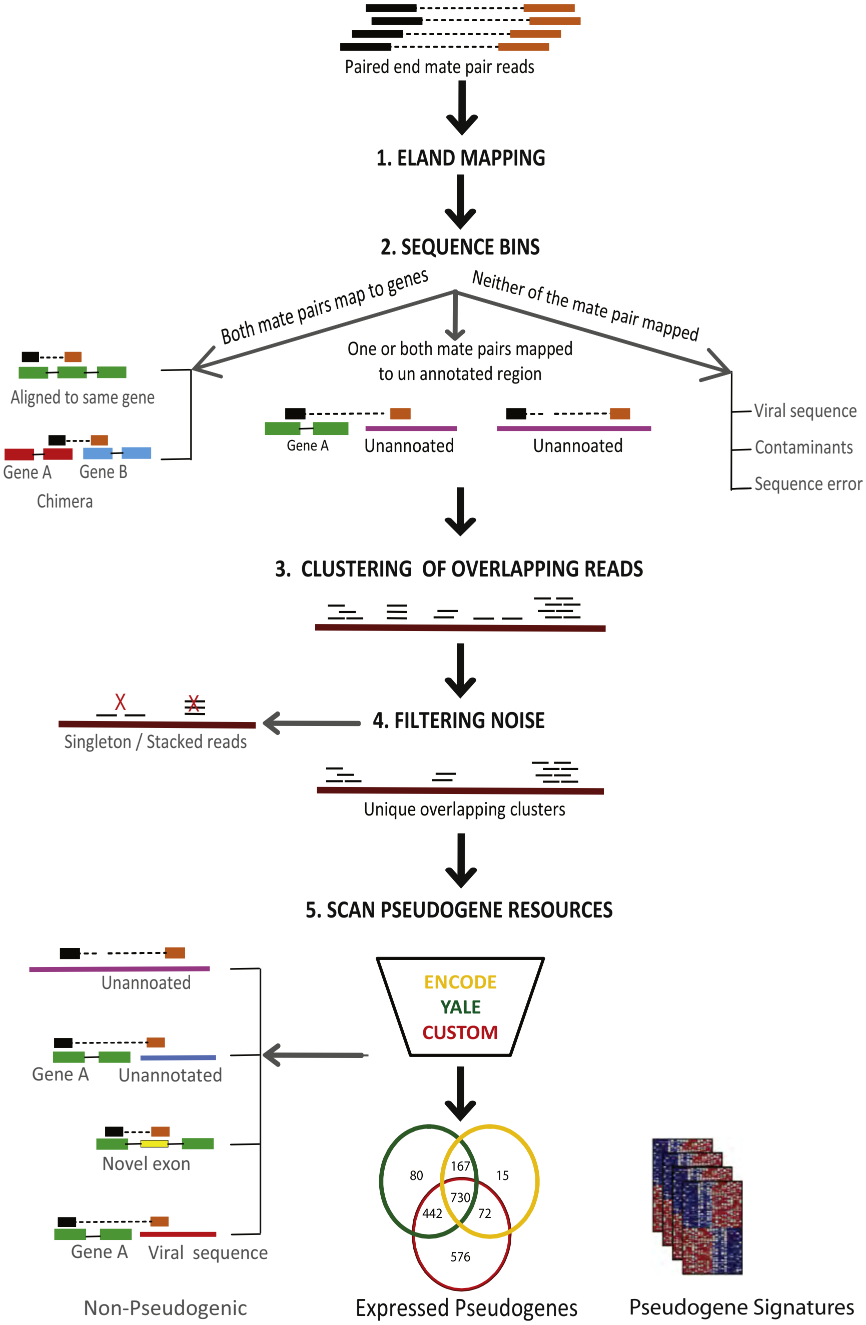


**Figure 1. Pseudogene Expression Analysis Pipeline**

The bioinformatics pipeline for analyzing pseudogene transcription involved the following steps: (1) Paired-end transcriptome sequencing reads were mapped to the human genome and UCSC Genes using ELAND. (2) Passed purity (PF) filter reads were assigned into three sequence bins as indicated. (3) Paired reads with one or both partners mapping to unannotated genomic regions were clustered based on overlapping alignments. (4) Clusters were filtered to remove singleton, stacked, and duplicate reads. (5) To determine a consensus pseudogene annotation, clusters were scanned through the Yale and ENCODE pseudogene databases as well as analyzed with a BLAT-based custom homology search. Data from individual samples were then compared to generate pseudogene expression signatures. Clusters not assigned at this stage were categorized as other potentially nonpseudogene transcripts.

See also Figures S1, S2, and S3 and Tables S1 and S2.

evidence of pseudogene transcripts obtained from disparate gene expression platforms, including public mRNA and EST databases, cap analysis gene expression (CAGE) studies, and gene identification signature-paired end tags (GIS-PET) (Ruan et al., 2007). Given the essentially anecdotal observations of pseudogene expression, only 160 expressed human pseudogenes are currently documented in ENCODE. Though this could be due to a general lack of transcription of pseudogenes, as generally presumed, it may also be reflective of an insufficient and uneven depth of coverage afforded by early gene expression analysis tools.

In this context, the recent maturation of next-generation high-throughput sequencing platforms provides unprecedented access to genome-wide expression analyses previously not achievable (Han et al., 2011a; Morozova et al., 2009). Here, we analyzed a compendium of RNA-Seq transcriptome data specifically focusing on pseudogene transcripts from a total of 293 samples encompassing 13 different tissue types, including 248 cancer and 45 benign samples. In order to carry out a systematic analysis of pseudogene expression, we developed a bioinformatics pipeline focused on detecting pseudogene transcription. This integrative approach provided evidence of expression for 2,082 distinct pseudogenes, which displayed lineage-specific, cancer-specific, as well as ubiquitous expression patterns. Taken together, this Resource nominates a multitude of expressed pseudogenes that merit further investigation to determine their roles in biology and in human disease.

## RESULTS

### Development of a Bioinformatics Platform for the Analysis of Pseudogene Transcription

Paired-end RNA-Seq data from a compendium of 293 samples, representing both cancer and benign samples from 13 different tissue types recently generated in our laboratory, was utilized to build a pseudogene analysis pipeline (Figure 1 and Figure S1 and Table S1 available online). Sequencing reads were mapped to the human genome (hg18) and University of California Santa Cruz (UCSC) Genes using Efficient Alignment of Nucleotide Databases (ELAND) software of the Illumina Genome Analyzer Pipeline (Table S2). Reads showing mismatches to the reference genes but mapping perfectly to unannotated regions elsewhere in the genome were used as the primary data for pseudogene expression analysis. Two or more unique, high-quality overlapping reads nucleating at the loci of differences between wild-type genes and pseudogenes were used to define de novo "clusters" (ranging from 40 to 5,000 bp). These clusters were employed for gene expression analyses in a way analogous to the "probes" used in microarray gene expression studies, though unlike predesigned and fixed probes used in microarrays, the sequence clusters used here were formed de novo, solely based on the presence (and levels) of transcripts. Thus, one or more clusters (like one or more probes in microarrays) represented a transcript, whereas the number of reads mapping to a cluster (analogous to fluorescence intensity due to probe hybridization on microarrays) provided a measure of expression of the corresponding (pseudo)genes. For example, Figure 2 shows a schematic representation of the cluster alignments for two representative pseudogenes, *ATP8A2-Ψ* (Figure 2A) and *CXADR-Ψ* (Figure 2B). As can be seen, mutation-dense regions in the reference sequence provide foci of pseudogene-specific cluster formation. Naturally, pseudogenes with sparse and dispersed mutations nucleate fewer clusters and require higher depth of coverage for reliable detection.

Overall, 2,156 unique pseudogene transcript clusters were identified, and their genomic coordinates (start and end points) were compared with the coordinates of pseudogenes annotated in the ENCODE (Zheng et al., 2007) and Yale pseudogene resources (http://www.pseudogene.org) (Karro et al., 2007), the two most comprehensive pseudogene annotation databases. Genomic coordinates of 934 unique pseudogene transcript clusters in our data set were found to overlap with the pseudogene coordinates annotated in *both* Yale and ENCODE databases. In addition, 585 clusters overlapped with Yale and 92 with ENCODE databases, displaying a high degree of overall concordance between our data and the authentic resources and highlighting a level of difference between the two reference databases (that necessitated our consideration of both resources). Further, as multiple clusters can sometimes represent one distinct pseudogene transcript, the 2,156 transcript clusters provided evidence for 2,082 distinct transcripts. Of these, 1,506 transcripts overlap with the genomic coordinates of pseudogenes in Yale and/or ENCODE, and up to 576 transcripts are potentially novel (described below) (Figure S2A). The 2,082 pseudogene transcripts, in turn, correspond to 1,437 wild-type genes, clearly indicating that the transcripts of multiple pseudogenes arisen from the same wild-type genes are also detected in our compendium. Taken together, our study provides evidence of widespread transcription of pseudogenes unraveled by high-throughput transcriptome sequencing (Table S3).

Pseudogene clusters across the sample-wise compendium reveal that pseudogenes of housekeeping genes such as ribosomal proteins are widely expressed across tissue types. Additionally, pseudogene transcripts corresponding to *CALM2* (calmodulin 2 phosphorylase kinase, delta), *TOMM40* (translocase of outer mitochondrial membrane 40), *NONO* (non-POU domain-containing, octamer-binding), *DUSP8* (dual-specificity phosphatase 8), *PERP* (TP53 apoptosis effector), and *YES* (v-yes-1 Yamaguchi sarcoma viral oncogene homolog 1), etc. were observed in more than 50 samples each, which were further validated by pseudogene-specific RT-PCR followed by Sanger sequencing (Table S4).

Further, because our RNA-Seq compendium comprises 35- to 45-mer short sequence reads that largely generated short sequence clusters not optimal for available pseudogene analysis tools such as Pseudopipe (Zhang et al., 2006) and Pseudofam (Lam et al., 2009) used in generating ENCODE and Yale databases, we carried out a direct query of individual clusters against the human genome (hg18) using the BLAT tool from UCSC, which is ideally suited for short sequence alignment searches (Kent, 2002). Based on this "custom" analysis, or simply BLAT (Figure S2A), we were able to independently assign 1,888 clusters representing 1,820 unique pseudogenes to unique genomic locations.

## Detection of Potentially Novel Pseudogene Transcripts

Comparing the genomic locations of the pseudogene clusters identified by BLAT analysis to those identified by Yale and ENCODE databases (Figure S2A), 762 clusters were found to be common to all three resources, but a remarkably large set of 585 clusters was uniquely defined by BLAT analysis alone. Some of the pseudogene transcripts thus identified included *BAT1*, *BTBD1*, *COX7A2L*, *CTNND1*, *EIF5*, *PAPOLA*, *PARP11*, *SYT*, *ZBTB12*, and others (n = 25) and were validated by Sanger sequencing (Table S4). Thus, analysis of RNA-Seq data provided a reliable assessment of expressed pseudogenes.

Though designating the BLAT-based pseudogene clusters as novel pseudogenes must await further sequence characterization (such as analysis of ORF structure and potential genesis of novel protein-coding gene family members, etc.), a small subset of clusters was seen to be localized in the vicinity of known pseudogenes. Thus, we found 92 clusters that resided adjacent (within 5 kb) to previously annotated pseudogenes (Figure S2B, left), and we hypothesize that these may represent pseudogenes with inaccurate annotations in the current databases. For example, the chromosomal coordinates of *CENTG2-Ψ* (OTTHUMT00000085288, Havana processed pseudogene) are defined in ENCODE as Chr1:177822463-177824935. As expected, we observed a cluster mapping to this locus; however, interestingly, we also observed a distinct cluster (Chr1:177825028-177826295) less than 100 base pairs away. Although unannotated in the current databases, the sequence of this adjacent locus shows a high degree of homology to the *CENTG2* parental gene (Figure S2B, right), strongly suggesting that this cluster represents an extension of the existing genomic coordinates of *CENTG2-Ψ* annotation. Similar observations were made with *HNRNPA1* and the *HNRNPA1-Ψ* on Chr6q27 (Figure S2B, right). 493 BLAT derived clusters that were not in close proximity to annotated pseudogenes likely represent putative pseudogenes currently missing in the database annotations (Table S3B).

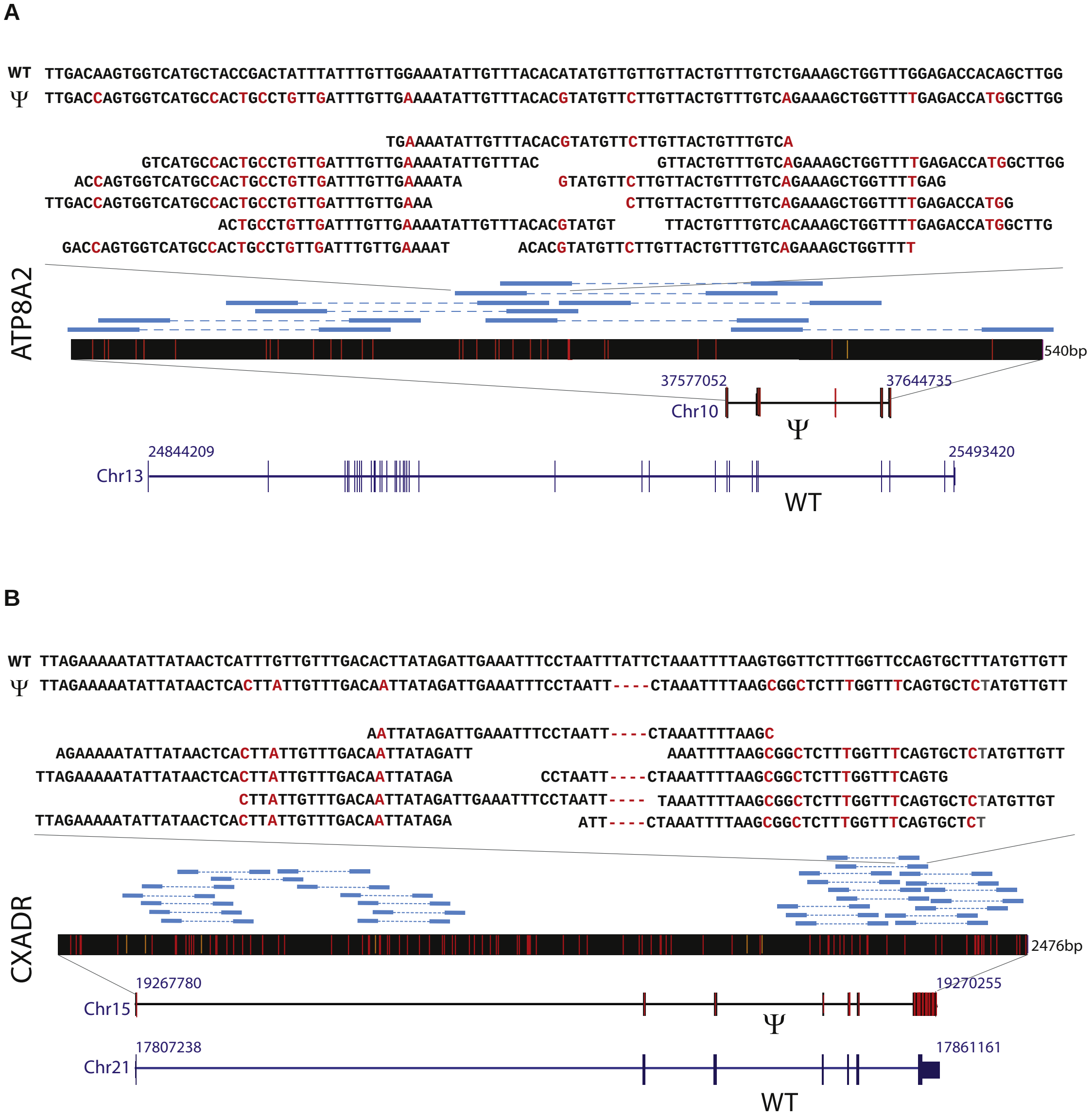

A
WT TTGACAAGTGGTCATGCTACCGACTATTTATTTGTTGGAAATATTGTTTACACATATGTTGTTGTTACTGTTTGTCTGAAAGCTGGTTTGGAGACCACAGCTTGG
Ψ TTGACCAGTGGTCATGCCACTGCCTGTTGATTTGTTGAAAATATTGTTTACACGTATGTTCTTGTTACTGTTTGTCAGAAAGCTGGTTTTGAGACCATGGCTTGG
TGAAAATATTGTTTACACGTATGTTCTTGTTACTGTTTGTCA
GTCATGCCACTGCCTGTTGATTTGTTGAAAATATTGTTTAC
GTTACTGTTTGTCAGAAAGCTGGTTTTGAGACCATGGCTTGG
ACCAGTGGTCATGCCACTGCCTGTTGATTTGTTGAAAATA
GTATGTTCTTGTTACTGTTTGTCAGAAAGCTGGTTTTGAG
TTGACCAGTGGTCATGCCACTGCCTGTTGATTTGTTGAAA
CTTGTTACTGTTTGTCAGAAAGCTGGTTTTGAGACCATGG
ACTGCCTGTTGATTTGTTGAAAATATTGTTTACACGTATGT
TTACTGTTTGTCACAAAGCTGGTTTTGAGACCATGGCTTG
GACCAGTGGTCATGCCACTGCCTGTTGATTTGTTGAAAAT
ACACGTATGTTCTTGTTACTGTTTGTCAGAAAGCTGGTTTT
ATP8A2
540bp
37577052
37644735
Chr10
Ψ
24844209
25493420
Chr13
WT
B
WT TTAGAAAAATATTATAACTCATTTGTTGTTTGACACTTATAGATTGAAATTTCCTAATTTATTCTAAATTTTAAGTGGTTCTTTGGTTCCAGTGCTTTATGTTGTT
Ψ TTAGAAAAATATTATAACTCACTTATTGTTTGACAATTATAGATTGAAATTTCCTAATT----CTAAATTTTAAGCGGCTCTTTGGTTTCAGTGCTCTATGTTGTT
AATTATAGATTGAAATTTCCTAATT----CTAAATTTTAAGC
AGAAAAATATTATAACTCACTTATTGTTTGACAATTATAGATT
AAATTTTAAGCGGCTCTTTGGTTTCAGTGCTCTATGTTGTT
TTAGAAAAATATTATAACTCACTTATTGTTTGACAATTATAGA
CCTAATT----CTAAATTTTAAGCGGCTCTTTGGTTTCAGTG
CTTATTGTTTGACAATTATAGATTGAAATTTCCTAATT----
TAAATTTTAAGCGGCTCTTTGGTTTCAGTGCTCTATGTTGT
TTAGAAAAATATTATAACTCACTTATTGTTTGACAATTATAGA
ATT----CTAAATTTTAAGCGGCTCTTTGGTTTCAGTGCTCT
CXADR
2476bp
19267780
19270255
Chr15
Ψ
17807238
17861161
Chr21
WT

**Figure 2. Schematic Representation of Cluster Alignments with Pseudogene Transcripts**

(A and B) The relative genomic structures of the parental genes are shown aligned to the respective pseudogenes, with their chromosomal locations indicated on the sides, (A) *ATP8A2*-Ψ and (B) *CXADR*-Ψ. The sequencing alterations distinguishing the pseudogene from the parental gene are indicated in red. The pseudogene transcripts are illustrated as black bars with red hatches, which indicate divergence from the parental sequence, and the length of the transcript in base pairs is shown on the side. These representations are then overlaid with schematics of paired-end reads used to form pseudogene clusters (in blue), followed by overlapping sequences in a zoomed-in region of the cluster. A comparative representation of the parental (WT) and pseudogene (Ψ) sequences for the specified region is shown on top.

See also Figure S4.

Next, we assessed the technical and analytical factors influencing the yield of pseudogene transcripts. As may be expected, a positive correlation was observed between the sequencing depth and total number of pseudogene transcripts (correlation coefficient, +0.65) (Figure S3A). However, no significant correlation was observed between the absolute measure of percent similarity between pseudogene-WT pairs and pseudogene yield. Importantly, the metric of overall percent similarity accounts for gap penalty and mismatches in BLAT search, but it is the "distribution" of the mismatches that is critical in resolving pseudogenes from nearly identical wild-type sequences; for example, a few mismatches, accumulated in a small stretch, are more effective in confidently distinguishing pseudogene expression from wild-types as compared to a higher number of mismatches that are scattered over long stretches of sequence (Figure 2). Thus, three primary factors determine the detection of pseudogene transcription by RNA-Seq: (1) the level of expression of the pseudogenes (i.e., the higher the level of expression, the higher the likelihood of detection), (2) the depth of RNA sequencing, and (3) overall distribution of mismatches with respect to the wild-type.

To explore the loci of transcription regulatory elements associated with pseudogene transcription, we carried out ChIP-Seq analysis of a breast cancer cell line MCF7 probed with H3K4me3, a histone mark associated with transcriptionally active chromosomal loci, and integrated the results with the MCF7 pseudogene transcript data. Interestingly, we observed a statistically significant enrichment of H3K4me3 peaks at expressed pseudogene loci as compared to nonexpressed pseudogenes (p = 0.0054) (Figure S3B), suggesting that the pseudogene transcripts observed by RNA-Seq are associated with transcriptionally active genomic loci. Interestingly, the pseudogene transcripts associated with H3K4me3 peaks encompass both unprocessed and processed pseudogenes, with no discernible differences in the pattern of expression. Considering the role of 3′ UTRs with MREs in ceRNA regulatory networks, we also looked at the frequency of 3′ UTR sequences retained in our set of pseudogene transcripts and observed that at least 71% of all pseudogene transcripts retain distinct 3′ UTR sequences similar to their cognate wild-type genes (Figure S3C). Interestingly, comparing the pseudogene transcripts with a list of genes implicated in ceRNA networks (Han et al., 2011b; Tay et al., 2011), we observed more than 400 overlapping transcripts (Table S5). The presence of noncoding pseudogene transcripts with similar 3′ UTRs (and MREs) adds a further level of complexity to ceRNA regulatory networks.

Next, we assessed a potential correlation between the expression of pseudogenes present within the introns of unrelated, expressed genes with their ‘‘host’’ genes. Interestingly, no significant association was observed, suggesting that pseudogenes are likely subject to independent regulatory mechanisms even when residing within other transcriptionally active genes. Further, our observations with the breast-specific unprocessed pseudogene *ATP8A2* (likely arisen from duplication of wild-type *ATP8A2*, thus likely harboring similar promoter elements) also indicate that there is no apparent correlation between the pseudogene expression with the wild-type gene that is expressed ubiquitously (described later). Thus, in summary, although it is tempting to speculate that pseudogene expression may be regulated by the promoter elements from the cognate gene or the host genes, our data suggest that more complex/indirect factors may be at play. Next, we assessed a possible correlation between the expression of pseudogenes with that of cognate wild-type genes, and intriguingly, no significant pattern of correlation was observed (Figure S3D).

Focusing on the pseudogenes whose genomic coordinates are annotated in the reference databases, we next analyzed the expression profiles of the 1,056 unique transcripts.

## Patterns of Pseudogene Expression in Human Tissues

Analyzing the expression data from 248 cancer and 45 benign samples from 13 different tissue types (total 293 samples), we observed broad patterns of pseudogene expression, including 1,056 pseudogenes that were detected in multiple samples (Table S6), which supports the hypothesis that transcribed pseudogenes contribute to the typical transcriptional repertoire of cells. In addition, we identified distinct patterns of pseudogene expression, akin to that of protein-coding genes, including 154 highly tissue/lineage-specific and 848 moderately tissue/lineage-specific (or enriched) pseudogenes (Figure 3A). Moreover, we found 165 pseudogenes exhibiting expression in more than 10 of the 13 tissue types examined, and these we classified as ubiquitous pseudogenes whose transcription is characteristic of most cell types (Figure 3A, bottom).

Of the 165 ubiquitous pseudogenes, a majority belonged to housekeeping genes, such as glyceraldehyde 3-phosphate dehydrogenase (*GAPDH*), ribosomal proteins, several cytokeratins, and other genes widely expressed in most cell types. This is expected, as these genes are known to have numerous pseudogenes, and it is likely that several of these pseudogenes retain the capacity for widespread transcription, mimicking their protein-coding counterparts.

A second set of pseudogenes exhibited near ubiquitous expression but were frequently transcribed at lower levels in most tissues and robustly transcribed in one or two tissues. These pseudogenes were termed ‘‘nonspecific,’’ and this group harbors more than 870 pseudogenes, comprising a large portion of our data set (Figure 3A, middle). Many of the pseudogenes previously shown to be expressed were found in this category, including some pseudogenes reported as tissue specific, such as *CYP4Z2P*, a pseudogene previously reported to be expressed only in breast cancer tissues (Rieger et al., 2004). Other candidates observed in this category include pseudogenes derived from *Oct-4* (Kastler et al., 2010), *Connexin-43* (Bier et al., 2009; Kandouz et al., 2004), and *BRAF* (Zou et al., 2009), among others (Table S6).

Though powerful, our approach is nevertheless limited to pseudogene transcripts that are expressed above the current threshold of detection by RNA-Seq and possess distinct stretches of sequence mismatches compared with their protein-coding parental genes. Thus, for example, *PTENP1*, a pseudogene of *PTEN* recently implicated in the biology of the phosphatidylinositol 3-kinase (PI3K) signaling pathway, was not detected in our compendium possibly due to the preponderance of cancer samples in our cohort, which tend to show low expression or deletion of this pseudogene (Poliseno et al., 2010).

## Lineage- and Cancer-Specific Pseudogene Expression Signatures

Lineage-specific pseudogene transcripts may have the potential for lineage-specific functions and may represent novel elements that facilitate biological characteristics that are unique to distinct tissue types. In this regard, we observed 154 pseudogenes with highly specific expression patterns, including pseudogenes derived from *AURKA* (kidney samples), *RHOB* (colon samples), and *HMGB1* (myeloproliferative neoplasms [MPNs]) (Figure 3A, top). Interestingly, however, lineage-specific pseudogenes tended to represent a small fraction of all pseudogenes expressed in a given tissue type, and the total number of lineage-specific pseudogenes observed in a tissue type did not show a correlation with the total number of samples analyzed. For example, B-lymphocyte cells (n = 19) and MPNs (n = 9) showed more lineage-specific pseudogenes than breast (n = 64) or prostate (n = 89). Conversely, we did observe more pseudogene

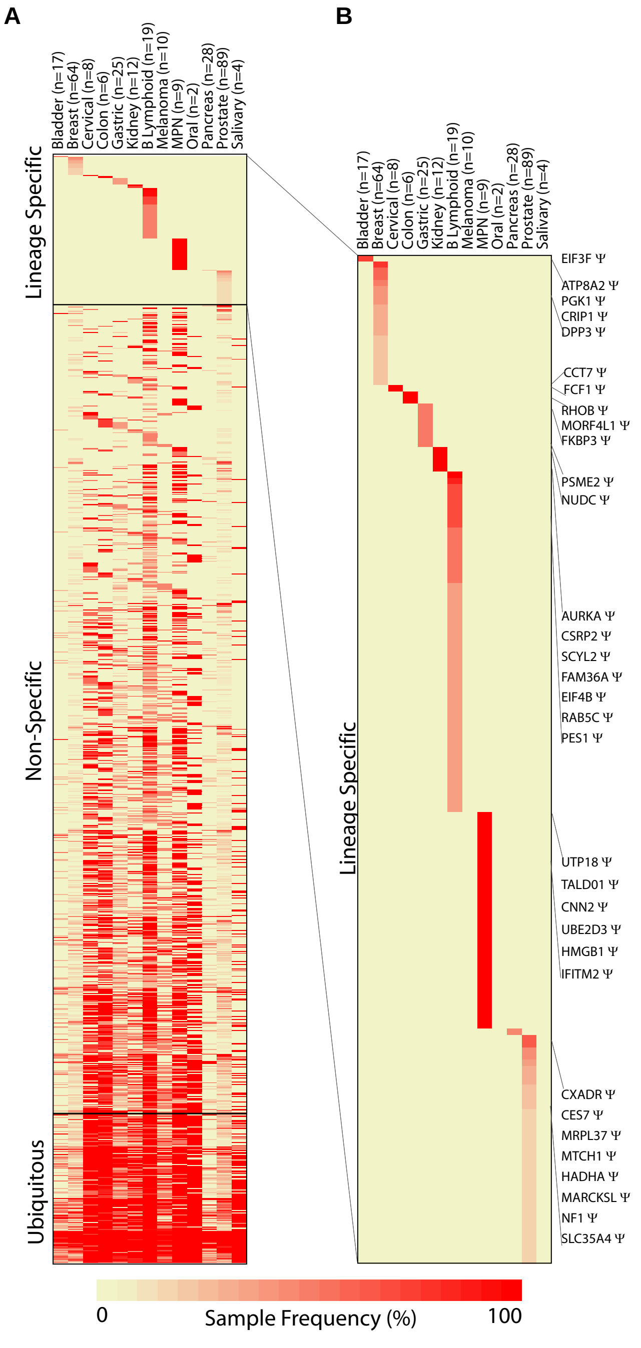


**Figure 3. Tissue/Lineage-Specific Pseudogene Expression Profiles**

(A) Heatmap of pseudogene expression sorted on the basis of tissue-specific expression displays tissue-specific (top), tissue-enriched/nonspecific (middle), and ubiquitously expressed pseudogenes (bottom).

(B) Zoomed-in version of the top panel displaying tissue-specific expressed pseudogenes. The columns represent different tissues, with the number of samples in parentheses. The rows represent individual clusters mapping to specific pseudogenes. The color intensity represents the frequency (%) of samples in a tissue type showing expression of a given pseudogenes (according to the scale indicated at the bottom). The key clusters are labeled with their corresponding parental gene symbols. MPN, myeloproliferative neoplasms.

See also Table S6.

transcripts in samples with longer read lengths and deeper coverage, as expected. Together, these data both confirm and formalize previous anecdotal observations of lineage-specific pseudogene expression patterns by exploiting the power of RNA-Seq to resolve individual transcripts (Figure 3B) (Bier et al., 2009; Lu et al., 2006; Rieger et al., 2004; Zou et al., 2009).

Because our sample compendium has a substantial number of cancer samples, we next focused on pseudogenes with cancer-specific expression. Though a majority of the pseudogenes examined were found in both cancer and benign samples, we observed 218 pseudogenes expressed only in cancer samples, of which 178 were observed in multiple cancers and 40 were found to have highly specific expression in a single cancer type only (Figure 4A and Table S7). Consistent with our previous results (Figure 3), we found that the number of cancer-type-specific pseudogenes did not correlate with the number of samples sequenced in a given cancer type. These results suggest that cancer samples harbor transcriptional patterns of pseudogenes that are both lineage and cancer specific.

Among the cancer-specific pseudogenes, a few noteworthy examples included pseudogenes derived from the eukaryotic translation initiation factors *EIF4A1* and *EIF4H*, the heterogeneous nuclear ribonucleoprotein *HNRPH2*, and the small nuclear ribonucleoprotein *SNRPG* (Figure 4B). Moreover, we observed pseudogenes corresponding to known cancer-associated genes, including *RAB-1*, a Ras-related protein; *VDAC1*, a type-1 voltage-dependent anion-selective channel/porin; *RCC2*, a regulator of chromosome condensation 2; and *PTMA*, prothymosin alpha. Interestingly, the parental protein-coding *PTMA* gene has given rise to five processed pseudogenes that retain consensus TATA elements, individual transcriptional start sites, and intact open reading frames that may potentially code for proteins closely related to the parental PTMA protein. Importantly, we find expression of *PTMA*-derived pseudogenes in more than 30 cancer samples, but not in any benign cells, and these data suggest that *PTMA*-derived pseudogenes may not only contribute transcripts to cancer cell biology but potentially proteins as well, warranting further study of these pseudogenes in tumorigenesis.

## Prostate Cancer Pseudogenes

To investigate individual pseudogenes in greater detail, we focused on pseudogenes associated with prostate and breast cancer, as our compendium has a substantial number of these two cancer types represented. Analysis of lineage-specific pseudogenes restricted to prostate cancers identified numerous pseudogenes, including several derived from parental genes known to be altered or dysregulated in cancer; for example, *NDUFA9*, which encodes an *NADH* oxidoreductase component of mitochondrial complex I that is reported to be upregulated in testicular germ cell tumors (Dormeyer et al., 2008); *EPCAM*, an epithelial cell adhesion molecule involved in cancer and stem cells signaling (Munz et al., 2009); and *CES7*, known to be expressed only in the male reproductive tract (Gang et al., 2011) (Figure 3B and Table S6). Among the prostate cancer specific pseudogenes, *CXADR*-Ψ, a processed pseudogene on chromosome 15, was of immediate interest, as the parental *CXADR* protein demonstrates putative tumor suppressor functions and its loss is implicated in α-catenin silencing (Pong et al., 2003). We therefore selected this pseudogene for further study in prostate cancer and first evaluated custom Taqman assays that could distinguish *CXADR*-Ψ from parental *CXADR*. The expression levels showed strong correlation with the RNA-Seq data (Figure S3E). *CXADR*-Ψ expression was found to be upregulated in ∼25% of prostate cancer tissues, with minimal expression seen in benign prostate samples and nonprostate tissues (Figure 5A). No correlation was observed between *CXADR*-Ψ and parental *CXADR* expression, although parental *CXADR* also had some proclivity for prostate cancer-specific expression (Figure 5B). Interestingly, *CXADR*-Ψ expression was nearly restricted to prostate cancers lacking an ETS gene fusion, with few ETS-positive samples exhibiting expression of this pseudogene. By contrast, parental *CXADR* gene expression was found in both ETS-positive and ETS-negative samples (Figure 5C). Finally, we interrogated *CXADR*-Ψ and *CXADR* parental gene expression in a set of six prostate patients with matched cancer and benign tissues (including four ETS-negative and two ETS-positive pairs). Again, ETS-negative prostate cancer samples displayed marked upregulation of *CXADR*-Ψ compared to the ETS-positive patients, with parental *CXADR* expression being fairly constant between this set of patients (Figure 5D). To establish the expression of *CXADR*-Ψ transcript, we were able to clone *CXADR*-Ψ cDNA from two RNA-Seq-positive prostate cancer samples (Figure S5A), and as predicted, these clones showed perfect sequence similarity to the pseudogene *CXADR*-Ψ and only 84% to *CXADR* wild-type gene (Figure S5B).

In the course of these analyses, we also identified a prostate-cancer-specific readthrough transcript involving *KLK4*, an androgen-induced gene, and *KLKP1*, an adjacent pseudogene. This chimeric RNA transcript *KLK4-KLKP1*, combining the first two exons of *KLK4* with the last two exons of *KLKP1*, retains an open reading frame incorporating 54 amino acids encoded by the *KLKP1* pseudogene in the putative chimeric protein (Figure S6A). Curiously, this readthrough was recently described in the prostate cancer cell line LNCaP as a *cis* sense-antisense chimeric transcript (Lai et al., 2010). Intriguingly, the *KLK4-KLKP1* transcript was highly expressed in 30%–50% of prostate cancer tissues, and this expression was lineage and cancer specific, with minimal expression seen in benign prostate and other tissues (Figure S6B). These data suggest that the *KLK4-KLKP1* may warrant further study as a potential biomarker of prostate cancer as well as a candidate protein implicated in the biological complexity of this disease.

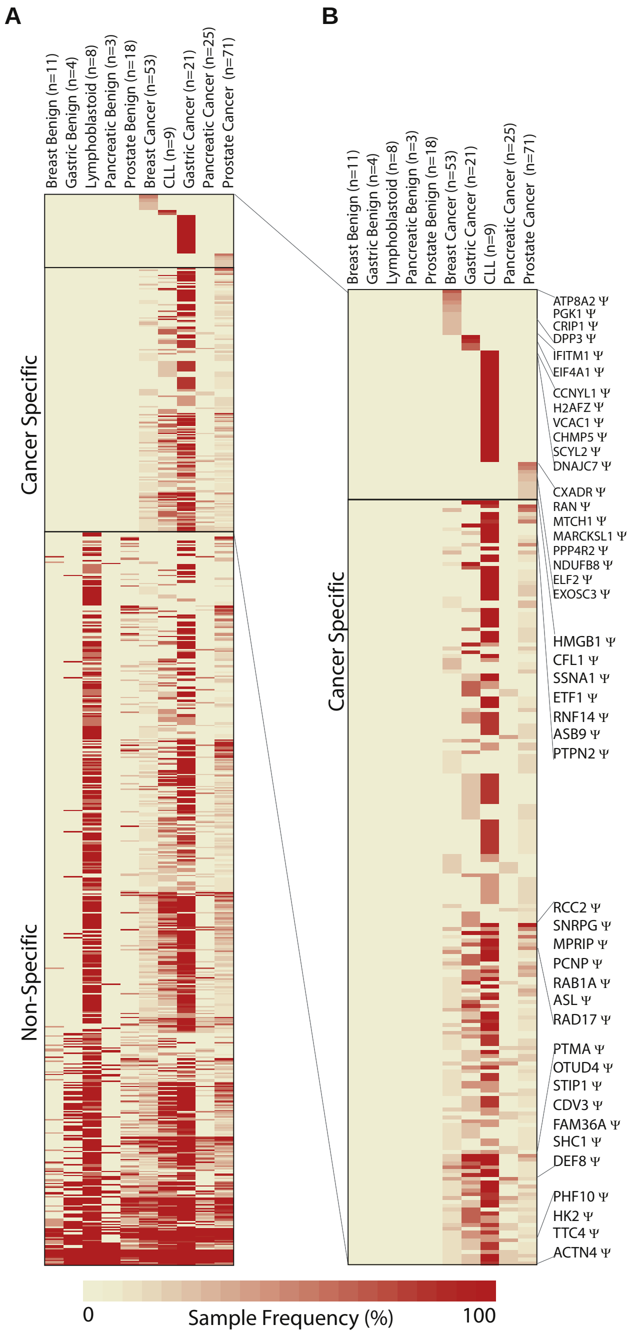


**Figure 4. Cancer-Specific Pseudogene Expression Profiles**

(A) Heatmap of pseudogene expression sorted according to cancer-specific expression patterns displays pseudogene transcripts specific to individual cancers (top), common across multiple cancers (tissue-enriched; middle), and nonspecific (bottom).

(B) Zoomed-in version of the top panel displaying individual cancer-specific expressed pseudogenes. The columns represent different tissues with the number of samples in parentheses. The rows represent individual clusters mapping to specific pseudogenes. The color intensity represents the frequency (%) of samples in a tissue type showing expression of a given pseudogenes (according to the scale indicated at the bottom). The key clusters are labeled with their corresponding parental gene symbols.

See also Figure S6 and Table S7.

## Breast Cancer Pseudogenes

Among the pseudogene candidates in breast cancer, we identified a unprocessed pseudogene cognate to *ATP8A2*, a LIM domain-containing protein speculated to be associated with stress response and proliferative activity (Khoo et al., 1997) (Figure 3A, top, and Table S3). Because *ATP8A2*-Ψ on chromosome 10 displays substantial sequence divergence from the cognate *ATP8A2*-WT gene on chromosome 13, it lends high confidence to our computational identification, and we selected this candidate for further validation. Taqman assays distinguishing *ATP8A2*-WT transcripts from *ATP8A2*-Ψ showed a strong correlation ($r^2$ = 0.98) with the expression pattern obtained

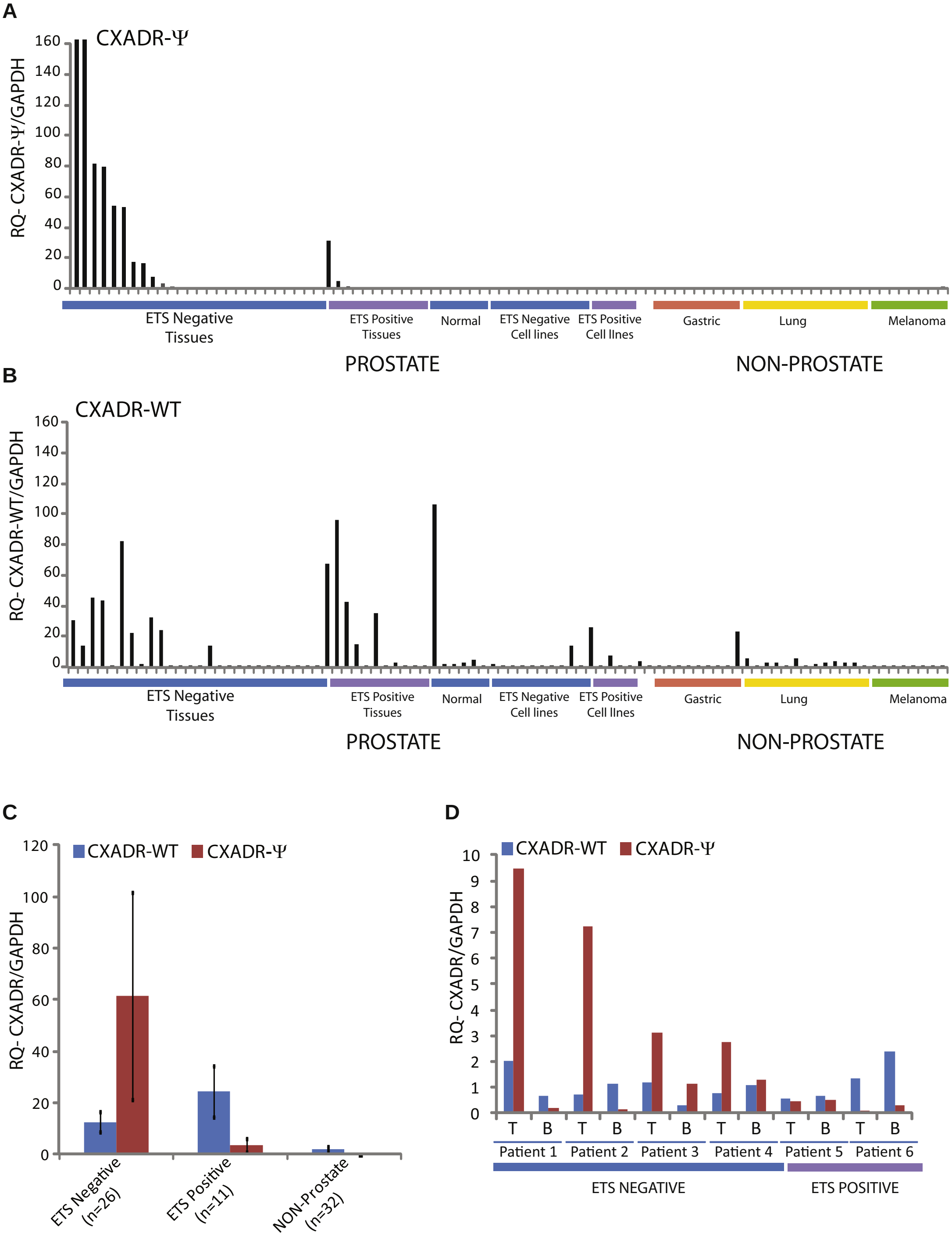


**Figure 5. Expression of *CXADR*-Ψ in Prostate Cancer**

(A and B) Histogram of expression values (y axis) of *CXADR*-Ψ (A) and *CXADR*-WT (B) across a panel of tissue samples (x axis). The order of samples on the x axis is identical in both graphs to facilitate a visual comparison.

(C) A summary histogram of the expression values of *CXADR*-Ψ and *CXADR*-WT in prostate cancers either harboring or lacking an ETS transcription factor gene fusion or in nonprostate samples.

(D) Expression of *CXADR*-Ψ and *CXADR*-WT in matched pairs of tumor and benign samples from prostate cancer patients. The patients' ETS status is indicated by the bar below.

T, prostate cancer; B, matched benign adjacent prostate. The expression values were normalized against *GAPDH*. Error bars represent means ± SE of the mean. See also Figure S5.

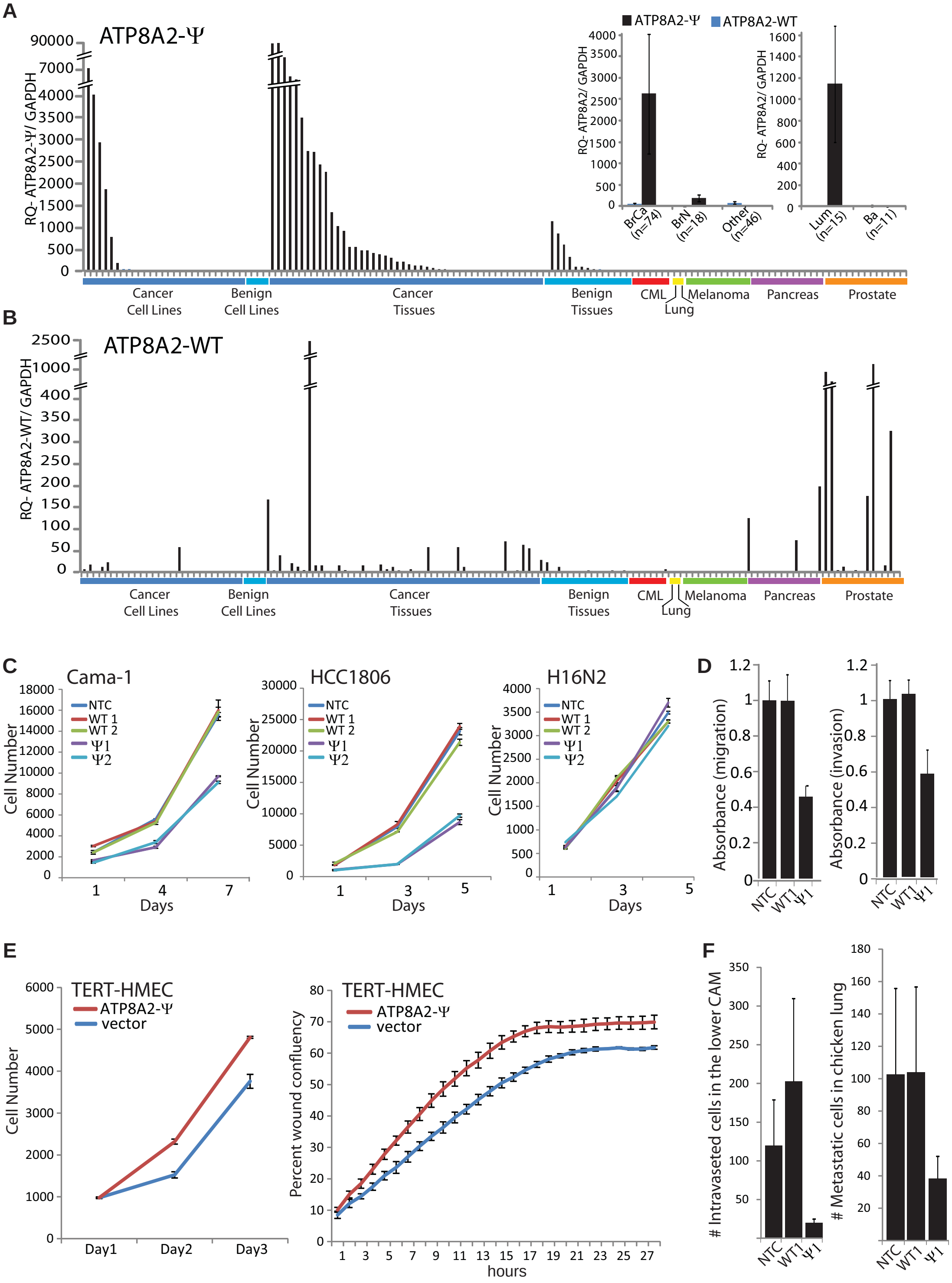


**Figure 6. Expression of *ATP8A2*-Ψ in Breast Cancer**

(A and B) Histogram of expression values (y axis) of *ATP8A2*-Ψ (A) and *ATP8A2*-WT (B) across a panel of tissue samples (x axis). The order of samples on the x axis is identical in both graphs to facilitate a visual comparison. (Inset) A summary histogram of the expression values of *ATP8A2*-Ψ and *ATP8A2*-WT in breast cancer samples relative to benign breast and other tissues (left) and luminal versus basal breast cancer subtypes (right). The expression values were normalized against *GAPDH*.

(C) Cell proliferation assays following siRNA knockdowns of *ATP8A2*-WT and -Ψ as indicated. NTC, nontargeting control; WT, siRNA against wild-type *ATP8A2*; Ψ, siRNA against *ATP8A2*-Ψ.

by RNA-Seq (Figure S3E), with *ATP8A2*-Ψ expression found to be restricted to breast samples, the highest levels seen in a subset of breast cancer tissues and cell lines (Figures 6A and 6B). By contrast, *ATP8A2*-WT expression was highly variable across different tissue types and showed no correlation with *ATP8A2*-Ψ expression (Figure 6B).

We were further intrigued by the pattern of *ATP8A2*-Ψ expression within breast tumors, where ∼25% of tumors demonstrate extremely high levels of this pseudogene, suggesting that *ATP8A2*-Ψ may contribute to a particular subtype of breast cancer. We therefore analyzed *ATP8A2*-Ψ expression with respect to luminal and basal breast subtypes, two prominent categories of breast cancer with distinct molecular and clinical characteristics. Unexpectedly, we found that *ATP8A2*-Ψ expression was restricted to tumors with luminal histology, whereas basal tumors showed minimal expression of this pseudogene (Figure 6A, right). The wild-type *ATP8A2* transcript did not display this pattern of expression.

To investigate a potential role of *ATP8A2*-Ψ expression in breast cancer, first we carried out siRNA-based knockdown of both the wild-type and pseudogene RNA in two independent breast cancer cell lines that expressed both the transcripts (Figure S7A). Knockdown of *ATP8A2*-Ψ with two independent siRNAs was found to specifically inhibit the proliferation of overexpressing cell lines Cama-1 and HCC1806 (Figure 6C), but not the cell lines with no detectable levels of *ATP8A2*-Ψ, for example, the benign breast epithelial cell line H16N2 (Figure 6C, right) and a pancreatic cancer cell line, BXPC3 (Figure S7D). Knockdown of *ATP8A2*-Ψ (but not *ATP8A2*-WT) also resulted in reduced cell migration and invasion seen in in vitro Boyden Chamber assays (Figure 6D) as well as in in vivo intravasation and metastasis in chicken chorioallantoic membrane xenograft assay (Figure 6F). In contrast, knockdown of wild-type *ATP8A2* had no effect on the proliferation of any of the cell lines tested, suggesting an unexpected growth regulatory role for *ATP8A2*-Ψ (Figure 6C). Surprisingly, though the knockdown of wild-type *ATP8A2* had a minimal effect on the pseudogene transcript levels, *ATP8A2*-Ψ-specific siRNAs, apart from reducing the *ATP8A2*-Ψ transcript, also reduced the wild-type protein levels (Figures S7C and S7E). Thus clearly, unlike *Oct4* and *BRAF* pseudogene transcripts having an inverse correlation with the wild-type transcript levels, *ATP8A2*-Ψ and wild-type *ATP8A2* transcripts (Figures 6A and 6B) and protein (Figure S7E) do not seem to be regulated in this manner. Subsequently, to assess the phenotypic effect of *ATP8A2*-Ψ overexpression in benign cells, we cloned and overexpressed the full-length *ATP8A2* pseudogene cDNA in benign breast epithelial cell line TERT-HMEC. Two independent pooled populations of *ATP8A2*-Ψ-overexpressing TERT-HMEC cells were found to undergo increased proliferation and migration (Figure 6E), indicating the potential oncogenic nature of this breast-specific pseudogene transcript.

## DISCUSSION

The recent advances in high-throughput transcriptome sequencing have revealed widespread expression of noncoding RNAs in the context of development and differentiation (Khachane and Harrison, 2010; Nagalakshmi et al., 2008; Pickrell et al., 2010; Prensner et al., 2011; Wilhelm et al., 2008). These studies, however, do not include pseudogene expression analyses in their purview, likely due to the challenge of extremely close sequence similarity with wild-type cognate genes. Here, we interrogated the potential of RNA-Seq data to unambiguously detect pseudogene transcripts and to assess whether pseudogene expression is more common in the transcriptome than previously realized. Surprisingly, we found evidence of a widespread expression of pseudogenes in our cancer transcriptome resource, including 1,500 pseudogenes annotated in the Yale and ENCODE databases, redefined the genomic coordinates of ∼100 pseudogenes in existing databases, and nominated more than 400 potentially novel pseudogenes. In aggregate, our analysis considerably expands the spectrum of expressed pseudogenes documented previously (Harrison et al., 2005; Yao et al., 2006; Zheng et al., 2007).

The extreme sequence similarity between pseudogenes and cognate wild-type genes suggests a functional role for pseudogene transcripts; indeed, pseudogene expression has been associated with both downregulation of cognate wild-type gene, such as *eNOS* in ovary, as well as a positive effect on the expression of the wild-type gene, as demonstrated recently, wherein *PTENP1* expression upregulates *PTEN* expression in prostate cells (Poliseno et al., 2010). Interestingly, a class of pseudogenes called ''unitary pseudogenes'' does not have extant cognate wild-type genes (Zhang et al., 2010). Nevertheless, as most pseudogenes do have distinct cognate wild-type genes, we assessed the correlation between expressed pseudogenes and their cognate wild-type genes across multiple samples (of the same tissue type or across diverse tissue types) and did not observe a statistically significant correlation. This is not surprising, partly because our data set is comprised of a heterogeneous set of samples representing diverse tissue types. Further, the sensitivity of detection of individual pseudogene transcripts is limited by the degree and distribution of dissimilarity with the wild-type gene that determines the ''effective'' depth of coverage; this limits the number of samples showing measurable expression of individual pseudogene-wild-type pairs, making it difficult to conduct robust statistical analyses. Future studies involving larger sample sets with higher depth of coverage and longer read length may be better able to resolve this question.

Taken together, our study provides a systematic approach to analyze expressed pseudogenes using RNA-Seq data, enabling comparisons of cancer versus benign tissues in multiple solid

(D) Boyden chamber assay showing cell migration (left) and invasion through matrigel (right).

(E and F) (E) The effect of *ATP8A2*-Ψ overexpression in TERT-HMEC cells on cell proliferation (left) and cell migration based on Incucyte wound confluency assay (right) and (F) chicken chorioallantoic membrane assay of HCC-1806 cells treated with nontargeting control siRNA, *ATP8A2*-WT, or *ATPA2*-Ψ siRNA showing relative number of cells intravasated in the lower CAM (left) and metastatic cells in chicken lung (right).

Error bars represent means ± SE of the mean.

tumors. Our efforts lend additional credence to the capacity of RNA-Seq to "re-define" the functional elements of the genome and "re-annotate" the population of pseudogenes implicated in human cell biology. Our approach overcomes the limitations of previous analyses of pseudogene expression, which were primarily anecdotal and heterogeneous in nature, and our methodologies suggest avenues to reconcile the difficulty in distinguishing pseudogene expression from parental protein-coding gene expression—a facet that is important for all RNA-Seq studies aiming to provide an accurate picture of gene expression. Finally, we describe *ATP8A2*-Ψ and *CXADR*-Ψ pseudogenes preferentially associated with distinct subsets of breast cancer and prostate cancer patients, respectively.

The recent description of intricate regulatory networks of protein-coding transcripts called competitive endogenous RNAs (ceRNAs) defined on the basis of coordinated regulation by common sets of microRNA response elements (MREs)—first intimated by Salmena et al. (Salmena et al., 2011) and subsequently supported by experimental results from multiple groups(Cesana et al., 2011; Han et al., 2011b; Karreth et al., 2011; Tay et al., 2011)—implicates potential noncoding functions for many protein-coding transcripts. In this context, pseudogene transcripts could provide an additional layer of complexity in conjunction with their cognate wild-type genes or independently.

The cancer/tissue-specific pseudogene expression signatures described here highlight the need to factor in pseudogene expression in all high-throughput gene expression studies and also show that pseudogene expression merits further exploration in its own right as an additional layer of transcriptional complexity. To facilitate further analyses, we provide here an extensive resource of RNA-Seq data of human cancer-related tissues and cell lines.

## EXPERIMENTAL PROCEDURES

### Data Set

Paired-end transcriptome sequence reads (2 × 40 and 2 × 80 base pairs) were obtained from a total of more than 293 samples from 13 tissue types (Figure S1 and Table S1). Each sample was sequenced on an Illumina Genome Analyzer I or II according to protocols provided by Illumina as described earlier (Palanisamy et al., 2010).

### Pseudogene Analysis Pipeline

Paired-end transcriptome reads were mapped to the human genome (NCBI36/hg18) and University of California Santa Cruz (UCSC) Genes using Efficient Alignment of Nucleotide Databases (ELAND) software of the Illumina Genome Analyzer Pipeline, using 32 bp seed length and allowing up to two mismatches; detailed mapping status is represented in Table S2. Passed purity filter reads obtained from Illumina export and extended output files (as described before) were parsed and binned into three major categories: (1) both of the paired reads map to annotated genes; (2) one or both of the paired reads map to unannotated regions in the genome; and (3) neither of the reads map (these include viral, bacterial, and other contaminant reads, as well as sequencing errors). The paired reads with one or both partners mapping to an unannotated region were clustered based on overlaps of aligned sequences using the chromosomal coordinates of the clusters. Singleton reads that did not cluster or stacked\duplicated reads with the same start and stop genomic coordinates (potential PCR artifacts) were filtered out. Passed filter clusters were defined as units of transcript expression (analogous to a "probe" on microarray platforms). These clusters were screened against two human pseudogene resources, Yale human pseudogene (Build 53, http://pseudogene.org/) (Karro et al., 2007) and Gencode (October 2009, http://genome.ucsc.edu/ENCODE/) (Zheng et al., 2007), to identify and annotate pseudogene clusters. The processed, duplicated, and fragmented categories of pseudogene entries from Yale and the entries corresponding to Level 1+2 (Manual Gene Annotations) and Level 3 (Automated Gene Annotations) from Gencode were used. The clusters were also subjected to homology search using the alignment tool BLAT (http://www.soe.ucsc.edu/~kent) (Kent, 2002) for an independent annotation. Sequence reads from individual samples were queried against the resultant clusters defined by the union of Yale, ENCODE, and BLAT output to assess the expression of pseudogenes (Figure 1 and Table S3). The cutoff value for pseudogene expression in a sample was set at five or more reads mapping to at least one cluster in a putative pseudogene transcript. Pseudogene transcripts (one or more probes overlapping with either Yale or ENCODE) detected in two or more samples in a tissue type and absent in all other tissue types were defined as tissue/lineage specific. Pseudogene probes detected in 10 out of 13 samples were designated as ubiquitous. All other cases were described as an intermediate category. Pseudogene transcripts detected in three or more cancer samples and absent in all benign samples were designated as cancer specific.

We carried out multiple correlation analyses (Figure S3), including: (1) passed filter reads (sequence yield) with total number of pseudogene transcripts observed per sequencing run (pseudogene transcript coverage); (2) expression of genes and pseudogenes carried out using 173 gene-pseudogene pairs in 64 samples that each show nonzero expression in at least ten samples across the data set; (3) expression levels of *ATP8A2* and *CXADR* pseudogenes obtained from RNA-Seq and qPCR; (4) ChIP-Seq analysis of a breast cancer cell line MCF7 that was probed with H3K4me3 and compared with MCF7 pseudogene transcript data; and (5) pseudogene transcripts with 3′ UTR sequences (± 2 kb) that were compared with 3′ UTR sequences of their cognate genes using BLAT.

Pseudogene transcripts showing an overlap with transcripts involved in ceRNA network genes reported previously were tabulated (Sumazin et al., 2011 and Tay et al., 2011) (Table S5). The entire sequence data set will be submitted to dbGAP after securing requisite approvals.

### RNA Isolation and cDNA Synthesis

Total RNA was isolated using Trizol and an RNeasy Kit (Invitrogen) with DNase I digestion according to the manufacturer's instructions. RNA integrity was verified on an Agilent Bioanalyzer 2100 (Agilent Technologies, Palo Alto, CA). cDNA was synthesized from total RNA using Superscript III (Invitrogen) and random primers (Invitrogen).

### Quantitative Real-Time PCR

Quantitative real-time PCR (qPCR) was performed using Taqman or SYBR green-based assays (Applied Biosystems, Foster City, CA) on an Applied Biosystems 7900HT Real-Time PCR System, according to standard protocols. The Taqman assays for *CXADR* and *ATP8A2* assays were custom designed based on regions of differences between the wild-type and pseudogene sequences (Figure S4). Oligonucleotide primers for SYBR green assays were obtained from Integrated DNA Technologies (Coralville, IA). The housekeeping gene *GAPDH* was used as a loading control. Fold changes were calculated relative to *GAPDH* and normalized to the median value of the benign samples.

*CXADR*-Ψ_F CGGTTTCAGTGCTCTATGTTGTTTG; *CXADR*-Ψ_R TAAATTTAGGATTACATGTTTCTAGAACA; *CXADR*-Ψ_M 6FAM ATGCCATCCAAAACCA; *ATP8A2*-Ψ_F CTGGTGTTCTTTGGCATCTACTCA; *ATP8A2*-Ψ_R CAGCTCAGGATCACAGTTGCT; *ATP8A2*-Ψ_M 6FAM CTGGTCCACCATTCTC; *ATP8A2*-WT_F ATCCTATTGAAGGAGGACTCTTTGGA; *ATP8A2*-WT_R CCAGCAAATTCCCAAGGTCAGT; *ATP8A2*-WT_M 6FAM AAGGGCAGCCATTACT; *KLK4-KLKP1*_F ATGGAAAACGAATTGTTCTG; and *KLK4-KLKP1*_R CAGTGTTCCGGGTGATGCAG.

Additionally, inventoried Taqman assays for *CXADR*-WT (Hs00154661_m1) and *ATP8A2*-WT (assay ID hs00185259_m1) were used.

### RT-PCR and Sanger Sequencing

Sequence stretches unique to pseudogene transcripts were identified by aligning the candidate pseudogene sequences with their corresponding wild-type genes. PCR primers specific to pseudogene transcripts (Table S4) were used to amplify pseudogene cDNAs from index samples followed by Sanger sequencing of the PCR products. The resultant sequences were analyzed using ClustalW to compare the identity between pseudogene and cognate wild-type sequences.

### Cell Proliferation Assays

Experimental cells were transfected with siRNAs using oligofectamine reagent (Life Sciences), and 3 days posttransfection, the cells were plated for proliferation assays. At the indicated times, cell numbers were measured using Coulter Counter.

### Wound Healing Assay Using Incucyte

For the wound healing assay, vector control or *ATP8A2* pseudogene-overexpressing cells were plated at high density, and 6 hr later, uniform scratch wounds were made using Woundmaker (Incucyte). Relative migration potential of the cells was assessed by confluence measurements at regular time intervals as indicated over the wound area.

### *ATP8A2* Pseudogene Overexpression Studies

The *ATP8A2* pseudogene cDNA from breast cancer cell line HCC1806 was cloned into pENTR-D-TOPO entry vector (Invitrogen) following manufacturer's instructions. Sequence-confirmed entry clones in correct orientation were recombined into Gateway pcDNA-DEST26 mammalian expression vector (Invitrogen) by LR Clonase II enzyme reaction following manufacturer's instructions. HMEC-TERT cells were transfected using Fugene 6, and polyclonal populations of cells expressing *ATP8A2* pseudogene cDNA or empty vector constructs were selected using geneticin. At the indicated times, cell numbers were measured using Coulter Counter.

### Chicken Chorioallantoic Membrane Assay

Chicken chorioallantoic membrane (CAM) assay for tumor growth was carried out as follows. Fertilized eggs were incubated in a humidified incubator at 38°C for 10 days, and then CAM was dropped by drilling two holes: a small hole through the eggshell into the air sac and a second hole near the allantoic vein that penetrates the eggshell membrane but not the CAM. Subsequently, a cutoff wheel (Dremel) was used to cut a 1 $cm^2$ window encompassing the second hole near the allantoic vein to expose the underlying CAM. When ready, CAM was gently abraded with a sterile cotton swab to provide access to the mesenchyme, and $2 \times 10^6$ cells in 50 μl volume were implanted on top. The windows were subsequently sealed and the eggs returned to the incubator. After 7 days, extraembryonic tumors were isolated and weighed. Five to ten eggs per group were used in each experiment.

## SUPPLEMENTAL INFORMATION

Supplemental Information includes seven figures and seven tables and can be found with this article online at doi:10.1016/j.cell.2012.04.041.

## ACKNOWLEDGMENTS

The authors thank Ram-Shankar Mani, Kalpana Ramnarayanan, Joseph Mierwaza, and Sooryanarayana Varambally for technical help and Javed Siddiqui for providing samples. This work was supported, in part, by the National Institutes of Health (R01CA132874), Early Detection Research Network grant UO1 CA111275, Prostate SPORE grant P50CA69568, and the Department of Defense Era of Hope grant BC075023 (A.M.C.). A.M.C. is supported by the Doris Duke Charitable Foundation Clinical Scientist Award and the Prostate Cancer Foundation. A.M.C. is an American Cancer Society Research Professor and A. Alfred Taubman Scholar.

S.K.-S., C.K.-S., and A.M.C. conceived the experiments. S.K.-S. designed and developed the pseudogene analysis pipeline and carried out bioinformatics analyses. C.K.-S., D.R.R., N.P., X.C., and Y.-M.W. performed transcriptome sequencing. S.K.-S. and T.B. carried out ELAND mapping and provided database support. C.K.-S., V.K., and S.M.D. carried out pseudogene validations. S.S., C.K.-S., S.M.D., and I.A.A. performed the experiments. J.R.P and A.S. helped in manuscript review. S.K.-S. and R.J.L. carried out gene-pseudogene correlation analyses. S.K.-S. and M.K.I. performed ChIP-Seq analysis. C.K.-S., S.K.-S., and A.M.C. wrote the manuscript, which was reviewed by all authors.

Received: June 6, 2011
Revised: January 10, 2012
Accepted: April 26, 2012
Published: June 21, 2012

## REFERENCES

Bier, A., Oviedo-Landaverde, I., Zhao, J., Mamane, Y., Kandouz, M., and Batist, G. (2009). Connexin43 pseudogene in breast cancer cells offers a novel therapeutic target. Mol. Cancer Ther. *8*, 786–793.

Cesana, M., Cacchiarelli, D., Legnini, I., Santini, T., Sthandier, O., Chinappi, M., Tramontano, A., and Bozzoni, I. (2011). A long noncoding RNA controls muscle differentiation by functioning as a competing endogenous RNA. Cell *147*, 358–369.

Dormeyer, W., van Hoof, D., Braam, S.R., Heck, A.J., Mummery, C.L., and Krijgsveld, J. (2008). Plasma membrane proteomics of human embryonic stem cells and human embryonal carcinoma cells. J. Proteome Res. *7*, 2936–2951.

Gang, L., Janecka, J.E., and Murphy, W.J. (2011). Accelerated evolution of *CES7*, a gene encoding a novel major urinary protein in the Cat family. Mol. Biol. Evol. *28*, 911–920.

Han, H., Nutiu, R., Moffat, J., and Blencowe, B.J. (2011a). SnapShot: High-throughput sequencing applications. Cell *146*, 1044.

Han, Y.J., Ma, S.F., Yourek, G., Park, Y.D., and Garcia, J.G. (2011b). A transcribed pseudogene of MYLK promotes cell proliferation. FASEB J. *25*, 2305–2312.

Harrison, P.M., Zheng, D., Zhang, Z., Carriero, N., and Gerstein, M. (2005). Transcribed processed pseudogenes in the human genome: an intermediate form of expressed retrosequence lacking protein-coding ability. Nucleic Acids Res. *33*, 2374–2383.

Kandouz, M., Bier, A., Carystinos, G.D., Alaoui-Jamali, M.A., and Batist, G. (2004). Connexin43 pseudogene is expressed in tumor cells and inhibits growth. Oncogene *23*, 4763–4770.

Karreth, F.A., Tay, Y., Perna, D., Ala, U., Tan, S.M., Rust, A.G., DeNicola, G., Webster, K.A., Weiss, D., Perez-Mancera, P.A., et al. (2011). In vivo identification of tumor- suppressive PTEN ceRNAs in an oncogenic BRAF-induced mouse model of melanoma. Cell *147*, 382–395.

Karro, J.E., Yan, Y., Zheng, D., Zhang, Z., Carriero, N., Cayting, P., Harrrison, P., and Gerstein, M. (2007). Pseudogene.org: a comprehensive database and comparison platform for pseudogene annotation. Nucleic Acids Res. *35* (Database issue), D55–D60.

Kastler, S., Honold, L., Luedeke, M., Kuefer, R., Möller, P., Hoegel, J., Vogel, W., Maier, C., and Assum, G. (2010). POU5F1P1, a putative cancer susceptibility gene, is overexpressed in prostatic carcinoma. Prostate *70*, 666–674.

Katoh, M., and Katoh, M. (2003). IGSF11 gene, frequently up-regulated in intestinal-type gastric cancer, encodes adhesion molecule homologous to CXADR, FLJ22415 and ESAM. Int. J. Oncol. *23*, 525–531.

Kent, W.J. (2002). BLAT—the BLAST-like alignment tool. Genome Res. *12*, 656–664.

Khachane, A.N., and Harrison, P.M. (2010). Mining mammalian transcript data for functional long non-coding RNAs. PLoS ONE *5*, e10316.

Khoo, C., Blanchard, R.K., Sullivan, V.K., and Cousins, R.J. (1997). Human cysteine-rich intestinal protein: cDNA cloning and expression of recombinant protein and identification in human peripheral blood mononuclear cells. Protein Expr. Purif. *9*, 379–387.

Lai, J., Lehman, M.L., Dinger, M.E., Hendy, S.C., Mercer, T.R., Seim, I., Lawrence, M.G., Mattick, J.S., Clements, J.A., and Nelson, C.C. (2010). A variant of the KLK4 gene is expressed as a cis sense-antisense chimeric transcript in prostate cancer cells. RNA *16*, 1156–1166.

Lam, H.Y., Khurana, E., Fang, G., Cayting, P., Carriero, N., Cheung, K.H., and Gerstein, M.B. (2009). Pseudofam: the pseudogene families database. Nucleic Acids Res. *37* (Database issue), D738–D743.

Lu, W., Zhou, D., Glusman, G., Utleg, A.G., White, J.T., Nelson, P.S., Vasicek, T.J., Hood, L., and Lin, B. (2006). KLK31P is a novel androgen regulated and transcribed pseudogene of kallikreins that is expressed at lower levels in prostate cancer cells than in normal prostate cells. Prostate *66*, 936–944.

Morozova, O., Hirst, M., and Marra, M.A. (2009). Applications of new sequencing technologies for transcriptome analysis. Annu. Rev. Genomics Hum. Genet. *10*, 135–151.

Munz, M., Baeuerle, P.A., and Gires, O. (2009). The emerging role of EpCAM in cancer and stem cell signaling. Cancer Res. *69*, 5627–5629.

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., and Snyder, M. (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. Science *320*, 1344–1349.

Palanisamy, N., Ateeq, B., Kalyana-Sundaram, S., Pflueger, D., Ramnarayanan, K., Shankar, S., Han, B., Cao, Q., Cao, X., Suleman, K., et al. (2010). Rearrangements of the RAF kinase pathway in prostate cancer, gastric cancer and melanoma. Nat. Med. *16*, 793–798.

Pickrell, J.K., Marioni, J.C., Pai, A.A., Degner, J.F., Engelhardt, B.E., Nkadori, E., Veyrieras, J.B., Stephens, M., Gilad, Y., and Pritchard, J.K. (2010). Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature *464*, 768–772.

Poliseno, L., Salmena, L., Zhang, J., Carver, B., Haveman, W.J., and Pandolfi, P.P. (2010). A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. Nature *465*, 1033–1038.

Pong, R.C., Lai, Y.J., Chen, H., Okegawa, T., Frenkel, E., Sagalowsky, A., and Hsieh, J.T. (2003). Epigenetic regulation of coxsackie and adenovirus receptor (CAR) gene promoter in urogenital cancer cells. Cancer Res. *63*, 8680–8686.

Prensner, J.R., Iyer, M.K., Balbin, O.A., Dhanasekaran, S.M., Cao, Q., Brenner, J.C., Laxman, B., Asangani, I.A., Grasso, C.S., Kominsky, H.D., et al. (2011). Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression. Nat. Biotechnol. *29*, 742–749.

Rieger, M.A., Ebner, R., Bell, D.R., Kiessling, A., Rohayem, J., Schmitz, M., Temme, A., Rieber, E.P., and Weigle, B. (2004). Identification of a novel mammary-restricted cytochrome P450, CYP4Z1, with overexpression in breast carcinoma. Cancer Res. *64*, 2357–2364.

Ruan, Y., Ooi, H.S., Choo, S.W., Chiu, K.P., Zhao, X.D., Srinivasan, K.G., Yao, F., Choo, C.Y., Liu, J., Ariyaratne, P., et al. (2007). Fusion transcripts and transcribed retrotransposed loci discovered through comprehensive transcriptome analysis using Paired-End diTags (PETs). Genome Res. *17*, 828–838.

Salmena, L., Poliseno, L., Tay, Y., Kats, L., and Pandolfi, P.P. (2011). A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language? Cell *146*, 353–358.

Sasidharan, R., and Gerstein, M. (2008). Genomics: protein fossils live on as RNA. Nature *453*, 729–731.

Solovyev, V., Kosarev, P., Seledsov, I., and Vorobyev, D. (2006). Automatic annotation of eukaryotic genes, pseudogenes and promoters. Genome Biol. *7* (*Suppl 1*), S10.1–S10.12.

Sumazin, P., Yang, X., Chiu, H.S., Chung, W.J., Iyer, A., Llobet-Navas, D., Rajbhandari, P., Bansal, M., Guarnieri, P., Silva, J., and Califano, A. (2011). An extensive microRNA-mediated network of RNA-RNA interactions regulates established oncogenic pathways in glioblastoma. Cell *147*, 370–381.

Tam, O.H., Aravin, A.A., Stein, P., Girard, A., Murchison, E.P., Cheloufi, S., Hodges, E., Anger, M., Sachidanandam, R., Schultz, R.M., and Hannon, G.J. (2008). Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. Nature *453*, 534–538.

Tay, Y., Kats, L., Salmena, L., Weiss, D., Tan, S.M., Ala, U., Karreth, F., Poliseno, L., Provero, P., Di Cunto, F., et al. (2011). Coding-independent regulation of the tumor suppressor PTEN by competing endogenous mRNAs. Cell *147*, 344–357.

Watanabe, T., Totoki, Y., Toyoda, A., Kaneda, M., Kuramochi-Miyagawa, S., Obata, Y., Chiba, H., Kohara, Y., Kono, T., Nakano, T., et al. (2008). Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes. Nature *453*, 539–543.

Wilhelm, B.T., Marguerat, S., Watt, S., Schubert, F., Wood, V., Goodhead, I., Penkett, C.J., Rogers, J., and Bähler, J. (2008). Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. Nature *453*, 1239–1243.

Yao, A., Charlab, R., and Li, P. (2006). Systematic identification of pseudogenes through whole genome expression evidence profiling. Nucleic Acids Res. *34*, 4477–4485.

Zhang, Z., Carriero, N., Zheng, D., Karro, J., Harrison, P.M., and Gerstein, M. (2006). PseudoPipe: an automated pseudogene identification pipeline. Bioinformatics *22*, 1437–1439.

Zhang, Z., and Gerstein, M. (2004). Large-scale analysis of pseudogenes in the human genome. Curr. Opin. Genet. Dev. *14*, 328–335.

Zhang, Z.D., Frankish, A., Hunt, T., Harrow, J., and Gerstein, M. (2010). Identification and analysis of unitary pseudogenes: historic and contemporary gene losses in humans and other primates. Genome Biol. *11*, R26.

Zheng, D., and Gerstein, M.B. (2006). A computational approach for identifying pseudogenes in the ENCODE regions. Genome Biol. *7* (*Suppl 1*), S13, 1–10.

Zheng, D., Frankish, A., Baertsch, R., Kapranov, P., Reymond, A., Choo, S.W., Lu, Y., Denoeud, F., Antonarakis, S.E., Snyder, M., et al. (2007). Pseudogenes in the ENCODE regions: consensus annotation, analysis of transcription, and evolution. Genome Res. *17*, 839–851.

Zhou, B.S., Beidler, D.R., and Cheng, Y.C. (1992). Identification of antisense RNA transcripts from a human DNA topoisomerase I pseudogene. Cancer Res. *52*, 4280–4285.

Zou, M., Baitei, E.Y., Alzahrani, A.S., Al-Mohanna, F., Farid, N.R., Meyer, B., and Shi, Y. (2009). Oncogenic activation of MAP kinase by BRAF pseudogene in thyroid tumors. Neoplasia *11*, 57–65.

NEOPLASIA
www.neoplasia.com

# Gene Fusions Associated with Recurrent Amplicons Represent a Class of Passenger Aberrations in Breast Cancer[1,2]

**Shanker Kalyana-Sundaram[*,†,‡], Sunita Shankar[*,†], Scott DeRoo[*,†], Matthew K. Iyer[*,§], Nallasivam Palanisamy[*,†], Arul M. Chinnaiyan[*,†,§,¶,3] and Chandan Kumar-Sinha[*,†,3]**

[*]Michigan Center for Translational Pathology, University of Michigan, Ann Arbor, MI; [†]Department of Pathology, University of Michigan, Ann Arbor, MI; [‡]Department of Environmental Biotechnology, Bharathidasan University, Tiruchirappalli, India; [§]Comprehensive Cancer Center, University of Michigan Medical School, Ann Arbor, MI; [¶]Howard Hughes Medical Institute, University of Michigan Medical School, Ann Arbor, MI

## Abstract

Application of high-throughput transcriptome sequencing has spurred highly sensitive detection and discovery of gene fusions in cancer, but distinguishing potentially oncogenic fusions from random, "passenger" aberrations has proven challenging. Here we examine a distinctive group of gene fusions that involve genes present in the loci of chromosomal amplifications—a class of oncogenic aberrations that are widely prevalent in breast cancers. Integrative analysis of a panel of 14 breast cancer cell lines comparing gene fusions discovered by high-throughput transcriptome sequencing and genome-wide copy number aberrations assessed by array comparative genomic hybridization, led to the identification of 77 gene fusions, of which more than 60% were localized to amplicons including 17q12, 17q23, 20q13, chr8q, and others. Many of these fusions appeared to be recurrent or involved highly expressed oncogenic drivers, frequently fused with multiple different partners, but sometimes displaying loss of functional domains. As illustrative examples of the "amplicon-associated" gene fusions, we examined here a recurrent gene fusion involving the mediator of mammalian target of rapamycin signaling, *RPS6KB1* kinase in BT-474, and the therapeutically important receptor tyrosine kinase *EGFR* in MDA-MB-468 breast cancer cell line. These gene fusions comprise a minor allelic fraction relative to the highly expressed full-length transcripts and encode chimera lacking the kinase domains, which do not impart dependence on the respective cells. Our study suggests that amplicon-associated gene fusions in breast cancer primarily represent a by-product of chromosomal amplifications, which constitutes a subset of passenger aberrations and should be factored accordingly during prioritization of gene fusion candidates.

*Neoplasia (2012) 14, 702–708*

Address all correspondence to: Assistant Professor Chandan Kumar-Sinha, PhD, Department of Pathology, University of Michigan Medical School, 1400 E Medical Center Dr 5316 CCGC, Ann Arbor, MI 48109-0602. E-mail: chakumar@med.umich.edu

[1] This project was supported in part by Department of Defense Breast Cancer Research Program (W81XWH-08-0110), American Association for Cancer Research Stand Up to Cancer (SU2C) award, the National Functional Genomics Center (W81XWH-11-1-0520) from Department of Defense to A.M.C., and, in part, by the US National Institutes of Health through the University of Michigan's Cancer Center Support grant 5 P30 CA46592. A.M.C. is supported by the US National Cancer Institute's Early Detection Research Network (U01 CA111275). A.M.C. is supported by the Doris Duke Charitable Foundation Clinical Scientist Award and the Burroughs Welcome Foundation Award in Clinical Translational Research. A.M.C. is an American Cancer Society Research professor and Taubman Scholar. S.D. is supported by Howard Hughes Medical Institute Medical Student Fellowship.

[2] This article refers to supplementary materials, which are designated by Tables W1 and W2 and Figures W1 to W4 and are available online at www.neoplasia.com.

[3] These authors share senior authorship.

Received 6 June 2012; Revised 9 July 2012; Accepted 10 July 2012

DOI 10.1593/neo.12914

## Introduction

Chromosomal amplifications and translocations are among the most common somatic aberrations in cancers [1,2]. Gene amplification is an important mechanism for oncogene overexpression and activation. Numerous recurrent loci of chromosomal amplifications have been characterized in breast cancer, which result in gain of copy number and overexpression of oncogenes such as *ERBB2* on 17q12 (the definitive molecular aberration in 20%-30% of all breast cancers) [3,4], as well as many other oncogenic drivers including *Myc* [5], *EGFR* [6], *FGFR1* [7], *CyclinD1* [8], *RPS6KB1* [9], and others [10]. Chromosomal translocations leading to generation of gene fusions represent another prevalent mechanism for the expression of oncogenes in epithelial cancers [11]. Recently, we described the discovery and characterization of recurrent gene fusions in breast cancer involving MAST family serine threonine kinases and Notch family of transcription factors [12]. Interestingly, we also observed a large number of gene fusions, including some recurrent fusions involving known oncogenes localized at loci of chromosomal amplifications.

Here we carried out a systematic analysis of the association between gene fusions and genomic amplification by integrating RNA-Seq data with array comparative genomic hybridization (aCGH)–based whole-genome copy number profiling from a panel of breast cancer cell lines. We examined a set of “amplicon-associated gene fusions” that refer to all the fusions where one or both gene partners are localized to a site of chromosomal amplification. Specifically, we assessed the functional relevance of two amplicon-associated fusion genes involving oncogenic kinases *EGFR* and *RPS6KB1* in the context of prioritizing fusion candidates important in tumorigenesis. Our results suggest that recurrent gene fusions localized to recurrent amplicons, displaying allelic imbalance between the fusion partners, may represent an epiphenomenon of genomic amplification cycles not essential for cancer development.

## Materials and Methods

### Gene Fusion Data Set

Chimeric transcript candidates were primarily obtained from paired-end transcriptome sequencing of breast cancer from a total of more than 49 cell lines and 40 tissue samples described previously [12]. aCGH data were generated using Agilent Human Genome 244A CGH Microarrays (Agilent Technologies, Santa Clara, CA) according to the manufacturer’s instructions, and data were analyzed using CGH Analytics (Agilent Technologies). Copy number alterations were assessed using ADM-2, with the threshold a setting of 6.0 and a bin size of 10.

### RNA Isolation and Complementary DNA Synthesis

Total RNA was isolated using TRIzol and RNeasy Kit (Invitrogen, Carlsbad, CA) with DNase I digestion according to the manufacturer’s instructions. RNA integrity was verified on an Agilent Bioanalyzer 2100 (Agilent Technologies). Complementary DNA was synthesized from total RNA using Superscript III (Invitrogen) and random primers (Invitrogen).

### Quantitative Real-time Polymerase Chain Reaction

Primers for validation of candidate gene fusions were designed using the National Center for Biotechnology Information Primer Blast (http://www.ncbi.nlm.nih.gov/tools/primer-blast/), with primer pairs spanning exon junctions amplifying 70- to 110-bp products for every chimera tested. Quantitative polymerase chain reaction (QPCR) was performed using SYBR Green MasterMix (Applied Biosystems, Carlsbad, CA) on an Applied Biosystems StepOne Plus Real-Time PCR System. All oligonucleotide primers were obtained from Integrated DNA Technologies and are listed in Table W1. *GAPDH* was used as endogenous control. All assays were performed twice, and results were plotted as average fold change relative to *GAPDH*.

### Cell Proliferation Assays

Cells were transfected with small interfering RNAs (siRNAs) using Oligofectamine reagent (Life Sciences, Carlsbad, CA), and 3 days after transfection, the cells were plated for proliferation assays. At the indicated times, cell numbers were counted using Coulter Counter (Indianapolis, IN).

### Western Blot

Cell pellets were sonicated in NP-40 lysis buffer (50 mM Tris-HCl, 1% NP-40, pH 7.4; Sigma, St. Louis, MO), complete protease inhibitor mixture (Roche, Indianapolis, IN), and phosphatase inhibitor (EMD Bioscience, San Diego, CA). Immunoblot analysis was carried out using antibodies for *ERBB2* (MS-730-PABX; Thermo Scientific, Fremont, CA) and *RPS6KB1* (2708S; Cell Signaling, Danvers, MA). Human β-actin antibody (Sigma, St. Louis, MO) was used as a loading control.

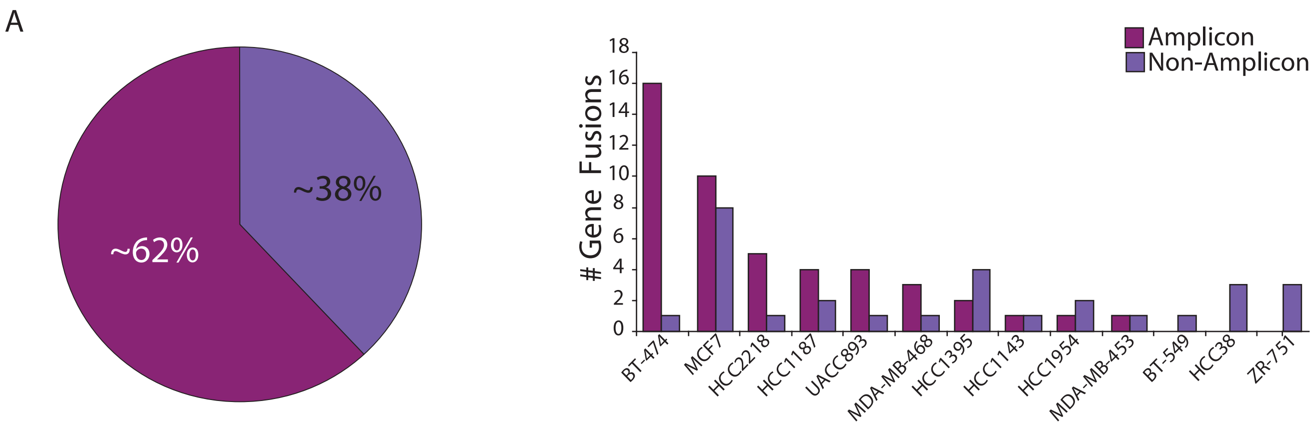
### Knockdown Assays

Short hairpin RNAs (shRNAs; Table W1) were transduced in presence of 1 μg/ml polybrene. All siRNA transfections were performed using Oligofectamine reagent (Life Sciences). For siRNA knockdown experiments, multiple custom siRNA sequences targeting the *ARID1A-MAST2* fusion (Thermo, Lafayette, CO) were used [12].

## Results

Paired-end transcriptome sequencing of breast cancer cell lines and tissues led to the identification of an average of more than four gene fusions per breast cancer sample [12]. Interestingly, we observed that some of the cell lines with the largest number of gene fusions also harbored many well-known chromosomal amplifications, prompting us to examine a likely association between genomic amplifications and gene fusions. To assess copy number alterations at the chromosomal coordinates of the fusion genes, we analyzed aCGH (244K Agilent array) data in a set of 14 cell lines (Table W2) and observed that as many as 62% of the total number of fusions were associated with regions of amplifications (Figure 1*A*). The genes involved in fusions were found to be significantly associated with their genomic amplification status based on Fisher exact *t* test ($P < .0004$), in four of six cell lines with the maximum number of fusions, including BT-474, MCF7, HCC2218, and UACC893 (Figure 1*B*).

Examining the distribution of fusion genes in individual samples revealed that a majority of the gene fusions were associated with 17q12 amplicon harboring *ERBB2* and 17q23 amplicon that includes genes such as *BCAS3*, *RPS6KB1*, and *TMEM49*, 20q13 amplicon with *BCAS4* and the chr8q amplicon commonly found amplified in breast cancer (Table W2 and Figures 2 and W1). Interestingly, the breast cancer cell line BT-474 that harbors both the chr17 amplicons and the chr20 amplicon and MCF7 with prominent amplifications in chr17, chr20, and chr8q showed the maximum number of gene fusions observed in a sample, accounting for as many as 26 gene fusions associated with amplicons compared against only 9 in unamplified loci (Figures 1 and 2 and Table W2).



B

| Sample | Fusion Partners | Number of Amplifications | Total Measureable Genes | Amplified Fusion Partners | Not-amplified Fusion Partners | Amplified without Fusion Partner | Unaffected Genes | p-value |
|---|---|---|---|---|---|---|---|---|
| MCF7 | 27 | 426 | 17691 | 11 | 16 | 415 | 17249 | 1.28E-11 |
| HCC2218 | 9 | 446 | 17691 | 7 | 2 | 439 | 17243 | 2.13E-10 |
| BT-474 | 31 | 5648 | 17691 | 25 | 6 | 5623 | 12038 | 3.19E-08 |
| UACC-893 | 8 | 2902 | 17691 | 6 | 2 | 2896 | 14787 | 0.0004016 |
| HCC1187 | 10 | 780 | 17691 | 2 | 8 | 778 | 16903 | 0.06908 |
| HCC1395 | 12 | 2111 | 17691 | 1 | 11 | 2110 | 15569 | 0.783 |

**Figure 1.** Distribution of gene fusions across breast cancer cell lines. (A) Pie chart representation of the relative proportion of gene fusions associated with loci of genomic amplifications compared to unamplified loci (left) and bar graph representation of the relative distribution of gene fusions across different breast cancer cell lines (right). (B) Table summarizing the statistical significance of association between gene fusions and chromosomal amplifications in breast cancer cell lines with the highest number of gene fusions in A (using Fisher exact *t* test, sorted by *P* value).

In the backdrop of a large number of somatic aberrations seen in cancers, any "recurrent" events observed across samples are generally regarded as potentially "driving" tumorigenesis. Interestingly, among the more than 380 gene fusions reported in our compendium of breast cancer fusions [12], as many as 62 genes were found to be recurrent partners (appear at least twice). Among these, whereas the *MAST* and *Notch* fusions were shown to be functionally recurrent and potentially driving aberrations in up to 5% to 7% of breast cancers, 33 of other recurrent gene fusions were found to be associated with known frequent amplicons, including *ERBB2*, *BCAS3/4*, and chr8q. Among these, three fusions each involved the ikaros family zinc finger protein 3 transcription factor (*IKZF3* on chr17q12 amplicon) and breast carcinoma amplified sequence 3 (*BCAS3* on chr17q23 amplicon) as 3′ partners—all with different 5′ partners. Similarly, tripartite motif containing 37 (*TRIM37* on chr17q23) was a common 5′ partner in three distinct gene fusions with different 3′ partners (Table W2). To further expand our integrative analysis of copy number aberrations and gene fusions, next we used the breast cancer aCGH data [13,14] and observed gene fusion–associated amplicons in MCF7, BT-474, and MDA-MB-468, HCC-1187 as seen in our data as well as in an additional panel of cell lines, including ZR-75-30, SUM190, MDA-MB-361, HCC-1428, and HCC-1569 (Figure W2). Clearly, apart from triggering overexpression of constituent genes, our observations strongly suggest that the loci of chromosomal amplifications also serve as "hotspots" for the generation of recurrent gene fusions.

Next, to assess whether amplicon-associated gene fusions impart oncogenic phenotypes on the cells, we examined the open reading frames (ORFs), functional domains/motifs, and conservation of fusion architecture across different samples. Among recurrent fusion candidates within amplicons, we focused on known cancer-associated partner genes such as kinases, oncogenes, tumor suppressors, or known fusion partners in the Mitelman Database of chromosomal aberrations in cancer [15] and observed several functionally plausible gene fusions. Here we describe our observations with two specific examples of gene fusions involving oncogenic kinases.

The triple-negative breast cancer cell line MDA-MB-468 is known to show an overexpression of epidermal growth factor receptor (EGFR) [16]. In our transcriptome sequencing compendium of 89 breast cancer cell lines and tissues, the highest expression of *EGFR* is observed in MDA-MB-468 (Figure 3*A*), potentially resulting from a focal amplification at chr7p12 (Figure 2). In addition, we detected an *EGFR* fusion transcript (*EGFR-POLD1*) in this cell line, encoding the N-terminal portion of EGFR, completely devoid of the tyrosine kinase domain (Figure 3*A*, *inset*). However, the uniform read-coverage observed across the full length of the *EGFR* transcript in this sample (Figure 3*B*), precluded the existence of any exon imbalance, suggesting that even as the kinase domain is lost in the fusion, the full-length EGFR protein is expressed in this cell line. Further, we observed a remarkable mismatch between the copy numbers of *EGFR* and its fusion partner *POLD1* (Figure 3*C*) that supports a predominant expression of full-length *EGFR* compared with the *EGFR-POLD1* chimera. This is unlike the observation in case of *MAST* kinase fusions in breast cancer characterized in our previous study [12], in which case a marked exon imbalance in coverage was observed (Figure W3). Considering that the

MDA-MB-468 harbors both *MAST2* and *EGFR* fusions, we were intrigued to assess its relative "dependence" on both the kinases. Surprisingly, a profound reduction in cell proliferation was observed on siRNA knockdown of *MAST2*, whereas *EGFR* knockdown showed little effect (Figure 3*D*). Next, testing the possibility of *EGFR* amplicon potentially cooperating with *MAST2*, we found that the effect of combined knockdown of *EGFR* and *MAST2* was comparable with that of *MAST2* knockdown alone (Figure 3*D*), further suggesting that *EGFR* amplification does not signify a driver aberration. In this context, the EGFR fusion transcript that represents a miniscule fraction of overall EGFR expression and encodes only the N-terminal portion lacking the kinase domain was reckoned to be inconsequential.

Next, we looked at recurrent gene fusions involving oncogenic serine threonine kinase ribosomal protein S6 kinase on chr17q23 frequently amplified in breast cancers [17–20] identified in BT-474 (*RPS6KB1-SNF8*) and MCF7 (*RPS6KB1-VMP1*). Both of these cell lines harbor amplifications at the *RPS6KB1* locus and express the highest levels of *RPS6KB1* among all the samples examined (Figure 4*A*). Both the chimeric transcripts retain only the first exon of *RPS6KB1* and the respective open reading frames show a complete loss of the kinase domain (Figure 4*A*, *inset*). We also observed an even read coverage across the *RPS6KB1* transcript in both fusion-positive cell lines, similar to a representative benign mammary epithelial cell line, albeit at a much higher level, indicating that full-length *RPS6KB1* protein is encoded in these samples (Figures 4*B* and W4*A*). Further, the difference between the copy number observed between the fusion partners in both the *RPS6KB1* fusions (Figures 4*C* and W4*B*) indicates an allelic imbalance between the full-length and the putative fusion genes. Next, considering that BT-474 is an *ERBB2*-positive cell line, we tested potential dependence of these cells on the RPS6KB1 protein. Surprisingly,

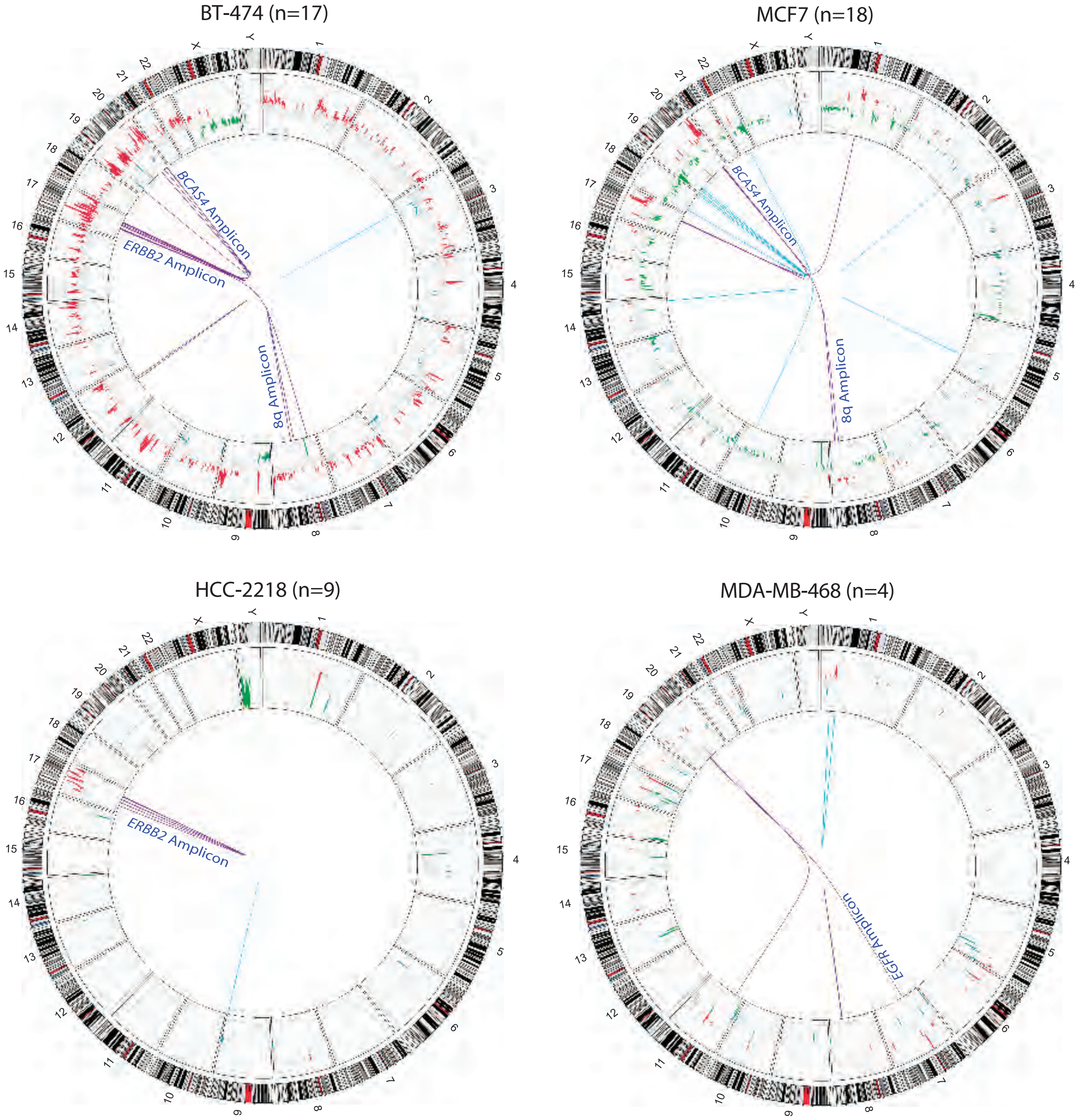

BT-474 (n=17)
MCF7 (n=18)
HCC-2218 (n=9)
MDA-MB-468 (n=4)

**Figure 2.** Graphical representation of integrative analysis of gene fusions with copy number analysis. Circos plots of the genome-wide distribution of gene fusions along with status of copy number alterations. Red and green peaks represent amplifications and deletions; purple and cyan lines represent the fusions associated with amplicons and nonamplicons, respectively. "*n*" refers to the total number of fusions identified.

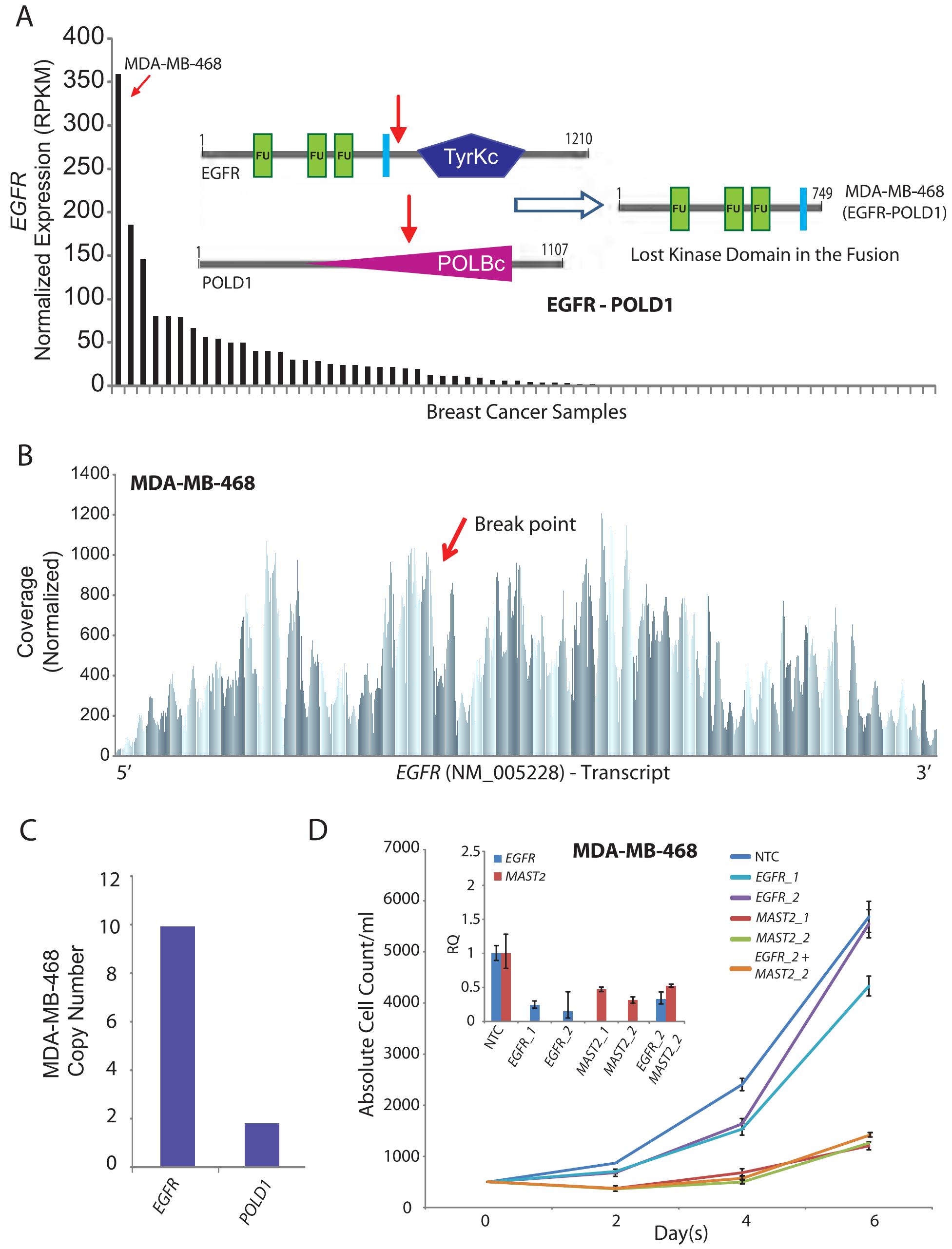


**Figure 3.** (A) Normalized expression (RPKM) of *EGFR* in descending order of expression in a panel of breast cancer samples obtained from RNA-Seq. Schematic representation of wild-type EGFR and POLD1 proteins with putative breakpoints indicated by red arrows and the domain structure of the putative fusion protein (inset). (B) Plot of normalized coverage of *EGFR* transcript in MDA-MB-468 cell line showing the location of the breakpoint (indicated by red arrow). (C) Bar graph representing the copy number of *EGFR* and *POLD1* in MDA-MB-468. (D) Proliferation assay showing absolute cell count (*y* axis) over a time course (*x* axis) after knockdown with *EGFR* and/or *MAST2* siRNAs in MDA-MB-468. QPCR assessment of knockdown efficiencies relative to nontargeted control (NTC; inset).

similar to our observations with *EGFR* knockdown in MDA-MB-468 cells, here we observed only a small effect on cell proliferation after shRNA knockdown of *RPS6KB1*, in dramatic contrast to the effect of *ERBB2* knockdown (Figure 4*D*). Notably, the shRNA knockdown of RPS6KB1 led to a significant depletion of the full-length protein yet it did not affect cell proliferation compared with ERBB2 protein depletion (Figure 4*D*, *inset*). Therefore, BT-474 cells do not display a dependence on RPS6KB1 protein, and considering that the RPS6KB1 fusion product is completely devoid of all functional domains of RPS6KB1, including the kinase domain, this fusion also likely represents a passenger event.

## Discussion

In our systematic search for gene fusions in breast cancer using high-throughput transcriptome sequencing, we observed a notably large number of fusion genes associated with many well characterized recurrent amplicons, including 17q12, 17q23, 20q13, and 8q, among others. Amplicon-associated gene fusions were found to involve complex and cryptic rearrangements, involving one or both partners within the amplicon site, with the chimeric transcript expression apparently concealed in the backdrop of highly expressed wild-type genes. The gene fusions considered here include only "expressed" chimeric transcripts derived from known/annotated fusion partners. Chromosomal rearrangements that do not express chimeric transcripts or that involve unannotated fusion partners are excluded from this analysis. This likely accounts for the variability observed in the number of gene fusions scored across multiple samples with known amplicons. Because many of the fusions at the amplicons appeared to be recurrent, although frequently fused with multiple different partners, it led us to examine whether the recurrence was incidentally associated with recurrent amplicons or signified functionally important aberrations.

MDA-MB-468 represents a prototype triple-negative breast cancer cell line with a "basal-like" gene expression profile that shows an overexpression of the oncogenic kinase *EGFR* due to a focal amplification at chr7p12. Here we discovered a chimeric transcript involving *EGFR*. However, careful examination of this transcript revealed that the fusion encodes N-terminal *EGFR* protein, without the kinase domain. Transcriptome sequencing did not show evidence of fusion-associated exon imbalance in *EGFR* expression, suggesting that full-length *EGFR* is expressed in this cell line. In addition, the significantly higher genomic copy number of *EGFR* compared to its fusion partner *POLD1* suggests that a minor allelic fraction of the *EGFR* is involved in fusion with *POLD1*, whereas other amplified copies of the gene

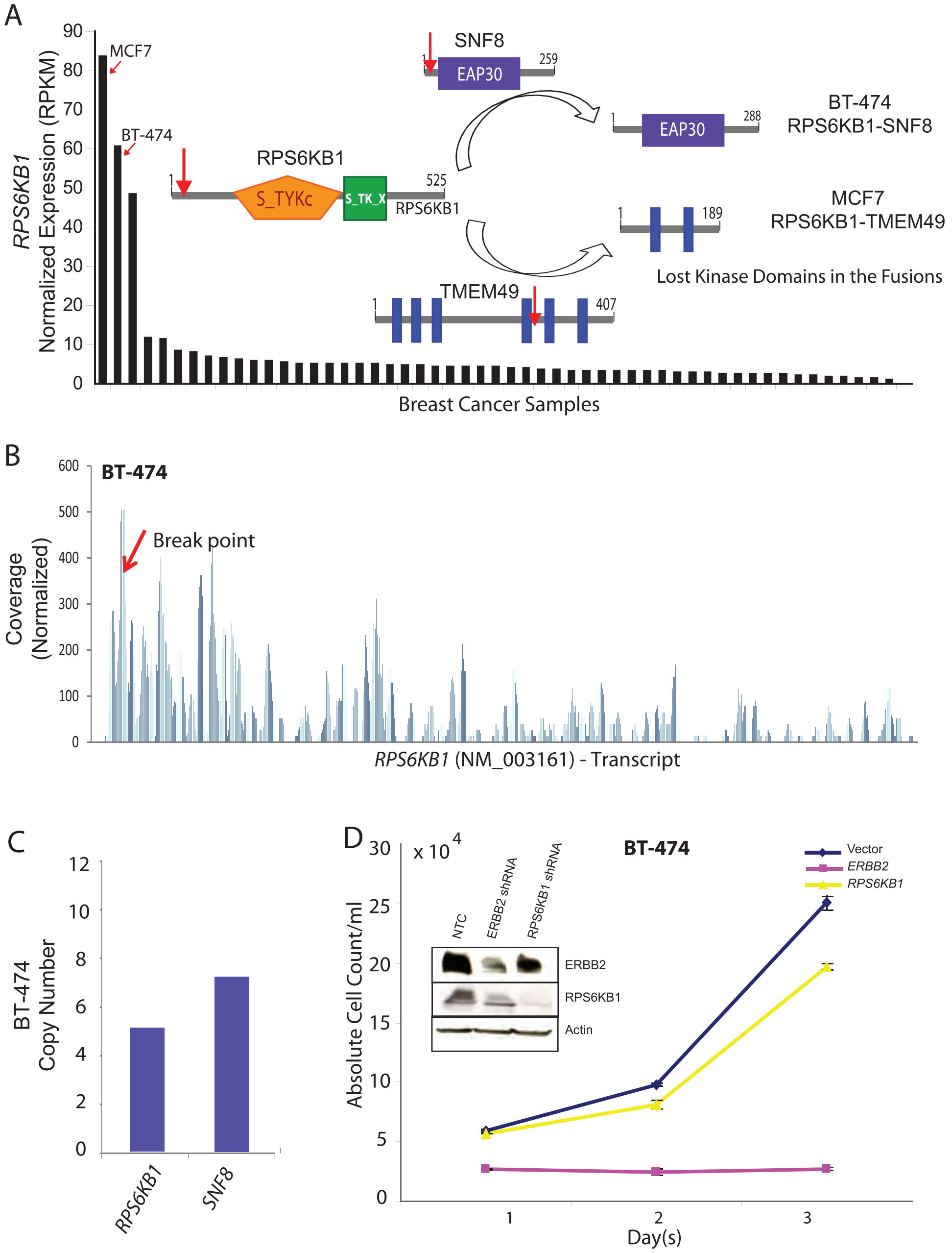


**Figure 4.** (A) Normalized expression (RPKM) of *RPS6KB1* in descending order of expression in a panel of breast cancer samples obtained from RNA-Seq. Schematic representation of wild-type RPS6KB1, TMEM49, and SNF8 proteins with putative breakpoints indicated by red arrows and the domain structure of the putative fusion proteins in BT-474 and MCF7 (inset). (B) Plot of normalized coverage of *RPS6KB1* transcript in BT-474 cell line showing the location of the breakpoint (indicated by red arrow). (C) Bar graph representing the copy number of *RPS6KB1* and *SNF8* in BT-474 (D) Proliferation assay showing absolute cell count (*y* axis) over a time course (*x* axis) after knockdown with *RPS6KB1* and/or *ERBB2* shRNAs in BT-474. Western blot assessment of the knockdown efficiency relative to nontargeted control (NTC). Actin was used as a loading control (inset).

express the full-length molecule. Technically, the detection and monitoring of the *EGFR* fusion transcript in the backdrop of extremely high levels of wild-type *EGFR* transcript is challenging; therefore, we chose to assess the dependency imparted by full-length *EGFR*. Interestingly, the knockdown of *EGFR* had only a slight effect on the proliferation of MDA-MB-468 cells, whereas a profound reduction in cell proliferation was observed on the knockdown the fusion gene *MAST2*. Combined knockdown of *MAST2* and *EGFR* produced the same effect as that by MAST2 alone, further calling into question the credentials of *EGFR* as a driver aberration in MDA-MB-468 cells. Interestingly, MDA-MB-468 is known to be insensitive to EGFR inhibitors like erlotinib [21] and gefitinib [22].

Similarly, the recurrent gene fusions involving *RPS6KB1* retain only the first exon, and the chimeric ORFs show a complete loss of the kinase domain in breast cancer cell lines BT-474 and MCF7. Similar to the *EGFR* fusion, DNA copy number analysis and RNA-Seq data provided the evidence that full-length RPS6KB1 protein is encoded in both these cell lines. Notably, both BT-474 and MCF7 have been shown to express high levels of full-length RPS6KB1 protein [23], suggesting that these cells exhibit elevated activity of RPS6KB1 as a result of amplification, independent of the fusion. Again, similar to *EGFR* knockdown in MDA-MB-468, *RPS6KB1* knockdown in BT-474 (an ERBB2-positive cell line) showed an insignificant effect on cell proliferation compared to *ERBB2* knockdown. Interestingly, in a previous study, knockdown of *RPS6KB1* was found to have no effect on apoptosis in both BT-474 and MCF7 breast cancer cells [24].

In the light of our observations, we surmise that repeated breaks and rejoining of chromosomes during chromosomal amplifications led to the generation of amplicon-associated gene fusions. Loci of recurrent genomic amplifications thus engender "pseudo" recurrent gene fusions that may largely represent passenger aberrations involving random breakpoints. The two cell lines with established drivers—*ERBB2* in BT-474 and *MAST2* in MDA-MB-468—made it possible for us to assess the relative importance of amplicon fusions involving *RPS6KB1* and *EGFR*, respectively. In cases where a driver is not clearly apparent, a more careful examination of all plausible fusion candidates will be required. Importantly, even as our study primarily pertains to breast cancers based on available data and a well-documented preponderance of copy number aberrations in breast cancers [10], we expect the association between amplicons and gene fusions to be consistent in other cancers as well. We argue here for a measure of caution in considering the functional implications of recurrent gene fusions associated with amplifications because these may be simply a result of massive chromosomal upheaval at the amplicons, not representing clonally selected oncogenic events.

## Acknowledgments

The authors thank Vishal Kothari and Nikita Consul for technical help.

## References

[1] Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, Rahman N, and Stratton MR (2004). A census of human cancer genes. *Nat Rev Cancer* **4**, 177–183.

[2] Santarius T, Shipley J, Brewer D, Stratton MR, and Cooper CS (2010). A census of amplified and overexpressed human cancer genes. *Nat Rev Cancer* **10**, 59–64.

[3] Slamon DJ, Clark GM, Wong SG, Levin WJ, Ullrich A, and McGuire WL (1987). Human breast cancer: correlation of relapse and survival with amplification of the HER-2/*neu* oncogene. *Science* **235**, 177–182.

[4] Slamon DJ, Leyland-Jones B, Shak S, Fuchs H, Paton V, Bajamonde A, Fleming T, Eiermann W, Wolter J, Pegram M, et al. (2001). Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2. *N Engl J Med* **344**, 783–792.

[5] Deming SL, Nass SJ, Dickson RB, and Trock BJ (2000). C-*myc* amplification in breast cancer: a meta-analysis of its occurrence and prognostic relevance. *Br J Cancer* **83**, 1688–1695.

[6] Bhargava R, Gerald WL, Li AR, Pan Q, Lal P, Ladanyi M, and Chen B (2005). EGFR gene amplification in breast cancer: correlation with epidermal growth factor receptor mRNA and protein expression and HER-2 status and absence of EGFR-activating mutations. *Mod Pathol* **18**, 1027–1033.

[7] Elbauomy Elsheikh S, Green AR, Lambros MB, Turner NC, Grainge MJ, Powe D, Ellis IO, and Reis-Filho JS (2007). FGFR1 amplification in breast carcinomas: a chromogenic *in situ* hybridisation analysis. *Breast Cancer Res* **9**, R23.

[8] Elsheikh S, Green AR, Aleskandarany MA, Grainge M, Paish CE, Lambros MB, Reis-Filho JS, and Ellis IO (2008). CCND1 amplification and cyclin D1 expression in breast cancer and their relation with proteomic subgroups and patient outcome. *Breast Cancer Res Treat* **109**, 325–335.

[9] Inaki K, Hillmer AM, Ukil L, Yao F, Woo XY, Vardy LA, Zawack KF, Lee CW, Ariyaratne PN, Chan YS, et al. (2011). Transcriptional consequences of genomic structural aberrations in breast cancer. *Genome Res* **21**, 676–687.

[10] Curtis C, Shah SP, Chin S-F, Turashvili G, Rueda OM, Dunning MJ, Speed D, Lynch AG, Samarajiwa S, Yuan Y, et al. (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**, 346–352.

[11] Chinnaiyan AM and Palanisamy N (2010). Chromosomal aberrations in solid tumors. *Prog Mol Biol Transl Sci* **95**, 55–94.

[12] Robinson DR, Kalyana-Sundaram S, Wu YM, Shankar S, Cao X, Ateeq B, Asangani IA, Iyer M, Maher CA, Grasso CS, et al. (2011). Functionally recurrent rearrangements of the MAST kinase and Notch gene families in breast cancer. *Nat Med* **17**, 1646–1651.

[13] Pollack JR, Sorlie T, Perou CM, Rees CA, Jeffrey SS, Lonning PE, Tibshirani R, Botstein D, Borresen-Dale AL, and Brown PO (2002). Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proc Natl Acad Sci USA* **99**, 12963–12968.

[14] Greshock J, Naylor TL, Margolin A, Diskin S, Cleaver SH, Futreal PA, deJong PJ, Zhao S, Liebman M, and Weber BL (2004). 1-Mb resolution array-based comparative genomic hybridization using a BAC clone set optimized for cancer gene analysis. *Genome Res* **14**, 179–187.

[15] Mitelman F, Johansson B, and Mertens F (2010). *Mitelman Database of Chromosome Aberrations and Gene Fusions in Cancer*. Cancer Genome Anatomy Project. Available at: http://cgap.nci.nih.gov/Chromosomes/Mitelman. Accessed March 2012.

[16] Hyatt DC and Ceresa BP (2008). Cellular localization of the activated EGFR determines its effect on cell growth in MDA-MB-468 cells. *Exp Cell Res* **314**, 3415–3425.

[17] Barlund M, Monni O, Kononen J, Cornelison R, Torhorst J, Sauter G, Kallioniemi O-P, and Kallioniemi A (2000). Multiple genes at 17q23 undergo amplification and overexpression in breast cancer. *Cancer Res* **60**, 5340–5344.

[18] Couch FJ, Wang XY, Wu GJ, Qian J, Jenkins RB, and James CD (1999). Localization of PS6K to chromosomal region 17q23 and determination of its amplification in breast cancer. *Cancer Res* **59**, 1408–1411.

[19] Monni O, Barlund M, Mousses S, Kononen J, Sauter G, Heiskanen M, Paavola P, Avela K, Chen Y, Bittner ML, et al. (2001). Comprehensive copy number and gene expression profiling of the 17q23 amplicon in human breast cancer. *Proc Natl Acad Sci USA* **98**, 5711–5716.

[20] Sinclair CS, Rowley M, Naderi A, and Couch FJ (2003). The 17q23 amplicon and breast cancer. *Breast Cancer Res Treat* **78**, 313–322.

[21] Bartholomeusz C, Yamasaki F, Saso H, Kurisu K, Hortobagyi GN, and Ueno NT (2011). Gemcitabine overcomes erlotinib resistance in EGFR-overexpressing cancer cells through downregulation of Akt. *J Cancer* **2**, 435–442.

[22] Maiello MR, D'Alessio A, De Luca A, Carotenuto A, Rachiglio AM, Napolitano M, Cito L, Guzzo A, and Normanno N (2007). AZD3409 inhibits the growth of breast cancer cells with intrinsic resistance to the EGFR tyrosine kinase inhibitor gefitinib. *Breast Cancer Res Treat* **102**, 275–282.

[23] Yamnik RL, Digilova A, Davis DC, Brodt ZN, Murphy CJ, and Holz MK (2009). S6 kinase 1 regulates estrogen receptor alpha in control of breast cancer cell proliferation. *J Biol Chem* **284**, 6361–6369.

[24] Heinonen H, Nieminen A, Saarela M, Kallioniemi A, Klefstrom J, Hautaniemi S, and Monni O (2008). Deciphering downstream gene targets of PI3K/mTOR/p70S6K pathway in breast cancer. *BMC Genomics* **9**, 348.

**Table W1.** Primer Sequences and siRNA/shRNA Clone Details.

| Gene Symbol | Clone ID |
|---|---|
| *EGFR* | LU-003114-00-0002 |
| *ERBB2* | SHCLNV-NM_004448 |
| *RPS6KB1* | SHCLNV-NM_003161 |

| Primer | Sequence |
|---|---|
| EGFR-f1 | GGGCCAGGTCTTGAAGGCTGT |
| EGFR-r1 | ATCCCCAGGGCCACCACCAG |
| EGFR-f2 | ACACCCTGGTCTGGAAGTACGCA |
| EGFR-r2 | AGTGGGAGACTAAAGTCAGACAGTGAA |
| EGFR-f3 | CCGAGGCAGGGAATGCGTGG |
| EGFR-r3 | TGGCCTGAGGCAGGCACTCT |
| ERBB2-f1 | TGCGCAGGCAGTGATGAGAGT |
| ERBB2-r1 | TCTCGGGACTGGCAGGGAGC |
| ERBB2-f2 | TCCTCCTCGCCCTCTTGCCC |
| ERBB2-r2 | TCTCGGGACTGGCAGGGAGC |
| RPS6KB1-f1 | TGCTGACTGGAGCACCCCCA |
| RPS6KB1-r1 | GCTTCTTGTGTGAGGTAGGGAGGC |
| GAPDH-f1 | GGCTGAGAACGGGAAGCTTGTCA |
| GAPDH-r1 | TCTCCATGGTGGTGAAGACGCCA |
| MAST2_f1 | GAAGTGAGTGAGGATGGCTGCCTT |
| MAST2_r1 | GAGCCGCTCCATGCTGCTGTAC |
| MAST2_f2 | ATTGAGGGCCATGGGGCATCT |
| MAST2_r2 | CCCCATAGGCGCCATTGCTGATG |

**Table W2.** List of Gene Fusions Identified in 14 Breast Cancer Cell Lines, along with Their Copy Number Status.

| Sample Name | 5′ Gene | 3′ Gene | Type | Sequencing Platform | No. Reads | Validation | Chromosomal Location | | aCGH Data (5′ and 3′) | | | | Amplicon Status |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | Fusion QPCR | 5′ Gene | 3′ Gene | No. Probe | Average Log Ratio | No. Probe | Average Log Ratio | |
| BT-474 | *RPS6KB1* | SNF8 | Intra | GA II | 92 | Y | chr17:55325224-55382568 | chr17:44362457-44377153 | 5 | 2.890 | 2 | 3.557 | Yes |
| BT-474 | *STX16* | RAE1 | Intra | GA II | 79 | Y | chr20:56659733-56687988 | chr20:55360024-55386926 | 4 | 2.910 | 4 | 2.910 | Yes |
| BT-474 | *ZMYND8* | CEP250 | Intra | GA II | 77 | Y | chr20:45271787-45418881 | chr20:33506636-33563217 | 15 | 3.650 | 5 | 1.876 | Yes |
| BT-474 | *TRPC4AP* | MRPL45 | Inter | GA II | 30 | | chr20:33053867-33144279 | chr17:33706516-33732628 | 11 | 3.290 | 4 | 3.452 | Yes |
| BT-474 | *MED1* | STXBP4 | Intra | GA II | 28 | Y | chr17:34814063-34861053 | chr17:50401124-50596448 | 4 | 4.029 | 21 | 2.507 | Yes |
| BT-474 | *TOB1* | AP1GBP1 | Intra | GA II | 16 | | chr17:46294585-46296412 | chr17:32949013-33043559 | 1 | 2.787 | 10 | 2.556 | Yes |
| BT-474 | *ACACA* | STAC2 | Intra | GA II | 15 | | chr17:32516039-32841015 | chr17:34620314-34635566 | 35 | 2.556 | 3 | 4.029 | Yes |
| BT-474 | *MED13* | BCAS3 | Intra | GA II | 13 | Y | chr17:57374747-57497425 | chr17:56109953-56824981 | 13 | 1.012 | 73 | 1.934 | Yes |
| BT-474 | *VAPB* | IKZF3 | Inter | GA II | 13 | Y | chr20:56397580-56459562 | chr17:35174724-35273967 | 7 | 3.404 | 10 | 3.701 | Yes |
| BT-474 | *RAB22A* | MYO9B | Inter | GA II | 9 | Y | chr20:56318176-56375969 | chr19:17047590-17185104 | 6 | 3.404 | 13 | 2.122 | Yes |
| BT-474 | *GLB1* | CMTM7 | Intra | GA II | 7 | | chr3:33013103-33113698 | chr3:32408166-32471337 | 11 | −0.425 | 6 | 0.428 | |
| BT-474 | *NCOA2* | ZNF704 | Intra | GA II | 7 | Y | chr8:71186820-71478574 | chr8:81703240-81949571 | 35 | 0.916 | 26 | 0.640 | Yes |
| BT-474 | *BCAS3* | MED13 | Intra | GA II | 6 | | chr17:56109953-56824981 | chr17:57374747-57497425 | 73 | 1.934 | 13 | 1.012 | Yes |
| BT-474 | *PIP4K2B* | RAD51C | Intra | GA II | 6 | | chr17:34175469-34209684 | chr17:54124961-54166691 | 6 | 4.813 | 5 | 1.700 | Yes |
| BT-474 | *PPP1R12A* | MGAT4C | Intra | GA II | 6 | | chr12:78691473-78853366 | chr12:84897167-85756812 | 19 | 1.218 | 90 | −0.397 | Yes |
| BT-474 | *STARD3* | DOCK5 | Inter | GA II | 6 | | chr17:35046858-35073980 | chr8:25098203-25326536 | 5 | 4.821 | 27 | 0.076 | Yes |
| BT-474 | *TRIM37* | MYO19 | Intra | GA II | 6 | | chr17:54414781-54539048 | chr17:31925711-31965418 | 14 | 2.244 | 6 | 2.344 | Yes |
| BT-483 | *SMARCB1* | MARK3 | Inter | GA II | 7 | Y | chr22:22459149-22506705 | chr14:102921453-103039919 | 8 | 1.170 | 17 | 0.381 | Yes |
| BT-549 | *CLTC* | TMEM49 | Intra | GA II | 18 | Y | chr17:55051831-55129099 | chr17:55139644-55272734 | 9 | −0.283 | 18 | −1.185 | |
| HCC1143 | *C18orf45* | HM13 | Inter | GA II | 25 | Y | chr18:19129977-19271923 | chr20:29565901-29591257 | 18 | 1.280 | 2 | 1.403 | Yes |
| HCC1143 | *C2ORF48* | RRM2 | Intra | GA II | 23 | Y | chr2:10198959-10269307 | chr2:10180145-10188997 | 8 | 0.134 | 2 | 0.134 | |
| HCC1187 | *PUM1* | TRERF1 | Inter | GA II | 38 | Y | chr1:31176939-31311151 | chr6:42300646-42527761 | 14 | 1.648 | 27 | 0.336 | Yes |
| HCC1187 | *SEC22B* | NOTCH2 | Intra | GA II | 30 | Y | chr1:143807763-143828279 | chr1:120255698-120413799 | 2 | 1.557 | 11 | 0.253 | Yes |
| HCC1187 | *CTAGE5* | SIP1 | Intra | GA II | 15 | | chr14:38806079-38890148 | chr14:38653238-38675928 | 9 | 0.940 | 4 | 0.235 | Yes |
| HCC1187 | *MCPH1* | AGPAT5 | Intra | GA II | 11 | | chr8:6251520-6488548 | chr8:6553285-6606429 | 29 | 0.495 | 5 | 0.738 | |
| HCC1187 | *KLK5* | CDH23 | Inter | GA II | 5 | | chr19:56138370-56148156 | chr10:73225533-73245710 | 3 | 0.888 | 1 | 0.953 | Yes |
| HCC1187 | *BC041478* | EXOSC10 | Inter | GA II | 3 | | chr19:42434668-42446354 | chr1:11049262-11082525 | 1 | 0.816 | 4 | 0.156 | |
| HCC1395 | *EIF3K* | CYP39A1 | Inter | GA II | 13 | Y | chr19:43801561-43819435 | chr6:46625403-46728482 | 2 | 0.852 | 11 | 0.611 | |
| HCC1395 | *HNRNPUL2* | AHNAK | Intra | GA II | 13 | Y | chr11:62238795-62251397 | chr11:62039949-62070908 | 2 | 0.629 | 5 | 1.172 | Yes |
| HCC1395 | *RAB7A* | LRCH3 | Intra | GA II | 6 | | chr3:129927668-130016331 | chr3:199002541-199082853 | 10 | 0.755 | 11 | −0.615 | |
| HCC1395 | *ERO1L* | FERMT2 | Intra | GA II | 5 | | chr14:52178354-52232169 | chr14:52395955-52487565 | 7 | 0.934 | 14 | 0.934 | Yes |
| HCC1395 | *FOSL2* | BRE | Intra | GA II | 5 | | chr2:28469282-28491020 | chr2:27966985-28415271 | 3 | 0.480 | 51 | 0.480 | |
| HCC1395 | *BCAR3* | ABCA4 | Intra | GA II | 4 | | chr1:93799936-93919973 | chr1:94230981-94359293 | 13 | 0.849 | 13 | 0.849 | |
| HCC1954 | *C6orf106* | SPDEF | Intra | GA II | 24 | Y | chr6:34663048-34772603 | chr6:34613557-34632069 | 13 | 0.036 | 3 | 0.374 | |
| HCC1954 | *INTS1* | PRKAR1B | Intra | GA II | 22 | Y | chr7:1476438-1510544 | chr7:555359-718687 | 4 | 1.034 | | | Yes |
| HCC1954 | *GALNT7* | ORC4L | Inter | GA II | 9 | | chr4:174326478-174481693 | chr2:148408201-148494933 | 15 | 0.409 | 7 | 0.504 | |
| HCC2218 | *SEC16A* | NOTCH1 | Intra | GA II | 14 | Y | chr9:138454368-138497328 | chr9:138508716-138560059 | 6 | 0.000 | 7 | −0.967 | |
| HCC2218 | *POLDIP2* | BRIP1 | Intra | GA II | 8 | | chr17:23697785-23708730 | chr17:57111328-57295702 | 3 | 1.113 | 19 | 3.925 | Yes |
| HCC2218 | *INTS2* | ZNF652 | Intra | GA II | 7 | | chr17:57297509-57360159 | chr17:44721566-44794834 | 9 | 3.925 | 6 | 2.649 | Yes |
| HCC2218 | *INTS2* | TMEM49 | Intra | GA II | 5 | | chr17:57297509-57360159 | chr17:55139644-55272734 | 9 | 3.925 | 18 | 3.202 | Yes |
| HCC2218 | *LRRC59* | NEUROD2 | Intra | GA II | 5 | | chr17:45813592-45829913 | chr17:35013546-35017701 | 3 | 2.649 | 1 | 3.451 | Yes |
| HCC2218 | *PERLD1* | PPM1D | Intra | GA II | 4 | Y | chr17:35082579-35097833 | chr17:56032335-56096818 | 2 | 3.451 | 7 | 3.340 | Yes |
| MCF7 | *BCAS4* | BCAS3 | Inter | GA II | 2788 | | chr20:48844873-48927121 | chr17:56109953-56824981 | 7 | 2.107 | 73 | 2.653 | Yes |
| MCF7 | *ARFGEF2* | SULF2 | Intra | GA II | 305 | Y | chr20:46971681-47086637 | chr20:45719556-45848215 | 11 | 0.823 | 13 | 3.398 | Yes |
| MCF7 | *RPS6KB1* | TMEM49 | Intra | GA II | 78 | Y | chr17:55325224-55382568 | chr17:55139644-55272734 | 5 | 3.412 | 18 | 2.197 | Yes |
| MCF7 | *STK11* | MIDN | Intra | GA II | 25 | | chr19:1156797-1179434 | chr19:1199551-1210142 | 4 | −1.367 | 2 | −0.279 | |
| MCF7 | *PAPOLA* | AK7 | Intra | GA II | 16 | Y | chr14:96038472-96103201 | chr14:95928200-96024865 | 7 | 0.343 | 13 | 0.343 | |
| MCF7 | *AHCYL1* | RAD51C | Inter | GA II | 12 | Y | chr1:110328830-110367887 | chr17:54124961-54166691 | 4 | −0.063 | 5 | 2.788 | Yes |
| MCF7 | *EIF3H* | FAM65C | Inter | GA II | 11 | | chr8:117726235-117837243 | chr20:48636052-48686833 | 12 | 0.456 | 5 | 1.554 | Yes |
| MCF7 | *BC017255* | TMEM49 | Intra | GA II | 10 | | chr17:54538741-54550409 | chr17:55139644-55272734 | 1 | 3.515 | 18 | 2.197 | Yes |
| MCF7 | *ADAMTS19* | SLC27A6 | Intra | GA II | 9 | | chr5:128824001-129102275 | chr5:128329108-128397234 | 30 | 0.051 | 8 | 0.051 | |
| MCF7 | *ARHGAP19* | DRG1 | Inter | GA II | 8 | Y | chr10:98971919-99042403 | chr22:30125538-30160172 | 8 | 0.387 | 5 | −0.420 | |

**Table W2.** (*continued*)

| Sample Name | 5′ Gene | 3′ Gene | Type | Sequencing Platform | No. Reads | Validation | Chromosomal Location | | aCGH Data (5′ and 3′) | | | | Amplicon Status |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | Fusion QPCR | 5′ Gene | 3′ Gene | No. Probe | Average Log Ratio | No. Probe | Average Log Ratio | |
| MCF7 | *MYO9B* | FCHO1 | Intra | GA II | 8 | Y | chr19:17047590-17185104 | chr19:17719526-17760377 | 13 | −1.126 | 4 | −0.529 | |
| MCF7 | *HSPE1* | PREI3 | Intra | GA II | 6 | Y | chr2:198072965-198076432 | chr2:198089016-198125760 | 1 | −0.361 | 4 | −0.361 | |
| MCF7 | *PARD6G* | C18ORF1 | Intra | GA II | 6 | | chr18:76016105-76106388 | chr18:13601664-13642753 | 10 | −0.674 | 5 | −0.407 | |
| MCF7 | *TRIM37* | TMEM49 | Intra | GA II | 6 | Y | chr17:54414781-54539048 | chr17:55139644-55272734 | 14 | 3.515 | 18 | 2.197 | Yes |
| MCF7 | *SMARCA4* | CARM1 | Intra | GA II | 5 | Y | chr19:10955827-11033958 | chr19:10843252-10894448 | 8 | 0.041 | 6 | 0.041 | |
| MCF7 | *BCAS4* | ZMYND8 | Intra | GA II | 4 | Y | chr20:48844873-48927121 | chr20:45271787-45418881 | 7 | 2.107 | 15 | 3.860 | Yes |
| MCF7 | *PVT1 (BC041065)* | MYC | Intra | GA II | 4 | Y | chr8:128875961-129182681 | chr8:128817496-128822862 | 27 | 1.186 | 3 | 1.186 | Yes |
| MCF7 | *TRIM37* | RNFT1 | Intra | GA II | 3 | | chr17:54414781-54539048 | chr17:55384504-55396899 | 14 | 3.515 | 2 | 3.412 | Yes |
| MDA-MB-361 | *TMEM104* | CRKRS | Intra | GA II | 18 | Y | chr17:70284216-70347517 | chr17:34871265-34944326 | 9 | 2.327 | 7 | 1.529 | Yes |
| MDA-MB-361 | *TANC1* | MTMR4 | Inter | GA II | 12 | Y | chr2:159533391-159797416 | chr17:53921891-53950250 | 27 | 0.000 | 6 | 1.658 | Yes |
| MDA-MB-361 | *TOX3* | GNAO1 | Intra | GA II | 7 | | chr16:51029418-51139215 | chr16:54782751-54939612 | 10 | −0.157 | 19 | 0.281 | |
| MDA-MB-453 | *MECP2* | TMLHE | Intra | GA II | 8 | | chrX:152948879-153016382 | chrX:154375389-154495816 | 8 | 1.611 | 11 | 1.602 | Yes |
| MDA-MB-453 | *MYO15B* | MAP3K3 | Intra | GA II | 4 | | chr17:71095733-71134522 | chr17:59053532-59127402 | 3 | 0.543 | 10 | 0.494 | |
| MDA-MB-468 | *UBR5* | SLC25A32 | Intra | GA II | 8 | | chr8:103334744-103493671 | chr8:104480041-104496644 | 18 | 0.070 | 4 | 0.927 | Yes |
| MDA-MB-468 | *ARID1A* | MAST2 | Intra | GA II | 5 | Y | chr1:26895108-26981188 | chr1:46041871-46274383 | 10 | 0.266 | 23 | 0.818 | |
| MDA-MB-468 | *EGFR* | POLD1 | Inter | GA II | 5 | | chr7:55054218-55203822 | chr19:55579404-55613083 | 17 | 4.944 | 4 | 0.732 | Yes |
| MDA-MB-468 | *RDH13* | FBXO3 | Inter | GA II | 3 | | chr19:60247503-60266397 | chr11:33724866-33752647 | 2 | 0.853 | 3 | 1.507 | Yes |
| UACC-893 | *FBXL20* | CRKRS | Intra | GA II | 31 | Y | chr17:34662422-34811435 | chr17:34871265-34944326 | 17 | 2.069 | 7 | 4.175 | Yes |
| UACC-893 | *CCDC6* | ANK3 | Intra | GA II | 27 | Y | chr10:61218511-61336420 | chr10:61458164-61570752 | 17 | 0.890 | 13 | 0.890 | |
| UACC-893 | *grb7V* | PPP1R1B | Intra | GA II | 23 | Y | chr17:35152031-35157064 | chr17:35038278-35046404 | 1 | 4.843 | 2 | 4.843 | Yes |
| UACC-893 | *MED1* | IKZF3 | Intra | GA II | 9 | Y | chr17:34814063-34861053 | chr17:35174724-35273967 | 4 | 3.908 | 10 | 4.843 | Yes |
| UACC-893 | *EIF2AK3* | PRKD3 | Intra | GA II | 5 | | chr2:88637373-88708209 | chr2:37331149-37397726 | 8 | 1.213 | 8 | 1.278 | Yes |
| ZR-75-1 | *FOXJ3* | CAMTA1 | Intra | GA II | 10 | | chr1:42414796-42573490 | chr1:6767970-6854694 | 17 | −0.380 | 10 | −0.089 | |
| ZR-75-1 | *GPATCH3* | CAMTA1 | Intra | GA II | 10 | | chr1:27089565-27099549 | chr1:6767970-6854694 | 3 | −0.225 | 10 | −0.089 | |
| ZR-75-1 | *C1ORF151* | RCC2 | Intra | GA II | 9 | | chr1:19796057-19828901 | chr1:17605837-17637644 | 4 | −0.013 | 4 | −0.225 | |

Fusions with a recurrent partner are highlighted in yellow.

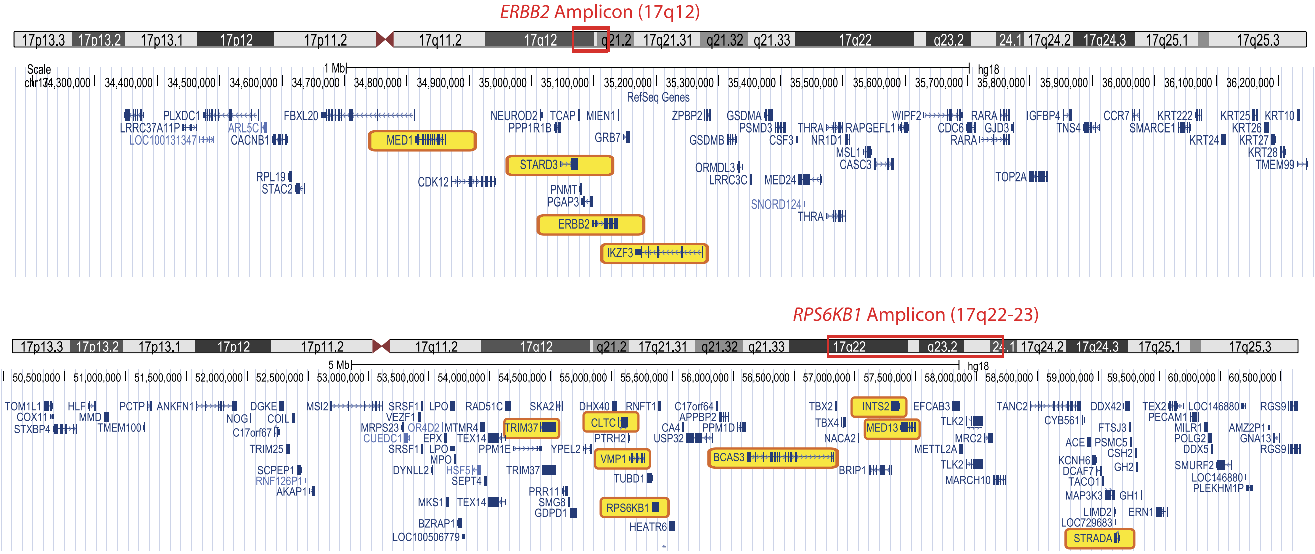


**Figure W1.** UCSC tracks displaying the *ERRB2* and *RPS6KB1* amplicons, with fusion genes highlighted in yellow.

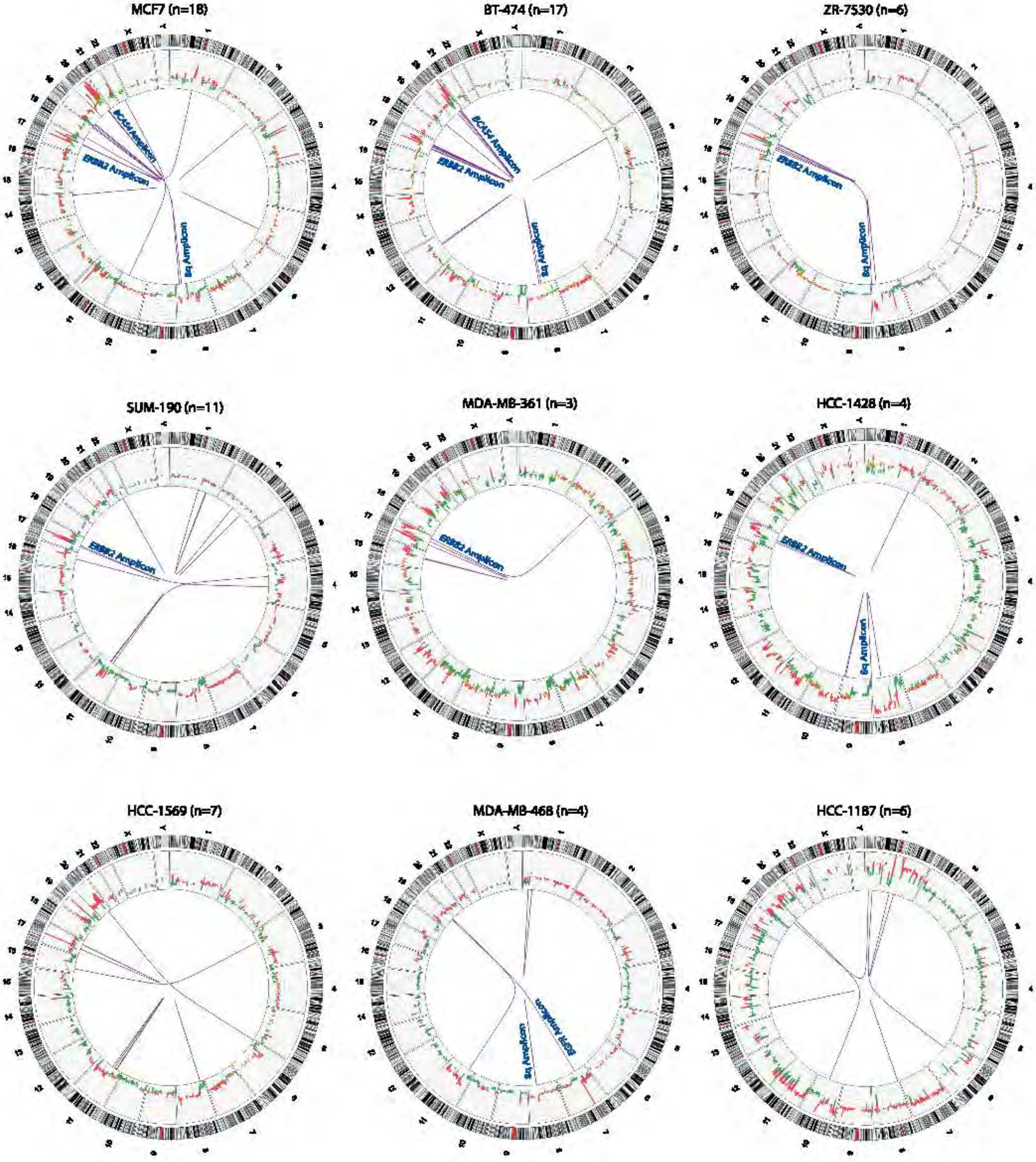


**Figure W2.** Graphical representation of integrative analysis of gene fusions with copy number analysis. Circos plots of the genome-wide distribution of gene fusions along with status of copy number alterations. Red and green peaks represent amplifications and deletions; purple line represents the fusions associated with amplicons and nonamplicons, respectively. "*n*" refers to the total number of fusions identified.

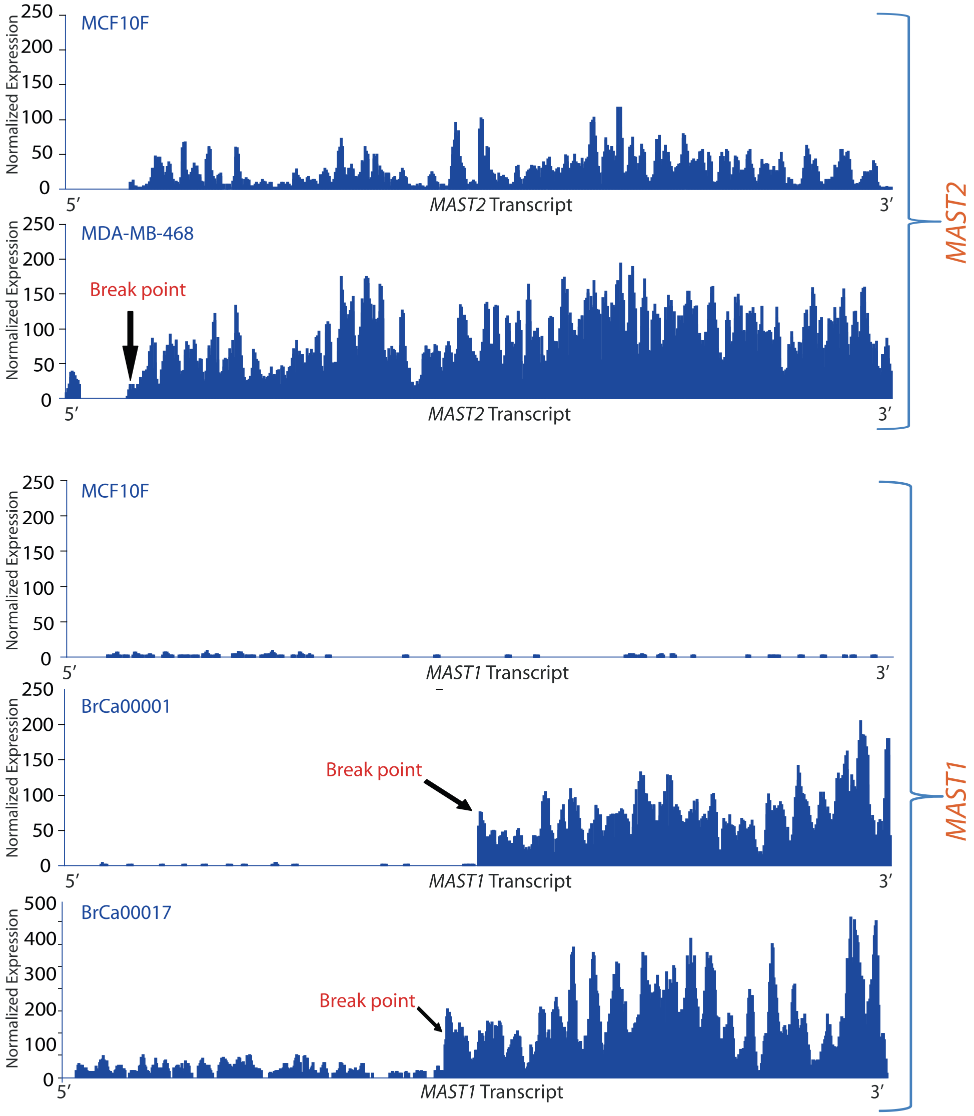


**Figure W3.** Plot of normalized coverage of *MAST1* and *MAST2* transcripts in *MAST* fusion-positive samples (breakpoint indicated by arrow).

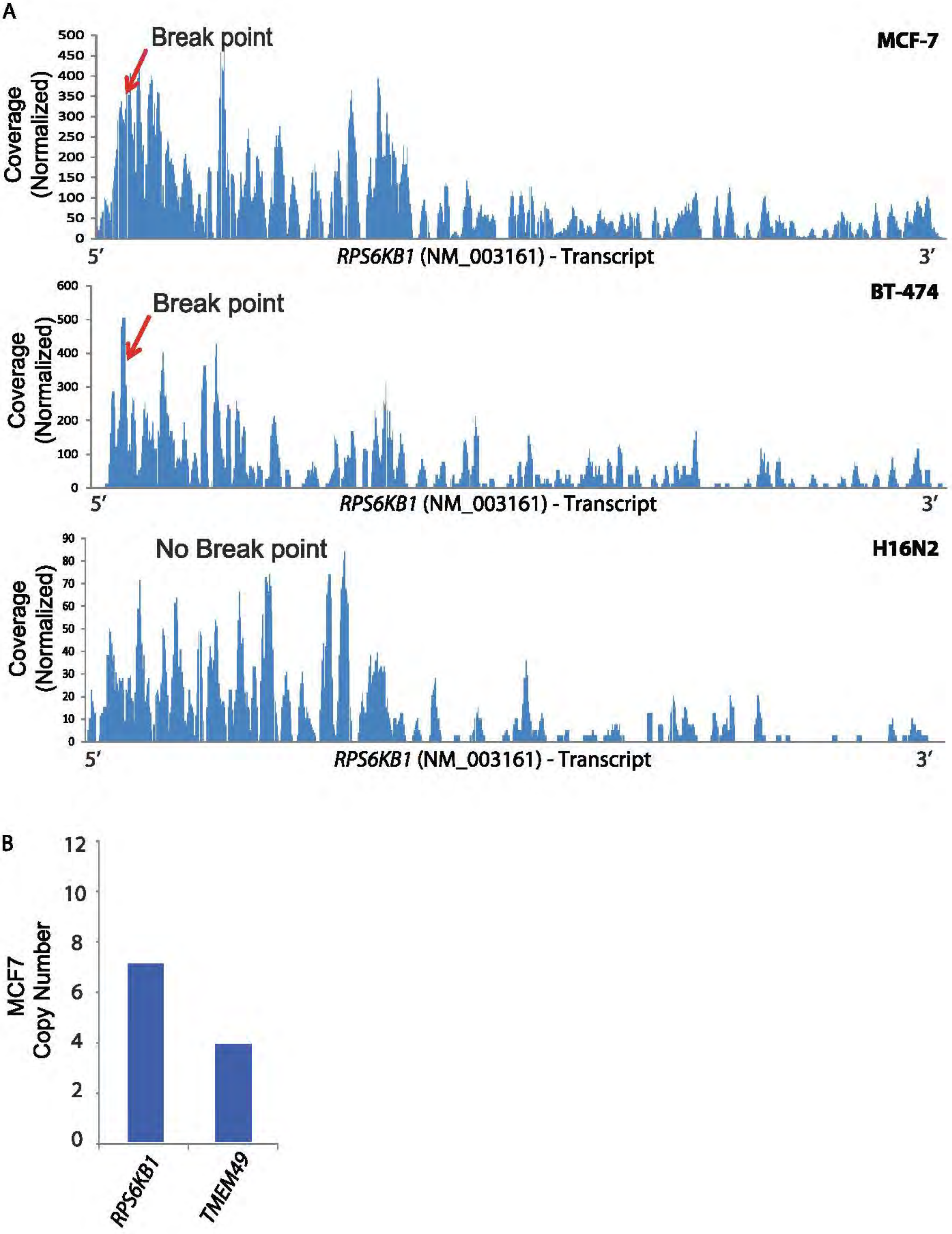


**Figure W4.** (A) Plot of normalized coverage of *RPS6KB1* transcript in BT-474, MCF7, and H16N2 cell lines. (B) Bar graph representing the copy number of *RPS6KB1* and *TMEM49* in MCF7.

RESEARCH ARTICLE

# Outlier Kinase Expression by RNA Sequencing as Targets for Precision Therapy

Vishal Kothari[1], Iris Wei[2], Sunita Shankar[1,3], Shanker Kalyana-Sundaram[1,3,8], Lidong Wang[2], Linda W. Ma[1], Pankaj Vats[1], Catherine S. Grasso[1], Dan R. Robinson[1,3], Yi-Mi Wu[1,3], Xuhong Cao[7], Diane M. Simeone[2,4,5], Arul M. Chinnaiyan[1,3,4,6,7], and Chandan Kumar-Sinha[1,3]

**ABSTRACT** Protein kinases represent the most effective class of therapeutic targets in cancer; therefore, determination of kinase aberrations is a major focus of cancer genomic studies. Here, we analyzed transcriptome sequencing data from a compendium of 482 cancer and benign samples from 25 different tissue types, and defined distinct "outlier kinases" in individual breast and pancreatic cancer samples, based on highest levels of absolute and differential expression. Frequent outlier kinases in breast cancer included therapeutic targets like *ERBB2* and *FGFR4*, distinct from *MET*, *AKT2*, and *PLK2* in pancreatic cancer. Outlier kinases imparted sample-specific dependencies in various cell lines, as tested by siRNA knockdown and/or pharmacologic inhibition. Outlier expression of polo-like kinases was observed in a subset of *KRAS*-dependent pancreatic cancer cell lines, and conferred increased sensitivity to the pan-PLK inhibitor BI-6727. Our results suggest that outlier kinases represent effective precision therapeutic targets that are readily identifiable through RNA sequencing of tumors.

**SIGNIFICANCE:** Various breast and pancreatic cancer cell lines display sensitivity to knockdown or pharmacologic inhibition of sample-specific outlier kinases identified by high-throughput transcriptome sequencing. Outlier kinases represent personalized therapeutic targets that could improve combinatorial therapy options. *Cancer Discov; 3(3); 280–93. ©2013 AACR.*

*See related commentary by Yegnasubramanian and Maitra, p. 252.*

**Authors' Affiliations:** [1]Michigan Center for Translational Pathology, Departments of [2]Surgery and [3]Pathology, University of Michigan Medical School; [4]Comprehensive Cancer Center, Departments of [5]Molecular and Integrative Physiology and [6]Urology, University of Michigan Medical Center; [7]Howard Hughes Medical Institute, Ann Arbor, Michigan; and [8]Department of Environmental Biotechnology, Bharathidasan University, Tiruchirappalli, India

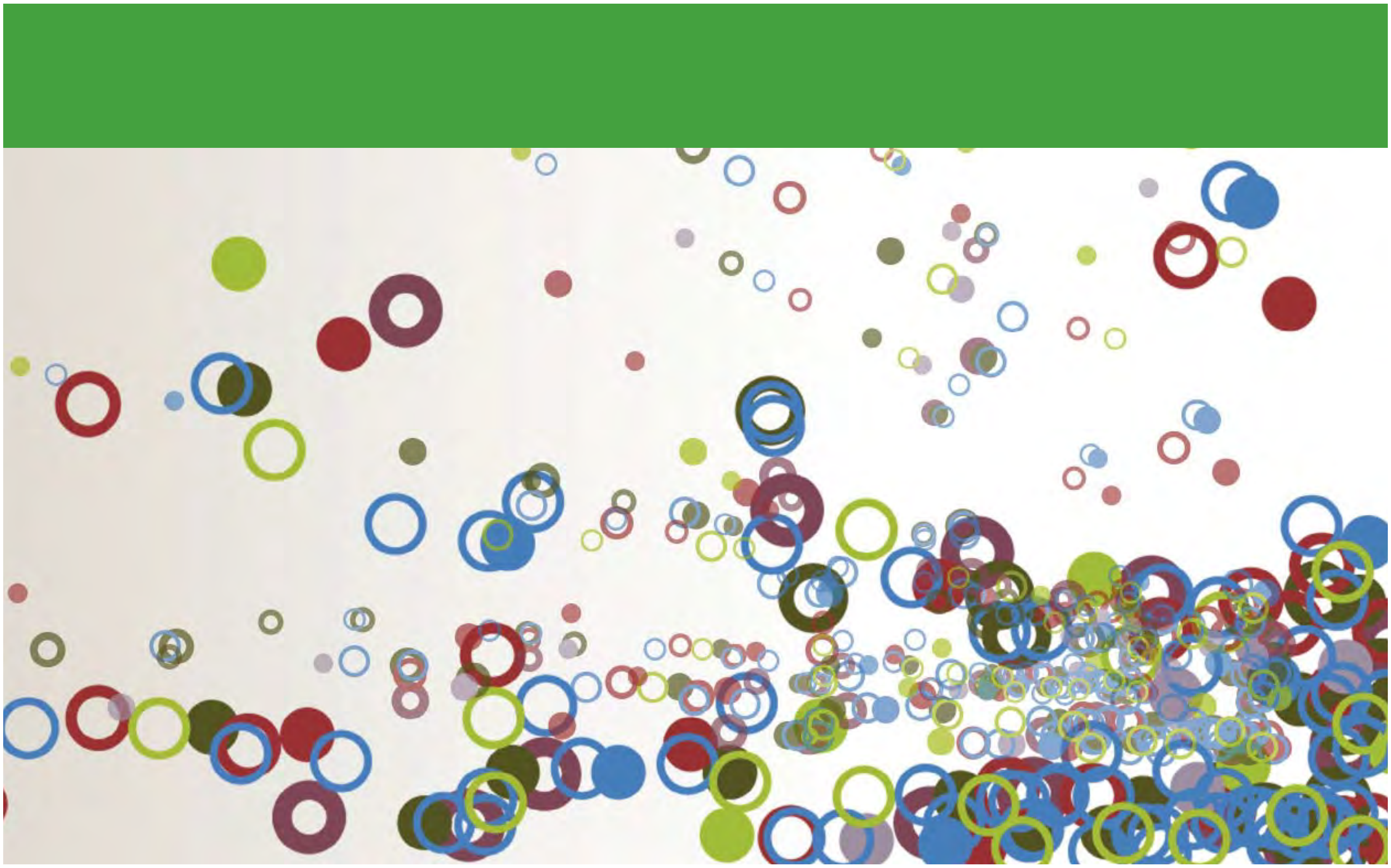
**Note:** Supplementary data for this article are available at Cancer Discovery Online (http://cancerdiscovery.aacrjournals.org/).

V. Kothari, I. Wei, and S. Shankar contributed equally to this work.

A.M. Chinnaiyan and C. Kumar-Sinha shared senior authorship.

**Corresponding Author:** Chandan Kumar-Sinha, Michigan Center for Translational Pathology, Department of Pathology, University of Michigan Medical School, 1400 East Medical Center Drive 5316 CCGC, Ann Arbor, MI 48109. Phone: 734-936-2592; Fax: 734-615-4055; E-mail: chakumar@med.umich.edu

**doi:** 10.1158/2159-8290.CD-12-0336



## INTRODUCTION

The dependence of cancers on a primary driver, most often a kinase (1, 2), forms the guiding principle of targeted therapy that has had some notable clinical successes, such as imatinib for *BCR-ABL*–positive chronic myeloid leukemia, trastuzumab and lapatinib for *ERBB2*-positive breast cancers, gefitinib for lung cancers with kinase domain mutations in *EGFR* (3, 4), and, more recently, crizotinib for lung cancers with *ALK* gene fusions (5). Thus, protein kinases are the mainstay of a majority of the current targeted therapeutic strategies for cancers, and inhibitors of several oncogenic kinases such as AKT, BRAF, CDKs, KIT, RET, SRC, MAPKs, MET, PIK3CA, PLKs, AURKs, S6Ks, and VEGFR are in various stages of clinical use, trials, or development (4, 6). While activating somatic mutations are associated with a few of these genes, overexpression of kinases (resulting from genomic amplification or other underlying somatic aberrations) is often a strong indicator of aberrant activity that may impart dependence of cancer cells.

Pancreatic cancer is the fourth leading cause of cancer-related deaths in the United States, with the worst prognosis (5-year survival <3%) of all major malignancies (7), due to diagnosis of the disease at an advanced, unresectable stage and poor responsiveness to chemo-/radiotherapy (8, 9). The overarching oncogenic driver of pancreatic cancer is mutant *KRAS*, which has eluded therapeutic interventions (10, 11), spurring the search for alternative targets (11). The identification of distinct kinases in independent screens for synthetic lethal interactors of *KRAS* (12–14) led us to systematically explore the expression profiles of all 468 human kinases (the kinome) to identify and test "personalized kinase targets" in a panel of pancreatic cancer cell lines.

Next-generation sequencing of transcriptomes offers significant advantages over microarrays in terms of throughput, elimination of probe bias, and simultaneous monitoring of diverse components of transcriptome biology (15), including gene expression (15–18), alternative splicing (19, 20), chimeric/read-through transcripts (21, 22), and noncoding transcripts (23, 24). Furthermore, transcriptome sequencing affords a direct and quantitative readout of transcript abundance, facilitating sample-wise gene expression analyses using a digital metric of normalized fragment reads, which are not possible using microarrays. Here, we set out to use transcriptome data from a compendium of 482 cancer and benign samples from 25 different tissue types to carry out gene expression profiling of the complete complement of kinases in the human genome, the kinome, to identify "individual sample-specific outlier kinases" inspired by the concept of cancer outlier profile analysis (COPA; refs. 25, 26). Importantly, while COPA analysis was used to identify subsets of "samples displaying outlier expression of candidate genes," here, we interrogated subsets of "outlier genes in individual samples," focusing on kinases that display the highest levels of absolute expression among all the kinases in a sample and the highest levels of differential expression compared with the median level of expression of the respective gene(s) across the compendium. As proof-of-concept, we observed outlier expression of the therapeutic target *ERBB2* specifically in all the breast cancer cell lines analyzed that are known to be *ERBB2* positive. Thus, we hypothesized that specific outlier kinases in other samples may also impart "dependence" owing to clonal selection for extremely high expression and may thereby represent personalized therapeutic targets.

Here, we analyzed kinome expression profiles of breast and pancreatic cancer samples to identify sample-specific outlier kinases. Next, focusing on cell lines displaying outlier expression of kinases with available therapeutics or pharmacologic inhibitors, we tested their dependence on specific outlier kinases compared with nonspecific targets using short hairpin RNA (shRNA) or siRNA and/or small-molecule inhibitors to assess their effects on cell proliferation. Using this approach, we identified several cell line–specific dependencies as well as kinase targets showing enhanced effects with *ERBB2* inhibition in breast and *KRAS* knockdown in pancreatic cancer cells.

## RESULTS

### Delineation of Cancer-Specific Kinome Outlier Profiles Using Transcriptome Sequencing Data

Taking advantage of the direct and unbiased readout of gene expression in terms of defined RNA sequencing (RNA-Seq) reads, we carried out a systematic analysis of the human kinome expression in cancer. RNA-Seq–based, normalized read-counts of all 468 kinases available in our transcriptome compendium, composed of 482 samples from 25 different tissue types, revealed distinct kinases expressed at very high levels—both in absolute terms and in the context of their typical range of expression levels—in virtually all the samples examined (Supplementary Table S1).

Querying individual breast cancer samples (43 cell lines and 67 tissues) for kinases that display the highest levels of absolute expression [>20 reads per kb transcript per million total reads in the given sequencing run (RPKM)] among all the kinases in an individual sample and the highest levels of differential expression compared with the median level of expression of the respective gene across the compendium (>5-fold), we identified outlier kinases across the cohort of breast cancer samples (Fig. 1A and Supplementary Table S2). In addition, each of the outliers was assessed for significant Mahalanobis distance from the center of the scatter plot distribution ($\chi^2$ test, $P < 0.05$) to prioritize sample-specific kinase outliers. For example, in the breast cancer cell line BT-474, *ERBB2* is the predominant outlier kinase (Fig. 1A, inset). Remarkably, with this approach, all breast cancer cell lines known to be *ERBB2*-positive were scored as displaying an outlier expression of *ERBB2*. Interestingly, many *ERBB2*-positive cell lines also displayed outlier expression of additional kinase genes like *CDK12* (Fig. 1A, inset), *FGFR4*, and/or *RET*, among others (Supplementary Table S2). Similar to the well-known case of *ERBB2*, we hypothesized that, in general, outlier kinases specific to individual cancer samples could represent additional therapeutic avenues and were thus explored further.

Likewise, kinome expression data from 22 pancreatic cancer cell lines and 13 pancreatic tissue samples also revealed a set of outlier kinases specifically overexpressed in pancreatic cancers (Fig. 1B and Supplementary Table S3), with the outlier kinase profile of a representative pancreatic cancer cell line AsPC-1 depicted in the inset (Fig. 1B). Assessment of outlier kinases in pancreatic and breast cancer cohorts revealed distinct outlier kinase profiles between the 2 diseases. For example, common outlier kinases in breast cancer included *ERBB2*, *FGFR4*, and *RET*, whereas kinases displaying outlier expression across multiple pancreatic cancer samples included *EPHA2*, *MET*, *PLK2*, *MST1R*, and *AKT2*. Interestingly, *AXL* and *EGFR* showed outlier expression in both pancreatic and breast cancer samples.

Before proceeding to test outlier kinase–specific dependencies in individual cell lines, we validated the gene expression readout provided by the RNA-Seq data. First, comparing the gene expression profiles of a prostate cancer cell line DU145 across 4 independent RNA-Seq runs, we observed a robust correlation ($R^2 > 0.96$) between the technical replicates (Supplementary Fig. S1A). Next, we analyzed the variance across RNA-Seq data from a breast cancer cell line, MCF-7, treated with estrogen (0, 3, and 6 hours) as biologic quasi-replicates. Interestingly, we observed an overall high correlation ($R^2 > 0.91$) here also, albeit less than the technical replicates (Supplementary Fig. S1B). Next, we validated the expression profiles of kinase genes derived from RNA-Seq by quantitative reverse-transcription PCR (qRT-PCR) and Western blot analyses. As an example, a strong correlation ($R^2 > 0.88$) was observed between the levels of *MET* expression by RNA-Seq and qRT-PCR, over a range of expression values across a panel of samples (Fig. 2A). In addition, individual samples showing outlier expression of *MET* by RNA-Seq showed distinctly higher expression by qRT-PCR, compared with nonoutlier samples (Fig. 2B). Similarly, we conducted qRT-PCR validation of RNA-Seq data from multiple samples for 8 additional kinases, again showing strong, statistically significant correlations with overall gene expression levels (Supplementary Fig. S2) as well as outlier calls (Supplementary Fig. S3). Furthermore, extending the correlation of outlier expression to protein levels, cell lines with outlier expression of *MET* were found to display higher levels of total as well as phosphorylated MET, compared with cells without outlier expression of *MET* (Fig. 2C). Finally, to assess the feasibility of identifying outlier kinases in cancer tissue samples in the backdrop of underlying benign stromal, vascular, and immune cells, we observed a strong correlation between the RNA-Seq data and outlier calls between a primary tumor-derived xenograft tissue, DS-08-947, and its derivative cell line (Supplementary Fig. S4A and Supplementary Table S4). Similar correlation was observed between BxPC-3 and PANC-1 cell lines and xenograft tissues derived from them (Supplementary Fig. S4B).

### A Subset of *ERBB2*-Positive Breast Cancer Cell Lines Display Outlier Expression of *FGFR4*

Among the *ERBB2*-positive breast cancer cell lines analyzed by RNA-Seq, ZR-75-30 exhibited singular outlier kinase expression of *ERBB2*, whose knockdown resulted in a strong growth inhibition (Fig. 3). However, knockdown of *RPS6KB1*, another oncogenic kinase on chromosome 17 located near the *ERBB2* amplicon and overexpressed in 40% to 50% of breast cancers, did not affect the proliferation rate of ZR-75-30 cells, which do not show outlier expression of *RPS6KB1* (Fig. 3). Many other *ERBB2*-positive cell lines, however, displayed outlier expression of additional kinases, frequently including *FGFR4*, such as MDA-MB-361 and MDA-MB-453 (Fig. 3), as well as MDA-MB-330, HCC202, and HCC1419 (Supplementary Table S2). To assess the dependence on the outlier expression of *FGFR4* in the backdrop of *ERBB2* overexpression, multiple shRNA-encoding lentiviral constructs were used to knock down *FGFR4* in MDA-MB-361 and MDA-MB-453 cells exhibiting outlier expression

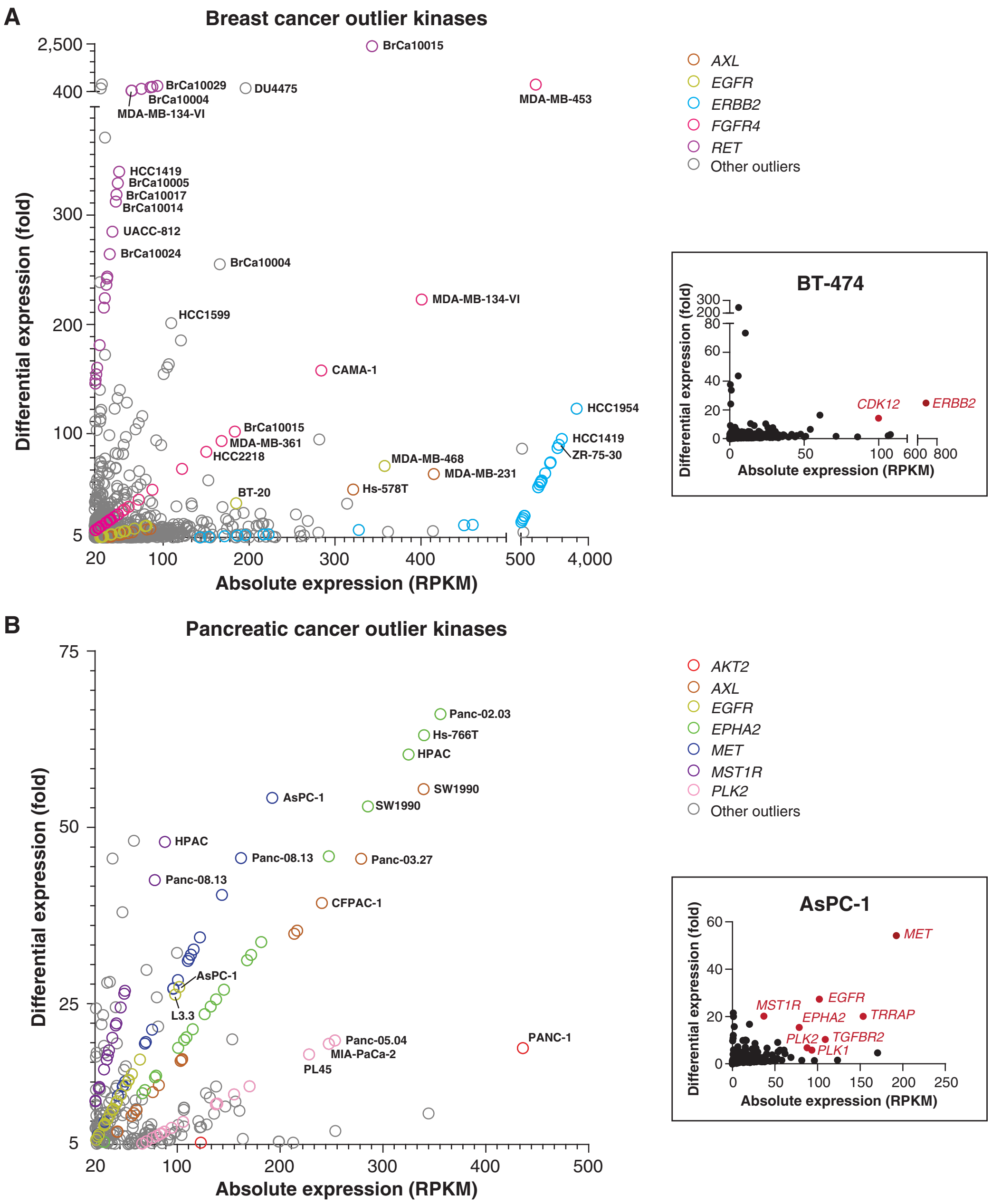


**Figure 1.** Scatter plot representation of outlier kinases in (**A**) breast and (**B**) pancreatic cancer samples. Kinases displaying an absolute expression >20 RPKM and differential expression >5-fold (versus median value across the compendium) were designated as outliers. The colored circles represent salient kinases displaying outlier expression in multiple samples. Examples of sample-specific kinome profiles are shown in the insets (BT-474 breast cancer and AsPC-1 pancreatic cancer cell lines); kinases with statistically significant outlier expression (absolute expression >20 RPKM, differential expression >5 fold, and $P < 0.05$) are highlighted in red.

of both *ERBB2* and *FGFR4*, as well as in CAMA-1, with outlier expression of *FGFR4* but not *ERBB2*. Target knockdown for all siRNA and shRNA experiments were assessed by qRT-PCR and/or Western blot analysis (Supplementary Fig. S5A–S5H). Remarkably, knockdown of *FGFR4* resulted in decreased cell proliferation in all 3 cell lines with *FGFR4* outlier expression (Fig. 3), whereas treatment of these cells with ERBB2-targeting trastuzumab had no effect on the proliferation of CAMA-1 and MDA-MB-361 cells. In contrast, MDA-MB-453 cells showed diminished cell proliferation rates independently upon *FGFR4* knockdown as well as trastuzumab treatment and showed an additive effect upon combined treatment.

To further examine the dependence of a subset of *ERBB2*-positive cells on *FGFR4*, we generated trastuzumab-resistant

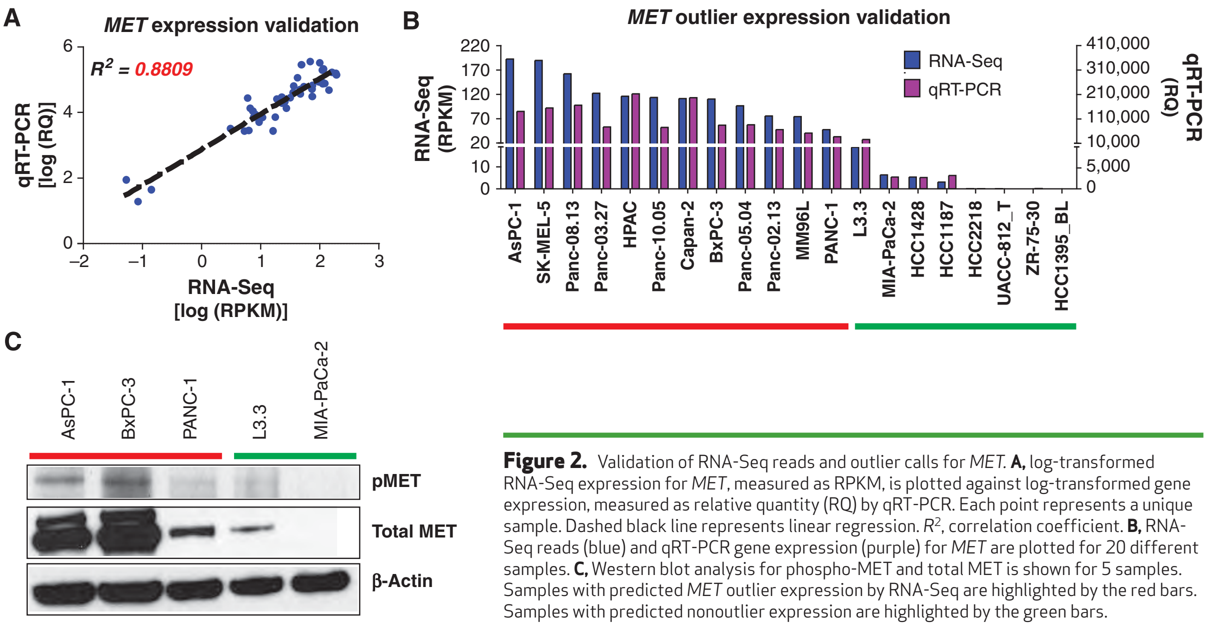


**Figure 2.** Validation of RNA-Seq reads and outlier calls for *MET*. **A,** log-transformed RNA-Seq expression for *MET*, measured as RPKM, is plotted against log-transformed gene expression, measured as relative quantity (RQ) by qRT-PCR. Each point represents a unique sample. Dashed black line represents linear regression. $R^2$, correlation coefficient. **B,** RNA-Seq reads (blue) and qRT-PCR gene expression (purple) for *MET* are plotted for 20 different samples. **C,** Western blot analysis for phospho-MET and total MET is shown for 5 samples. Samples with predicted *MET* outlier expression by RNA-Seq are highlighted by the red bars. Samples with predicted nonoutlier expression are highlighted by the green bars.

sublines of MDA-MB-453 and BT-474, an *ERBB2*-positive breast cancer cell line that does not exhibit *FGFR4* outlier expression (Fig. 4A). Consistent with the experiments involving trastuzumab and shRNA-mediated knockdown of *FGFR4* (Fig. 3), MDA-MB-453 cells were found to be independently responsive to both trastuzumab and PD173074, a small-molecule inhibitor of FGFR, whereas a combined treatment with both of these reagents provided the strongest effect on cell proliferation (Fig. 4B, left). Interestingly, MDA-MB-453 cells, grown to be resistant to trastuzumab, continued to display responsiveness to PD173074 (Fig. 4B, right), suggesting that *FGFR4* represents an independent therapeutic target in a subset of *ERBB2*-positive breast cancer cells. Similar results were obtained with another FGFR inhibitor, dovitinib, which significantly decreased cell proliferation in both the MDA-MB-453 parental and trastuzumab-resistant subline (Fig. 4C, left) but did not affect the BT-474 parental or trastuzumab-resistant subline, neither of which displays *FGFR4* outlier expression (Fig. 4C, right). Next, we carried out dose–response experiments using specific pharmacologic inhibitors against outlier kinases (Supplementary Fig. S6A–S6C). Cell lines exhibiting outlier expression of *FGFR*s displayed a dose-dependent response to PD173074 and dovitinib, with significantly lower $IC_{50}$ values, as compared with cell lines without outlier expression (Supplementary Fig. S6A and S6B). Taken together, these results suggest that a subset of *ERBB2*-positive breast cancers that display outlier expression of *FGFR4* may specifically respond to combined treatment with ERBB2 and FGFR inhibitors more effectively than to *ERBB2*-directed therapy alone.

## Pancreatic Cancer Cell Lines Are Sensitive to Knockdown of Cell-Specific Outlier Kinases

We next extended our kinome outlier analysis to pancreatic cancer, a tumor type critically lacking in rational therapeutic options, particularly in the realm of actionable kinases. Kinome expression profiles of individual pancreatic cancer cell lines were used to identify sample-specific outlier kinases (Fig. 5, left). The pancreatic cancer cell lines were then tested for effects on cell proliferation following siRNA-based knockdown of sample-specific outlier and nonoutlier kinases. Knockdown of the sample-specific outlier kinases—for example, *EGFR* in L3.3, *PLK2* in MIA-PaCa-2, *MET* in BxPC-3, and *AKT2* in PANC-1 cells—inhibited the proliferation of respective cells (Fig. 5, middle). A similar growth inhibition was observed following knockdown of *MET* in HPAC and *AXL* in Panc-08.13 and PL45 cells (Supplementary Fig. S7). Conversely, knockdown of the nonoutlier kinases *AXL* in L3.3, *MET* in MIA-Paca-2, *PLK2* in BxPC-3, and PANC-1 cells did not significantly affect cell growth (Fig. 5, right). Also, L3.3 cells remained unaffected by knockdown of the nonoutlier *PLK2* (Supplementary Fig. S7). These observations strongly support the notion that outlier kinases represent specific therapeutic targets in individual cancer samples.

Notably, knockdown of the outlier kinase *PLK2* in MIA-PaCa-2 cells did not have as profound an effect on cell proliferation as outlier kinase targeting in many other samples (Fig. 5, middle). We hypothesized that this could be due to a pervasive influence of oncogenic *KRAS* activity in these cells. To test this idea next, we analyzed the effect of *KRAS* knockdown in pancreatic cancer cell lines with *PLK* outlier expression.

## Outlier Expression of Polo-Like Kinases Marks a Subset of *KRAS*-Dependent Pancreatic Cancer Cells

A panel of pancreatic cancer cell lines with and without *PLK* outlier expression was stably transduced with 2 independent inducible shRNAs against *KRAS* and assessed for sensitivity to *KRAS* knockdown and/or the PLK inhibitor BI-6727 (Fig. 6). Following induction by doxycycline, the cells expressing *KRAS* shRNAs were distinguished by red fluorescence, resulting from the red fluorescence protein (RFP) tag coexpressed with the shRNA (Fig. 6, middle). *KRAS* knockdown efficiency of approximately 50% or more was

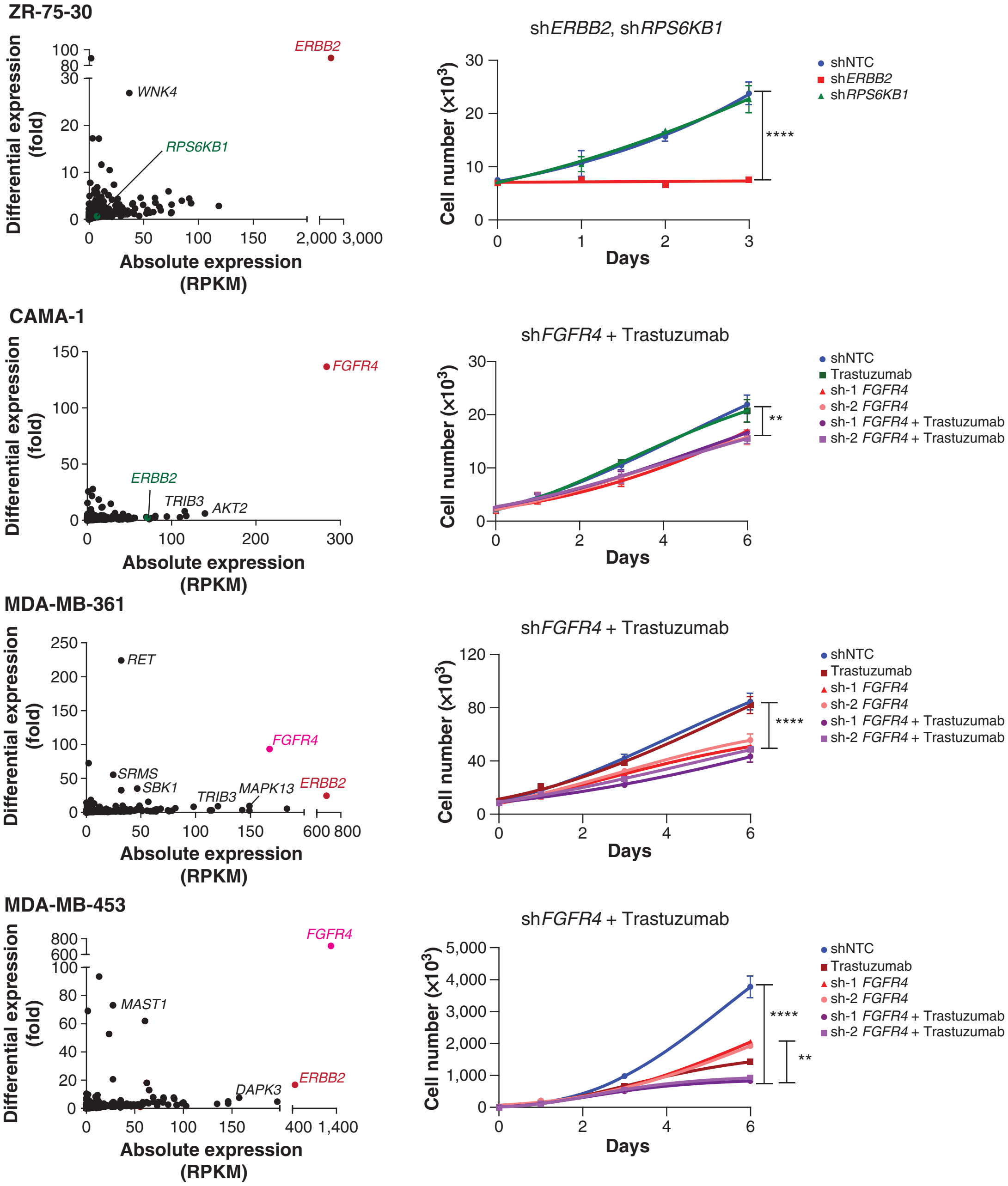


**Figure 3.** Sample-wise outlier kinases in *ERBB2*-positive breast cancer cell lines. Left, the scatter plots display kinome expression profiles of individual breast cancer cell lines. Kinases with (red/pink) and without (green) outlier expression that were targeted for knockdown are shown in color. Labels in black denote additional kinases with outlier expression. Right, growth curves show the effect of targeting outlier (*ERBB2*) versus nonoutlier (*RPS6KB1*) kinases in ZR-75-30 cells and the effects of trastuzumab and/or knockdown of the outlier *FGFR4* in CAMA-1, MDA-MB-361, and MDA-MB-453 cells. Values represent mean ± SD. **, $P < 0.01$; ****, $P < 0.0001$.

obtained in all the cells tested (Supplementary Fig. S5H). Of the cell lines tested, knockdown of *KRAS* significantly inhibited the proliferation of L3.3, MIA-PaCa-2, and Panc-03.27, which all harbor oncogenic mutations in *KRAS* and were therefore designated as *KRAS* dependent (Fig. 6A). BxPC-3 cells, which have wild-type *KRAS*, as well as HPAC and PANC-1 cells, which have mutant *KRAS*, were not affected by *KRAS* knockdown and were therefore categorized as *KRAS* independent (Fig. 6B). Incidentally, all 3 *PLK* outlier cell lines tested here—L3.3, MIA-PaCa-2, and Panc-03.27—were found to be in the *KRAS*-dependent category based on their reduced proliferation upon *KRAS* knockdown (Fig. 6A). Furthermore, treatment with the PLK inhibitor BI-6727 significantly inhibited proliferation in cell lines with *PLK* outlier expression

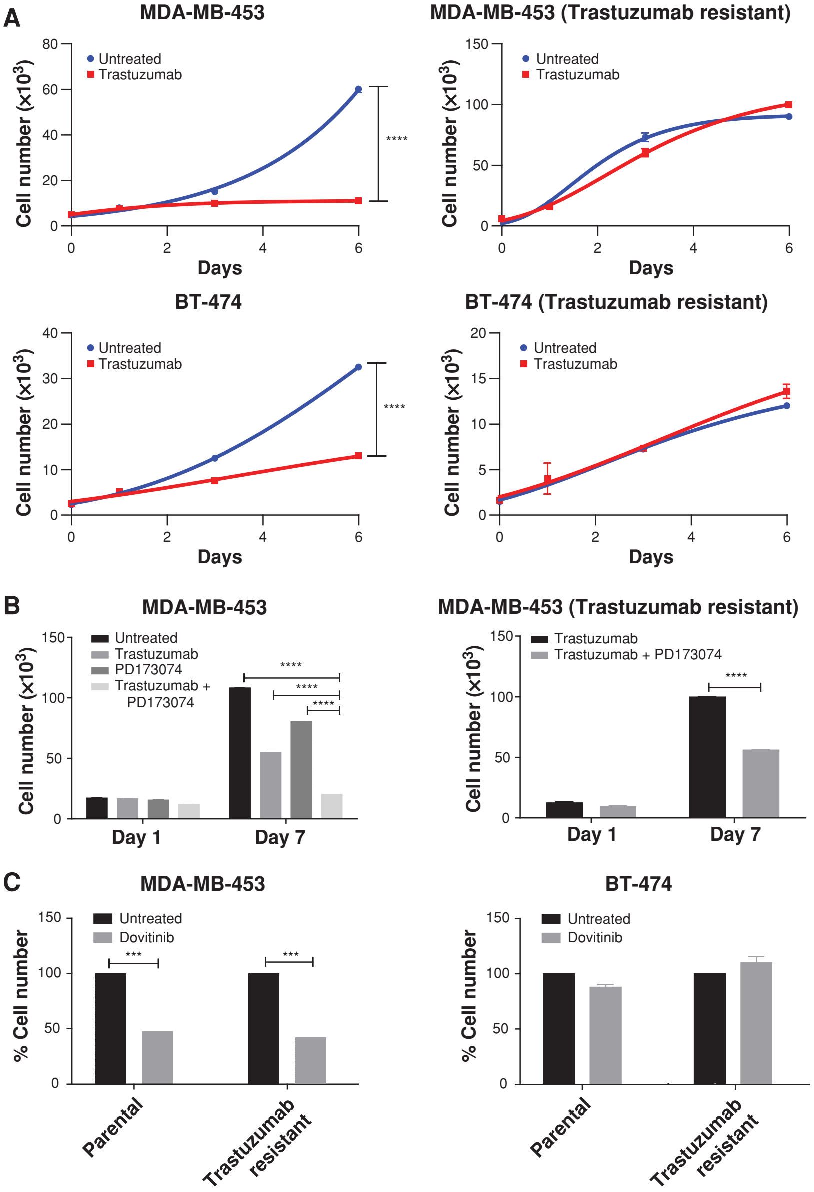


**Figure 4.** Trastuzumab-resistant cell lines respond to targeting of the outlier kinase *FGFR4*. **A,** the growth curves show the effect of trastuzumab treatment on MDA-MB-453 and BT-474 (left) and their trastuzumab-resistant sublines (right). **B,** the bar graphs show the individual and combined effects of trastuzumab and the FGFR inhibitor PD173074 on cell proliferation in MDA-MB-453 (left) and its trastuzumab-resistant subline (right). **C,** the bar graphs display the effect of the FGFR inhibitor dovitinib on parental and trastuzumab-resistant sublines of MDA-MB-453 (with *FGFR4* outlier expression) and BT-474 (without *FGFR4* outlier expression) on day 5. Values represent mean ± SD. ***, $P < 0.001$; ****, $P < 0.0001$.

(Fig. 6A, right) but had no effect in cell lines without *PLK* outlier expression (Fig. 6B, right). The decrease in cell proliferation following BI-6727 treatment was associated with increased apoptosis, as measured by the flow cytometry of Annexin V/propidium iodide–stained cells (Supplementary Fig. S8A). Finally, treatment with BI-6727 in combination with knockdown of *KRAS* enhanced the inhibition of cell proliferation in the *KRAS*-dependent, *PLK* outlier cells (Fig. 6A, right) but had no effect in the *KRAS*-independent cells without *PLK* outlier expression (Fig. 6B, right). Investigating the likely reason for the lack of sensitivity to *KRAS* knockdown in a subset of pancreatic cancer cells harboring oncogenic *KRAS*, we observed that following *KRAS* knockdown, the levels of phospho-ERK, one of the major downstream effector proteins in the *RAS* signaling pathway, were reduced in the *KRAS*-dependent cell lines L3.3 and MIA-PaCa-2, but not in the *KRAS*-independent cell line PANC-1 (Supplementary Fig. S8B), suggesting that ERK activity in PANC-1 cells may be sustained by other convergent pathways. Notably, the *KRAS*-independent cell lines BxPC-3 and PANC-1 did respond to inhibition of their respective outlier kinases, both *in vitro* (Fig. 5, middle) and *in vivo*, as described below.

## Inhibition of Outlier Kinases Inhibits the Growth of Pancreatic Cancer Cell Line Xenografts

To test the effect of inhibiting sample-specific outlier kinases *in vivo*, we treated orthotopic tumor xenografts of 2 *KRAS*-independent pancreatic cancer cell lines, BxPC-3 and PANC-1, established in nonobese diabetic/severe combined immunodeficient (NOD/SCID) mice, with the MET inhibitor

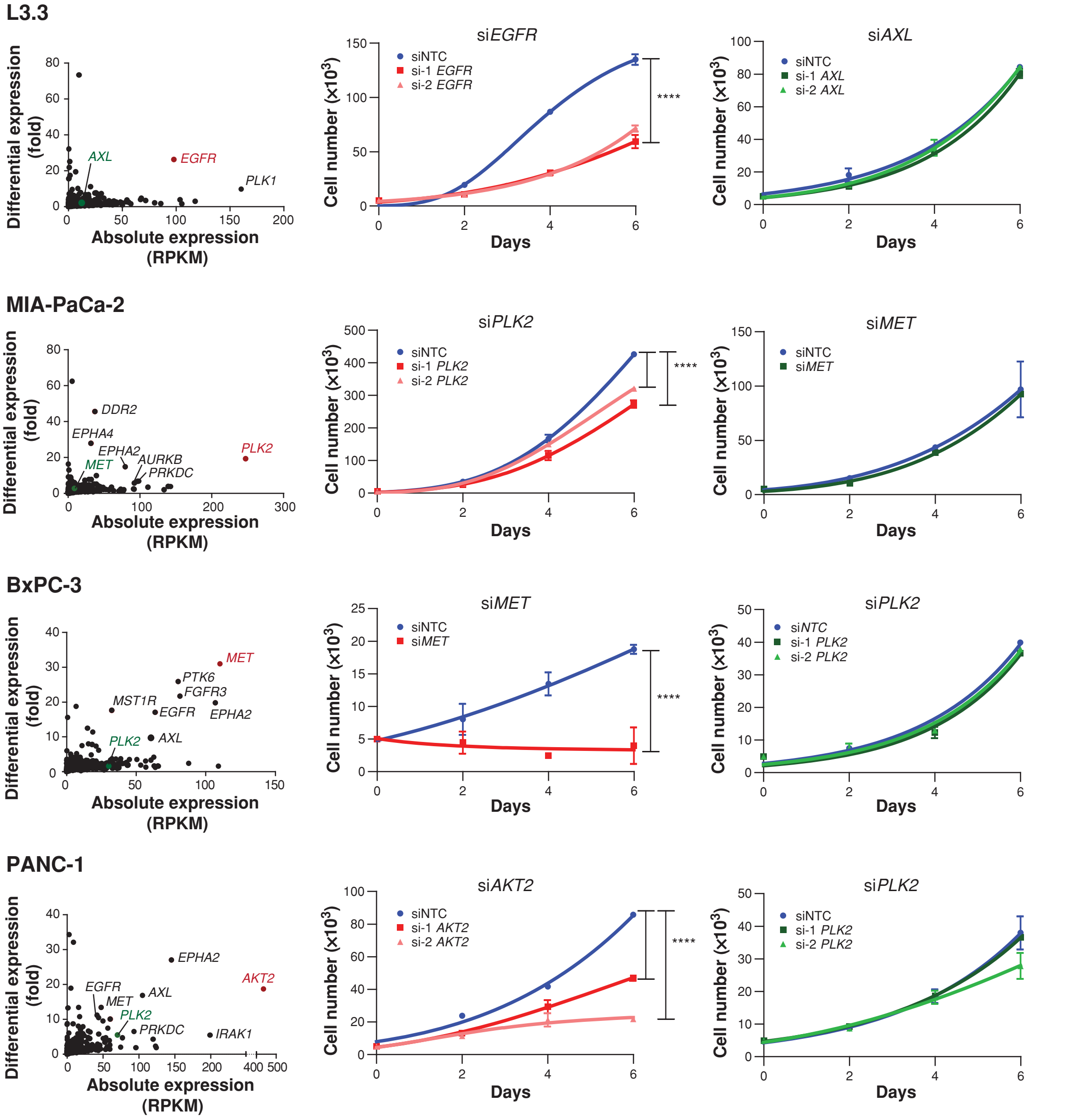


**Figure 5.** Pancreatic cancer cell lines are sensitive to knockdown of outlier kinases. Left, scatter plots display kinome profiles of select pancreatic cancer cell lines; kinases targeted for knockdown are shown in color (red, outliers; green, nonoutliers). Labels in black denote additional kinases with outlier expression. The growth curves display the effects of siRNA-mediated knockdown of sample-specific outliers (middle) and nonoutliers (right) for each cell line. Values represent mean ± SD. ****, $P < 0.0001$.

XL184. BxPC-3 cells and, to a lesser but significant degree, PANC-1 cells, were found to have *MET* outlier expression by RNA-Seq, which was validated by qRT-PCR and Western blot analyses (Fig. 2). Notably, both of these cell lines also displayed a dose-dependent response to XL184 *in vitro*, with significantly lower $IC_{50}$ values compared with the L3.3 cell line that does not have outlier expression of *MET* (Supplementary Fig. S6C). Consistent with our hypothesis of dependence on outlier kinases, growth of both BxPC-3 and PANC-1 xenografts was also significantly inhibited by treatment with XL184, as measured by tumor volume and weight (Fig. 7A–C). Of note, no significant difference was found in the body weight of XL184-treated and untreated mice, suggesting that the effective dose of the inhibitor caused no measurable toxicity *in vivo* (Fig. 7D).

The specificity of response to the MET inhibitor XL184 was analyzed by Western blot analysis, which showed a sharp decrease in phospho-MET levels in BxPC-3 and to a relatively lesser extent in PANC-1 cells following treatment with XL184 (Fig. 7E). Considering that *AKT2* represents the predominant outlier kinase in PANC-1 cells (Mahalanobis distance 217.6, *P* ~ 0; Supplementary Table S3), lending

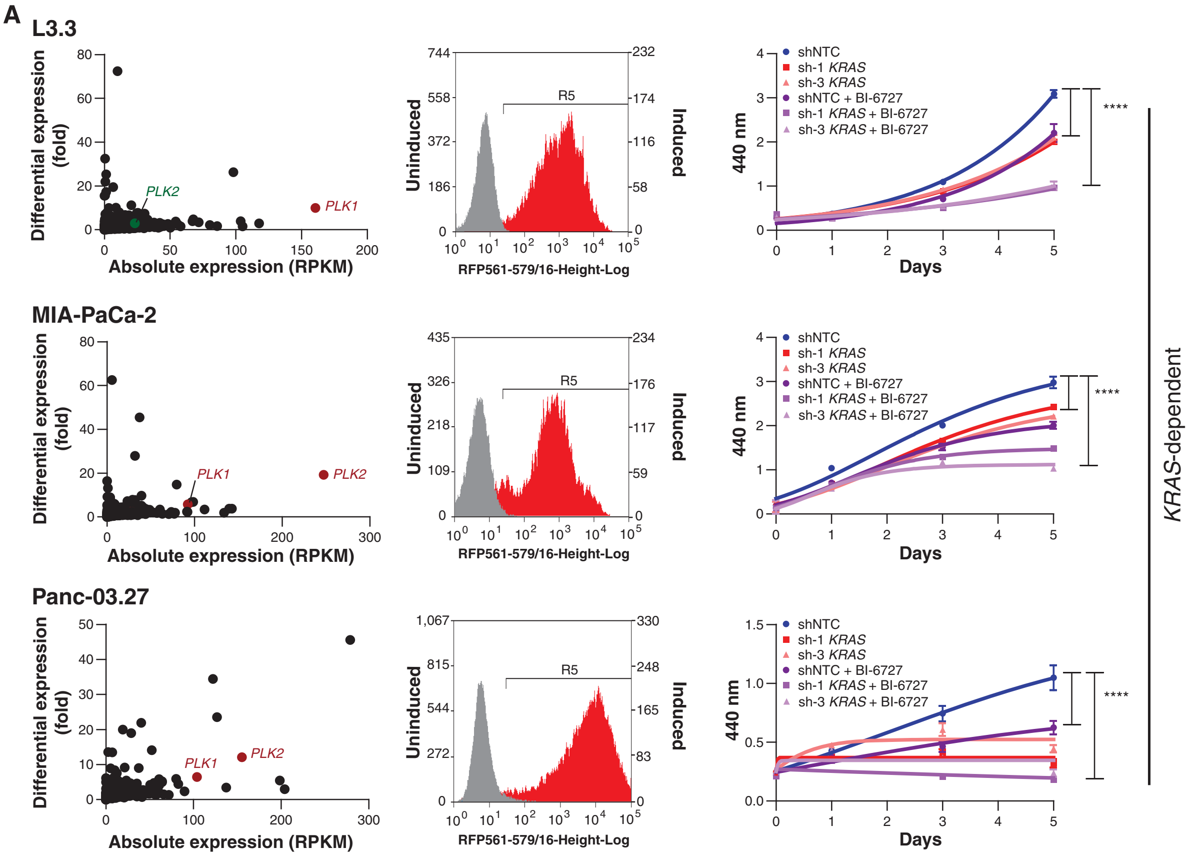


**Figure 6.** Knockdown of *KRAS* combined with PLK inhibition reduces cell proliferation in indicated *KRAS*-dependent cell lines (**A**) but not in *KRAS*-independent cell lines (**B**). (*continued on following page*)

significant dependence on *AKT2* (Fig. 5), we queried whether the profound inhibitory effect of XL184 on PANC-1 xenografts was also mediated through nonspecific targeting of AKT. Western blot analysis of PANC-1 xenograft tumor lysates revealed a markedly decreased level phospho-AKT following XL184 treatment (Fig. 7F). This supports the notion that XL184 suppresses PANC-1 proliferation through inhibition of both AKT and MET signaling. Thus, PANC-1 represents an example of a cancer sample showing dependency on multiple actionable outliers that may respond to a combinatorial therapeutic option or appropriate pan-kinase inhibitors.

## DISCUSSION

The advent of high-throughput sequencing enables a comprehensive characterization of the genomic and transcriptomic landscape of individual cancer samples, inexorably leading to the challenge of defining and prioritizing clinically relevant findings to translate into improved diagnostic and therapeutic options (27, 28). Clinical sequencing of cancers aims to identify actionable genomic aberrations and match patients with available therapies. Protein kinases, being central to biologic and disease processes, including cancer, and being therapeutically targetable, constitute a large proportion of available and potential targets; thus, any novel disease-specific kinase aberrations are of great clinical interest. This study proposes and tests the hypothesis that specific kinases showing outlier expression in individual cancer samples impart "dependence" on the cells, which may be targeted in combination with existing treatment modalities. Importantly, a case is made for considering the entire profile of kinome aberrations to prioritize potentially effective targets.

The "sample-centric" analysis of kinome expression revealed unique profiles of outlier kinases that were tested for dependency. The receptor tyrosine kinase ERBB2 overexpressed in 20% to 30% of breast cancers confers a more aggressive phenotype, increased metastasis, and worse patient prognosis (29, 30). In our outlier kinase analysis, several well-known "*ERBB2*-positive" breast cancer cell lines, including MDA-MB-361 and MDA-MB-453, were found to display outlier expression of *ERBB2*, as expected, but frequently also an outlier expression of the therapeutic target *FGFR4*. Notably, a survey of microarray-based gene expression data in Oncomine (31, 32) also displayed a subset of *ERBB2*-positive primary breast cancer samples with outlier expression of *FGFR4* (data not shown), emphasizing the clinical relevance of our observations. Targeting outlier *FGFR4* in *ERBB2*-positive breast cancer samples was found to confer independent as well as additive inhibitory effects

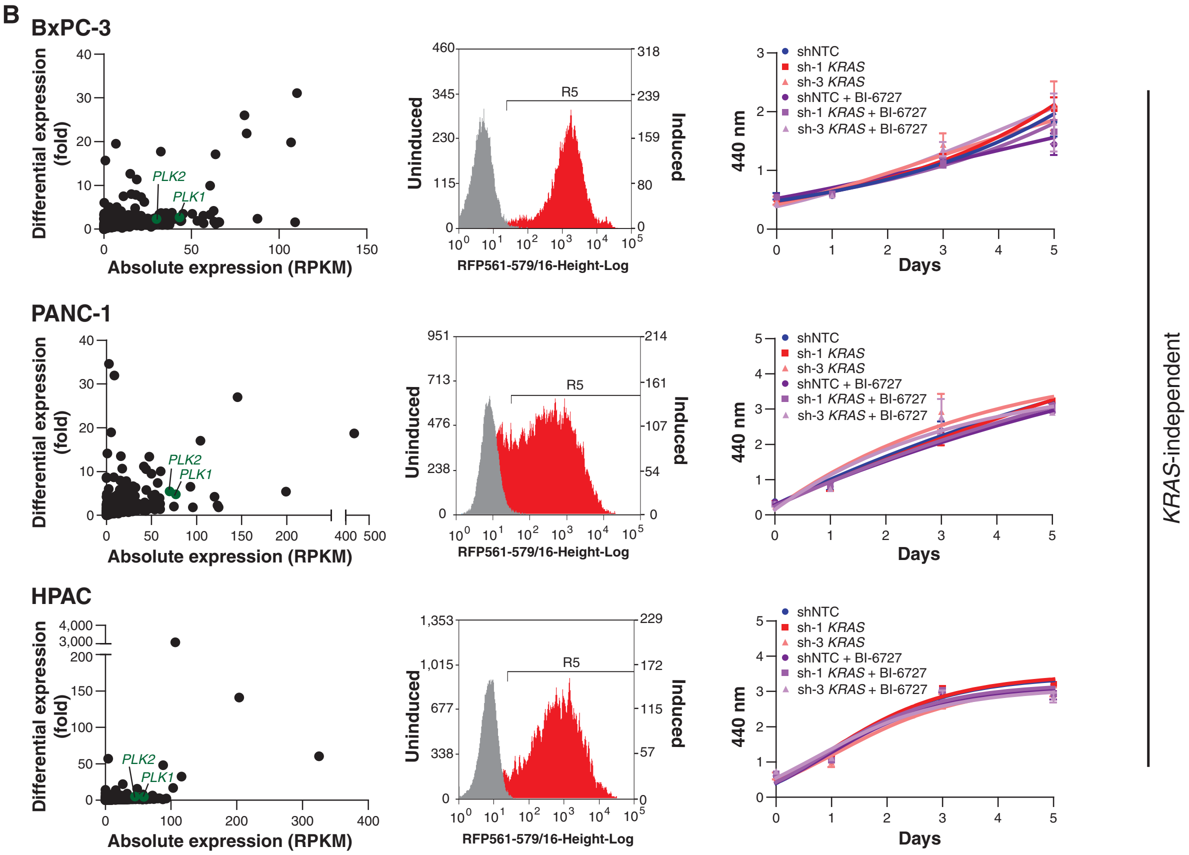


**Figure 6.** ***(Continued)*** The scatter plots show the absolute and differential expressions of *PLK1* and *PLK2* for each cell line (left). The flow cytometric profiles of doxycycline-induced cells expressing *KRAS* shRNA with RFP expression (red) versus uninduced cells (gray) are displayed (middle).The growth curves show the individual and combined effects of *KRAS* shRNA and the PLK inhibitor BI-6727, using WST-1 assay measured at 440 nm absorbance (right). Values represent mean ± SD. ****, *P* < 0.0001.

upon their combined knockdown (Fig. 3), highlighting the potential of combining 2 or more outlier kinase targets in treating cancer, even in cases with a predominant driver such as *ERBB2*. Interestingly, we also observed that the *ERBB2*-positive MDA-MB-453 cells grown resistant to trastuzumab treatment continued to remain dependent on *FGFR4* and responded to FGFR inhibitors (Fig. 4). In clinical trials with *ERBB2*-positive metastatic breast cancer, 50% to 74% patients have been reported as not responsive to trastuzumab monotherapy or in combination with chemotherapy (33, 34). Our results suggest that the *ERBB2*-positive breast cancers may be partly dependent on additional drivers, such as FGFR4, RET, EGFR, and MET, which may sustain these cancers following therapeutic abrogation of ERBB2 activity. Another important corollary to our observations is that combinatorial targeting of ERBB2 and additional outlier kinases at the outset may be much more effective than approaching a single target at a time, a concept that warrants further study. Furthermore, each cancer sample needs to be investigated individually to rationally determine patient-specific unique target combinations.

Next, we extended the approach of nominating sample-specific outlier kinases to pancreatic cancer, which is characterized by a bleak prognosis due to presentation at an advanced stage and resistance to traditional chemotherapy and radiation in the setting of its pancreatic cancer sanctuary, encompassing tumor stroma, extracellular matrix, tumor-infiltrating immune cells, and cancer stem cells. Given the paucity of effective targets in pancreatic cancer, the strong response of pancreatic cancer cell lines to knockdown or inhibition of *a priori* designated outlier kinases is a promising lead. Our results also underscore the importance of matching sample-specific actionable targets with the appropriate therapeutics. For example, targeting MET was found to be more effective in pancreatic cancer cell lines with *MET* outlier expression than in nonoutlier samples. Notably, many of our experimental results are consistent with several anecdotal studies using kinase inhibitors against EGFR, MET, and AKT2 (35–39).

We also examined the effect of targeting sample-specific outlier kinases in conjunction with the oncogenic *KRAS* mutation that is present in virtually all cases of pancreatic cancer. Consistent with previous reports (40–42), we observed that only a subset of *KRAS*-mutant cells display *KRAS* dependency. Using tetracycline (tet)–sh*KRAS* stable cell lines, we determined L3.3, MIA-PaCa-2, and Panc-03.27 cells

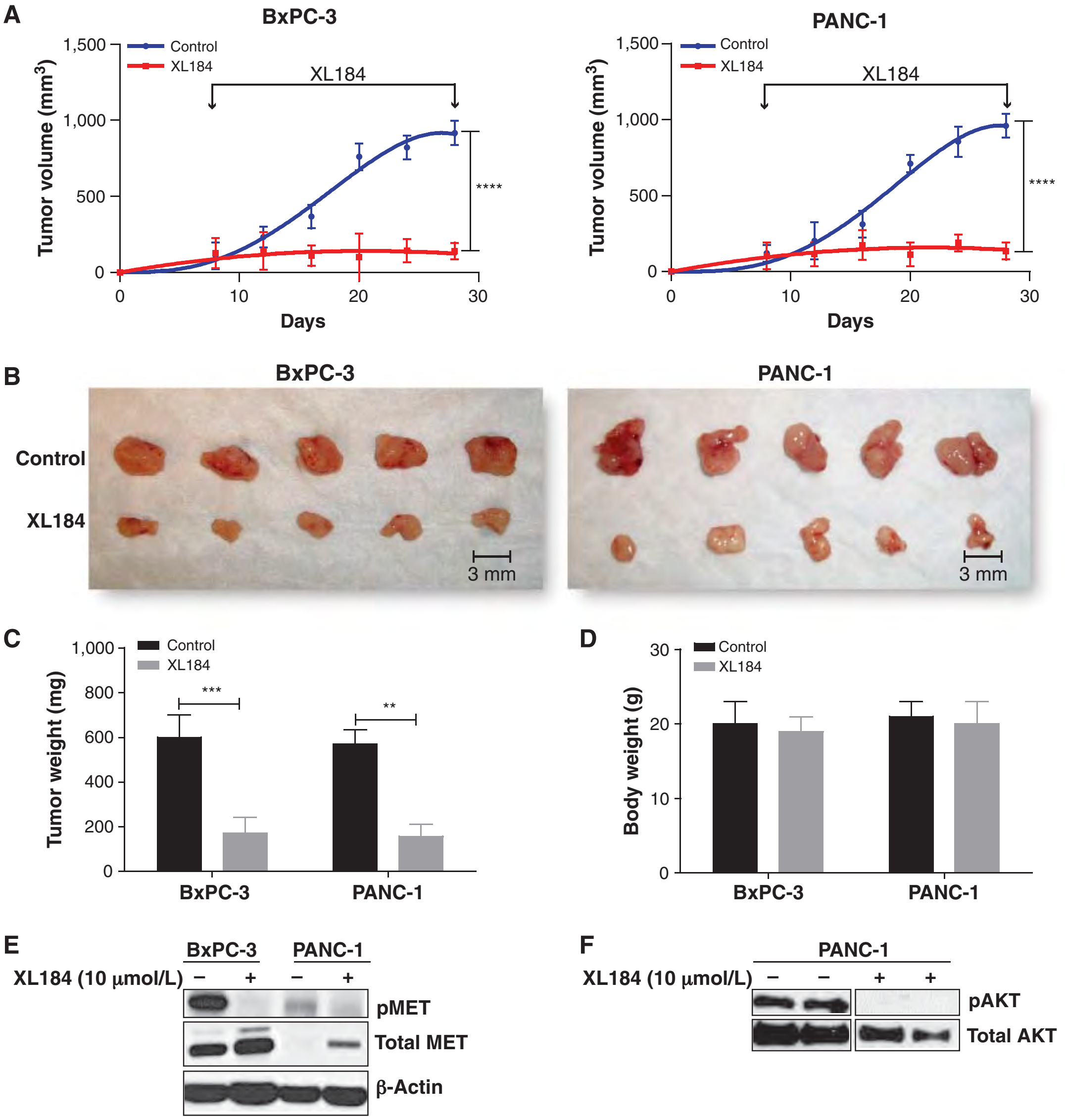


**Figure 7.** XL184 treatment suppresses tumor growth in BxPC-3 and PANC-1 pancreatic cancer xenografts. **A,** The growth curves show the effect of the MET inhibitor XL184 on tumor growth in BxPC-3 and PANC-1 xenografts. **B,** BxPC-3 and PANC-1 xenograft tumors after 3 weeks of XL184 treatment are shown, as compared with the controls. The bar graphs display tumor weight (**C**) and total body weight (**D**) after 3 weeks of XL184 treatment. Values represent mean ± SE. **, $P < 0.01$; ***, $P < 0.001$; ****, $P < 0.0001$. **E,** immunoblot results showing the effect of XL184 treatment on phospho-MET (pMET) in BxPC-3 and PANC-1 cells. **F,** immunoblot results showing the effect of XL184 treatment on phospho-AKT (pAKT) level in the PANC-1 orthotopic xenograft.

to be *KRAS* dependent, whereas BxPC-3 cells (the only pancreatic cancer cell line in our panel with wild-type *KRAS*) as well as PANC-1 and HPAC were *KRAS* independent. Interestingly, comparing our results with the published literature, we noted a general lack of consensus in the "*KRAS* dependence" status of pancreatic cancer cell lines (10, 14, 40–45). For example, whereas 2 prior studies using siRNA-mediated knockdown of *KRAS* in the *KRAS*-mutant cell line MIA-PaCa-2 designated it as *KRAS* dependent, based on reduced cellular proliferation, invasion, and colony formation assays (10, 44), more recently, Collisson and colleagues (40) observed no significant effect on proliferation in MIA-PaCa-2 cells transduced with sh*KRAS* lentivirus. Similarly, PANC-1 was identified as *KRAS* dependent in 4 different studies by both siRNA- and shRNA-mediated knockdowns, as assessed by cellular proliferation, colony formation, invasion, and xenograft tumor growth (10, 14, 43, 44), whereas 3 studies found PANC-1 to be *KRAS* independent by shRNA-mediated knockdown and farnesyl transferase inhibitor treatment using similar *in vitro* assays (40–42). Conversely, the *KRAS* wild-type cell line BxPC-3 has been consistently reported to be *KRAS* independent (14, 44), similar to our findings. Interestingly, HPAC was described as *KRAS* dependent by Collisson and colleagues (40) but was found to be *KRAS* independent in our assays. No published references were found for L3.3 and Panc-03.27, which we report as *KRAS* dependent.

Several *KRAS* synthetic lethal screens and DNA microarray analyses have been used to describe genes and gene signatures associated with *KRAS* dependence (12–14, 40, 41, 46) and include kinase genes such as *PLK1*, *MST1R*, and *SYK* (12, 40, 41). Interestingly, we observed outlier expression of *PLK* to be restricted to *KRAS*-dependent cells, and these cells showed higher sensitivity to the pan-PLK inhibitor BI-6727 both alone and in combination with *KRAS* knockdown, as compared with *KRAS*-independent cells. Previously, Luo and colleagues identified *PLK1* as a *RAS* synthetic lethal interactor in a lung and a colorectal cancer cell line, although they did not test any pancreatic cancer cell lines (12). Our results additionally show that cells respond to the pan-PLK inhibitor BI-6727 only if they have outlier expression of either *PLK1* or *PLK2* (Fig. 6A and B). This finding highlights the importance of using therapeutic targets in a sample-specific manner.

Overall, our study provides a generalizable metric to define and prioritize personalized target spectra specific to individual tumors. The recent report of a remarkably successful treatment of a patient with acute lymphoblastic leukemia with sunitinib targeting "wildly active" expression of *FLT3* kinase identified by RNA-Seq when whole-genome sequencing failed to identify any actionable aberrations (47), provides an anecdotal yet powerful illustration of the potential application of the systematic identification of outlier kinases proposed in our study.

## METHODS

### Kinome Analysis

Transcriptome sequencing data from 482 cancer and benign samples from 25 different tissue types previously generated on Illumina GA and GAII platforms were mapped using Bowtie (48) against University of California Santa Cruz (Santa Cruz, CA) Genome Browser genes in the hg18 human genome assembly (49). Unique best-match hit sequences normalized for the number of RPKM (16) were used to generate a gene expression data matrix for the entire compendium (24). The expression data for the complete list of kinase genes (50) were used to identify "outlier kinases" in individual samples based on their absolute expression within the sample and differential expression (defined as absolute expression divided by median expression level of that gene across the compendium). GraphPad Prism software was used to generate kinome expression profiles for each sample, plotting absolute expression versus differential expression for all kinases.

Statistical significance of outlier expression was quantified using a Mahalanobis distance metric [$D^2 = (x - \mu)'\Sigma^{-1}(x - \mu)$; $\Sigma$ = covariance matrix, $D$ = Mahalanobis distance of the point $x$ to the mean $\mu$; refs. 51, 52), to measure the "distance" of each kinase's absolute and differential expression from the center of the scatter plot distribution. *P* values were calculated assuming a $\chi^2$ distribution, with 2 degrees of freedom. Kinases with absolute expression of more than 20 RPKM, differential expression of more than 5-fold, and $P < 0.05$ were nominated as having "outlier expression." R language (53) was used to conduct statistical analysis.

### Cell Culture

All human breast and pancreatic cancer and benign epithelial cell lines were purchased from the American Type Culture Collection (ATCC), except the benign immortalized pancreatic epithelial cell line HPDE and the xenograft cell lines derived from primary pancreatic adenocarcinoma tissues, which were provided by D.M. Simeone (University of Michigan, Ann Arbor, MI). The pancreatic adenocarcinoma cell line L3.3 was obtained from the University of Texas MD Anderson Characterized Cell Line Core (Houston, TX). All cell lines were grown in recommended culture media and maintained at 37°C in 5% $CO_2$. To ensure cellular identities, a panel of cell lines was genotyped at the University of Michigan Sequencing Core using Profiler Plus (Applied Biosystems) and compared with the short tandem repeat (STR) profiles of respective cell lines available in the STR Profile Database (ATCC).

### Transcript Knockdowns and Cell Proliferation Assays

ON-TARGETplus siRNA against *AKT2*, *AXL*, *EGFR*, *MET*, and *PLK2*, and nontargeting control (siNTC) from Dharmacon (Supplementary Table S5A) were used at 100 nmol/L. Cells were transfected in 6-well plates at a density of 50,000 cells per well using Oligofectamine (Invitrogen), according to the manufacturer's protocol. Transfection was repeated 24 hours later; the cells were grown for an additional 48 hours and replated at a density of 5,000 cells per well in 24-well plates. Cells were counted over a period of 1 to 6 days using a Beckman Coulter cell counter. Transient transductions with shRNA against *ERBB2*, *RPS6KB1*, and *FGFR4*, or nontargeting control (shNTC), were carried out in 6-well plates in the presence of 8 μg/mL hexadimethrine bromide (Polybrene; Sigma). For trastuzumab (Herceptin; Roche) experiments, cells were grown for 3 days in 24-well plates with and without trastuzumab (100 μg/mL), in combination with the FGFR inhibitor PD173074 (TOCRIS Bioscience) at 1 μmol/L or TKI-258 (dovitinib; Selleck Chemicals) at 0.1 μmol/L. Trastuzumab-resistant cell lines were generated from MDA-MB-453 and BT-474 by maintaining the cells in the continuous presence of 100 μg/mL trastuzumab over 1 month. Cell proliferation assays were carried out over a period of 1 to 7 days, using a Beckman Coulter cell counter, and growth curves were plotted using GraphPad Prism software. Statistical comparisons were conducted using one-way ANOVA.

### Generation of Stable Cell Lines with Doxycycline-Inducible KRAS-*shRNA* Lentiviral Constructs

Doxycycline-inducible shRNAmir-TRIPZ lentiviral constructs targeting *KRAS* or nontargeting control (Open Biosystems) tagged with RFP were used to transduce a panel of pancreatic cell lines in the presence of 8 μg/mL Polybrene (Supplementary Table S5A). Forty-eight hours after transduction, cells were selected in medium containing 1 μg/mL puromycin (Invitrogen) for 4 days. The shRNA expression was induced by growing cells in medium containing 1 μg/mL doxycycline (Sigma) for 72 hours. The enrichment of stable cells and efficiency of shRNA induction were assessed by measuring the percentage of cells displaying red fluorescence by flow cytometry (FACSAria Cell Sorter; BD Biosciences). Experiments with stable cell lines were carried out in the presence of 1 μg/mL doxycycline, refreshed daily. Experiments with the PLK inhibitor BI-6727 (volasertib; Selleck Chemicals) were carried out with cells plated in 96-well culture plates at a density of 3,000 to 4,000 cells per well and treated with 10 nmol/L BI-6727 or dimethyl sulfoxide (DMSO). This concentration was selected on the basis of $IC_{50}$ values calculated from prior proliferation assays using 1 to 500 nmol/L BI-6727 (data not shown). At 0, 1, 3, and 5 days following drug treatment, viable cells were quantified using WST-1 reagent (Roche) and absorbance was measured at 440 nm, per the manufacturer's protocol. Growth curves were plotted using GraphPad Prism software. Statistical comparisons were conducted using one-way ANOVA.

### Western Blot Analysis

Cell or tissue lysates were separated on 4% to 12% SDS polyacrylamide gels (Novex) and blotted on polyvinylidene difluoride membranes (Amersham) by semi-dry transfer. Antibodies to FGFR4 (Santa Cruz), phospho-AKT, total AKT, phospho-ERK, total ERK, phospho-MET, and total MET (Cell Signaling Technology) were used at 1:1,000 dilutions for standard immunoblotting and detection by enhanced chemiluminescence (ECL Prime), per the manufacturer's protocol.

For phospho-MET blots, cells treated with 10 μmol/L XL184 for 12 hours were stimulated with 100 ng/mL human recombinant hepatocyte growth factor (Invitrogen) for 1 hour before harvesting in radioimmunoprecipitation assay RIPA buffer.

### Quantitative RT-PCR Assay

RNA was isolated from cell lysates by the RNeasy Micro Kit (Qiagen), and cDNA was synthesized from 1 μg RNA using SuperScript III (Invitrogen) and Random Primers (Invitrogen), per the manufacturer's protocol. qRT-PCR was carried out on the StepOne Real-Time PCR system (Applied Biosystems) using gene-specific primers designed with Primer-BLAST (Supplementary Table S5B and S5C) and synthesized by IDT Technologies. Validation of RNA-Seq results was carried out using TaqMan Universal PCR Master Mix II with uracil-*N*-glycosylase (Applied Biosystems) and Universal ProbeLibrary System probes (Roche), following the manufacturer's protocol. Validation of siRNA- and shRNA-mediated knockdown was carried out using Fast SYBR Green Master Mix (Invitrogen), per the manufacturer's protocol. qRT-PCR data were analyzed using the relative quantification method and plotted as average fold-change compared with the control. *Glyceraldehyde-3-phosphate dehydrogenase* (*GAPDH*) was used as an internal reference. For qRT-PCR validation studies, GraphPad Prism software was used to conduct linear regression and calculate $R^2$ correlation coefficients.

### Dose Response

Experiments with the FGFR inhibitors PD173074 and dovitinib and the MET inhibitor XL184 were carried out with cells seeded at a density of 3,000 to 4,000 cells per well, plated in 96-well culture plates, and treated with concentrations from 100 to 0.1 μmol/L. WST-1 assay (Roche) was conducted after 72 hours, and readings were recorded at 440 nm. GraphPad Prism software was used to generate nonlinear regression curves and calculate $IC_{50}$ values.

### Apoptosis Assay

The apoptosis assay was carried out using ApoScreen Annexin V Apoptosis Kit (Southern Biotech), per the manufacturer's protocol. Briefly, cells treated for 48 hours with DMSO or increasing concentrations of BI-6727 were washed with cold PBS, suspended in cold 1× binding buffer, stained with Annexin V and propidium iodide, and subjected to flow cytometry by FACSAria Cell Sorter (BD Biosciences). Results were analyzed and plotted using Summit 6.0 Software (Beckman Coulter).

### In Vivo *Tumorigenicity Assay*

Six-week-old male NOD/SCID mice (Taconic) were housed under pathogen-free conditions approved by the American Association for Accreditation of Laboratory Animal Care in accordance with current regulations and standards of the U.S. Department of Agriculture and Department of Health and Human Services. Animal experiments were approved by the University of Michigan Animal Care and Use Committee and carried out in accordance with established guidelines. Mice anesthetized with an intraperitoneal injection of xylazine (9 mg/kg) and ketamine (100 mg/kg body weight) were implanted with $1 \times 10^6$ BxPC-3 or PANC-1 cells suspended in 50 μL 1:1 mixture of Media 199 and Matrigel (BD Biosciences) injected subcutaneously into their flanks using a 30-gauge needle. When tumor size reached 0.4 mm, mice were randomized into control and treatment groups ($n = 8$ per group). The MET inhibitor XL184 (Exelixis Chemicals) was orally administered at 30 mg/kg body weight twice per week for 3 weeks. Tumor growth was monitored weekly. Tumor caliper measurements were converted into tumor volumes using the formula ½[length × $(\text{width})^2$] $\text{mm}^3$ and plotted using GraphPad Prism software. At 3 weeks of treatment, mice were weighed and euthanized and the tumors harvested. Statistical comparisons were conducted using one-way ANOVA.

## Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

## Authors' Contributions

**Conception and design:** D.R. Robinson, X. Cao, D.M. Simeone, A.M. Chinnaiyan, C. Kumar-Sinha
**Development of methodology:** I. Wei, S. Shankar, L.W. Ma, D.R. Robinson, Y.-M. Wu, D.M. Simeone, C. Kumar-Sinha
**Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.):** V. Kothari, I. Wei, S. Shankar, S. Kalyana-Sundaram, L. Wang, L.W. Ma, D.R. Robinson, X. Cao, D.M. Simeone, A.M. Chinnaiyan, C. Kumar-Sinha
**Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis):** V. Kothari, I. Wei, S. Shankar, S. Kalyana-Sundaram, L. Wang, L.W. Ma, P. Vats, C.S. Grasso, D.M. Simeone, C. Kumar-Sinha
**Writing, review, and/or revision of the manuscript:** V. Kothari, I. Wei, S. Kalyana-Sundaram, D.M. Simeone, A.M. Chinnaiyan, C. Kumar-Sinha
**Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases):** I. Wei, S. Shankar, S. Kalyana-Sundaram, L.W. Ma, Y.-M. Wu, D.M. Simeone, C. Kumar-Sinha
**Study supervision:** D.M. Simeone, A.M. Chinnaiyan, C. Kumar-Sinha

## Acknowledgments

The authors thank Terrence Barrette, Michael Quist, Robert Lonigro, and Sheeba Powar for bioinformatics help; Mark Hynes for help with animal work; and Irfan A. Asangani and Filip Bednar for useful discussions. Trastuzumab (Herceptin; Roche) was kindly provided by Dr. Max Wicha (University of Michigan Cancer Center).

## Grant Support

This work was supported in part by NIH 5-R21-CA-155992-02 (to C. Kumar-Sinha), NIH 2T32CA009672-21 (to I. Wei), NIH R01CA131045-01 and NIH P50CA130810-1A (to D.M. Simeone), and the Department of Defense Era of Hope grant W81XWH-08-0110 (to A.M. Chinnaiyan). D.M. Simeone is also supported by the Rich Rogel Fund for Pancreatic Cancer Research; A.M. Chinnaiyan is supported by the Doris Duke Charitable Foundation Clinical Scientist Award and is an American Cancer Society Research Professor and A. Alfred Taubman Scholar; and C. Kumar-Sinha is supported by University of Michigan Gastrointestinal (GI) Specialized Programs of Research Excellence (SPORE) Career Development Award and is a recipient of Lustgarten Foundation Award.

Received July 16, 2012; revised December 5, 2012; accepted December 6, 2012; published OnlineFirst February 5, 2013.

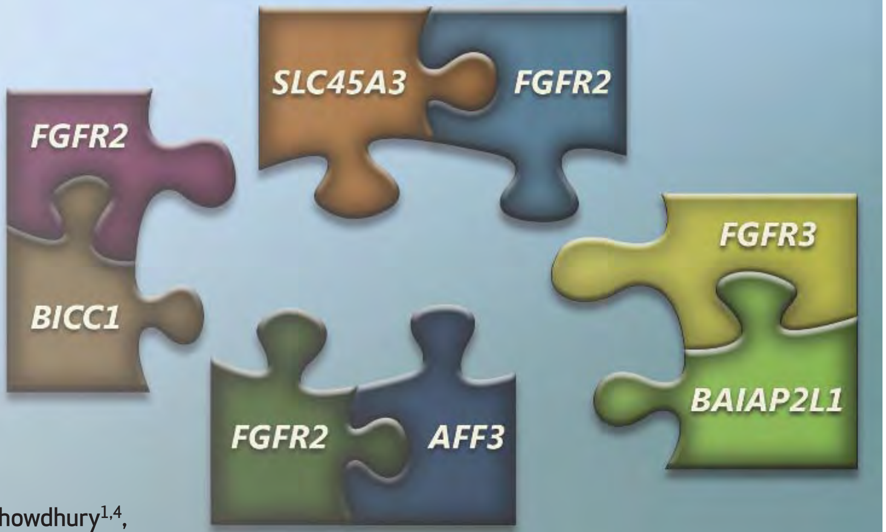
## REFERENCES

1. Weinstein IB, Joe A. Oncogene addiction. Cancer Res 2008;68:3077–80.
2. Baselga J, Arribas J. Treating cancer's kinase 'addiction.' Nat Med 2004;10:786–7.
3. Knight ZA, Lin H, Shokat KM. Targeting the cancer kinome through polypharmacology. Nat Rev Cancer 2010;10:130–7.
4. Zhang J, Yang PL, Gray NS. Targeting cancer with small molecule kinase inhibitors. Nat Rev Cancer 2009;9:28–39.
5. Camidge DR, Doebele RC. Treating ALK-positive lung cancer—early successes and future challenges. Nat Rev Clin Oncol 2012;9:268–77.
6. Manning BD. Challenges and opportunities in defining the essential cancer kinome. Sci Signal 2009;2:pe15.
7. Maitra A, Hruban RH. Pancreatic cancer. Annu Rev Pathol 2008;3:157–88.

8. Zanini N, Masetti M, Jovine E. The definition of locally advanced pancreatic cancer. Br J Cancer 2010;102:1306–7.
9. Cardenes HR, Chiorean EG, Dewitt J, Schmidt M, Loehrer P. Locally advanced pancreatic cancer: current therapeutic approach. Oncologist 2006;11:612–23.
10. Fleming JB, Shen GL, Holloway SE, Davis M, Brekken RA. Molecular consequences of silencing mutant K-ras in pancreatic cancer cells: justification for K-ras-directed therapy. Mol Cancer Res 2005;3:413–23.
11. Strimpakos A, Saif MW, Syrigos KN. Pancreatic cancer: from molecular pathogenesis to targeted therapy. Cancer Metastasis Rev 2008;27:495–522.
12. Luo J, Emanuele MJ, Li D, Creighton CJ, Schlabach MR, Westbrook TF, et al. A genome-wide RNAi screen identifies multiple synthetic lethal interactions with the Ras oncogene. Cell 2009;137:835–48.
13. Barbie DA, Tamayo P, Boehm JS, Kim SY, Moody SE, Dunn IF, et al. Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. Nature 2009;462:108–12.
14. Scholl C, Frohling S, Dunn IF, Schinzel AC, Barbie DA, Kim SY, et al. Synthetic lethal interaction between oncogenic KRAS dependency and STK33 suppression in human cancer cells. Cell 2009;137:821–34.
15. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev 2009;10:57–63.
16. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 2008;5:621–8.
17. Sultan M, Schulz MH, Richard H, Magen A, Klingenhoff A, Scherf M, et al. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. Science 2008;321:956–60.
18. Li H, Lovci MT, Kwon YS, Rosenfeld MG, Fu XD, Yeo GW. Determination of tag density required for digital transcriptome analysis: application to an androgen-sensitive prostate cancer model. Proc Natl Acad Sci U S A 2008;105:20179–84.
19. Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet 2008;40:1413–5.
20. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 2009;25:1105–11.
21. Maher CA, Palanisamy N, Brenner JC, Cao X, Kalyana-Sundaram S, Luo S, et al. Chimeric transcript discovery by paired-end transcriptome sequencing. Proc Natl Acad Sci U S A 2009;106:12353–8.
22. Maher CA, Kumar-Sinha C, Cao X, Kalyana-Sundaram S, Han B, Jing X, et al. Transcriptome sequencing to detect gene fusions in cancer. Nature 2009;458:97–101.
23. Prensner JR, Iyer MK, Balbin OA, Dhanasekaran SM, Cao Q, Brenner JC, et al. Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression. Nat Biotechnol 2011;29:742–9.
24. Kalyana-Sundaram S, Kumar-Sinha C, Shankar S, Robinson DR, Wu Y-M, Cao X, et al. Expressed pseudogenes in the transcriptional landscape of human cancers. Cell 2012;149:13.
25. MacDonald JW, Ghosh D. COPA–cancer outlier profile analysis. Bioinformatics 2006;22:2950–1.
26. Tomlins SA, Rhodes DR, Perner S, Dhanasekaran SM, Mehra R, Sun XW, et al. Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. Science 2005;310:644–8.
27. Hutchinson L. Personalized cancer medicine: era of promise and progress. Nat Rev Clin Oncol 2011;8:121.
28. Hood L, Friend SH. Predictive, personalized, preventive, participatory (P4) cancer medicine. Nat Rev Clin Oncol 2011;8:184–7.
29. Slamon DJ, Clark GM, Wong SG, Levin WJ, Ullrich A, McGuire WL. Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene. Science 1987;235:177–82.
30. Yu D, Hung MC. Overexpression of ErbB2 in cancer and ErbB2-targeting strategies. Oncogene 2000;19:6115–21.
31. Rhodes DR, Kalyana-Sundaram S, Mahavisno V, Varambally R, Yu J, Briggs BB, et al. Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles. Neoplasia 2007;9:166–80.
32. Rhodes DR, Yu J, Shanker K, Deshpande N, Varambally R, Ghosh D, et al. ONCOMINE: a cancer microarray database and integrated data-mining platform. Neoplasia 2004;6:1–6.
33. Slamon DJ, Leyland-Jones B, Shak S, Fuchs H, Paton V, Bajamonde A, et al. Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2. N Engl J Med 2001;344:783–92.
34. Vogel CL, Cobleigh MA, Tripathy D, Gutheil JC, Harris LN, Fehrenbacher L, et al. Efficacy and safety of trastuzumab as a single agent in first-line treatment of HER2-overexpressing metastatic breast cancer. J Clin Oncol 2002;20:719–26.
35. Buck E, Eyzaguirre A, Haley JD, Gibson NW, Cagnoni P, Iwata KK. Inactivation of Akt by the epidermal growth factor receptor inhibitor erlotinib is mediated by HER-3 in pancreatic and colorectal tumor cell lines and contributes to erlotinib sensitivity. Mol Cancer Ther 2006;5:2051–9.
36. Ali S, El-Rayes BF, Sarkar FH, Philip PA. Simultaneous targeting of the epidermal growth factor receptor and cyclooxygenase-2 pathways for pancreatic cancer therapy. Mol Cancer Ther 2005;4:1943–51.
37. Morgan MA, Parsels LA, Kollar LE, Normolle DP, Maybaum J, Lawrence TS. The combination of epidermal growth factor receptor inhibitors with gemcitabine and radiation in pancreatic cancer. Clin Cancer Res 2008;14:5142–9.
38. Cheng JQ, Ruggeri B, Klein WM, Sonoda G, Altomare DA, Watson DK, et al. Amplification of AKT2 in human pancreatic cells and inhibition of AKT2 expression and tumorigenicity by antisense RNA. Proc Natl Acad Sci U S A 1996;93:3636–41.
39. Miwa W, Yasuda J, Murakami Y, Yashima K, Sugano K, Sekine T, et al. Isolation of DNA sequences amplified at chromosome 19q13.1-q13.2 including the AKT2 locus in human pancreatic cancer. Biochem Biophys Res Commun 1996;225:968–74.
40. Collisson EA, Sadanandam A, Olson P, Gibb WJ, Truitt M, Gu S, et al. Subtypes of pancreatic ductal adenocarcinoma and their differing responses to therapy. Nat Med 2011;17:500–3.
41. Singh A, Greninger P, Rhodes D, Koopman L, Violette S, Bardeesy N, et al. A gene expression signature associated with "K-Ras addiction" reveals regulators of EMT and tumor cell survival. Cancer Cell 2009;15:489–500.
42. Sepp-Lorenzino L, Ma Z, Rands E, Kohl NE, Gibbs JB, Oliff A, et al. A peptidomimetic inhibitor of farnesyl:protein transferase blocks the anchorage-dependent and -independent growth of human tumor cell lines. Cancer Res 1995;55:5302–9.
43. Ji Z, Mei FC, Xie J, Cheng X. Oncogenic KRAS activates hedgehog signaling pathway in pancreatic cancer cells. J Biol Chem 2007;282:14048–55.
44. Nakada Y, Saito S, Ohzawa K, Morioka CY, Kita K, Minemura M, et al. Antisense oligonucleotides specific to mutated K-ras genes inhibit invasiveness of human pancreatic cancer cell lines. Pancreatology 2001;1:314–9.
45. Shen YM, Yang XC, Yang C, Shen JK. Enhanced therapeutic effects for human pancreatic cancer by application K-ras and IGF-IR antisense oligodeoxynucleotides. World J Gastroenterol 2008;14:5176–85.
46. Bild AH, Yao G, Chang JT, Wang Q, Potti A, Chasse D, et al. Oncogenic pathway signatures in human cancers as a guide to targeted therapies. Nature 2006;439:353–7.
47. Kolata G. Genetic gamble, new approaches to fighting cancer: in treatment for leukemia, glimpses of the future. The New York Times. 2012 July 7.
48. Langmead B. Aligning short sequencing reads with Bowtie. Curr Protoc Bioinformatics 2010; Chapter 11:Unit 11 7.
49. UCSC Genome Browser. Available from: http://genome.ucsc.edu.
50. Manning G, Whyte DB, Martinez R, Hunter T, Sudarsanam S. The protein kinase complement of the human genome. Science 2002;298:1912–34.
51. Mahalanobis PC. On the generalised distance in statistics. Proc Natl Inst Sci India 1936;2:49–55.
52. De Maesschalck R, Jouan-Rimbaud D, Massart DL. The Mahalanobis distance. Chemom Intell Lab Syst 2000;50:1–18.
53. The R Project for Statistical Computing. Available from: http://www.r-project.org.

RESEARCH BRIEF

# Identification of Targetable FGFR Gene Fusions in Diverse Cancers



Yi-Mi Wu[1,2], Fengyun Su[1,2],
Shanker Kalyana-Sundaram[1,2],
Nickolay Khazanov[10], Bushra Ateeq[1,2],
Xuhong Cao[1,7], Robert J. Lonigro[1,8],
Pankaj Vats[1,2], Rui Wang[1,2], Su-Fang Lin[11],
Ann-Joy Cheng[12], Lakshmi P. Kunju[1,2],
Javed Siddiqui[1,2], Scott A. Tomlins [1,2],
Peter Wyngaard [10], Seth Sadis[10], Sameek Roychowdhury[1,4],
Maha H. Hussain[3], Felix Y. Feng[1,4,8], Mark M. Zalupski[3],
Moshe Talpaz[3], Kenneth J. Pienta[1,3,6,8], Daniel R. Rhodes[1,2,5,10],
Dan R. Robinson[1,2], and Arul M. Chinnaiyan[1,2,6,7,8,9]

**ABSTRACT** Through a prospective clinical sequencing program for advanced cancers, four index cases were identified which harbor gene rearrangements of *FGFR2*, including patients with cholangiocarcinoma, breast cancer, and prostate cancer. After extending our assessment of FGFR rearrangements across multiple tumor cohorts, we identified additional FGFR fusions with intact kinase domains in lung squamous cell cancer, bladder cancer, thyroid cancer, oral cancer, glioblastoma, and head and neck squamous cell cancer. All FGFR fusion partners tested exhibit oligomerization capability, suggesting a shared mode of kinase activation. Overexpression of FGFR fusion proteins induced cell proliferation. Two bladder cancer cell lines that harbor FGFR3 fusion proteins exhibited enhanced susceptibility to pharmacologic inhibition *in vitro* and *in vivo*. Because of the combinatorial possibilities of FGFR family fusion to a variety of oligomerization partners, clinical sequencing efforts, which incorporate transcriptome analysis for gene fusions, are poised to identify rare, targetable FGFR fusions across diverse cancer types.

**SIGNIFICANCE:** High-throughput sequencing technologies facilitate defining the mutational landscape of human cancers, which will lead to more precise treatment of patients with cancer. Here, through integrative sequencing efforts, we identified a variety of FGFR gene fusions in a spectrum of human cancers. FGFR fusions are active kinases. Cells harboring FGFR fusions showed enhanced sensitivity to the FGFR inhibitors PD173074 and pazopanib, suggesting that patients with cancer with FGFR fusions may benefit from targeted FGFR kinase inhibition. *Cancer Discov; 3(6); 636–47. ©2013 AACR.*

*See related commentary by Sabnis and Bivona, p. 607.*

**Authors' Affiliations:** [1]Michigan Center for Translational Pathology; Departments of [2]Pathology, [3]Internal Medicine, [4]Radiation Oncology, [5]Computational Medicine and Bioinformatics, and [6]Urology; [7]Howard Hughes Medical Institute; [8]Comprehensive Cancer Center; and [9]Center for Computational Medicine and Biology, University of Michigan Medical School; [10]Compendia Bioscience, Ann Arbor, Michigan; [11]National Institute of Cancer Research, National Health Research Institutes, Miaoli County; and [12]Department of Medical Biotechnology and Laboratory Science, Chang Gung University, Taiwan

**Note:** Supplementary data for this article are available at Cancer Discovery Online (http://cancerdiscovery.aacrjournals.org/).

**Corresponding Authors:** Arul M. Chinnaiyan, Michigan Center for Translational Pathology, University of Michigan Medical School, 1400 E. Medical Center Drive 5316 CCGC, Ann Arbor, MI 48109-5940. Phone: 734-615-4062; Fax: 734-615-4498; E-mail: arul@umich.edu; and Dan R. Robinson, Michigan Center for Translational Pathology, University of Michigan Medical School, 1400 E. Medical Center Drive 5410 CCGC, Ann Arbor, MI 48109-5940. E-mail: danrobi@umich.edu

**doi:** 10.1158/2159-8290.CD-13-0050

## INTRODUCTION

Advances in next-generation sequencing technologies have refined the molecular taxonomy of a spectrum of human diseases and facilitated a move toward "precision medicine" (1, 2). With regard to oncology, defining the mutational landscape of an individual patient's tumor will lead to more precise treatment and management of patients with cancer. Comprehensive clinical sequencing programs for patients with cancer have been initiated at a variety of medical centers, including our own (3, 4). In addition to the potential for identifying "actionable" therapeutic targets in patients with cancer, these clinical sequencing efforts may lead to the identification of novel "driver" mutations that may be rare in a common cancer type or newly revealed in relatively rare cancer types.

Recurrent gene fusions are an important class of "driver" mutation in cancer, as exemplified by the *BCR–ABL* gene fusion that characterizes chronic myeloid leukemia (CML; ref. 5). Importantly, virtually all patients with CML harbor the BCR–ABL kinase fusion and respond to the small-molecule kinase inhibitor, imatinib, representing one of the earliest examples of precision medicine in practice (6). In 2005, it was discovered that more than 50% of prostate cancers harbor recurrent gene fusions of the androgen-regulated gene *TMPRSS2* with ETS transcription factors (7), suggesting that gene fusions/translocations may play a significant role in common epithelial tumors, similar to hematologic malignancies and sarcomas. Subsequently, recurrent gene rearrangements have been identified in carcinomas of the lung, breast, colon, and thyroid, among other epithelial tissues (8–12). Of these, the *EML4–ALK* gene fusion, which characterizes 1% to 5% of lung adenocarcinomas, has gained the most traction in the context of precision therapy, as patients with this gene fusion respond to the kinase inhibitor crizotinib (13, 14). Recently, *FGFR1* and *FGFR3* fusions with *TACC1* and *TACC3*, respectively, have been identified in approximately 3% of the tumor glioblastoma multiforme (GBM; ref. 15), and *FGFR3–TACC3* fusions were identified in a subset of bladder carcinomas (16). Preclinical studies suggest that patients with GBM with *FGFR–TACC* gene fusions may benefit from targeted FGFR kinase inhibition (17, 18).

## RESULTS

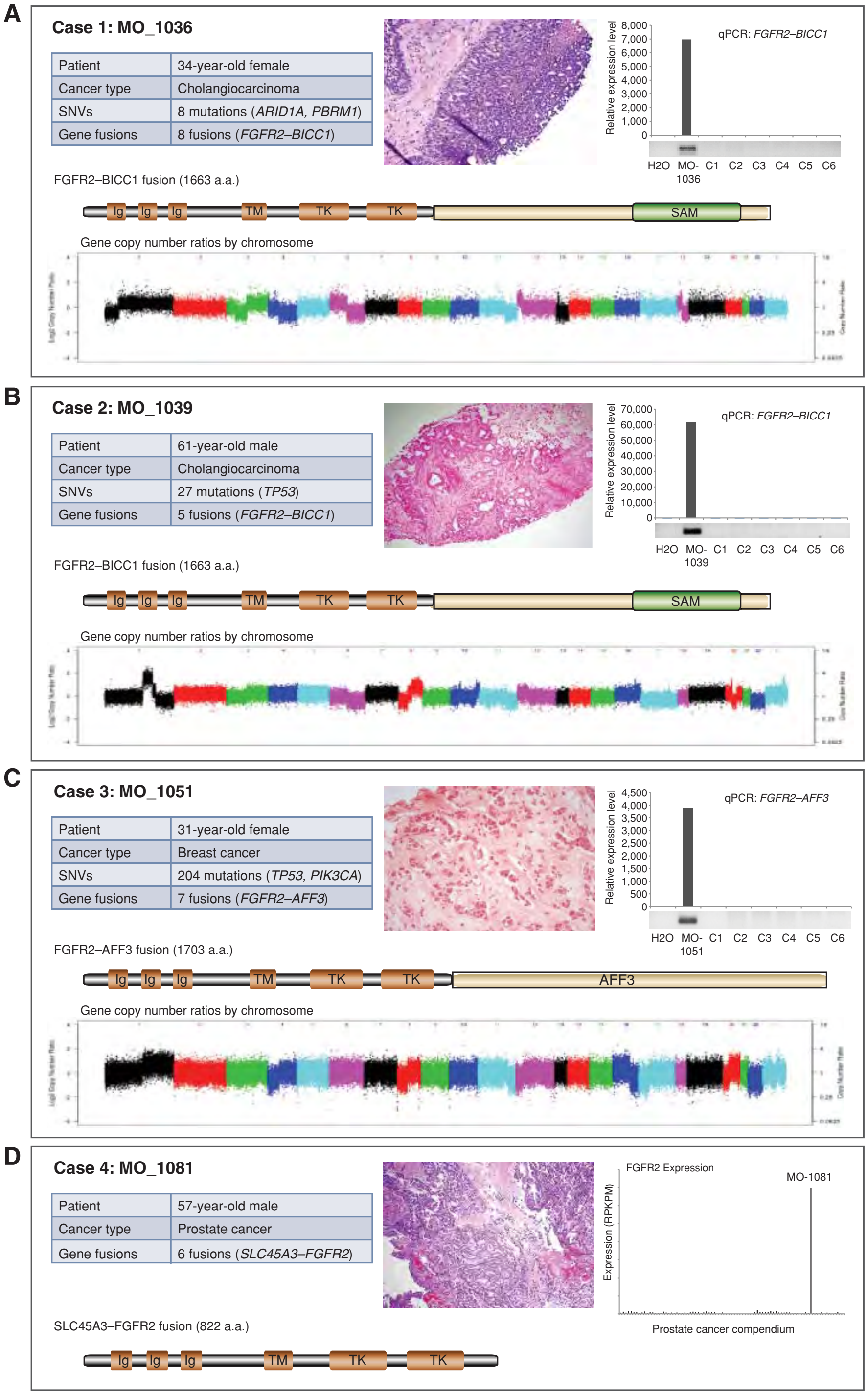
Our Institutional Review Board (IRB)-approved clinical sequencing program, called the Michigan Oncology Sequencing Program (MI-ONCOSEQ), enrolls patients with advanced cancer across all histologies (3). Since April 2011, we have enrolled more than 100 patients on this program, which involves obtaining a current tumor biopsy with matched normal samples (blood and/or buccal swab). The samples are then subjected to integrative sequencing that includes whole-exome sequencing of the tumor and matched normal tissue, transcriptome sequencing, and, as needed, low pass genome sequencing (3). This combination of DNA and RNA sequencing technologies allows one to be relatively comprehensive with regard to the mutational landscape of coding genes, including point mutations, indels, amplifications, deletions, gene fusions/translocations, and outlier gene expression. These results are generated within a 5 to 7 week time frame and are presented at an institutional "precision tumor board" (previously called sequencing tumor board) to deliberate upon potentially actionable findings.

In this study, 4 MI-ONCOSEQ patients who harbored gene fusions of *FGFR2* by transcriptome sequencing were prospectively identified (Fig. 1). The first patient (MO_1036) was a 34-year-old female diagnosed with metastatic cholangiocarcinoma. By whole-exome sequencing of the tumor relative to the matched normal, we detected 8 nonsynonymous somatic point mutations (Supplementary Table S1). The most interesting of these in terms of tumor biology was the inactivation of the SWI/SNF chromatin remodeling complex through mutation of *ARID1A* (Q1573*) and *PBRM1* (C736*). The SWI/SNF complex has been implicated as a tumor suppressor, and inactivating somatic mutations of *ARID1A* and *PBRM1* have been identified in renal cell carcinoma, breast cancer, and ovarian cancer (19). The copy number landscape for MO_1036 as determined by whole-exome sequencing is shown in Fig. 1A and Supplementary Table S2. Interestingly, by paired-end RNA sequencing, we detected an intrachromosomal fusion that resulted in the in-frame fusion of the *FGFR2* to *BICC1* (Fig. 1A). Although 7 additional chimeric RNAs were detected (Supplementary Table S3), only the *FGFR2–BICC1* fusion exhibited a combination of high supporting reads ($n$ = 259), predicted in-frame fusion protein, and predicted potential therapeutic actionability via kinase inhibition. The *FGFR2–BICC1* fusion was confirmed by quantitative PCR (qPCR) analysis (Fig. 1A). Neither copy number aberrations nor point mutations were observed in *FGFR2* or *BICC1*.

The second MI-ONCOSEQ patient with an *FGFR2* fusion (MO_1039) was a 61-year-old male with metastatic cholangiocarcinoma. Like the first patient, this individual's tumor expressed an *FGFR2–BICC1* fusion of identical configuration (Fig. 1B and Supplementary Table S4). This fusion was similarly validated by qPCR (Fig. 1B). In contrast, however, this cholangiocarcinoma case exhibited 27 nonsynonymous somatic point mutations, including an inactivating mutation of *TP53* (R267W; Supplementary Table S5) and a distinct copy number landscape (Fig. 1B and Supplementary Table S6). Neither point mutations nor copy number changes in FGFR genes were identified in this patient.

The third patient with an *FGFR2* fusion was a 31-year-old woman with metastatic breast cancer (MO_1051). RNA sequencing revealed an in-frame interchromosomal fusion of *FGFR2* with *AFF3*, which had a functional structure analogous to the FGFR2 kinase fusions found in cholangiocarcinoma (Fig. 1C). In addition to the *FGFR2–AFF3* fusion, which was detected with 138 supporting reads and validated by qPCR (Fig. 1C), 6 additional gene fusions with a lower number of reads were identified (Supplementary Table S7). This breast cancer case also harbored 204 nonsynonymous point mutations, including mutation of *TP53* (G199E) and a known activating mutation of *PIK3CA* (H1047R; Supplementary Table S8). While this breast cancer case exhibited a number of amplifications and deletions (Supplementary Table S9), as expected (based on past clinical pathology data), this patient was negative for the *ERBB2* amplification.

The fourth patient (MO_1081) with an *FGFR2* fusion was a 57-year-old male with Gleason score 5+4 metastatic prostate cancer. Transcriptome sequencing of a brain metastasis revealed an interchromosomal fusion of *SLC45A3* with *FGFR2* in which the *SLC45A3* noncoding exon 1 was fused to the intact coding region of *FGFR2* (Fig. 1D and Supplementary Table S10).



Case 1: MO_1036

| | |
|---|---|
| Patient | 34-year-old female |
| Cancer type | Cholangiocarcinoma |
| SNVs | 8 mutations (*ARID1A, PBRM1*) |
| Gene fusions | 8 fusions (*FGFR2–BICC1*) |

Case 2: MO_1039

| | |
|---|---|
| Patient | 61-year-old male |
| Cancer type | Cholangiocarcinoma |
| SNVs | 27 mutations (*TP53*) |
| Gene fusions | 5 fusions (*FGFR2–BICC1*) |

Case 3: MO_1051

| | |
|---|---|
| Patient | 31-year-old female |
| Cancer type | Breast cancer |
| SNVs | 204 mutations (*TP53, PIK3CA*) |
| Gene fusions | 7 fusions (*FGFR2–AFF3*) |

Case 4: MO_1081

| | |
|---|---|
| Patient | 57-year-old male |
| Cancer type | Prostate cancer |
| Gene fusions | 6 fusions (*SLC45A3–FGFR2*) |

**Figure 1.** Integrative sequencing and mutational analysis of 4 index cancer patients found to harbor FGFR fusions. A computed tomography-guided biopsy was used to obtain tumor specimens from patients with cancer enrolled in the MI-ONCOSEQ protocol. A sample of their normal tissue (blood or buccal swab) was also obtained for germline studies. The samples were subjected to integrative sequencing and analyzed for mutations. For each patient, a diagram summarizing the cancer type, histopathology, number of nonsynonymous somatic point mutations and gene fusions detected, and gene copy number landscape is presented. The predicted structure of the FGFR fusion protein identified in each case is illustrated. FGFR gene fusions were validated by quantitative real-time PCR followed by gel electrophoresis or by outlier expression assessed by RNA-seq. The four index cases shown are MO_1036, cholangiocarcinoma (**A**), MO_1039, cholangiocarcinoma (**B**), MO_1051, breast cancer (**C**), and MO_1081, prostate cancer (**D**). qPCR results for each case are compared with a set of 6 cDNA controls from unrelated patient tumors (C1–C6). For the patient with prostate cancer, expression of *FGFR2* is shown (in reads per kilobase per million reads) relative to a compendium of 84 prostate cancer samples. SNVs, single-nucleotide variants.

As *SLC45A3* is a prostate-specific, androgen-regulated gene (20), the *SLC45A3–FGFR2* fusion is predicted to drive overexpression of wild-type FGFR2. Importantly, *FGFR2* exhibited outlier expression in the index case relative to our compendium of prostate cancer tissues (*n* = 84; Fig. 1D), and a similar rare case of *FGFR2* outlier expression was identified in the Glinsky and colleagues (21) prostate cancer cohort (Supplementary Fig. S1A and B).

As we had identified novel *FGFR2* gene fusions in cholangiocarcinoma, breast cancer, and prostate cancer, we next asked whether FGFR family fusions are present across carcinomas of different histologies. To address this, we analyzed RNA-seq data generated from an internal cohort of diverse tumors (*n* = 322) and The Cancer Genome Atlas (TCGA) effort (*n* = 2,053; Supplementary Table S11) for gene fusions using several bioinformatics approaches (see Methods). Including the initial 4 index cases, we identified 24 tumors or cell lines with *FGFR1*, *2*, and *3* fusions (Fig. 2 and Supplementary Tables S12, S13, and S14). All of the gene fusions nominated expressed an FGFR family member as a 5′ or 3′ fusion partner with intact kinase domains suggesting potential actionability. 5′ FGFR fusions to *BICC1*, *AFF3*, *CASP7*, *CCDC6*, *KIAA1967*, *OFD1*, *BAIAP2L1*, and *TACC3* (multiple exons) were identified and 3′ FGFR fusions to *SLC45A3*, *BAG4*, and *ERLIN2* were identified. Cancer types harboring FGFR fusions were quite diverse and included cholangiocarcinoma (*n* = 2), breast cancer (*n* = 4), prostate cancer (*n* = 1), thyroid cancer (*n* = 1), lung squamous cell carcinoma (*n* = 6), bladder cancer (*n* = 5), oral cancer (*n* = 1), head and neck squamous cell carcinoma (*n* = 2), and glioblastoma (*n* = 2). FGFRs are known to exhibit tissue-specific splicing, resulting in IIIb and IIIc isoforms (22). Both IIIb and IIIc isoforms of *FGFR2* and *FGFR3* were evident in the RNA-seq data of the fusion cases, depending on cancer type (Supplementary Table S12).

As most of the diverse FGFR fusion partners contribute domains with known dimerization motifs, including coiled-coil, SAM, LisH, BAR, SPFH, and caspase (23–29), we hypothesized that oligomerization may serve as the common mechanism of activation of FGFR fusion proteins. Thus, we expressed selected epitope-tagged versions of the FGFR fusions in HEK 293T cells and looked for protein oligomerization by coimmunoprecipitation. For example, whereas FGFR3–BAIAP2L1, FGFR3–TACC3, FGFR2–BICC1, and FGFR2–CCDC6 interacted *in vitro*, wild-type FGFR2 and FGFR3 did not in the absence of FGF ligands (Fig. 3A and Supplementary Fig. S2). We also show that the isolated fusion domains provided by BAIAP2L1, TACC3, KIAA1967, CCDC6, and BICC1 interact *in vitro* as oligomerization domains (Supplementary Fig. S3), further supporting the notion of oligomerization-induced activation of FGFR kinase fusions. We additionally showed dimerization capability of the coiled-coil domain present in the FGFR2–CIT fusion identified recently in a lung adenocarcinoma by Seo and colleagues (ref. 30; Supplementary Fig. S3).

Unlike wild-type FGFR2 and FGFR3, overexpression of selected examples of FGFR fusions, including FGFR2–BICC1, FGFR3–BAIAP2L1, and FGFR3–TACC3, in 293T cells induced morphologic changes characterized by rounding up of cells (Supplementary Fig. S4). Overexpression of these FGFR fusion proteins also enhanced cell proliferation based on real-time cell imaging (Fig. 3B). To further show that FGFR fusion kinases are biologically active, we stably expressed FGFR fusions in benign immortalized TERT-HME cell lines. Stable lines harboring the FGFR3–BAIAP2L1, FGFR3–TACC3, and FGFR2–CCDC6 fusions showed expression of active FGFR fusion kinases (as shown by tyrosine phosphorylation of the fusion kinases) and enhanced proliferation of the cells (Fig. 3C–E). Activation of downstream mitogen-activated protein kinase ERK1/2 and the transcription factor STAT1 was also observed in the stable lines (Supplementary Fig. S5). In addition, the ERLIN2–FGFR1 fusion also produced an active FGFR kinase, as shown by tyrosine phosphorylation of the expressed fusion construct (Supplementary Fig. S6).

To evaluate the effects of pharmacologic inhibition of cells naturally harboring FGFR fusions, we assessed the sensitivity of bladder cancer cell lines to an FGFR small-molecule kinase inhibitor PD173074 (31). SW780 cells were characterized to have a fusion of *FGFR3–BAIAP2L1* in this study and a study by Williams and colleagues (ref. 16; Supplementary Fig. S7A), whereas J82 and HT-1197 cells harbor activating point mutations of *FGFR3* [K652E and S249C respectively (32), Catalog of Somatic Mutations in Cancer (COSMIC)]. Importantly, while the FGFR fusion-positive cell line SW780 was sensitive to nanomolar concentrations of PD173074, the *FGFR3*-mutant cell lines used here were not (Fig. 4A), suggesting that FGFR fusions may exhibit sensitivity to FGFR inhibitors, whereas some FGFR mutations are known to be resistant (33). Inhibition of proliferation was also shown with a second FGFR inhibitor, pazopanib, again showing sensitivity of the FGFR fusion-positive lines SW780 and RT4 (Supplementary Fig. S7B). PD173074 exerted a cell-cycle arrest effect on fusion-positive SW780 cells, but not on fusion-negative HT-1197 cells (Supplementary Fig. S8). Similar results for FGFR fusion-positive lines were obtained *in vivo*. SW780 xenografts exhibited decreased tumor growth with increasing doses of PD173074, whereas J82 xenografts did not (Fig. 4B). Expression of the *FGFR3–BAIAP2L1* fusion *in vitro* induced ERK1/2 activation (Supplementary Fig. S5), and, similarly, fusion-positive SW780 xenografts exhibited strong ERK1/2 activation, which could be abolished by treatment with the FGFR inhibitor PD173074 (Fig. 4C). The RT4 urothelial carcinoma line harboring *FGFR3-TACC3* fusion also exhibited sensitivity to FGFR inhibition in a xenograft model (Fig. 4B). Toxicity of PD173074 was monitored by assessment of mouse body weight (Supplementary Fig. S7C).

Further experiments using siRNA knockdown show the central role of *FGFR3–BAIAP2L1* fusion in SW780 cell proliferation. Knockdowns using either *FGFR3* of *BAIAP2L1* siRNAs resulted in a dramatic reduction in cell proliferation in fusion-positive SW780 cells. In contrast, knockdown of *FGFR3* or *BAIAP2L1* did not have significant effects on cell proliferation in either fusion-negative cell line J82 or HT-1197 (Supplementary Fig. S9).

## DISCUSSION

Sequencing and analysis of each of the 4 FGFR fusion-positive patients described in this study were carried out in a time frame of 5 to 7 weeks. The sequencing results were each presented at our bimonthly multidisciplinary precision tumor board for discussion and deliberation. The first patient with cholangiocarcinoma, MO_1036, who harbored the *FGFR2–BICC1*

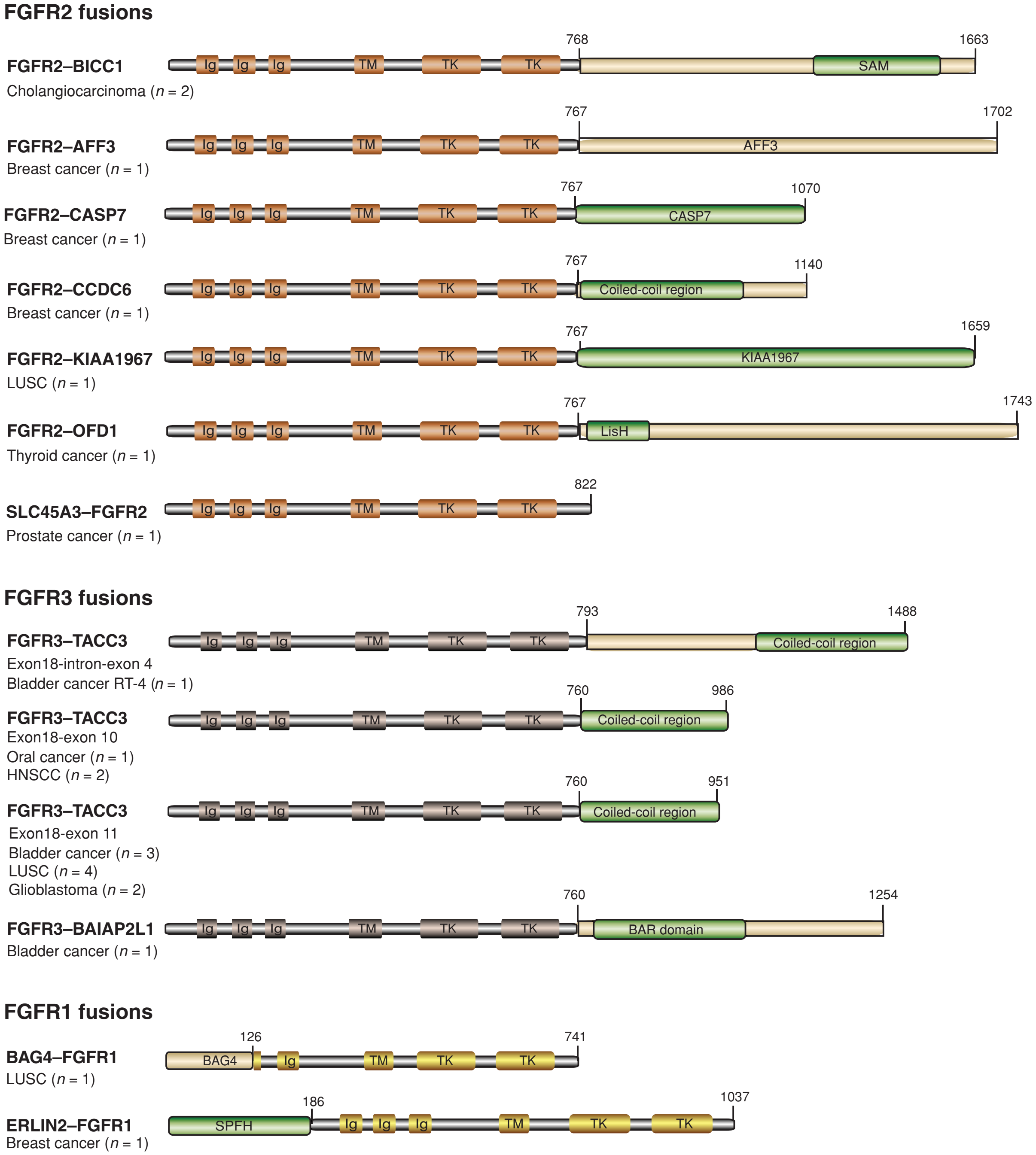


**Figure 2.** Schematic representations of the predicted FGFR fusions identified by transcriptome sequencing of human cancers. Data used include RNA sequencing results from the 4 index patients, our internal tumor cohort, and the TCGA compendium. Out of 4 FGFR receptor family members, *FGFR1*, *FGFR2*, and *FGFR3* are involved in gene fusions with various partners located on several chromosomes. Eleven distinct fusion partners of FGFRs were identified. Exon and codon numberings are based on the reference accessions in Supplementary Table S13. LUSC, lung squamous cell carcinoma; HNSCC, head and neck squamous cell carcinoma.

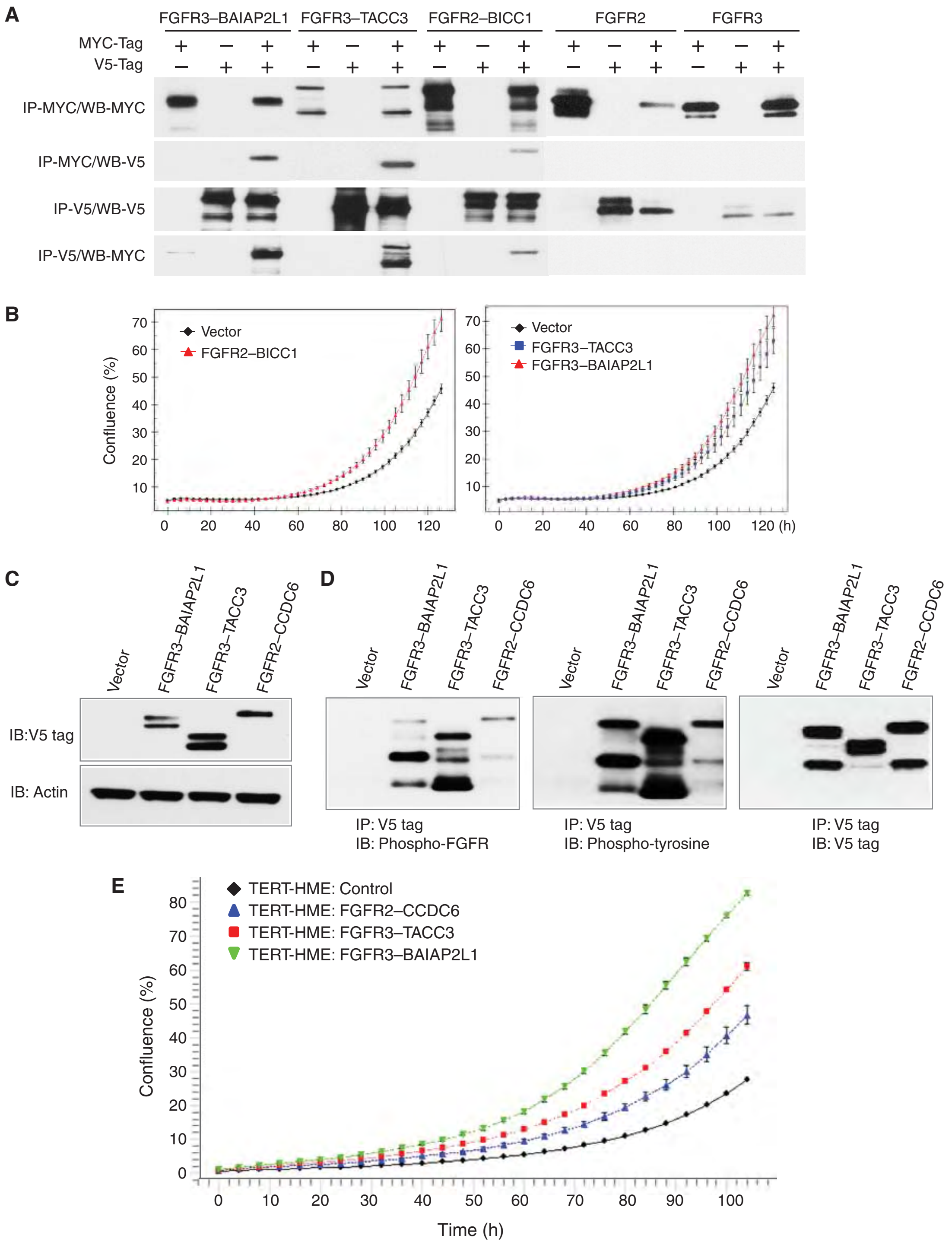


**Figure 3.** Functional characterization of FGFR fusion proteins. **A,** oligomerization of FGFR fusion proteins shown by immunoprecipitation (IP)-Western blotting (WB). HEK 293T cells were transfected with respective MYC- and V5-tagged FGFR wild-type or fusion proteins and reciprocal IP-WBs were carried out. **B,** cell proliferation assays as determined by live-cell imaging of 293T cells overexpressing various FGFR fusion proteins. Data shown are cell confluence versus time at 3-hour intervals. Each data point is the mean of quadruplicates. **C,** stable expression of FGFR fusion proteins in TERT-HME cells. Cell lysates were prepared from various stable lines and expression of chimeric proteins was detected by anti-V5 antibody. **D,** FGFR fusion protein activity in TERT-HME cells. Cell lysates from various stable lines were immunoprecipitated and immunoblotted (IB) with the antibodies indicated. **E,** overexpression of FGFR fusions induces cell proliferation in TERT-HME cells. Cell proliferation assays were conducted by IncuCyte live-cell imaging. Data shown are cell confluence versus time at 3-hour intervals. Each data point is the mean of quadruplicates.

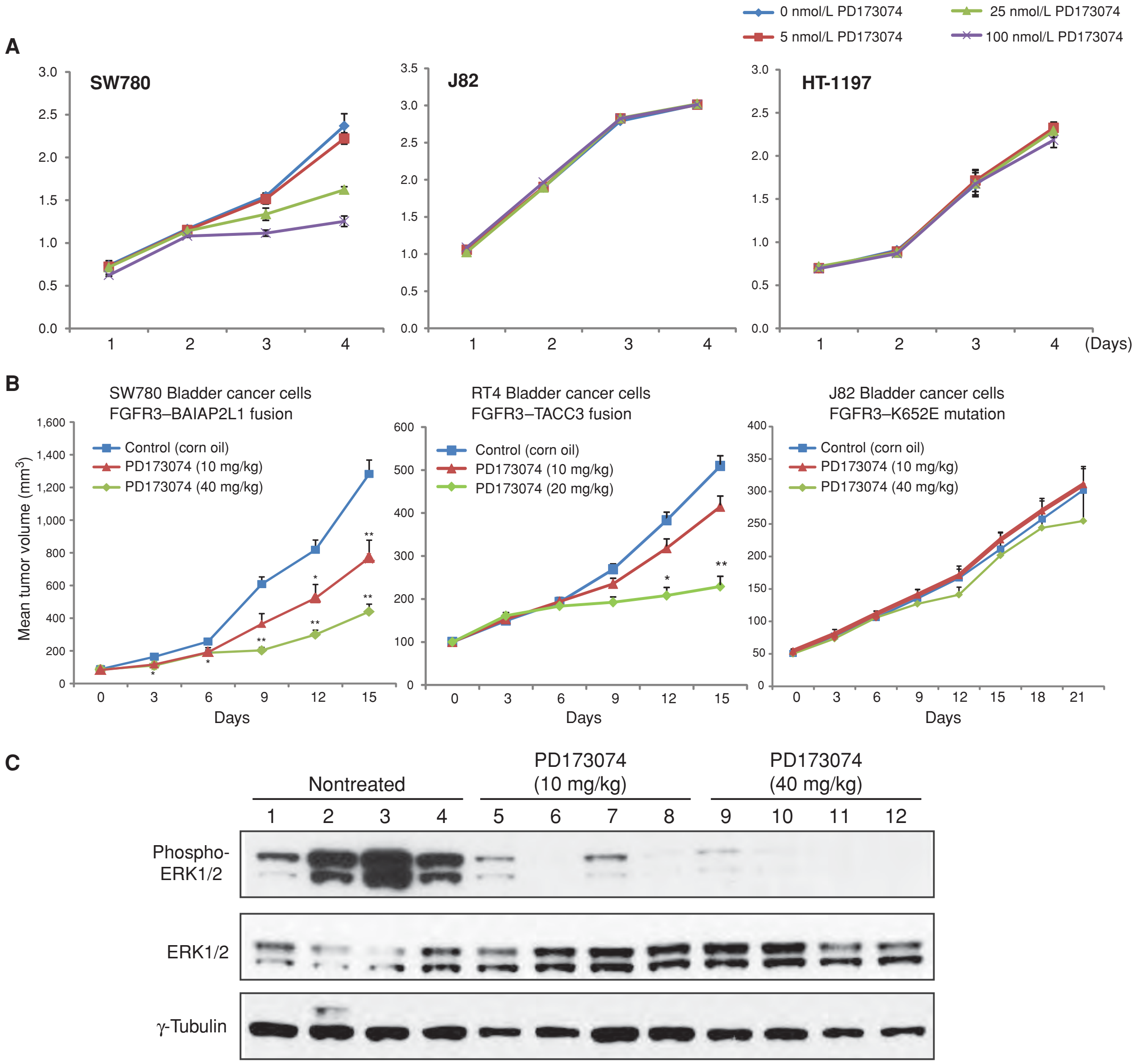


**Figure 4.** Inhibition of FGFR fusion kinase activity repressed tumor growth in a mouse xenograft model. **A,** inhibition of cell proliferation by the FGFR inhibitor PD173074. The *FGFR3–BAIAP2L1* bladder cell line SW780, and 2 control bladder cell lines J82 (K652E mutation) and HT-1197 (S249C mutation), were tested for the effects of PD173074 at 3 concentrations on cell proliferation, assessed by the WST-1 method at the indicated times. Data shown are the means of triplicates. **B,** differential sensitivity of FGFR fusion-positive versus FGFR-mutant bladder cancer xenograft growth to PD173074. Mice xenografted with bladder cancer SW780 cells (*FGFR3–BAIAP2L1* fusion), RT4 (*FGFR3–TACC3* fusion), or J82 cells (K652E mutation) were treated daily with PD173074 after tumors were formed. The tumor size was monitored over a time course of 3 weeks. *, $P < 0.05$; **, $P < 0.005$. **C,** inhibition of the FGFR signaling pathway by the FGFR inhibitor PD173074 in mouse xenograft tumors. Bladder cancer SW780 cells were implanted in mice and treated with PD173074 after tumor formation as shown in **B**. Protein lysates of tumor tissues were prepared and immunoblotted with antibodies against phospho-ERK1/2, pan-ERK1/2, and γ-tubulin.

fusion, underwent a conventional chemotherapy regimen in which her cancer progressed, chose not to pursue FGFR-directed therapy, and died 3 months after enrollment on this protocol. The second patient with cholangiocarcinoma, MO_1039, also harboring an *FGFR2–BICC1* fusion, underwent conventional chemotherapy but did not show tumor shrinkage and was enrolled on an FGFR inhibitor clinical trial. The patient with metastatic breast cancer, MO_1051, harboring the *FGFR2–AFF3* fusion, died of end-stage disease before the sequencing results were available. The patient with metastatic prostate cancer underwent irradiation of the brain (after brain metastasis resection) and continues to be maintained on hormonal treatment. Because of his brain metastasis, the prostate cancer patient was not eligible for an FGFR clinical trial.

Activating point mutations of *FGFR1*, *FGFR2*, *FGFR3*, or *FGFR4* have been identified in a variety of cancers, including gliomas, bladder cancer, multiple myeloma, and rhabdomyosarcomas (34). Studies of hematologic diseases led to the identification of 3′ gene fusions of *FGFR1* in myeloproliferative disorder (35) and 3′ *FGFR3* fusions in peripheral T-cell lymphoma (36) and multiple myeloma (35). As described earlier, 5′ gene fusions of *FGFR1* and *FGFR3* with *TACC1* and *TACC3* have recently been identified in GBM in 2 studies (15, 37). Here, we identify potentially actionable 5′ and 3′ FGFR rearrangements across a diverse array of both common and rare solid tumors. Ten novel FGFR fusion partners were identified. In the Singh and colleagues (15) GBM study, the mechanism of activation of the FGFR fusions was proposed to be through mislocalization to mitotic spindle poles mediated by the coiled-coil domain of the TACC fusion partner. This presumably leads to mitotic and chromosomal segregation defects, triggering aneuploidy. In the Parker and colleagues (37) GBM study, increased expression through loss of the *FGFR3* 3′ UTR and *miR-99a* regulation was hypothesized as an activating mechanism. While these may be potential mechanisms in the specific case of the FGFR3–TACC3 fusion proteins in GBM, this likely does not explain the diverse array of fusion partners identified for FGFRs in this study. We propose a different, potentially more inclusive, model in which the FGFR fusion partners (e.g., BICC1, TACC3, CCDC6, BAIAP2L1, KIAA1967, CASP7, CIT, and OFD1) mediate oligomerization, which triggers activation of the respective FGFR kinase. Of note, we have not detected any FGFR fusions that result in simple truncation of the FGFR protein, despite prior investigations suggesting that 3′ truncating splicing isoforms encode activated FGFR2 proteins (38). The FGFR fusions detected have persistently exhibited substantial dimerization domain contributions from the 3′ fusion partner.

The *SLC45A3–FGFR2* gene fusion identified in the index prostate cancer is quite interesting, as its pathogenic role is likely through a mechanism that is distinct from fusion protein oligomerization (shared by the other gene fusions tested). The entire open reading frame of *FGFR2* is expressed under the control of an androgen-regulated promoter of *SLC45A3*, leading to the marked overexpression of FGFR2. The *SLC45A3–FGFR2* fusion is analogous to the previously characterized *TMPRSS2–ETS* gene fusions characterized in more than 50% of prostate cancers (7). One would predict that this patient should respond to second-generation antiandrogens, such as MDV3100 (39), as well as FGFR inhibition. Another interesting observation in this study is the enhanced sensitivity to the FGFR inhibitor PD173074 of cell lines harboring an *FGFR3* fusion relative to those that have an activating point mutation of *FGFR3*. While beyond the scope of this study, additional FGFR inhibitors and larger panels of FGFR fusions and FGFR-mutant cell lines will need to be studied to determine the broader applicability of these results. Clinical trials for several FGFR inhibitors are underway or in late-stage preclinical development (33, 40, 41). It will be important to enrich these early-stage clinical trials with patients harboring FGFR gene fusions, similar to the successful development of the small-molecule kinase inhibitor crizotonib in patients with lung cancer harboring the *EML4–ALK* gene fusion. The wide range of cancers in which FGFR rearrangements were detected in this study suggests that development of FGFR rearrangements is lineage-independent and emphasizes the importance of developing mutation-enriched clinical trials rather than trials based on tissue of origin. While each individual type of genetic aberration may occur at low frequency, the integrated sequencing approach identifies a wide range of informative genetic aberrations, potentially guiding the enrollment into numerous trials of diverse therapeutics.

In this study, we identified 4 patients with FGFR family gene fusions through an established clinical sequencing project called MI-ONCOSEQ. Combining these index patients with an analysis of transcriptome data from our internal tumor cohorts as well as the TCGA identified FGFR fusions in a wide array of cancers, including cholangiocarcinoma, GBM, squamous lung cancer, bladder cancer, breast cancer, thyroid cancer, oral cancer, head and neck squamous cell carcinoma, and prostate cancer. In addition to *TACC1* and *TACC3*, we identified 10 additional FGFR fusion partners, as well as implicated 3 out of 4 FGFR family members (*FGFR1*, *2*, and *3*) in gene rearrangements. We also suggest a common mechanism of activation of these fusion proteins and show that FGFR gene fusion-positive cancers have enhanced susceptibility to FGFR inhibitors over activating point mutations of FGFR.

## METHODS

### Clinical Study and Specimen Collection

Sequencing of clinical samples was conducted under IRB-approved studies at the University of Michigan (Ann Arbor, MI). Patients were enrolled and consented for integrative tumor sequencing, MI-ONCOSEQ (IRB# HUM00046018; ref. 3). Medically qualified patients 18 years or older with advanced or refractory cancer are eligible for the study. Informed consent details the risks of integrative sequencing and includes upfront genetic counseling. Biopsies were arranged for safely accessible tumor sites. Needle biopsies were snap frozen in optimum cutting temperature compound and a longitudinal section was cut. Hematoxylin and eosin-stained frozen sections were reviewed by pathologists to identify cores with highest tumor content. Remaining portions of each needle biopsy core were retained for nucleic acid extraction.

### Cell Lines and Antibodies

Cell lines were purchased from the American Type Culture Collection and verified by next-generation transcriptome sequencing methods to identify known somatic mutations (COSMIC database). Oral cancer cell lines were obtained from their originating lab (A.-J. Cheng) and are not verified. Cells were grown in specified media supplemented with FBS and antibiotics (Invitrogen). Anti-c-MYC antibody was purchased from Sigma. Anti-V5 antibody was purchased from Life Technologies. Anti-FGFR3 antibodies were purchased from Epitomics and Cell Signaling. Antisera for phospho-FGFR, phospho-ERK1/2, pan-ERK1/2, phospho-STAT1, and pan-STAT1 were purchased from Cell Signaling. Anti-phosphotyrosine antibody clone 4G10 is from Millipore.

### DNA/RNA Isolation and cDNA Synthesis

Genomic DNA from frozen needle biopsies and blood was isolated using the Qiagen DNeasy Blood & Tissue Kit, according to the manufacturer's instructions. Total RNA was extracted from frozen needle biopsies using the Qiazol reagent with disruption using a 5-mm bead on a TissueLyser II (Qiagen), and purified using a miRNeasy Kit (Qiagen) with DNase I digestion, according to the manufacturer's instructions. Total RNA was isolated from cancer cell lines using the TRIzol reagent (Life Technologies). RNA integrity was verified on an Agilent 2100 Bioanalyzer using RNA Nano reagents (Agilent Technologies). cDNA was synthesized from total RNA using SuperScript

III (Invitrogen) and random primers (Invitrogen) for quantitative real-time PCR (qRT-PCR) analysis.

### Preparation of Next-Generation Sequencing Libraries

Transcriptome libraries were prepared following Illumina's TruSeq RNA protocol, using 1–2 μg of total RNA. Poly(A)$^+$ RNA was isolated using Sera-Mag oligo(dT) beads (Thermo Scientific) and fragmented with the Ambion Fragmentation Reagents kit (Ambion). cDNA synthesis, end-repair, A-base addition, and ligation of the Illumina indexed adapters were conducted according to Illumina's protocol. Libraries were size-selected for 250–300 bp cDNA fragments on a 3% Nusieve 3:1 (Lonza) agarose gel, recovered using QIAEX II gel extraction reagents (Qiagen), and PCR-amplified using Phusion DNA polymerase (New England Biolabs) for 14 PCR cycles. The amplified libraries were purified using AMPure XP beads. Library quality was measured on an Agilent 2100 Bioanalyzer for product size and concentration. Paired-end libraries were sequenced with the Illumina HiSeq 2000 (2 × 100 nucleotide read length). Reads that passed the chastity filter of Illumina BaseCall software were used for subsequent analysis.

Exome libraries of matched pairs of tumor/normal genomic DNAs were generated using the Illumina TruSeq DNA Sample Prep Kit, following the manufacturer's instructions. Three micrograms of each genomic DNA was sheared using a Covaris S2 to a peak target size of 250 bp. Fragmented DNA was concentrated using AMPure XP beads (Beckman Coulter), followed by end-repair, A-base addition, and ligation of the Illumina indexed adapters according to Illumina's protocol. The adapter-ligated libraries were electrophoresed on 3% Nusieve 3:1 (Lonza) agarose gels and fragments between 300 and 350 bp were recovered using QIAEX II gel extraction reagents (Qiagen). Recovered DNA was amplified using Illumina index primers for 8 cycles. The amplified libraries were purified using AMPure XP beads and the DNA concentration was determined using a Nanodrop spectrophotometer. One microgram of the libraries were hybridized to the Agilent SureSelect Human All Exon V4 at 65°C for 60 hours following the manufacturer's protocol (Agilent). The targeted exon fragments were captured on Dynal M-280 streptavidin beads (Invitrogen), and enriched by amplification with the Illumina index primers for 9 additional cycles. After purification of the PCR products with AMPure XP beads, the quality and quantity of the resulting exome libraries were analyzed using an Agilent 2100 Bioanalyzer and DNA 1000 reagents.

We used the publicly available software FastQC to assess sequencing quality. For each lane, we examined per-base quality scores across the length of the reads. Lanes were deemed passing if the per-base quality score boxplot indicated that more than 75% of the reads had >Q20 for bases 1–80. In addition to the raw sequence quality, we also assessed alignment quality using the Picard package. This allows monitoring of duplication rates and chimeric reads that may result from ligation artifacts and provides crucial statistics for interpreting the results of copy number and structural variant analysis.

### Nomination of Gene Fusions

To identify gene fusions, paired-end transcriptome reads passing filter were mapped to the human reference genome and UCSC genes, allowing up to 2 mismatches, with Illumina ELAND software (Efficient Alignment of Nucleotide Databases) and Bowtie (42). Sequence alignments were subsequently processed to nominate gene fusions using the method described earlier (9). In brief, paired-end reads were processed to identify those that either contained or spanned a fusion junction. Encompassing paired reads refer to those in which each read aligns to an independent transcript, thereby encompassing the fusion junction. Spanning mate pairs refer to those in which one sequence read aligns to a gene and its paired-end spans the fusion junction. Both categories undergo a series of filtering steps to remove false positives before being merged together to generate the final chimera nominations. Reads supporting each fusion were realigned using BLAT (UCSC Genome Browser) to reconfirm the fusion breakpoint.

### Mutation Analyses

We annotated the resulting somatic mutations using RefSeq transcripts. HUGO gene names were used. The impact of coding nonsynonymous amino acid substitutions on the structure and function of a protein was assessed using Blocks Substitution Matrix scores. We also assessed whether the somatic variant was previously reported in dbSNP135 or COSMIC v5668.

Tumor content for each tumor exome library was estimated from the sequence data by fitting a binomial mixture model with 2 components to the set of most likely single-nucleotide variant (SNV) candidates on 2-copy genomic regions. The set of candidates used for estimation consisted of coding variants that (i) exhibited at least 3 variant fragments in the tumor sample, (ii) exhibited zero variant fragments in the matched benign sample with at least 16 fragments of coverage, (iii) were not present in dbSNP, (iv) were within a targeted exon or within 100 base pairs of a targeted exon, (v) were not in homopolymer runs of 4 or more bases, and (vi) exhibited no evidence of amplification or deletion. To filter out regions of possible amplification or deletion, we used exon coverage ratios to infer copy number changes, as described below. Resulting SNV candidates were not used for estimation of tumor content if the segmented log-ratio exceeded 0.2 in absolute value. Candidates on the Y chromosome were also eliminated because they were unlikely to exist in 2-copy genomic regions. Using this set of candidates, we fit a binomial mixture model with 2 components using the R package flexmix, version 2.3-8. One component consisted of SNV candidates with very low variant fractions, presumably resulting from recurrent sequencing errors and other artifacts. The other component, consisting of the likely set of true SNVs, was informative of tumor content in the tumor sample. Specifically, under the assumption that most or all of the observed SNV candidates in this component are heterozygous SNVs, we expect the estimated binomial proportion of this component to represent one-half of the proportion of tumor cells in the sample. Thus, the estimated binomial proportion as obtained from the mixture model was doubled to obtain an estimate of tumor content.

Copy number aberrations were quantified and reported for each gene as the segmented normalized log2-transformed exon coverage ratios between each tumor sample and matched normal sample (43). To account for observed associations between coverage ratios and variation in GC content across the genome, locally weighted scatterplot smoothing (LOWESS) normalization was used to correct per-exon coverage ratios before segmentation analysis. Specifically, mean GC percentage was computed for each targeted region, and a LOWESS curve was fit to the scatterplot of log2-coverage ratios vs. mean GC content across the targeted exome using the LOWESS function in R (version 2.13.1) with smoothing parameter $f = 0.05$.

Somatic point mutations were identified in the tumor exome sequence data using the matched normal exome data to eliminate germline polymorphisms. Parameters and computational methods were as previously described (44).

For RNA-seq gene expression analysis, transcriptome data was processed as previously described. Genes were nominated as exhibiting potential "outlier" expression relative to a cohort of $n$ = 282 previously sequenced tissues using the following conditions: (i) the gene was required to have an expression value of at least 20 RPKM in the sample of interest; (ii) the gene was required to be at or above the 90th percentile relative to all previously sequenced tissues, of any type; (iii) the gene was required to have a fold change of at least 2 relative to the maximum reads per kilobase per million reads over all previously sequenced benign tissues; and (iv) the 25th percentile of the gene expression measurements over the previously sequenced tissues was required to be less than 50 RPKM. Collectively, these parameters target genes with (i) high absolute expression, (ii) high expression relative to previously sequenced tissues, (iii) high expression relative to all benign tissues, and (iv) expression that is not uniformly high across all tissues.

Partially redundant sequencing of areas of the genome affords the ability for cross-validation of findings. We cross-validated exome-based point mutation calls by manually examining the genomic and transcriptomic reads covering the mutation using the UCSC Genome Browser. Likewise, gene fusion calls from the transcriptome data can be further supported by structural variant detection in the genomic sequence data as well as copy number information derived from the genome and exome sequencing.

### Quantitative RT-PCR

For validation of fusion transcripts, qRT-PCR assays were conducted. Total cDNAs of index cases and negative control samples were synthesized using SuperScript III System according to the manufacturer's instructions (Invitrogen). Quantitative RT-PCR was conducted using fusion-specific primers (Supplementary Table S15) with SYBR Green Master Mix (Applied Biosystems) on the StepOne Real-Time PCR System (Applied Biosystems). The PCR products were further analyzed by agarose gel electrophoresis. Relative mRNA levels of the fusion transcripts were normalized to the expression of the housekeeping gene *GAPDH*.

### Inhibition of FGFR Receptors and Cell Proliferation Assay

Bladder cancer cells SW780, J82, and HT-1197 were seeded into 96-well plates in triplicate and allowed to attach before drug treatment. The FGFR inhibitor PD173074 (Selleck Chemicals) was added to the cultures at concentrations of 0, 5, 25, and 100 nmol/L. Relative cell numbers were measured by WST-1 assays at indicated time points following the manufacturer's instructions (Roche). To test the effects of the FGFR inhibitor pazopanib (Selleck Chemicals) on cell proliferation, SW780, RT4, J82, and HT-1197 cells were seeded into 24-well plates in quadruplicates and allowed to attach before drug treatment. Pazopanib was added to the cultures at concentrations of 0, 0.1, 0.5, and 1 μmol/L. Cell proliferation was determined by IncuCyte live-cell imaging system (Essen Biosciences).

### Cloning and Expression of FGFR Fusions

The FGFR fusion alleles were PCR amplified from cDNA of the index cases or cell lines using the primers listed in Supplementary Table S15 and the Expand High Fidelity protocol (Roche). PCR products were digested with restriction endonuclease and ligated into the pcDNA3.1 vector (Invitrogen), which had been modified to contain a C-terminal MYC-epitope tag or V5-epitope tag. Expression constructs were transfected into HEK 293T cells using FuGene HD transfection reagent (Promega). Cells were harvested 24 hours after transfection for protein analysis. For stable line establishment in TERT-HME cells, FGFR fusion alleles were cloned into the pCDH510B lentiviral vector (System Biosciences), which had been modified to contain a C-terminal V5 epitope tag. Lentiviruses were produced with the ViraPower packaging mix (Invitrogen) in 293T cells using FuGene HD transfection reagent (Roche). Benign TERT-HME cells at 30% confluence were infected at a multiplicity of infection of 20 with the addition of polybrene at 8 mg/mL, and the cells were selected by 20 μg/mL puromycin. Stable pools of resistant cells were obtained and analyzed for expression of the FGFR fusion proteins by Western blot analysis with anti-V5 antibody. Cell proliferation was measured by IncuCyte imaging system as described above.

For the cell proliferation assay, HEK 293T cells were transfected with control vector or FGFR fusion constructs. Twenty-four hours after transfection, cells were trypsinized, resuspended in Dulbecco's Modified Eagle Medium (DMEM) containing 2% FBS, and plated in quadruplicate at 12,000 cells per well in 24-well plates. The plates were incubated at 37°C and 5% $CO_2$ atmosphere using the IncuCyte live-cell imaging system (Essen Biosciences). Cell proliferation was assessed by kinetic imaging confluence measurements at 3-hour time intervals.

### Coimmunoprecipitation

HEK 293T cells were grown to approximately 70% confluence in DMEM supplemented with 10% FBS, followed by transfection with MYC-tagged or V5-tagged expression construct alone or in combination using FuGene6 reagent (Promega). Twenty-four hours after transfection, cell pellets were lysed in lysis buffer (58 mmol/L Na2HPO4, 17 mmol/L NaH2PO4, 68 mmol/L NaCl, 1% Triton X-100, 0.5% sodium deoxycholate, 0.1% SDS, and protease inhibitors), followed by immunoprecipitation with tag epitope-specific antibodies (Sigma) and protein-G Dynabeads (Invitrogen). Precipitates were washed 3 times with IP Wash buffer (20 mmol/L Tris, pH 8, 2 mmol/L EDTA, 150 mmol/L NaCl, 1% Triton X100) and eluted in SDS-PAGE loading buffer at 95°C for 5 minutes. Immunoprecipitated proteins were separated on SDS-PAGE and detected by Western blotting with tag epitope-specific antibodies (Sigma).

### siRNA Knockdown of FGFR3 and BAIAP2L1

SW780, J82, and HT-1197 bladder cancer cells were transfected twice with *FGFR3*-targeting siRNA, *BAIAP2L1*-targeting siRNA, or nontargeting siRNA (Thermo Scientific Dharmacon) using DharmaFECT1 reagent (Dharmacon). The siRNAs used were as follows: ON-TARGETplus FGFR3 L-003133-00-0005, ON-TARGETplus BAIAP2L1 L-018664-00-0005, and ON-TARGETplus Nontargeting pool. Twenty-four hours after transfection, cells were trypsinized and plated in triplicate at 8,000 cells per well in 24-well plates. The plates were incubated at 37°C with 5% $CO_2$ atmosphere in the IncuCyte live-cell imaging system (Essen Biosciences). Cell proliferation rate was assessed by kinetic imaging confluence measurements at 3-hour time intervals.

### Mouse Xenograft Models

Five week-old male C.B17/SCID mice were procured from a breeding colony at University of Michigan, maintained by Dr. Kenneth Pienta. Mice were anesthetized using a cocktail of xylazine (80 mg/kg, intraperitoneal) and ketamine (10 mg/kg, intraperitoneal) for chemical restraint. Bladder cancer cells SW780 (2 million cells for each implantation site) or J82 (5 million cells for each implantation site) were resuspended in 100 μL of 1× PBS with 20% Matrigel (BD Biosciences) and were implanted subcutaneously into flank region on both sides. Eight mice were included in each experimental group. All tumors were staged for 2 weeks (SW780 cells) and 3 weeks (J82 cells) before starting the drug treatment. Xenografted mice with palpable tumors were treated with a FGFR inhibitor PD173074 (Selleck Chemicals) dissolved in 5% ethanol in corn oil (intraperitoneal). Mice in control group received 5% ethanol in corn oil as vehicle control. Tumor growth was recorded weekly by using digital calipers, and tumor volumes were calculated using the formula $(\pi/6)$ $(L \times W2)$, where $L =$ length of tumor and $W =$ width. Any decrease in the body weight of mice was monitored biweekly during the course of the study. All experimental procedures involving mice were approved by the University Committee on Use and Care of Animals at the University of Michigan and conform to their relevant regulatory standards. Tumor tissues from xenografted SW780 cells were harvested and lysed in radioimmunoprecipitation assay buffer containing protease/phosphatase inhibitors for Western blot analysis.

## Disclosure of Potential Conflicts of Interest

A.M. Chinnaiyan is a consultant to Life Technologies, co-founder of Compendia Biosciences, which is now owned by Life Technologies, and advisor to Ventana/Roche and Gen-Probe/Hologic.

## Authors' Contributions

**Conception and design:** Y.-M. Wu, F. Su, X. Cao, S. Roychowdhury, M. Talpaz, D.R. Robinson, A.M. Chinnaiyan
**Development of methodology:** Y.-M. Wu, F. Su, S. Kalyana-Sundaram, B. Ateeq, A.-J. Cheng, D.R. Rhodes, D.R. Robinson, A.M. Chinnaiyan
**Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.):** F. Su, X. Cao, J. Siddiqui, S.A. Tomlins, S. Roychowdhury, M. Hussain, F.Y. Feng, M.M. Zalupski, M. Talpaz, K.J. Pienta
**Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis):** Y.-M. Wu, F. Su, S. Kalyana-Sundaram, N. Khazanov, B. Ateeq, R.J. Lonigro, P. Vats, L.P. Kunju, S.A. Tomlins, P. Wyngaard, S. Sadis, S. Roychowdhury, K.J. Pienta, D.R. Rhodes, D.R. Robinson
**Writing, review, and/or revision of the manuscript:** Y.-M. Wu, F. Su, R. Wang, L.P. Kunju, J. Siddiqui, M. Hussain, F.Y. Feng, M.M. Zalupski, M. Talpaz, D.R. Robinson, A.M. Chinnaiyan
**Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases):** F. Su, X. Cao, R. Wang, S.-F. Lin, A.-J. Cheng, J. Siddiqui, S. Roychowdhury
**Study supervision:** Y.-M. Wu, S. Roychowdhury, D.R. Rhodes, D.R. Robinson, A.M. Chinnaiyan

## Acknowledgments

The authors thank Terrance Barrette and Doug Gibbs for hardware and database management, Shanda Birkeland, Mary Pierce-Burlingame, and Karen Giles for assistance with sample and manuscript preparation, Rachell Stender for animal care, and Dan Hayes for referral of patients to the MI-ONCOSEQ program. The authors also thank the larger MI-ONCOSEQ team, including clinical research coordinator Lynda Hodges, Amy Gursky for patient sample procurement, cancer geneticists S. Gruber and J. Innis, bioethicists J. Scott Roberts and Scott Y. Kim, genetic counselors J. Everett, J. Long, and V. Raymond, and radiologists E. Higgins, E. Caoili, and R. Dunnick.

## Grant Support

This project is supported in part by the Prostate Cancer Foundation, the NCI Early Detection Research Network (U01 CA111275), NIH RO1CA132874-01, a Department of Defense Era of Hope Scholar Award, and the University of Michigan Comprehensive Cancer Center Prostate SPORE (P50CA 069568). A.M. Chinnaiyan is supported by a Stand Up To Cancer-Prostate Cancer Foundation Prostate Dream Team Translational Cancer Research Grant. Stand Up To Cancer is a program of the Entertainment Industry Foundation administered by the American Association for Cancer Research (SU2C-AACR-DT0712). A.M. Chinnaiyan is also supported by the Alfred A. Taubman Institute, the American Cancer Society, the Howard Hughes Medical Institute, and a Doris Duke Charitable Foundation Clinical Scientist Award.

Received February 5, 2013; revised April 3, 2013; accepted April 3, 2013; published OnlineFirst April 4, 2013.

## REFERENCES

1. Chin L, Andersen JN, Futreal PA. Cancer genomics: from discovery science to personalized medicine. Nat Med 2011;17:297–303.
2. Meyerson M, Gabriel S, Getz G. Advances in understanding cancer genomes through second-generation sequencing. Nat Rev Genet 2010;11:685–96.
3. Roychowdhury S, Iyer MK, Robinson DR, Lonigro RJ, Wu YM, Cao X, et al. Personalized oncology through integrative high-throughput sequencing: a pilot study. Sci Transl Med 2011;3:111ra21.
4. Welch JS, Westervelt P, Ding L, Larson DE, Klco JM, Kulkarni S, et al. Use of whole-genome sequencing to diagnose a cryptic fusion oncogene. JAMA 2011;305:1577–84.
5. Rowley JD. Letter: A new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and Giemsa staining. Nature 1973;243:290–3.
6. Druker BJ. Translation of the Philadelphia chromosome into therapy for CML. Blood 2008;112:4808–17.
7. Tomlins SA, Rhodes DR, Perner S, Dhanasekaran SM, Mehra R, Sun XW, et al. Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. Science 2005;310:644–8.
8. Perner S, Wagner PL, Demichelis F, Mehra R, Lafargue CJ, Moss BJ, et al. EML4-ALK fusion lung cancer: a rare acquired event. Neoplasia 2008;10:298–302.
9. Robinson DR, Kalyana-Sundaram S, Wu YM, Shankar S, Cao X, Ateeq B, et al. Functionally recurrent rearrangements of the MAST kinase and Notch gene families in breast cancer. Nat Med 2011;17:1646–51.
10. Santoro M, Melillo RM, Fusco A. RET/PTC activation in papillary thyroid carcinoma: European Journal of Endocrinology Prize Lecture. Eur J Endocrinol 2006;155:645–53.
11. Seshagiri S, Stawiski EW, Durinck S, Modrusan Z, Storm EE, Conboy CB, et al. Recurrent R-spondin fusions in colon cancer. Nature 2012;488:660–4.
12. Soda M, Choi YL, Enomoto M, Takada S, Yamashita Y, Ishikawa S, et al. Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer. Nature 2007;448:561–6.
13. Gerber DE, Minna JD. ALK inhibition for non-small cell lung cancer: from discovery to therapy in record time. Cancer Cell 2010;18:548–51.
14. Kwak EL, Bang YJ, Camidge DR, Shaw AT, Solomon B, Maki RG, et al. Anaplastic lymphoma kinase inhibition in non-small-cell lung cancer. N Engl J Med 2010;363:1693–703.
15. Singh D, Chan JM, Zoppoli P, Niola F, Sullivan R, Castano A, et al. Transforming fusions of FGFR and TACC genes in human glioblastoma. Science 2012;337:1231–5.
16. Williams SV, Hurst CD, Knowles MA. Oncogenic FGFR3 gene fusions in bladder cancer. Hum Mol Genet 2013;22:795–803.
17. Lamont FR, Tomlinson DC, Cooper PA, Shnyder SD, Chester JD, Knowles MA. Small molecule FGF receptor inhibitors block FGFR-dependent urothelial carcinoma growth *in vitro* and *in vivo*. Br J Cancer 2011;104:75–82.
18. Liang G, Liu Z, Wu J, Cai Y, Li X. Anticancer molecules targeting fibroblast growth factor receptors. Trends Pharmacol Sci 2012;33:531–41.
19. Wilson BG, Roberts CW. SWI/SNF nucleosome remodellers and cancer. Nat Rev Cancer 2011;11:481–92.
20. Tomlins SA, Laxman B, Dhanasekaran SM, Helgeson BE, Cao X, Morris DS, et al. Distinct classes of chromosomal rearrangements create oncogenic ETS gene fusions in prostate cancer. Nature 2007;448:595–9.
21. Glinsky GV, Glinskii AB, Stephenson AJ, Hoffman RM, Gerald WL. Gene expression profiling predicts clinical outcome of prostate cancer. J Clin Invest 2004;113:913–23.
22. Turner N, Grose R. Fibroblast growth factor signalling: from development to cancer. Nat Rev Cancer 2010;10:116–29.
23. Browman DT, Hoegg MB, Robbins SM. The SPFH domain-containing proteins: more than lipid raft markers. Trends Cell Biol 2007;17:394–402.
24. Chai J, Wu Q, Shiozaki E, Srinivasula SM, Alnemri ES, Shi Y. Crystal structure of a procaspase-7 zymogen: mechanisms of activation and substrate binding. Cell 2001;107:399–407.
25. Ishizaki T, Naito M, Fujisawa K, Maekawa M, Watanabe N, Saito Y, et al. p160ROCK, a Rho-associated coiled-coil forming protein kinase, works downstream of Rho and induces focal adhesions. FEBS Lett 1997;404:118–24.
26. Knight MJ, Leettola C, Gingery M, Li H, Bowie JU. A human sterile alpha motif domain polymerizome. Protein Sci 2011;20:1697–706.
27. Mateja A, Cierpicki T, Paduch M, Derewenda ZS, Otlewski J. The dimerization mechanism of LIS1 and its implication for proteins containing the LisH motif. J Mol Biol 2006;357:621–31.

28. Peter BJ, Kent HM, Mills IG, Vallis Y, Butler PJ, Evans PR, et al. BAR domains as sensors of membrane curvature: the amphiphysin BAR structure. Science 2004;303:495–9.
29. Tong Q, Li Y, Smanik PA, Fithian LJ, Xing S, Mazzaferri EL, et al. Characterization of the promoter region and oligomerization domain of H4 (D10S170), a gene frequently rearranged with the ret proto-oncogene. Oncogene 1995;10:1781–7.
30. Seo JS, Ju YS, Lee WC, Shin JY, Lee JK, Bleazard T, et al. The transcriptional landscape and mutational profile of lung adenocarcinoma. Genome Res 2012;22:2109–19.
31. Mohammadi M, Froum S, Hamby JM, Schroeder MC, Panek RL, Lu GH, et al. Crystal structure of an angiogenesis inhibitor bound to the FGF receptor tyrosine kinase domain. EMBO J 1998;17:5896–904.
32. Miyake M, Ishii M, Koyama N, Kawashima K, Kodama T, Anai S, et al. 1-tert-butyl-3-[6-(3,5-dimethoxy-phenyl)-2-(4-diethyl-amino-butylamino)-pyrido[2,3-d]pyrimidin-7-yl]-urea (PD173074), a selective tyrosine kinase inhibitor of fibroblast growth factor receptor-3 (FGFR3), inhibits cell proliferation of bladder cancer carrying the FGFR3 gene mutation along with up-regulation of p27/Kip1 and G1/G0 arrest. J Pharmacol Exp Ther 2010;332:795–802.
33. Guagnano V, Kauffmann A, Wohrle S, Stamm C, Ito M, Barys L, et al. FGFR genetic alterations predict for sensitivity to NVP-BGJ398, a selective pan-FGFR inhibitor. Cancer Discov 2012;2:1118–33.
34. Wesche J, Haglund K, Haugsten EM. Fibroblast growth factors and their receptors in cancer. Biochem J 2011;437:199–213.
35. Jackson CC, Medeiros LJ, Miranda RN. 8p11 myeloproliferative syndrome: a review. Hum Pathol 2010;41:461–76.
36. Yagasaki F, Wakao D, Yokoyama Y, Uchida Y, Murohashi I, Kayano H, et al. Fusion of ETV6 to fibroblast growth factor receptor 3 in peripheral T-cell lymphoma with a t(4;12)(p16;p13) chromosomal translocation. Cancer Res 2001;61:8371–4.
37. Parker BC, Annala MJ, Cogdell DE, Granberg KJ, Sun Y, Ji P, et al. The tumorigenic FGFR3-TACC3 gene fusion escapes miR-99a regulation in glioblastoma. J Clin Invest 2013;123:855–65.
38. Cha JY, Maddileti S, Mitin N, Harden TK, Der CJ. Aberrant receptor internalization and enhanced FRS2-dependent signaling contribute to the transforming activity of the fibroblast growth factor receptor 2 IIIb C3 isoform. J Biol Chem 2009;284:6227–40.
39. Scher HI, Beer TM, Higano CS, Anand A, Taplin ME, Efstathiou E, et al. Antitumour activity of MDV3100 in castration-resistant prostate cancer: a phase 1-2 study. Lancet 2010;375:1437–46.
40. Brooks AN, Kilgour E, Smith PD. Molecular pathways: fibroblast growth factor signaling: a new therapeutic opportunity in cancer. Clin Cancer Res 2012;18:1855–62.
41. Greulich H, Pollock PM. Targeting mutant fibroblast growth factor receptors in cancer. Trends Mol Med 2011;17:283–92.
42. Langmead B. Aligning short sequencing reads with Bowtie. Curr Protoc Bioinformatics 2010;Chapter 11:Unit 11.7.
43. Lonigro RJ, Grasso CS, Robinson DR, Jing X, Wu YM, Cao X, et al. Detection of somatic copy number alterations in cancer using targeted exome capture sequencing. Neoplasia 2011;13:1019–25.
44. Grasso CS, Wu YM, Robinson DR, Cao X, Dhanasekaran SM, Khan AP, et al. The mutational landscape of lethal castration-resistant prostate cancer. Nature 2012;487:239–43.

nature genetics

# Activating *ESR1* mutations in hormone-resistant metastatic breast cancer

Dan R Robinson[1,2,12], Yi-Mi Wu[1,2,12], Pankaj Vats[1,2], Fengyun Su[1,2], Robert J Lonigro[1,3], Xuhong Cao[1,4], Shanker Kalyana-Sundaram[1,2], Rui Wang[1,2], Yu Ning[1,2], Lynda Hodges[1], Amy Gursky[1,2], Javed Siddiqui[1,2], Scott A Tomlins[1,2], Sameek Roychowdhury[5], Kenneth J Pienta[6], Scott Y Kim[7], J Scott Roberts[8], James M Rae[3,9], Catherine H Van Poznak[9], Daniel F Hayes[9], Rashmi Chugh[9], Lakshmi P Kunju[1,2], Moshe Talpaz[9], Anne F Schott[9] & Arul M Chinnaiyan[1–4,10,11]

**Breast cancer is the most prevalent cancer in women, and over two-thirds of cases express estrogen receptor-α (ER-α, encoded by *ESR1*). Through a prospective clinical sequencing program for advanced cancers, we enrolled 11 patients with ER-positive metastatic breast cancer. Whole-exome and transcriptome analysis showed that six cases harbored mutations of *ESR1* affecting its ligand-binding domain (LBD), all of whom had been treated with anti-estrogens and estrogen deprivation therapies. A survey of The Cancer Genome Atlas (TCGA) identified four endometrial cancers with similar mutations of *ESR1*. The five new LBD-localized *ESR1* mutations identified here (encoding p.Leu536Gln, p.Tyr537Ser, p.Tyr537Cys, p.Tyr537Asn and p.Asp538Gly) were shown to result in constitutive activity and continued responsiveness to anti-estrogen therapies *in vitro*. Taken together, these studies suggest that activating mutations in *ESR1* are a key mechanism in acquired endocrine resistance in breast cancer therapy.**

Advances in high-throughput sequencing technologies are beginning to establish a molecular taxonomy for a spectrum of human diseases and has facilitated a move toward 'precision medicine' (refs. 1,2). With regard to oncology, defining the mutational landscape of a patient's tumor will lead to more precise treatment and management of individuals with cancer. Comprehensive clinical sequencing programs for cancer patients have been initiated at a variety of medical centers, including our own[3,4]. In addition to the potential for identifying 'actionable' therapeutic targets in cancer patients, these clinical sequencing efforts may also shed light on acquired resistance mechanisms developed against targeted therapies[5–7].

ER is the primary therapeutic target in breast cancer and is expressed in 70% of cases[8]. Drugs directly antagonizing ER, such as tamoxifen and fulvestrant, are a mainstay of breast cancer treatment; however, approximately 30% of ER-positive breast cancers exhibit *de novo* resistance, whereas 40% acquire resistance to these therapies[9]. In addition to anti-estrogen therapies, patients with ER-positive breast cancer are also treated with aromatase inhibitors such as letrozole and exemestane[10]. Aromatase inhibitors block the peripheral conversion of androgens into estrogen and, in post-menopausal women, lead to over a 98% decrease in circulating levels of estrogen. As with anti-estrogens, treatment with aromatase inhibitors results in the development of resistance, but this is presumably due to different mechanisms, as patients with breast cancer who develop resistance to aromatase inhibitors often still respond to anti-estrogen therapies[11]. The molecular mechanisms of endocrine resistance in ER-positive breast cancer continues to be an active area of research[12].

Our institutional review board (IRB)-approved clinical sequencing program, called MI-ONCOSEQ (the Michigan Oncology Sequencing Program), enrolls patients with advanced cancer across all histologies[3]. Since April 2011, we have enrolled over 200 patients in this program, which involves obtaining a current tumor biopsy with matched normal samples (blood and/or buccal swab). Samples are then subjected to integrative sequencing, which includes whole-exome sequencing of the tumor and matched normal sample, transcriptome sequencing and, as needed, low-pass whole-genome sequencing[3]. This combination of DNA and RNA sequencing technologies allows one to be relatively comprehensive with regard to the mutational landscape of coding genes, including analysis of point mutations, indels, amplifications, deletions, gene fusions or translocations, and outlier gene expression profiles. These results are generated within a 5- to 7-week time frame and are presented at an institutional 'precision medicine tumor board' to deliberate upon potentially actionable findings.

[1]Michigan Center for Translational Pathology, University of Michigan Medical School, Ann Arbor, Michigan, USA. [2]Department of Pathology, University of Michigan Medical School, Ann Arbor, Michigan, USA. [3]Comprehensive Cancer Center, University of Michigan Medical School, Ann Arbor, Michigan, USA. [4]Howard Hughes Medical Institute, University of Michigan Medical School, Ann Arbor, Michigan, USA. [5]Department of Internal Medicine, Ohio State University, Columbus, Ohio, USA. [6]Brady Urological Institute, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA. [7]Center for Bioethics and Social Science in Medicine, University of Michigan, Ann Arbor, Michigan, USA. [8]Department of Health Behavior & Health Education, University of Michigan, Ann Arbor, Michigan, USA. [9]Department of Internal Medicine, University of Michigan Medical School, Ann Arbor, Michigan, USA. [10]Department of Urology, University of Michigan Medical School, Ann Arbor, Michigan, USA. [11]Center for Computational Medicine and Biology, University of Michigan, Ann Arbor, Michigan, USA. [12]These authors contributed equally to this work. Correspondence should be addressed to A.M.C. (arul@umich.edu).

Received 22 July; accepted 10 October; published online 3 November 2013; doi:10.1038/ng.2823

**Table 1 Clinical sequencing of 11 metastatic ER-positive breast cancer cases**

| Case | Age (years) | ER/PR/ERBB2 | Treatment[a] | Number of SNVs/fusions | Genetic aberration[b] |
|---|---|---|---|---|---|
| MO_1031 | 41 | +/+/– | Tamoxifen, letrozole, fulvestrant | 266/18 | *ESR1* (p.Leu536Gln), gene copy gains of *FGFR1*, *FGFR2*, *CCND1* and *GNRHR* |
| MO_1051 | 31 | +/–/– | Oophorectomy, letrozole, fulvestrant | 248/5 | *ESR1* (p.Tyr537Ser), *PIK3CA* (p.His1047Arg), *TP53* (p.Gly199Glu), *FGFR2-AFF3* fusion |
| MO_1069 | 62 | +/+/– | Tamoxifen, letrozole, fulvestrant | 74/9 | *ESR1* (p.Asp538Gly), *ARID2* (p.Glu245*), gene copy losses of *TP53*, *BRCA1*, *RB1*, *ARID1A* and *SMARCA4* |
| MO_1129 | 44 | +/+/– | Tamoxifen, oophorectomy, anastrozole, fulvestrant, exemestane | 32/3 | *ESR1* (p.Tyr537Ser), *PIK3CA* (p.Glu542Lys), gene copy gains of *CCND1* and *PAK1* |
| MO_1030 | 78 | +/+/– | Tamoxifen (short), anastrozole, fulvestrant | 26/2 | *PIK3CA* (p.Glu545Ala), *TP53* copy loss |
| MO_1068 | 65 | +/–/– | Tamoxifen, anastrozole | 83/10 | *PIK3CA* (p.His1047Arg), *TP53* (p.Glu51*), *MSH2* copy loss |
| MO_1090 | 52 | +/+/– | Tamoxifen, anastrozole | 28/11 | No significant drivers identified |
| MO_1107 | 46 | +/+/– | Tamoxifen, oophorectomy, anastrozole, fulvestrant, exemestane | 63/12 | *BRCA1* (c.5385_5386insC), frameshift deletions in *TP53*, *SMARCA4* and *NF1* |
| MO_1167 | 60 | +/–/– | Tamoxifen, letrozole | 47/3 | *ESR1* (p.Asp538Gly) |
| MO_1185 | 58 | +/+/– | Tamoxifen, letrozole, fulvestrant, exemestane | 88/1 | *ESR1* (p.Tyr537Ser), *CDH1* (p.Gln641*), *NOTCH2* (frameshift deletion), *TP53* copy loss |
| TP_2004[c] | 52 | +/–/– | Tamoxifen (short) | 29/22 | *MDM2* gene amplification, gene copy losses of *CDKN2A* and *CDKN2B* |

PR, progesterone receptor; SNVs, single-nucleotide variants.

[a] Only anti-estrogen–related treatments are listed. Patients also received chemotherapies, radiation or mastectomy in the interim between diagnosis and MI-ONCOSEQ sequencing. [b] Amino acid substitutions caused by nonsynonymous somatic mutations are given in parentheses. [c] TP_2004 is male.

As part of the MI-ONCOSEQ program, we enrolled and sequenced 11 patients with metastatic ER-positive breast cancer (**Table 1** and **Supplementary Table 1**). A diverse array of aberrations were identified in individual patients, some of which are potentially actionable, including mutations in *PIK3CA* ($n$ = 4), *BRCA1* aberrations ($n$ = 2), *FGFR2* aberrations ($n$ = 2)[13], *NOTCH2* frameshift deletion ($n$ = 1), cyclin and associated cyclin-dependent kinase aberrations ($n$ = 3) and *MDM2* amplification and overexpression ($n$ = 1). Aberrations were also frequently found in the tumor suppressor *TP53* ($n$ = 6), the DNA mismatch repair gene *MSH2* ($n$ = 1) and in epigenetic regulators ($n$ = 2), including *ARID2*, *ARID1A* and *SMARCA4*, among others. The complete spectra of somatic mutations with associated alterations in expression levels and copy number in the index patients are given in **Supplementary Figure 1** and **Supplementary Tables 2** and **3**. Two of the index patients, MO_1031 and MO_1051, exhibited a high level of mutations consistent with 'signature B' identified in a whole-genome study of mutational processes in breast cancer[14]. There were 39 gene fusions identified in the 6 index patients, with 11 encoding in-frame fusion proteins (**Supplementary Fig. 2** and **Supplementary Tables 4** and **5**), including an activating *FGFR2-AFF3* fusion[13].

The most notable observation in the mutational landscapes of these treated patients with ER-positive breast cancer was the finding of nonsynonymous mutations in *ESR1* affecting the LBD ($n$ = 6). The six index patients MO_1031, MO_1051, MO_1069, MO_1129, MO_1167 and MO_1185 had mutations encoding p.Leu536Gln, p.Tyr537Ser, p.Asp538Gly, p.Tyr537Ser, p.Asp538Gly and p.Tyr537Ser alterations in the LBD, respectively. The respective mutation in each case was detected by whole-exome sequencing of the tumor relative to the matched normal sample and was corroborated by whole-transcriptome sequencing, as *ESR1* was expressed at moderate to high levels (**Supplementary Table 2**). The clinical histories of the index patients are depicted in timelines in **Figure 1**. For three of the patients (MO_1051, MO_1069 and MO_1129), we had access to primary diagnostic material and showed that the *ESR1* mutations were not present at an earlier stage, indicating that they were acquired after endocrine therapy (**Fig. 1** and **Supplementary Table 2**). Interestingly, all of the index patients were treated with anti-estrogens (tamoxifen and/or fulvestrant) and aromatase inhibitors (letrozole, anastrozole and/or exemestane). Two of the patients also had an oophorectomy. Comparison of the mutations present in each primary versus post-treatment pair showed a substantial number of shared mutations in both samples of the pair, including activating mutations in *PIK3CA* in two of the cases. Thus, it is clear that the index cases presented with recurrent disease of the original primary tumor surviving in an estrogen-deprived state and having acquired *ESR1* mutations. Of note, neither *ESR1* amplifications nor gene fusions were observed in these cases.

The five new LBD alterations of ESR1 identified in this study are depicted in **Figure 2**. Each occurred in the vicinity of the synthetic alterations of ESR1 that are inverted in response to tamoxifen and involve p.Met543Ala and p.Leu544Ala alterations (Inv-mut-AA2)[15] and served as a positive control for our subsequent *in vitro* studies. We next investigated the occurrence of *ESR1* mutations in a range of breast cancer types. Here we took advantage of data from the TCGA Project, which has generated whole-exome sequences for 27 tumor types across at least 4,000 individual samples. As expected, LBD-disrupting mutations of *ESR1* were not detected in the 390 ER-positive breast cancers sequenced by TCGA, as these were primary resection samples before hormonal treatment[16], nor did we detect *ESR1* mutations in a cohort of 80 triple-negative breast carcinoma transcriptomes (D.R.R., Y.-M.W., X.C., S.K.-S., A.M.C. *et al.*, unpublished data).

As the LBD-disrupting mutations of *ESR1* we identified were somatic and were acquired after treatment, we next assessed whether the encoded proteins were dependent on estrogen for activation. We cloned into expression vectors each of the five *ESR1* mutants identified in this study (encoding p.Leu536Gln, p.Tyr537Ser, p.Asp538Gly, p.Tyr537Cys and p.Tyr573Asn alterations) and subsequently cotransfected these constructs into HEK293T cells with an estrogen response element (ERE)-luciferase reporter system. We then exposed steroid hormone–deprived cells to β-estradiol for 24 h and assessed ERE reporter levels. Surprisingly, unlike wild-type ESR1, which had little ERE reporter activity in the absence of ligand, all five of the ESR1 mutants had strong constitutive activation of the ERE reporter that was not markedly enhanced with β-estradiol (**Fig. 3**). This finding suggested that each of the mutations developed in the context of evolution during

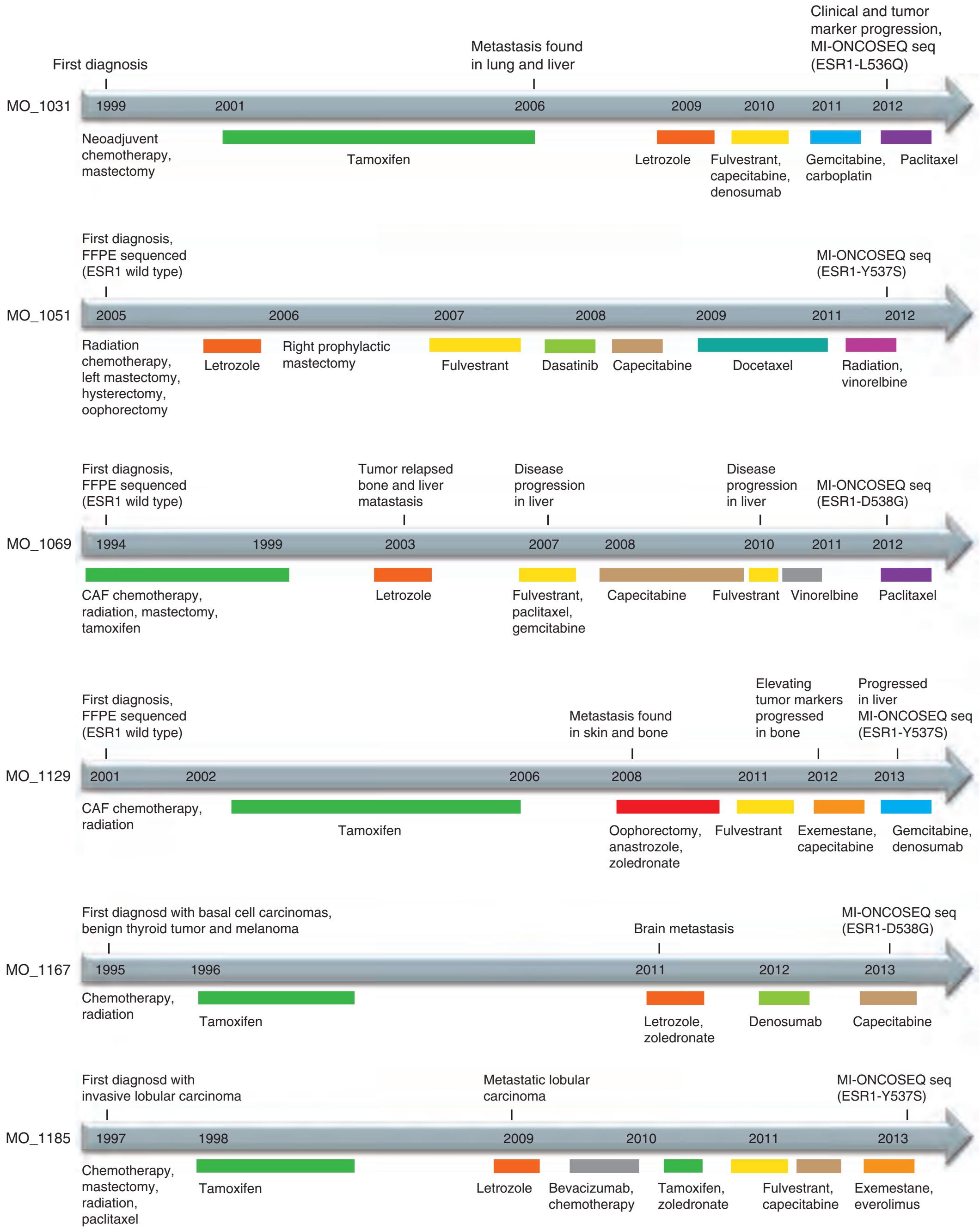


**Figure 1** Clinical timelines for the six index ER-positive metastatic breast cancer patients harboring *ESR1* mutations. Shown are patients' histories of clinical treatment from first diagnosis until enrollment in the MI-ONCOSEQ study. Each bar represents the timeframe of treatment. FFPE, formalin fixed, paraffin embedded; CAF, cyclophosphamide, doxorubicin, fluorouracil chemotherapy.

an estrogen-deprived state. Consistent with this idea, a whole-genome sequencing study of 46 patients with ER-positive breast cancer enrolled in 2 aromatase inhibitor trials did not identify any of these *ESR1* mutations in the pretreatment samples analyzed[17].

Next, we assessed whether anti-estrogen therapies affected the functional activity of the LBD mutants. As effects on inhibition can be influenced by the levels of ectopic ER expression, we performed a dose response study with expression plasmid and selected a dose of 50 ng for the following experiments[18] (**Supplementary Fig. 3**). As expected, wild-type ESR1 was inhibited in a dose-dependent fashion by the anti-estrogens 4-hydroxytamoxifen, fulvestrant and endoxifen (**Fig. 4** and **Supplementary Figs. 4–6**). In addition, the mutant corresponding to the synthetic *ESR1* mutation (Inv-mut-AA2) was activated in a dose-dependent fashion by these anti-estrogens (**Fig. 4**), which has been reported previously[15]. Interestingly, ESR1 with each of the five LBD alterations identified in this study was inhibited by tamoxifen and fulvestrant in a dose-dependent fashion and did not exhibit the inverted response to anti-estrogens that the synthetic Inv-mut-AA2 mutant did. One could speculate that the corresponding mutations did not arise under selective pressure of anti-estrogen treatment but rather in the context of an estrogen deprivation setting, such as treatment with aromatase inhibitors and/or oophorectomy. The $IC_{50}$ (half-maximal

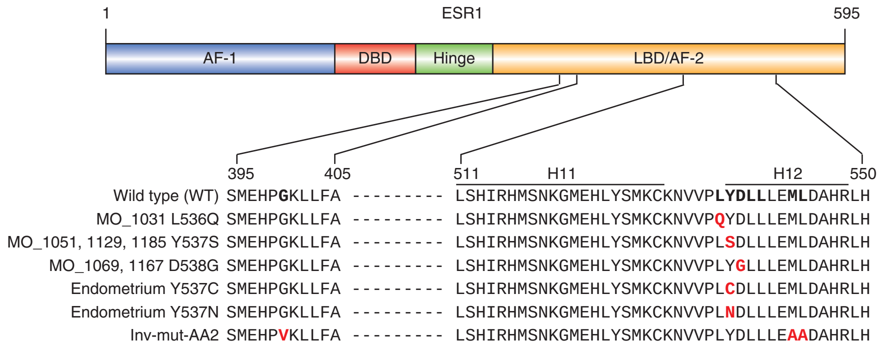

1
ESR1
595
AF-1
DBD
Hinge
LBD/AF-2
395
405
511
H11
H12
550
Wild type (WT) SMEHPGKLLFA --------- LSHIRHMSNKGMEHLYSMKCKNVVPLYDLLLEMLDAHRLH
MO_1031 L536Q SMEHPGKLLFA --------- LSHIRHMSNKGMEHLYSMKCKNVVPQYDLLLEMLDAHRLH
MO_1051, 1129, 1185 Y537S SMEHPGKLLFA --------- LSHIRHMSNKGMEHLYSMKCKNVVPLSDLLLEMLDAHRLH
MO_1069, 1167 D538G SMEHPGKLLFA --------- LSHIRHMSNKGMEHLYSMKCKNVVPLYGLLLEMLDAHRLH
Endometrium Y537C SMEHPGKLLFA --------- LSHIRHMSNKGMEHLYSMKCKNVVPLCDLLLEMLDAHRLH
Endometrium Y537N SMEHPGKLLFA --------- LSHIRHMSNKGMEHLYSMKCKNVVPLNDLLLEMLDAHRLH
Inv-mut-AA2 SMEHPVKLLFA --------- LSHIRHMSNKGMEHLYSMKCKNVVPLYDLLLEAADAHRLH

**Figure 2** Schematic of ESR1 alterations identified in this study. The structural domains of ESR1 are illustrated on top, including the transcription activation function-1 domain (AF-1), the DNA-binding domain (DBD), the hinge domain and the ligand-binding domain (LBD/AF-2). Altered residues identified in mutants are marked in red, and reference residues are shown in bold in the wild-type sequence. Endometrium p.Tyr537Cys and p.Tyr537Asn are two alterations discovered in endometrial cancer samples from the TCGA study. Inv-mut-AA2 represents a ligand activity inversion mutant of ESR1 that confers inverted responses to anti-estrogen and estrogen. H11, helix 11; H12, helix 12.

inhibitory concentration) values for both 4-hydroxytamoxifen and fulvestrant were two- to fourfold higher for all mutants compared to wild-type ESR1. Fulvestrant exhibited greater maximal inhibition than 4-hydroxytamoxifen for all the mutants tested (**Supplementary Figs. 4** and **5**).

The ESR1 alterations identified in this study cluster near the beginning of helix 12 (**Fig. 2**). Structural studies have demonstrated a key role for the position of helix 12 in the response of the ER to agonists and antagonists[19], and Tyr537 has been postulated to form a capping motif contributing to the activity of the receptor[20]. Specifically, the p.Tyr537Ser mutant has been reported to have higher affinity for estrogen than wild-type ESR1 and interacts with the SRC1 coactivator in the absence of ligand[21,22]. Several studies using experimental mutagenesis have implicated the same three residues identified here as critical determinants of the transcriptional activity of the receptor[21,23,24].

As estrogen therapy has been shown to have a positive effect in treating aromatase inhibitor–resistant advanced breast cancers, we tested the effect of low- to high-dose estrogen on the activity of the mutants in the transient luciferase reporter assays (**Supplementary Fig. 7**)[25,26]. The results did not suggest that the effectiveness of this therapy is mediated through direct control of the transcriptional activity of these mutants, if encoded by the responding patients.

Although the primary intent of our broad-based clinical sequencing program is to identify actionable and/or driver mutations in advanced cancers, this study demonstrates how such prospective, real-time sequencing efforts can also shed light on resistance mechanisms that develop against targeted therapies. A number of resistance mechanisms have been suggested to function in the evasion of endocrine treatment, including activation of the mTOR and phosphoinositide 3-kinase (PI3K) pathways, among others[9,27]. Although the total number of ER-positive breast cancers we have sequenced is modest, we have done so in a comprehensive fashion in terms of delineating mutational landscapes and incorporating both DNA and RNA sequencing. This analysis identified *de novo* driver mutations and/or potentially acquired mutations in breast cancer such as mutations resulting in PI3K activation, *PAK1* amplification and *FGFR* fusion and amplification, which have been described previously[13,28,29]. Among potential new mechanisms described, we identified profound focal amplification of *MDM2* (which encodes a negative regulator of p53 that is targetable) and copy gains of *GNRHR* (encoding gonadotropin-releasing hormone receptor), which may be related to past endocrine therapy.

As the LBD-disrupting mutations of *ESR1* identified in this study result in constitutive activity, the encoded mutant proteins can function in the absence of ligand and maintain ER signaling. In 1997, an *ESR1* mutation affecting the LBD, encoding a p.Tyr537Asn alteration, was detected in a single individual with stage IV metastatic breast cancer who had been treated with diethylstibesterol, but, since then, this mutation has been considered to be very rare[30]. With the advent of widespread aromatase inhibitor therapy, we suggest that alteration of the ESR1 LBD is likely a common mechanism of resistance that develops in low-estrogen states. Interestingly, LBD-disrupting mutations of *ESR1* were detected somatically in 4 of 373 cases of endometrial cancer[31]. We speculate that the four TCGA endometrial tumors that harbor LBD-affecting mutations likely came from patients with concurrent breast cancer, as tamoxifen treatment is known to be associated with higher incidence of this tumor type and such patients also often receive estrogen deprivation treatment[32].

Our study suggests that it is unlikely that these LBD alterations develop in the context of anti-estrogen treatment, as the mutated *ESR1* variants continue to be responsive to direct ER antagonists such as tamoxifen and fulvestrant. This finding is consistent with clinical reports showing that patients that develop resistance to aromatase inhibitors still respond to anti-estrogen treatment[11]. Although this prospective clinical sequencing study was not designed to characterize a specific cancer type or treatment resistance mechanism,

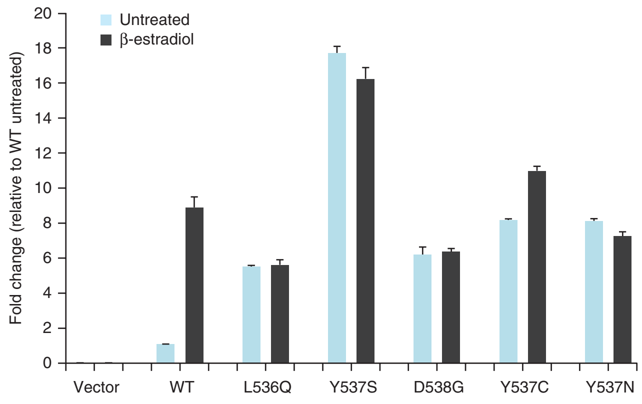


**Figure 3** *ESR1* with acquired mutations encodes constitutively active protein. HEK293T cells were cotransfected with an ERE–firefly luciferase reporter plasmid, a plasmid constitutively expressing *Renilla* luciferase as an internal control and various *ESR1* constructs (as illustrated in **Fig. 2**). Steroid hormone–deprived cells were either untreated or stimulated with 5 nM β-estradiol for 24 h. Firefly luciferase activities were normalized using corresponding *Renilla* luciferase activities for each condition. Fold change in ESR1-driven transcriptional activity was calculated using untreated wild type as a control for each condition. Data shown are the means from triplicate experiments. Amino acid changes in respective ESR1 mutants are indicated. WT, wild-type ESR1. Error bars, s.d.

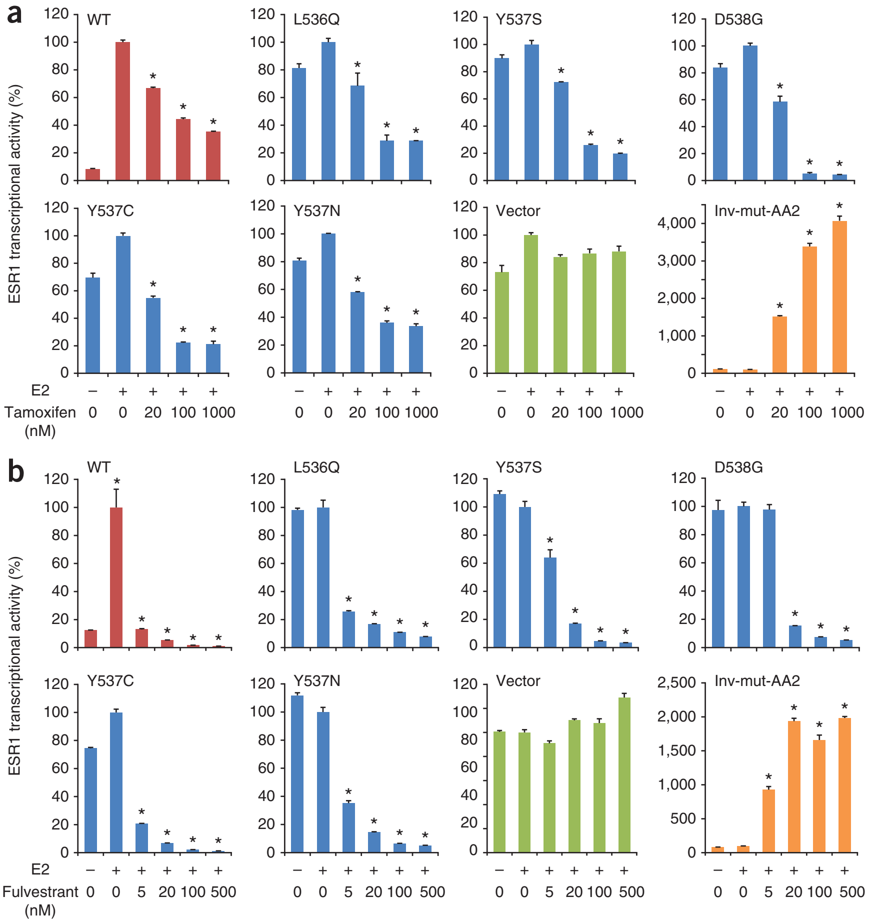


**Figure 4** Acquired *ESR1* mutations result in maintained sensitivity to anti-estrogen therapies. HEK293T cells were cotransfected with an ERE–firefly luciferase reporter plasmid, a plasmid constitutively expressing *Renilla* luciferase and various *ESR1* constructs as indicated. (**a**,**b**) Steroid hormone–deprived cells were either untreated or treated with increasing doses of the anti-estrogen drugs tamoxifen (**a**) and fulvestrant (**b**) in the presence of 5 nM β-estradiol (E2) for 24 h. Percentage change in ESR1-driven transcriptional activity was calculated using E2-treated cells as the control for each tested construct. Data shown are the means from triplicate experiments. Error bars, s.d. $*P < 0.001$. Red, wild type; blue, clinically identified mutants; orange, synthetic ligand-inversion mutant; green, vector control.

future studies comprising larger cohorts of breast cancer patients with disease that recurs after varied endocrine treatments will more precisely delineate the incidence of this acquired resistance mechanism. The focused nature of these mutations and their role in aromatase inhibitor resistance suggest the possibility of monitoring patients undergoing treatment using circulating tumor DNA methods[33,34]. In this manner, treatment could be shifted to head off evolving tumor resistance.

**URLs.** BLAT, http://genome.ucsc.edu/cgi-bin/hgBlat; ORF Finder, http://www.ncbi.nlm.nih.gov/gorf/gorf.html.

## METHODS

Methods and any associated references are available in the online version of the paper.

**Accession codes.** Sequence data have been deposited in the database of Genotypes and Phenotypes (dbGaP), which is hosted by the National Center for Biotechnology Information (NCBI), under accessions phs000602.v1.p1 and phs000673.v1.p1.

*Note: Any Supplementary Information and Source Data files are available in the online version of the paper.*

ACKNOWLEDGMENTS

The authors thank D. Miller, T. Barrette and D. Gibbs for hardware and database management, K. Giles for assistance with manuscript preparation, physicians M. Wicha, L. Pierce, D. Smith, K. Levin and F. Feng for referring patients, and C. Betts and J. Athanikar for assistance with tumor boards. We also thank the larger MI-ONCOSEQ team, including clinical research coordinator E. Williams, pathologist R. Mehra, genetic counselors J. Everett, S. Gustafson and V. Raymond, and radiologists E. Higgins, E. Caoili and R. Dunnick. This project is supported in part by the Prostate Cancer Foundation with funding for our sequencing infrastructure, the National Cancer Institute Early Detection Research Network (U01 CA111275), National Human Genome Research Institute Clinical Sequencing Exploratory Research (CSER) Consortium (1UM1HG006508), US Department of Defense contract W81XWH-12-1-0080 and a Department of Defense Era of Hope Scholar Award. A.M.C. is also supported by the Alfred A. Taubman Institute, the American Cancer Society, the Howard Hughes Medical Institute and a Doris Duke Charitable Foundation Clinical Scientist Award.

### AUTHOR CONTRIBUTIONS

D.R.R., Y.-M.W. and A.M.C. conceived the experiments. D.R.R., Y.-M.W., X.C., R.W., F.S. and Y.N. performed exome and transcriptome sequencing. P.V., R.J.L., S.K.-S. and D.R.R. carried out bioinformatics analysis of high-throughput sequencing data for somatic mutation, copy number and tumor content determination and performed gene expression and gene fusion analysis. D.R.R., Y.-M.W. and F.S. generated *ESR1* constructs and carried out *in vitro* experiments. L.H. coordinated patients for clinical research. J.S. and A.G. collected and processed clinical tissue samples for next-generation sequencing. L.P.K. and S.A.T. provided pathology review. J.M.R. provided experimental analysis. C.H.V.P., D.F.H., R.C. and A.F.S. enrolled patients and provided clinical data and consultation at tumor boards. D.R.R., X.C., Y.-M.W., P.V., R.J.L., S.K.-S., S.Y.K., J.S.R., S.R., M.T., K.J.P. and A.M.C. developed the integrated clinical sequencing protocol. D.R.R., Y.-M.W. and A.M.C. prepared the manuscript, which was reviewed by all authors.

### COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Chin, L., Andersen, J.N. & Futreal, P.A. Cancer genomics: from discovery science to personalized medicine. *Nat. Med.* **17**, 297–303 (2011).
2. Meyerson, M., Gabriel, S. & Getz, G. Advances in understanding cancer genomes through second-generation sequencing. *Nat. Rev. Genet.* **11**, 685–696 (2010).
3. Roychowdhury, S. *et al.* Personalized oncology through integrative high-throughput sequencing: a pilot study. *Sci. Transl. Med.* **3**, 111ra121 (2011).
4. Welch, J.S. *et al.* Use of whole-genome sequencing to diagnose a cryptic fusion oncogene. *J. Am. Med. Assoc.* **305**, 1577–1584 (2011).
5. Gorre, M.E. *et al.* Clinical resistance to STI-571 cancer therapy caused by *BCR-ABL* gene mutation or amplification. *Science* **293**, 876–880 (2001).
6. Korpal, M. *et al.* An F876L mutation in androgen receptor confers genetic and phenotypic resistance to MDV3100 (enzalutamide). *Cancer Discov.* **3**, 1030–1043 (2013).
7. Joseph, J.D. *et al.* A clinically relevant androgen receptor mutation confers resistance to 2nd generation anti-androgens enzalutamide and ARN-509. *Cancer Discov.* **3**, 1020–1029 (2013).
8. Ariazi, E.A., Ariazi, J.L., Cordera, F. & Jordan, V.C. Estrogen receptors as therapeutic targets in breast cancer. *Curr. Top. Med. Chem.* **6**, 181–202 (2006).
9. Riggins, R.B., Schrecengost, R.S., Guerrero, M.S. & Bouton, A.H. Pathways to tamoxifen resistance. *Cancer Lett.* **256**, 1–24 (2007).

10. Lønning, P.E. & Eikesdal, H.P. Aromatase inhibition 2013: clinical state of the art and questions that remain to be solved. *Endocr. Relat. Cancer* **20**, R183–R201 (2013).
11. Ingle, J.N. *et al.* Fulvestrant in women with advanced breast cancer after progression on prior aromatase inhibitor therapy: North Central Cancer Treatment Group Trial N0032. *J. Clin. Oncol.* **24**, 1052–1056 (2006).
12. Osborne, C.K. & Schiff, R. Mechanisms of endocrine resistance in breast cancer. *Annu. Rev. Med.* **62**, 233–247 (2011).
13. Wu, Y.M. *et al.* Identification of targetable *FGFR* gene fusions in diverse cancers. *Cancer Discov.* **3**, 636–647 (2013).
14. Nik-Zainal, S. *et al.* Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979–993 (2012).
15. Feil, R., Wagner, J., Metzger, D. & Chambon, P. Regulation of Cre recombinase activity by mutated estrogen receptor ligand-binding domains. *Biochem. Biophys. Res. Commun.* **237**, 752–757 (1997).
16. TCGA. Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70 (2012).
17. Ellis, M.J. *et al.* Whole-genome analysis informs breast cancer response to aromatase inhibition. *Nature* **486**, 353–360 (2012).
18. Huang, H.J., Norris, J.D. & McDonnell, D.P. Identification of a negative regulatory surface within estrogen receptor α provides evidence in support of a role for corepressors in regulating cellular responses to agonists and antagonists. *Mol. Endocrinol.* **16**, 1778–1792 (2002).
19. Shiau, A.K. *et al.* The structural basis of estrogen receptor/coactivator recognition and the antagonism of this interaction by tamoxifen. *Cell* **95**, 927–937 (1998).
20. Skafar, D.F. Formation of a powerful capping motif corresponding to start of "helix 12" in agonist-bound estrogen receptor-α contributes to increased constitutive activity of the protein. *Cell Biochem. Biophys.* **33**, 53–62 (2000).
21. Carlson, K.E., Choi, I., Gee, A., Katzenellenbogen, B.S. & Katzenellenbogen, J.A. Altered ligand binding properties and enhanced stability of a constitutively active estrogen receptor: evidence that an open pocket conformation is required for ligand interaction. *Biochemistry* **36**, 14897–14905 (1997).
22. Weis, K.E., Ekena, K., Thomas, J.A., Lazennec, G. & Katzenellenbogen, B.S. Constitutively active human estrogen receptors containing amino acid substitutions for tyrosine 537 in the receptor protein. *Mol. Endocrinol.* **10**, 1388–1398 (1996).
23. Pearce, S.T., Liu, H. & Jordan, V.C. Modulation of estrogen receptor α function and stability by tamoxifen and a critical amino acid (Asp-538) in helix 12. *J. Biol. Chem.* **278**, 7630–7638 (2003).
24. Zhao, C. *et al.* Mutation of Leu-536 in human estrogen receptor-α alters the coupling between ligand binding, transcription activation, and receptor conformation. *J. Biol. Chem.* **278**, 27278–27286 (2003).
25. Ellis, M.J. *et al.* Lower-dose vs high-dose oral estradiol therapy of hormone receptor–positive, aromatase inhibitor–resistant advanced breast cancer: a phase 2 randomized study. *J. Am. Med. Assoc.* **302**, 774–780 (2009).
26. Swaby, R.F. & Jordan, V.C. Low-dose estrogen therapy to reverse acquired antihormonal resistance in the treatment of breast cancer. *Clin. Breast Cancer* **8**, 124–133 (2008).
27. Sokolosky, M.L. *et al.* Involvement of Akt-1 and mTOR in sensitivity of breast cancer to targeted therapy. *Oncotarget* **2**, 538–550 (2011).
28. Kan, Z. *et al.* Diverse somatic mutation patterns and pathway alterations in human cancers. *Nature* **466**, 869–873 (2010).
29. Shrestha, Y. *et al. PAK1* is a breast cancer oncogene that coordinately activates MAPK and MET signaling. *Oncogene* **31**, 3397–3408 (2012).
30. Barone, I., Brusco, L. & Fuqua, S.A. Estrogen receptor mutations and changes in downstream gene expression and signaling. *Clin. Cancer Res.* **16**, 2702–2708 (2010).
31. Kandoth, C. *et al.* Integrated genomic characterization of endometrial carcinoma. *Nature* **497**, 67–73 (2013).
32. Fisher, B. *et al.* Endometrial cancer in tamoxifen-treated breast cancer patients: findings from the National Surgical Adjuvant Breast and Bowel Project (NSABP) B-14. *J. Natl. Cancer Inst.* **86**, 527–537 (1994).
33. Dawson, S.J. *et al.* Analysis of circulating tumor DNA to monitor metastatic breast cancer. *N. Engl. J. Med.* **368**, 1199–1209 (2013).
34. Diehl, F. *et al.* Circulating mutant DNA to assess tumor dynamics. *Nat. Med.* **14**, 985–990 (2008).

## ONLINE METHODS

**Clinical study and specimen collection.** Sequencing of clinical samples was performed under IRB-approved studies at the University of Michigan. Patients were enrolled and consented for integrative tumor sequencing in MI-ONCOSEQ (Michigan Oncology Sequencing Protocol, HUM00046018). Medically qualified patients 18 years or older with advanced or refractory cancer were eligible for the study. Informed consent detailed the risks of integrative sequencing and includes up-front genetic counseling. Informed consent was obtained from all subjects included in this study. Biopsies were arranged for safely accessible tumor sites. Needle biopsies were snap frozen in OCT (Optimal Cutting Temperature) compound, and a longitudinal section was cut. Frozen sections stained with hematoxylin and eosin were reviewed by pathologists to identify cores with the highest tumor content. Remaining portions of each needle biopsy core were retained for nucleic acid extraction.

**Extraction of DNA and RNA.** Genomic DNA from frozen needle biopsies and blood was isolated using the Qiagen DNeasy Blood and Tissue kit, according to the manufacturer's instructions. Total RNA was extracted from frozen needle biopsies using the Qiazol reagent with disruption using a 5-mm bead on a Tissuelyser II (Qiagen) and was purified using a miRNeasy kit (Qiagen) with DNase I digestion, according to the manufacturer's instructions. RNA integrity was verified on an Agilent 2100 Bioanalyzer using RNA Nano reagents (Agilent Technologies).

**Preparation of next-generation sequencing libraries.** Transcriptome libraries were prepared using 1–2 μg of total RNA. Polyadenylated RNA was isolated using Sera-Mag oligo(dT) beads (ThermoScientific) and fragmented with the Ambion Fragmentation Reagents kit. cDNA synthesis, end repair, A-base addition and ligation of the Illumina indexed adaptors were performed according to Illumina's TruSeq RNA protocol. Libraries were selected for DNA fragments of 250–300 bp in size on a 3% Nusieve 3:1 agarose gel (Lonza), recovered using QIAEX II gel-extraction reagents (Qiagen) and PCR amplified using Phusion DNA polymerase (New England BioLabs). Amplified libraries were purified using AMPure XP beads (Beckman Coulter). Library quality was measured on an Agilent 2100 Bioanalyzer by product size and concentration. Paired-end libraries were sequenced with the Illumina HiSeq 2000 platform (2 × 100-nucleotide read length). Reads that passed the chastity filter of Illumina BaseCall software were used for subsequent analysis.

Exome libraries of matched pairs of tumor and normal genomic DNA were generated using the Illumina TruSeq DNA Sample Prep kit, following the manufacturer's instructions. In brief, 1–3 μg of each genomic DNA sample was sheared using a Covaris S2 to a peak target size of 250 bp. Fragmented DNA was concentrated using AMPure XP beads, and end repair, A-base addition and ligation of Illumina indexed adaptors were performed. Adaptor-ligated libraries were electrophoresed on 3% Nusieve agarose gels, and fragments of 300–350 bp were recovered using QIAEX II gel-extraction reagents. Recovered DNA was amplified using Illumina index primers for eight cycles and purified using AMPure XP beads, and DNA concentration was determined using a Nanodrop spectrophotometer. Libraries (1 μg) were hybridized to the Agilent SureSelect Human All Exon v4 chip at 65 °C for 60 h, following the manufacturer's protocol (Agilent Technologies). Targeted exon fragments were captured on Dynal M-280 streptavidin beads (Invitrogen) and enriched by amplification with the Illumina index primers for nine additional PCR cycles. PCR products were purified with AMPure XP beads and analyzed for quality and quantity using an Agilent 2100 Bioanalyzer and DNA 1000 reagents.

We used the publicly available software FastQC to assess sequencing quality. For each lane, we examined per-base quality scores across the length of the reads. Lanes were deemed passing if the per-base quality score box plot indicated that >85% of the reads had >Q20 for bases 1–100. In addition to raw sequence quality, we also assessed alignment quality using the Picard package. This allows monitoring of duplication rates and chimeric reads that may result from ligation artifacts, crucial statistics for interpreting the results of copy number and structural variant analysis.

**Gene fusion detection.** Paired-end transcriptome sequencing reads were aligned to the human reference genome (GRCh37/hg19) using an RNA sequencing (RNA-seq) spliced read mapper Tophat2 (ref. 35) (Tophat 2.0.4) with the '–fusion-search' option turned on to detect potential gene fusion transcripts. In the initial process, Tophat2 internally deploys an ultrafast short-read alignment tool, Bowtie (Version 0.12.8), to map the transcriptome data. Potential false-positive fusion candidates were filtered out using the 'Tophat-Post-Fusion' module. Further, fusion candidates were manually examined for annotation and ligation artifacts. Junction reads supporting the fusion candidates were realigned using the BLAT alignment tool to confirm fusion breakpoints. Full-length sequence of each fusion gene was constructed on the basis of supporting junction reads and evaluated for potential ORFs using an ORF Finder. For gene fusions with robust ORFs, the amino acid sequences of the fused proteins were explored using the Simple Modular Architecture Research Tool (SMART) to examine the gain or loss of known functional domains in the fusion proteins.

**Gene expression.** BAM 'accepted_hits.bam' files, which were generated by the Tophat mapping module, were used to quantify the expression data through Cufflinks[36] (Version 2.0.2), an isoform assembly and RNA-seq quantification package. The structural features of 56,369 transcripts from the Ensembl resource (Ensembl 66) were used as an annotation reference to quantify the expression of individual transcripts and isoforms. The 'Max Bundle Length' parameter was set to '10000000', and 'multi-read-correct' was flagged on to perform an initial estimation procedure to more accurately weight reads mapping to multiple locations in the genome.

**Mutation analysis.** Whole-exome sequencing was performed on an Illumina HiSeq 2000 or HiSeq 2500 instrument in paired-end mode, and primary base call files were converted into FASTQ sequence files using the bcl2fastq converter tool bcl2fastq-1.8.4 in the CASAVA 1.8 pipeline. FASTQ sequence files were then processed through an in-house pipeline constructed for whole-exome sequence analyses of paired cancer and normal genomes. Sequencing reads were aligned to reference genome build hg19 (GRCh37) using Novoalign multithreaded (Version 2.08.02, Novocraft) and converted into BAM files using SAMtools (Version 0.1.18)[37]. Sorting and indexing of BAM files used Novosort threaded (Version 1.00.01), and duplicate reads were removed using Picard (Version 1.74). Mutation analysis was performed using VarScan2 algorithms (Version 2.3.2)[38] with the pileup files created by SAMtools mpileup for tumor and matched normal samples, simultaneously performing pairwise comparisons of base call and normalized sequence depth at each position. For SNV detection, filtering parameters including coverage, variant read support, variant frequency, *P* value, base quality, the presence of homopolymers and strandedness were applied. For indel analysis, Pindel (Version 0.2.4) was used on tumor and matched normal samples, and indels common to both samples were classified as germline, whereas indels present in tumor but not in normal samples were classified as somatic. Finally, a list of candidate indels as well as of somatic and/or germline mutations was generated by excluding synonymous SNVs. ANNOVAR[39] was used to functionally annotate the detected genetic variants, and positions are based on Ensembl 66 transcript sequences.

Tumor content for each tumor exome library was estimated from the sequence data by fitting a binomial mixture model with two components to the set of most likely SNV candidates from two-copy genomic regions. The set of candidates used for estimation consisted of coding variants that (i) were supported by at least 3 variant fragments in the tumor sample, (ii) were not supported by variant fragments in the matched benign sample, with at least 16 fragments of coverage, (iii) were not present in dbSNP, (iv) were within a targeted exon or within 100 bp of a targeted exon, (v) were not in homopolymer runs of 4 or more bases and (vi) exhibited no evidence of amplification or deletion. To filter out regions of possible amplification or deletion, we used exon coverage ratios to infer copy number changes, as described below. Resulting SNV candidates were not used for the estimation of tumor content if the segmented log ratio exceeded 0.2 in absolute value. Candidates on the Y chromosome were also eliminated because they were unlikely to exist in two-copy genomic regions. Using this set of candidates, we fit a binomial mixture model with two components using the R package flexmix, version 2.3-8. One component consisted of SNV candidates with very low variant fractions, presumably resulting from recurrent sequencing errors and other artifacts. The other component, consisting of the set of likely true SNVs, was informative of tumor content in the tumor sample. Specifically,

doi:10.1038/ng.2823

under the assumption that most or all of the observed SNV candidates in this component are heterozygous SNVs, we expect the estimated binomial proportion of this component to represent one-half of the proportion of tumor cells in the sample. Thus, the estimated binomial proportion obtained from the mixture model was doubled to obtain an estimate of tumor content.

Copy number aberrations were quantified and reported for each gene as the segmented, normalized, $\log_2$-transformed exon coverage ratio between each tumor sample and its matched normal sample[40]. To account for observed associations between coverage ratios and variation in GC content across the genome, lowess normalization was used to correct per-exon coverage ratios before segmentation analysis. Specifically, mean GC percentage was computed for each targeted region, and a lowess curve was fit to the scatterplot of $\log_2$ coverage ratios versus mean GC content across the targeted exome using the lowess function in R (version 2.13.1) with smoothing parameter $f = 0.05$.

Partially redundant sequencing of areas of the genome affords the ability for cross-validation of findings. We cross-validated exome-based point mutation calls by manually examining the genomic and transcriptomic reads covering the mutation using the UCSC Genome Browser. Likewise, gene fusion calls from the transcriptome data can be further supported by structural variant detection in the genomic sequence data, as well as by copy number information derived from genome and exome sequencing.

**Chemicals and reagents.** β-estradiol, (Z)-4-hydroxytamoxifen, (E/Z)-endoxifen hydrochloride hydrate and fulvestrant were purchased from Sigma-Aldrich.

**Plasmids and cloning.** cDNA for wild-type *ESR1* was PCR amplified from a breast cell line MCF7 (ATCC) with the introduction of a sequence encoding an N-terminal Flag tag. cDNAs encoding the relevant mutations of *ESR1* were generated by site-directed mutagenesis (QuikChange, Agilent Technologies), and full-length constructs were fully sequenced. All *ESR1* variants were placed in the lentiviral vector pCDH (System Biosciences) for eukaryotic expression.

**ERE-luciferase reporter assays.** For cell transfection experiments, HEK293T cells (ATCC) were plated at a density of 1–2 × $10^5$ cells per well (24-well plates) in phenol red–free DMEM containing 10% FBS and antibiotics. Once cells attached, the medium was replaced with DMEM containing 10% charcoal/dextran-treated FBS (HyClone), and cells were cultured overnight. The next day, cells were transiently cotransfected with *ESR1* expression plasmid (50 ng/well) and luciferase reporter constructs (25 ng/well; SABiosciences) using FuGene 6 reagent (Promega). The ER-responsive luciferase plasmid encoding the firefly luciferase reporter gene is driven by a minimal CMV promoter and tandem repeats of the estrogen transcriptional response element (ERE). A second plasmid constitutively expressing *Renilla* luciferase served as an internal control for normalizing transfection efficiencies (Cignal ERE Reporter, SABiosciences). After transfection for 18 h, cells were serum starved for a few hours before treatment with β-estradiol or anti-estrogen drugs. Cells were harvested 18 h after treatment, and luciferase activity was measured using the Dual-Luciferase Reporter Assay System (Promega). $IC_{50}$ values were computed using the GraphPad Prism application to fit a four-parameter dose response curve.

35. Kim, D. & Salzberg, S.L. TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. *Genome Biol.* **12**, R72 (2011).
36. Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **7**, 562–578 (2012).
37. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
38. Koboldt, D.C. *et al.* VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* **22**, 568–576 (2012).
39. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
40. Lonigro, R.J. *et al.* Detection of somatic copy number alterations in cancer using targeted exome capture sequencing. *Neoplasia* **13**, 1019–1025 (2011).