AD_____
(Leave blank)

Award Number: **W81XWH-08-1-0110**

(Enter Army Award number assigned to research, i.e., DAMD17-00-1-0296)

TITLE: A Search for Gene Fusions/Translocations in Breast Cancer
(Enter title of award)

PRINCIPAL INVESTIGATOR: **Arul M. Chinnaiyan, M.D., Ph.D.**
(Enter the name and degree of Principal Investigator and any Associates)

CONTRACTING ORGANIZATION: **Regents of the University of Michigan, Ann Arbor, Michigan 48109-1274**
(Enter the Name, City, State and Zip Code of the Contracting Organization)

REPORT DATE: **October 2009**
(Enter month and year, i.e., January 2001)

TYPE OF REPORT: **Annual**
(Enter type of report, i.e., annual, midterm, annual summary, final)

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

DISTRIBUTION STATEMENT: (Check one)

x Approved for public release; distribution unlimited

☐ Distribution limited to U.S. Government agencies only; report contains proprietary information

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
08/01/2009	Annual	09/01/2008 – 08/31/2009

4. TITLE AND SUBTITLE
A Search for Gene Fusions/Translocations in Breast Cancer

5a. CONTRACT NUMBER
W81XWH-08-1-0110

5b. GRANT NUMBER
BC075023

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Arul M. Chinnaiyan, M.D., Ph.D.

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

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

Regents of the 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 Materiel 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

Enter a brief (approximately 200 words) unclassified summary of the most significant finding during the research period). We have undertaken a systematic evaluation of breast cancer to map disease-specific, recurrent chromosomal or transcriptional chimeras in breast cancer towards development of novel biomarkers and therapeutic targets. Analysis of in-house and publicly available gene expression and array comparative genomic hybridization (aCGH) data lead to the discovery that a subset of estrogen receptor positive breast cancers overexpress angiotensin II receptor, type 1 (AGTR1). In experimental model systems- both in vitro as well as in xenografts in mice, AGTR1 overexpressing breast cancers are sensitive to losartan, an AGTR1 antagonist that is used to treat high blood pressure. These studies published recently in PNAS have generated much interest in the community and will be followed up by us and others. Towards systematic gene fusion discovery, we have developed paired end transcriptome sequencing protocols and bioinformatic pipeline to nominate gene fusion candidates, also published recently in PNAS. Promising gene fusion candidates from breast cancer cell lines and tissues will be followed up in recurrence screens and functional characterization. Another major advance this year has been the discovery of the role of micro RNA 101 in regulating the expression of histone methyltransferase EZH2 in aggressive breast and prostate cancer, published in Science.

15. SUBJECT TERMS
Angiotensin II receptor (AGTR1), biomarkers, therapeutic targets,

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			19b. TELEPHONE NUMBER *(include area code)*
U	U	U	UU	68	USAMRMC

Table of Contents

DOD Era of Hope Annual Report
A Search for Gene Fusions/Translocations in Breast Cancer

INTRODUCTION: Our laboratory reported the unexpected discovery of recurrent gene fusions in prostate cancer in October 2005(1) and since then we, and researchers around the world, have discovered and clinically characterized several recurrent gene fusions in prostate(2-5) and lung cancers(6, 7), strongly supporting the notion that gene fusions are prevalent in common solid cancers (and are not restricted to hematological malignancies, as was previously thought(8, 9)). Considering that the characterization of gene fusions potentially provides novel diagnostic and therapeutic markers, as exemplified by the successful application of BCR-ABL1 gene fusion in the diagnosis and therapy of chronic myeloid leukemia(10, 11), we embarked on a hunt for recurrent gene fusions in breast cancer, the most prevalent cancer of women in the United States and other developed countries. The recent technical breakthroughs in high throughput sequencing technologies now provide unprecedented depth and resolution of the DNA/ RNA aberrations in cancer cells, and we have successfully adopted these techniques in our search for gene fusions in common solid cancers(12).

In our ongoing project entitled **"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. We began with the analysis of in-house and publicly available gene expression and array comparative genomic hybridization (aCGH) data using our microarray data compendium, Oncomine that lead us to the discovery of a subset of breast cancers that overexpress angiotensin II receptor, type 1 (AGTR1) and are thus sensitive to losartan, an AGTR1 antagonist that is used to treat high blood pressure (13). In a more direct approach towards gene fusion discovery, we adopted next generation sequencing technologies to nominate gene fusion candidates by paired end transcriptome sequencing followed by fusion specific quantitative real time PCR validation(14). We have identified several promising gene fusion candidates from breast cancer cell lines and tissues that will be followed up in recurrence screens and functional characterization. Another major advance this year has been the discovery of the role of micro RNA 101 in regulating the expression of histone methyltransferase EZH2 in aggressive breast and prostate cancer(15).

A detailed, itemized report of the progress in work follows:

STATEMENT OF WORK

Task 1: Characterization of recurrent gene fusions in breast cancer

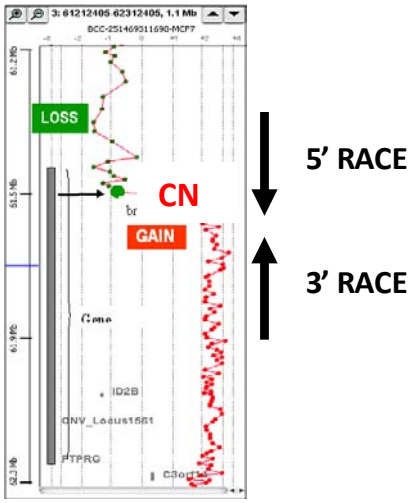
A. Integrative analysis of MCF7 cells to nominate gene fusions in breast cancer (Years 1-2)

-use of break point prediction based on array CGH data

- array CGH and gene expression analysis of at least 70 matched samples

B. RACE analysis and fusion PCR of candidates (Years 1-3)

Based on high resolution oligonucleotide based aCGH profiles of cancer genomic DNA, we have identified whole chromosome gains, losses, and many regions of gains and losses at sub-microscopic level in the size range of < 30kb. The boundaries of amplifications and deletions, defined as copy number transition (CNT) loci, that map to known intergenic regions (introns or exons) are nominated as candidate gene fusion partners (***Figure 1***). Further analysis of CNT loci by spectral karyotyping (SKY), fluorescence in situ hybridization (FISH) and rapid amplification of cDNA ends (RACE)-PCR, will be carried out to identify novel gene fusions in the proof-of- principle analysis on breast cancer cell line MCF7**.** This study is the first of its kind to nominate gene fusions through aCGH data analysis and is likely to find widespread application in the hunt for gene fusions in common solid cancers.



Strategy to isolate fusion gene from a Copy Number Transition (CNT region)

Identify CNT region within a gene → Confirm genomic rearrangement by FISH → Identify genomic interval of the CNT region → Design primer from the region present in at least one copy, and exons close to the CNT region → Decide on 5' or 3' RACE depending on the orientation of the gene → Clone PCR product and sequence → Confirm RACE- PCR results by fusion specific RT-PCR.

Figure 1: *Identification of gene fusion from a region of copy number transition.*

Identification of gene fusions in the commonly amplified regions in breast cancer

Characterization of amplifications in Breast Cancer

Chromosomal regions 17q23 (including *RPS6KB1, MUL, BCAS3, APPBP2,* and *TRAP240* genes) and 20q13 (including *EYA2, PRKCBP1, NCOA3, SULF2, PREX1, and ARFGEF2, AIB1, ZNF217, BCAS4, BTAK,* and *NABC1* genes) are frequently amplified in breast cancers(16-18). All the genes present in an amplicon do not display uniformly high expression, suggesting additional rearrangements. Earlier, through analysis of BAC clones, recurrent amplicons have been proposed as hot spots of genomic rearrangements(19), and now our study provides an independent and much higher resolution tool for a genome-wide analysis of such rearrangements.

To assess the genomic organization of the amplified regions in MCF7, we performed FISH analysis using a BAC clone for BRIP1 (RP11-482H10) gene within the amplified region at 17q23. FISH results indicated that the amplified sequences are inserted at many locations within the genome **(*Figure 2*)** confirming the added complexity of the rearrangements. The uneven distributions of signal intensity of the amplified signals at different locations indicate further rearrangements. Such cryptic rearrangements are not detectable even with high-resolution array CGH.

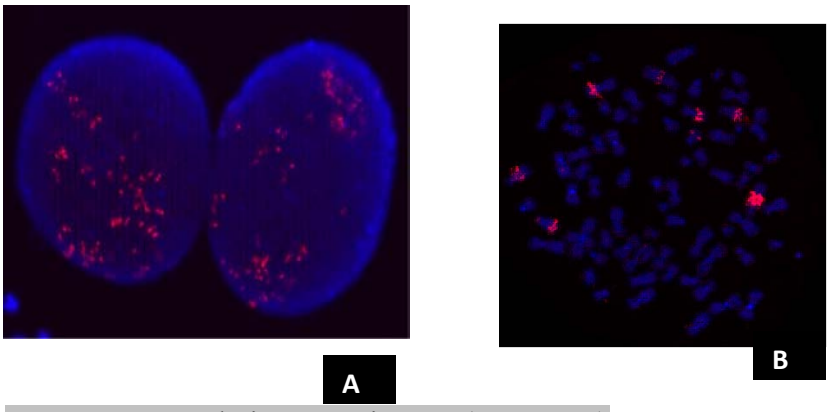


Figure 2: *FISH analysis of an amplified region on 17q23 showing insertion (red) of the amplified sequences in multiple locations in MCF7 genome. A. Interphase nuclei, B. Metaphase chromosomes.*

F. Estrogen regulation experiments (Years 1-3)

We have carried out a time course treatment with estrogen on three cancer cell lines (MCF7, T47D and BT474) and subjected them to ChIP-sequencing by next generation sequencing, to elucidate the genomic scale landscape of estrogen regulated genes. Based on the preliminary analyses, a large number of genes, both known and novel, were found to contain ER binding peaks in their upstream promoter regions, while some shared across the cell lines and others were often specific to cell types. Overall, we are geared to integrate this estrogen regulation data with our gene expression profiling results, and will use this information to annotate our gene fusion candidates as potentially estrogen regulated.

Task 2: Next generation sequencing analysis by Solexa

A. Whole transcriptome sequence analysis of 20 breast cancers (Years 1-2)

B. Whole genome paired-end sequence analysis of 20 breast cancers (Years 1-2)

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+ of these cell lines were cultured, and DNA, RNA and protein extracted from them. Breast cancer tissue samples, representing all of the 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.

Sequencing: RNA isolated from all experimental samples was assessed for quality and integrity through Bioanalyzer (Agilent) (RNA Integrity Number≥8) and 2 to 10 µg total RNA was used to prepare transcriptome sequencing libraries. Briefly, total RNA was passed over oligo-dT bearing magnetic beads to purify mRNA, which was then 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 libraries with a clean, single peak (representing approximately 300bp), which was applied on the flowcells for cluster generation (**Appendix 2**). Typically, we sequenced one sample over one lane of the flowcell; one sequencing slide bore eight lanes, which permitted the run of seven samples and a control phiX DNA library simultaneously. A typical paired end run takes five days to complete, followed by a two days for downloading of sequencing data from the instrument hardware, processing, filtering for quality, and mapping to the genome for sequence analysis. 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.***

Presently, we are carrying out whole transcriptome sequencing of **a panel of breast cancer cell lines** (including normal), **breast cancer tissues**, and **normal breast tissues**.

Sequence Analysis: Primary sequence analysis is focused on identifying novel gene fusions in each sample analyzed. In a proof of concept study by our group published recently in PNAS, we have reported successful implementation of a bioinformatic pipeline developed in-house to nominate gene fusions from paired end transcriptome sequence data(14).

In this study, we rediscovered the 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 3***).

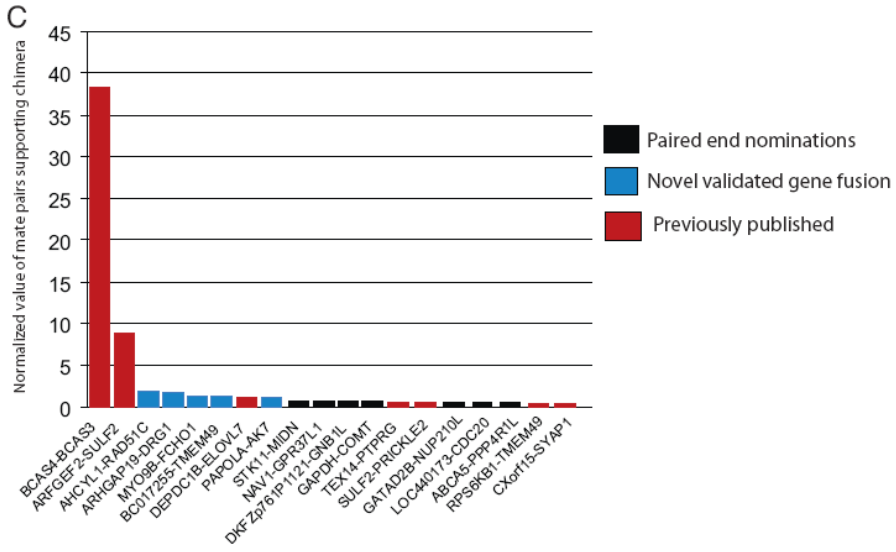


Figure 3. *Discovery of gene fusions in MCF7 by Paired End Transcriptome Sequencing*

A detailed description of the experimental and analytical methods is available in the enclosed **Appendix**.

Sequence analysis of breast cancer cell lines and tissues is underway according to our published protocols and candidate gene fusions are being nominated and examined.

Ongoing investigations are focused on screening large sample cohorts to identify recurrent gene fusions, as well as on the functional characterization of gene fusions in samples that harbor them. Considering that breast cancer cell lines provide useful

surrogates for clinical samples(20) we are sequencing a panel of cell lines representing the clinicopathological gamut of breast cancer that would serve as ready *in vitro* models of gene fusion biology.

Task 3. High-throughput FISH scanning for gene fusions

A. FISH split probe analysis on 50 top COPA candidates (Years 1-2)

B. FISH analysis on 30 ETS family members (Years 1-2)

We are carrying out fluorescence *in situ* hybridization (FISH) to perform split-signal analysis of the complete list of Ets family genes (total number 27) on tissue microarrays of approximately 100 breast cancer tissues corresponding to all major clinic pathological stages of breast cancer, analogous to our efforts in prostate cancer which led to the identification of several novel gene fusions(3).

C. FISH analysis on Mitelman cohort of 3' fusion partners (Years 2-4)

In addition to screening for ets gene aberrations in breast cancer, we are also performing fluorescence *in situ* hybridization (FISH) based split-signal analysis on the complete list of genes enumerated in the Mitelman Database of Chromosome Aberratins in Cancer (http://cgap.nci.nih.gov/Chromosomes/Mitelman) on tissue microarrays of approximately 100 breast cancer samples, encompassing the major clinic-pathological stages of breast cancer.

Task 4. AGTR as a COPA candidate in breast cancer.

In order to identify genes that display outlier expression in breast cancers, and therefore serve as potential gene fusion candidates, we employed our gene expression data compendium Oncomine 3.0 (www/oncomine.org)(21, 22) to perform Cancer Outlier Profile Analysis (COPA) as previously used for the discovery of gene fusions in prostate cancer(1, 23). Briefly, 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(24).

A. Integrative analysis with gene expression (Year 1)

Meta-Copa analysis of breast cancer datasets on 31 breast cancer profiling studies comprising 3,157 microarray experiments lead 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 (**Appendix**)(13). Potential genomic rearrangement of AGTR1 locus was investigated as a likely reason for overexpression.

B. FISH analysis of AGTR on tissue microarrays (Year 1)

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 to be associated with AGTR1 overexpression in 7 of 112 cases (6.25%) (***Figure 4***). 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.

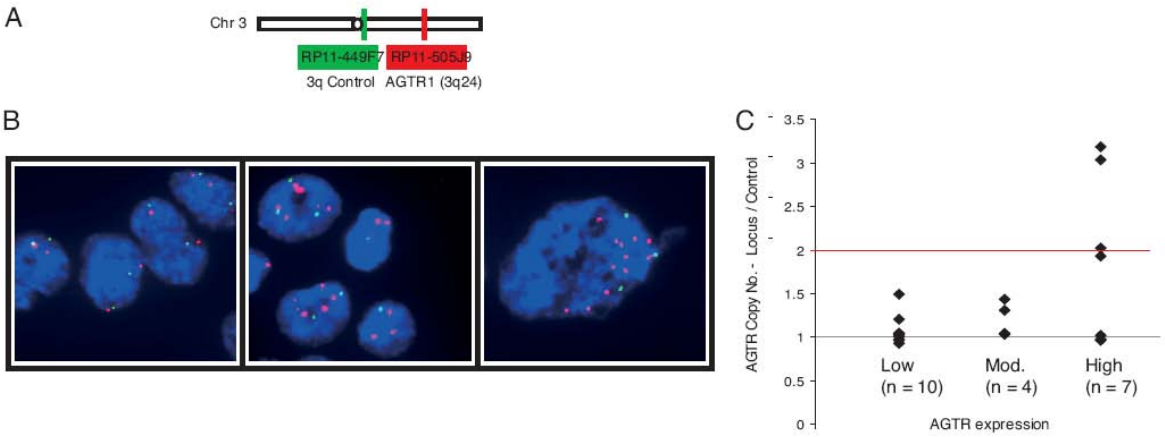


Figure 4. *Copy number analysis of the AGTR1 locus. (A) A schematic of probes used for FISH analysis- Control (green) and AGTR1 (red). (B) Representative images from FISH analysis- left, representative negative case, middle and right, cases with copy number gains of AGTR1. (C) Association ofAGTR1overexpression with copy number gain.*

C. Overexpression and knock-down of AGTR in breast cancer cell lines (Year 1)

Ectopic overexpression of AGTR1 in primary mammary epithelial cells, such as HMEC and H16N2, combined with angiotensin II stimulation, led to a highly invasive phenotype that was attenuated by the AGTR1 antagonist losartan. This indicated a possible functional role of AGTR1 in breast cancers (***Figure 5***).

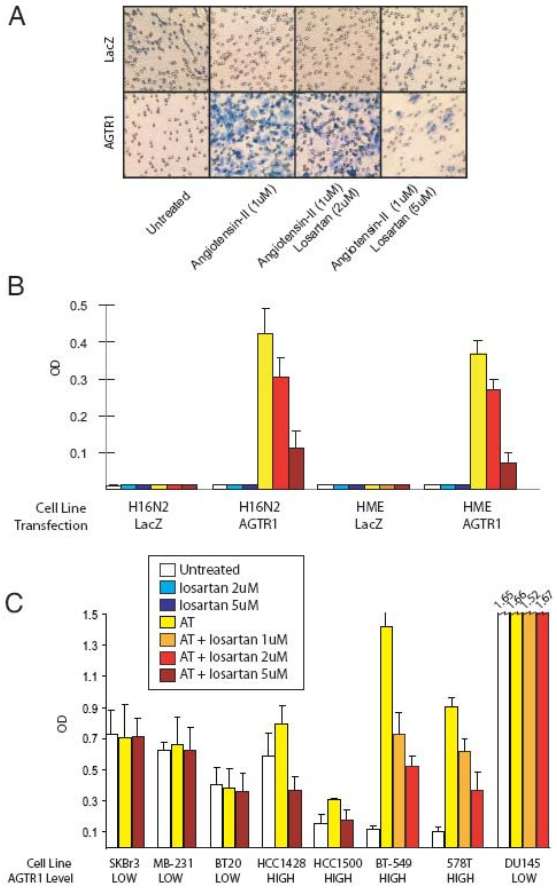


Figure 5. *AGTR1 overexpression and effect on cell invasion. (A) Matrigel invasion assays of Human Mammary Epithelial Cells (HMEC) or immortalized normal mammary epithelial cells, H16N2 overexpressing AGTR1 or LacZ. Cells cultured with and without agonist, angiotensin (AT) or antagonist, losartan. Similar 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.

D. Development of xenograft models of AGTR overexpression in breast cancer (Years 2-3)

Similar to the observations of *invitro* cell culture experiments, the AGTR inhibitor losartan exerted an inhibitory effect on AGTR1-positive breast cancer xenografts, reducing tumor growth by 30% (***Figure 6***).

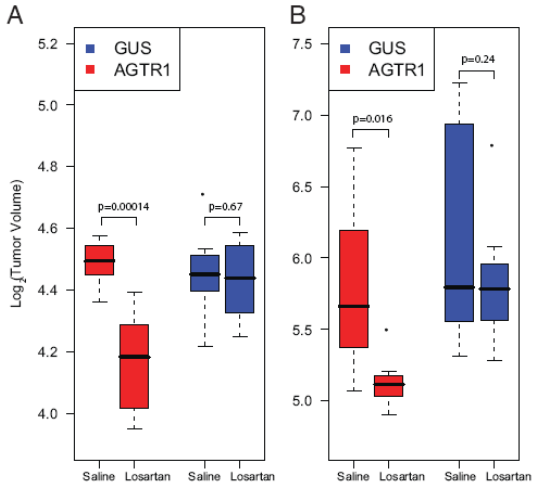


Figure 6*. Effect of losartan treatment on AGTR1 expressing MCF7 cell xenografts. (A) Xenograft tumor size at 2 weeks. (B) Xenograft tumor size at 8 weeks.*

E. Studies using losartan as an antagonist of AGTR (Years 1-3)

Both, *in vitro* studies using AGTR1 overexpression in normal mammary epithelial cells (C) and *in vivo* studies involving tumor xenografts of AGTR1 overexpressing breast cancer cells (D) indicated that a subpopulation of ER-positive, ERBB2-negative breast cancers, that overexpress AGTR1may benefit from targeted therapy with AGTR1 antagonists, such as losartan.
Future work would attempt to further characterize the role of AGTR1 in breast cancer progression and stimulate clinical trials using losartan in women with breast cancer that have high AGTR levels

Task 5. Study breast cancer microRNAs relative to gene fusion candidates

Enhancer of zeste homolog 2 (EZH2) is a mammalian histone methyltransferase that is overexpressed in aggressive solid tumors, including breast cancer(25) 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 lead to the significant discovery of genomic loss of mir101 accounting for increased expression of EZH2 in a cohort of aggressive prostate and breast cancers, that was recently published in **Science** (15) (***Figure 7***).

A. Evaluate mir101 in breast cancer (Years 1-2)

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 level was observed (Figure 8). These observations were later extended to other breast and prostate cancer samples.

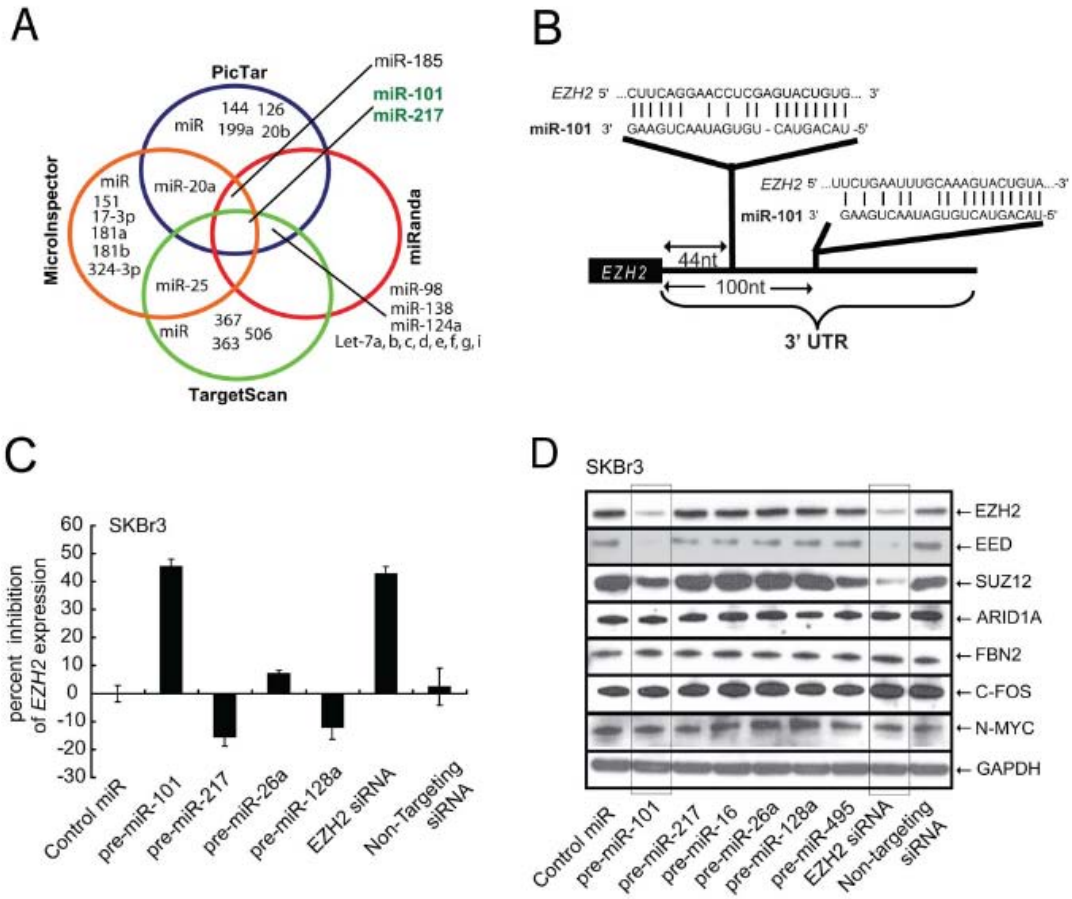


Figure 7. *miR-101 inhibits EZH2 transcript and protein expression in breast cancer cell line SKBR3.* ***(A)*** *Venn diagram displaying miRNAs computationally predicted to target EZH2 using different target prediction programs.* ***(B)*** *Schematic of two predicted miR-101 binding sites in the EZH2 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.*

B. Profile microRNAs in breast cancer samples (Years 2-4)

Spurred by our success in delineating the role of mir101 in breast and prostate cancers we plan to profile microRNA expression by next generation sequencing platform in a cohort of breast cancer samples in the coming year.

C. Study role of mir101 relative to epigenetic pathways (Years 1-3)

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 8A***), and this resulted in increased gene expression of the target genes (***Figure 8B***). Gene-expression array analysis of SKBr3 cells transfected with either miR-101 or EZH2 siRNA duplexes showed significant overlap in gene expression.

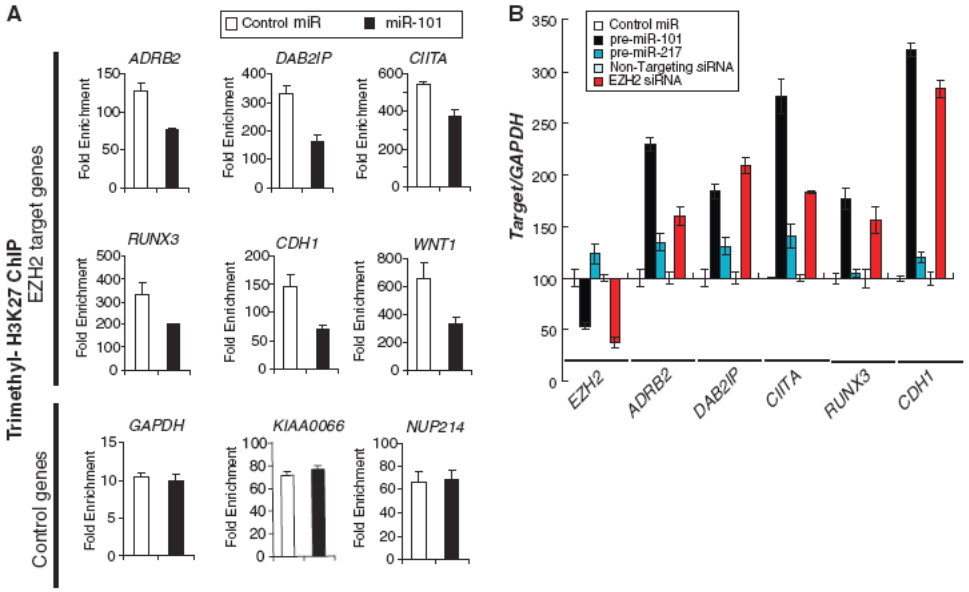


Figure 8. *miR-101 regulation of the cancer epigenome through EZH2 and H3K27 tri-methylation. (**A**) Chromatin immunoprecipitation (ChIP) assay of the trimethyl H3K27 histone mark when miR-101 is overexpressed. Known PRC2 repression targets were examined in SKBr3 cells. ChIP was performed to test H3K27 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.*

D. Role of mir101 in breast cancer development using in vitro and *in vivo* models (Yrs 2-5).

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 9***)

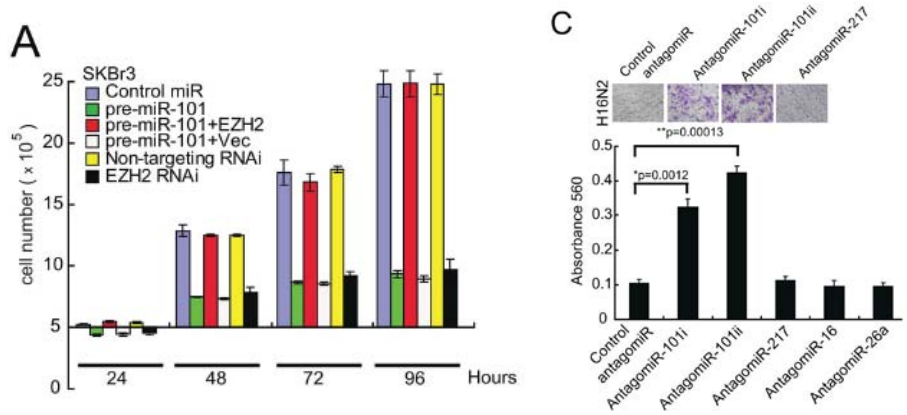


Figure 9. *The role of miR-101 in regulating cell proliferation, invasion and tumor growth.*
*(**C**) AntagomiRs to miR-101 induce invasion in benign immortalized H16N2 breast epithelial cells.*

KEY RESEARCH ACCOMPLISHMENTS: Bulleted list of key research accomplishments emanating from this research.

The current funding period for the first year was very productive and we accomplished the majority of the goals of the proposal and performed additional studies to lay the groundwork for the discovery of recurrent gene fusions and other important molecular aberrations in breast cancer.

- 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 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.

REPORTABLE OUTCOMES: *Provide a list of reportable outcomes that have resulted from this research to include: manuscripts, abstracts, presentations; patents and licenses applied for and/or issued; degrees obtained that are supported by this award; development of cell lines, tissue or serum repositories; informatics such as databases and animal models, etc.; funding applied for based on work supported by this award; employment or research opportunities applied for and/or received based on experience/training supported by this award.*

1. 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 [PubMed - indexed for MEDLINE]
2. 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 [PubMed - indexed for MEDLINE]
3. 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 [PubMed - as supplied by publisher]

CONCLUSION:

Here we have 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 have 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. Additionally, based on observations so far, several gene fusions appear to involve one 5' partner fused to different 3' partners or one 3' partner driven by different 5' partner genes. This presents a further level of complexity that we plan to delve in detail in the coming days.

"So what?": Gene fusions represent exquisitely specific cancer biomarkers as well as therapeutic targets, and the discovery of recurrent gene fusions in common solid cancers such as prostate and lung cancers proffers a unified genetic basis for the apparently dichotomous realms of liquid cancers (hematological and soft tissue malignancies) and solid cancers (epithelial cancers). In that context, it is imperative to 'smoke out' the gene fusions (almost certainly) driving breast cancers, one of the most common epithelial cancers. While most 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 anticipated meeting our aim of characterizing recurrent gene fusions in breast cancer…or make some other unexpected breakthrough discoveries in the process.

Our discovery of AGTR1 overexpressing subset of ER positive breast cancers that may respond to available drugs such as losartan, is one such unexpected discovery that may yet translate to novel prognostic and therapeutic options for this cohort. Likewise, the discovery of the role of mir101 as a negative regulator of the polycomb group protein EZH2, earlier discovered by our group as associated with metastatic breast and prostate cancers, marks another fundamental advance in our understanding of cancer biology, cutting across organ types.

REFERENCES:

1. 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 (New York, NY 2005;310(5748):644-8.
2. Helgeson BE, Tomlins SA, Shah N, Laxman B, Cao Q, Prensner JR, Cao X, Singla N, Montie JE, Varambally S, Mehra R, Chinnaiyan AM.. Characterization of TMPRSS2:ETV5 and SLC45A3:ETV5 gene fusions in prostate cancer. Cancer research 2008;68(1):73-80.
3. Han B, Mehra R, Dhanasekaran SM, Yu J, Menon A, Lonigro RJ, Wang X, Gong Y, Wang L, Shankar S, Laxman B, Shah RB, Varambally S, Palanisamy N, Tomlins SA, Kumar-Sinha C, Chinnaiyan AM.. A fluorescence in situ hybridization screen for E26 transformation-specific aberrations: identification of DDX5-ETV4 fusion protein in prostate cancer. Cancer research 2008;68(18):7629-37.
4. Tomlins SA, Laxman B, Dhanasekaran SM, Helgeson BE, Cao X, Morris DS, Menon A, Jing X, Cao Q, Han B, Yu J, Wang L, Montie JE, Rubin MA, Pienta KJ, Roulston D, Shah RB, Varambally S, Mehra R, Chinnaiyan AM. Distinct classes of chromosomal rearrangements create oncogenic ETS gene fusions in prostate cancer. Nature 2007;448(7153):595-9.
5. Kumar-Sinha C, Tomlins SA, Chinnaiyan AM. Recurrent gene fusions in prostate cancer. Nature reviews 2008;8(7):497-511.
6. Soda M, Choi YL, Enomoto M, Takada S, Yamashita Y, Ishikawa S, Fujiwara S, Watanabe H, Kurashina K, Hatanaka H, Bando M, Ohno S, Ishikawa Y, Aburatani H, Niki T, Sohara Y, Sugiyama Y, Mano H. Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer. Nature 2007;448(7153):561-6.
7. Rikova K, Guo A, Zeng Q, Possemato A, Yu J, Haack H, Nardone J, Lee K, Reeves C, Li Y, Hu Y, Tan Z, Stokes M, Sullivan L, Mitchell J, Wetzel R, Macneill J, Ren JM, Yuan J, Bakalarski CE, Villen J, Kornhauser JM, Smith B, Li D, Zhou X, Gygi SP, Gu TL, Polakiewicz RD, Rush J, Comb MJ. Global survey of phosphotyrosine signaling identifies oncogenic kinases in lung cancer. Cell 2007;131(6):1190-203.
8. Vogelstein B, Kinzler KW. Cancer genes and the pathways they control. Nature medicine 2004;10(8):789-99.
9. Mitelman F, Johansson B, Mertens F. The impact of translocations and gene fusions on cancer causation. Nature reviews 2007;7(4):233-45.
10. Druker BJ, Guilhot F, O'Brien SG, Gathmann I, Kantarjian H, Gattermann N, Deininger MW, Silver RT, Goldman JM, Stone RM, Cervantes F, Hochhaus A, Powell BL, Gabrilove JL, Rousselot P, Reiffers J, Cornelissen JJ, Hughes T, Agis H, Fischer T, Verhoef G, Shepherd J, Saglio G, Gratwohl A, Nielsen JL, Radich JP, Simonsson B, Taylor K, Baccarani M, So C, Letvak L, Larson RA; IRIS Investigators. Five-year follow-up of patients receiving imatinib for chronic myeloid leukemia. The New England journal of medicine 2006;355(23):2408-17.

11. Deininger M, Buchdunger E, Druker BJ. The development of imatinib as a therapeutic agent for chronic myeloid leukemia. Blood 2005;105(7):2640-53.
12. 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;458(7234):97-101.
13. 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. Proceedings of the National Academy of Sciences of the United States of America 2009;106(25):10284-9.
14. 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. Proceedings of the National Academy of Sciences of the United States of America 2009.
15. 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 (New York, NY 2008;322(5908):1695-9.
16. Bärlund M, Monni O, Weaver JD, Kauraniemi P, Sauter G, Heiskanen M, Kallioniemi OP, Kallioniemi A. Cloning of BCAS3 (17q23) and BCAS4 (20q13) genes that undergo amplification, overexpression, and fusion in breast cancer. Genes, chromosomes & cancer 2002;35(4):311-7.
17. Kallioniemi A, Kallioniemi OP, Piper J, Tanner M, Stokke T, Chen L, Smith HS, Pinkel D, Gray JW, Waldman FM. Detection and mapping of amplified DNA sequences in breast cancer by comparative genomic hybridization. Proceedings of the National Academy of Sciences of the United States of America 1994;91(6):2156-60.
18. Muleris M, Almeida A, Gerbault-Seureau M, Malfoy B, Dutrillaux B. Detection of DNA amplification in 17 primary breast carcinomas with homogeneously staining regions by a modified comparative genomic hybridization technique. Genes, chromosomes & cancer 1994;10(3):160-70.
19. Bignell GR, Santarius T, Pole JC, Butler AP, Perry J, Pleasance E, Greenman C, Menzies A, Taylor S, Edkins S, Campbell P, Quail M, Plumb B, Matthews L, McLay K, Edwards PA, Rogers J, Wooster R, Futreal PA, Stratton MR. Architectures of somatic genomic rearrangement in human cancer amplicons at sequence-level resolution. Genome research. 2007;17(9):1296-303.
20. Neve RM, Chin K, Fridlyand J, Yeh J, Baehner FL, Fevr T, Clark L, Bayani N, Coppe JP, Tong F, Speed T, Spellman PT, DeVries S, Lapuk A, Wang NJ, Kuo WL, Stilwell JL, Pinkel D, Albertson DG, Waldman FM, McCormick F, Dickson RB, Johnson MD, Lippman M, Ethier S, Gazdar A, Gray JW. A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. Cancer cell 2006;10(6):515-27.

21. 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 (New York, NY 2007;9(2):166-80.
22. 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 (New York, NY 2004;6(1):1-6.
23. Tomlins SA, Rhodes DR, Yu J, Varambally S, Mehra R, Perner S, Demichelis F, Helgeson BE, Laxman B, Morris DS, Cao Q, Cao X, Andrén O, Fall K, Johnson L, Wei JT, Shah RB, Al-Ahmadie H, Eastham JA, Eggener SE, Fine SW, Hotakainen K, Stenman UH, Tsodikov A, Gerald WL, Lilja H, Reuter VE, Kantoff PW, Scardino PT, Rubin MA, Bjartell AS, Chinnaiyan AM. The role of SPINK1 in ETS rearrangement-negative prostate cancers. Cancer cell 2008;13(6):519-28.
24. 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. Proceedings of the National Academy of Sciences of the United States of America 2004;101(25):9309-14.
25. 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. Proceedings of the National Academy of Sciences of the United States of America 2003;100(20):11606-11.

EFFORTS IN BREAST CANCER RESEARCH

W81XWH-08-0110 (PI: Chinnaiyan) 09/01/08 – 11/30/13 25%
Department of Defense $500,000/yr **25% Breast cancer**
A Search for Gene Fusions/Translocations in Breast Cancer
Specific Aims: 1) develop high-throughput adaptations of existing methodologies such as fluorescence in situ hybridization (FISH), 2) employ bioinformatics and associated analytical tools to elucidate recurrent gene fusions in breast cancers, 3) employ next generation whole transcriptome sequencing of breast tumors.
Contact Information at funding agency: Grants Officer: JenniferHayes, 301-619-6746, Jennifer.Hayes@us.army.mil

Effort to breast cancer: 25%

U01 CA111275 (PI: Chinnaiyan) 09/20/04-06/30/10 10%
NIH $404,077/yr **5% Breast cancer**
Grants Officer: Shane Woodward, 302-846-1017, woodwars@mail.nih.gov
EDRN Biomarker Development Lab
Goals:
Specific Aims: 1) to characterize and validate the humoral immune response to AMACR in different patient cohorts, 2) employ high-throughput phage epitope microarrays to identify candidate humoral response markers of cancer and 3) define and develop a multiplexed protein/epitope microarray to identify cancer based on humoral response.

Effort to breast cancer: 5% (5% to prostate cancer). While this grant has been focused on prostate cancer, in general it is a biomarker development lab and half of my effort can be designated to the development of breast cancer biomarkers including AGTR in ER+, erbB2 - patients

1 U54 DA021519-01A1 (PI: Athey) 09/25/05-08/31/10 3%
NIH $2,543,758/yr **3% Breast cancer**
National Center for Integrative Biomedical Informatics
Grants Officer: Catherine Mills, 301-443-6710, cmills@ngmsmtp.nida.nih.gov
Goals: Develop bioinformatics and computational approaches for high-throughput data.
Specific Aims: 1) Create an integrated model for cancer progression using microarray gene expression, MPSS transcript, proteomics, and protein-protein interaction data and text. Use Oncomine and Molecular Concepts Maps. 2) Explore at a systems level the roles of Polycomb Group (PcG)proteins in transcription, chromatin structure, histone protein interactions, and protein expression patterns in progression, invasion, and metastasis of cancers 3) Characterize translocations, including fusion genes important to etiology of cancers
Role: Co-Investigator

Effort to breast cancer: 3%

Project# 1005930 (PI: Chinnaiyan) 07/01/06-06/30/11 10%
Burroughs Wellcome Fund $150,000/yr **10% Breast cancer**
Grants Officer: Nancy Sung, 919-991-5100
Autoantibody Profiles for Cancer Diagnosis, Prognosis, and Therapy
Goals: Develop immunomic profiles for cancer and human disease.
Specific Aims: 1) Extend the autoantibody screening platform we have developed in prostate cancer to other solid tumors for the purpose of cancer diagnosis; 2) Determine whether autoantibody signatures can be used to classify cancers based on type and/or sub-type. The overall goal would be to develop a multi-

cancer classifier based on autoantibody profiles as well as develop prognostic and/or histopathologic classifiers based on autoantibody profiles.

Effort to breast cancer: 10%. There are no restrictions on the type of cancer focused on here and thus breast cancer will be the focus.

PI: Chinnaiyan 01/01/09 – 12/31/13 10%
Doris Duke Foundation $275,000/yr **10% Breast cancer**
Distinguished Clinical Scientist Award for Excellence in "Bench to Bedside" Research
Specific Aims: 1) Develop and employ high-throughput fluorescence in situ hybridization (FISH) in order to interrogate solid tumors for recurrent chromosomal aberrations including gene fusions and translocations; 2) Employ bioinformatics and associated analytical tools to elucidate recurrent gene fusions in common solid tumors;. 3) Employ next generation whole transcriptome and paired-end sequencing of common solid tumors to identify recurrent gene fusions and integrated non-human sequences that may represent pathogens.

Effort to breast cancer: 10%

W81XWH-09-2-0014 (PI: Wicha) 03/01/09 – 04/24/10 4%
Department of Defense $443,618/yr **4% Breast cancer**
National Functional Genomics Center
Goals: to develop a comprehensive approach to genetics, proteomics and bioinformatics that can help elucidate the mechanisms driving tumorigenesis. This research investigates the notion that cancer stem cells are the key cell component driving tumorigenesis, metastasis and treatment resistance.
Specific Aims: 1) To isolate and achieve molecular characterization of cancer stem cells from human breast, prostate, colon, pancreas, head and neck, brain, ovarian and melanomas. 2) To better define pathways that regulate cancer we will utilize the integrative oncogenomics approaches including HIMAP to elucidate the interacting pathways regulating cancer stem cells. 3) To identify novel genes regulating cancer stem cells we propose to utilize a high throughput siRNA approach to screen for genes which play a functional role in stem cell self-renewal.
Role: Co-Investigator

Effort to breast cancer: 4%

TOTAL EFFORT DEDICATED TO BREAST CANCER RESEARCH: 57%

COMPLETE LIST OF EXISTING AND PENDING SUPPORT

CHINNAIYAN, A.M.
ACTIVE

Howard Hughes Medical Institute (HHMI) 02/01/08 – 01/31/13 NA
Howard Hughes Medical Institute $700,000/yr
Investigator
Though HHMI supports Dr. Chinnaiyan as an HHMI Investigator, these funds are not awarded to a specific research proposal or project.

W81XWH-08-0110 (PI: Chinnaiyan) 09/01/08 – 11/30/13 25%
Department of Defense $500,000/yr **25% Breast cancer**
A Search for Gene Fusions/Translocations in Breast Cancer
Specific Aims: 1) develop high-throughput adaptations of existing methodologies such as fluorescence in situ hybridization (FISH), 2) employ bioinformatics and associated analytical tools to elucidate recurrent gene fusions in breast cancers, 3) employ next generation whole transcriptome sequencing of breast tumors.
Contact Information at funding agency: Grants Officer: JenniferHayes, 301-619-6746, Jennifer.Hayes@us.army.mil

P50 CA69568 (PI: Pienta) 06/01/08 - 05/31/13 8%
NCI $196,297/yr
SPORE in Prostate Cancer
Project 1 Title: *Role of gene fusions in prostate cancer*
Goals: determine the role of ETS family gene fusions in prostate cancer cell lines; characterize the phenotype of androgen-regulated ETS transgenic mice.
Specific Aims: Specific aims: 1) Characterization of Oncogenic ETS Gene Fusions in Prostate Cancer; 2) Determine the role of ETS family gene fusions in prostate cancer cell lines; 3) characterize the phenotype of androgen-regulated ETS transgenic mice.
Role: Co-Investigator
Contact Information at funding agency: Andrew Hruszkewycz, 301-496-8528, hruszkea@mail.nih.gov

P50 CA69568 (PI: Pienta) 06/01/08 – 05/31/13 5%
Core 3: Tissue/Informatics Core Director $335,726/yr
Goals: the goal of the Core is to collect biological material with associated clinical information to facilitate translational research.
Role: Core Director
Contact Information at funding agency: Andrew Hruszkewycz, 301-496-8528, hruszkea@mail.nih.gov

U01 CA111275 (PI: Chinnaiyan) 09/20/04-06/30/10 10%
NIH $404,077/yr **5% Breast cancer**
EDRN Biomarker Development Lab
Specific Aims: 1) to characterize and validate the humoral immune response to AMACR in different patient cohorts, 2) employ high-throughput phage epitope microarrays to identify candidate humoral response markers of cancer and 3) define and develop a multiplexed protein/epitope microarray to identify cancer based on humoral response.

Contact Information at funding agency: Shane Woodward, 302-846-1017, woodwars@mail.nih.gov

U01 CA113913 (PI: Wei) 03/29/05 – 02/28/10 1%
Beth Israel Hospital (NIH Prime) $100,172
Harvard-Michigan Prostate Cancer Biomarker Clinical Validation Center
Goals: Collect samples for the EDRN validation studies and early validation of EDRN biomarkers.
Role: Co-Investigator
Sponsor contact Information: Jennifer Sabbagh, Beth Israel Deaconess Medical Center, 330 Brookline Ave. ST8M-18, Boston, MA 02215. Email, jsabbagh@bidmc.harvard.edu

1 U54 DA021519-01A1 (PI: Athey) 09/25/05-08/31/10 3%
NIH $2,543,758/yr **3% Breast cancer**
National Center for Integrative Biomedical Informatics
Goals: Develop bioinformatics and computational approaches for high-throughput data.
Specific Aims: 1) Create an integrated model for cancer progression using microarray gene expression, MPSS transcript, proteomics, and protein-protein interaction data and text. Use Oncomine and Molecular Concepts Maps. 2) Explore at a systems level the roles of Polycomb Group (PcG)proteins in transcription, chromatin structure, histone protein interactions, and protein expression patterns in progression, invasion, and metastasis of cancers 3) Characterize translocations, including fusion genes important to etiology of cancers
Role: Co-Investigator
Contact Information at funding agency: Catherine Mills, 301-443-6710, cmills@ngmsmtp.nida.nih.gov

Project# 1005930 (PI: Chinnaiyan) 07/01/06-06/30/11 10%
Burroughs Wellcome Fund $150,000/yr **10% Breast cancer**
Autoantibody Profiles for Cancer Diagnosis, Prognosis, and Therapy
Goals: Develop immunomic profiles for cancer and human disease.
Specific Aims: 1) Extend the autoantibody screening platform we have developed in prostate cancer to other solid tumors for the purpose of cancer diagnosis; 2) Determine whether autoantibody signatures can be used to classify cancers based on type and/or sub-type. The overall goal would be to develop a multi-cancer classifier based on autoantibody profiles as well as develop prognostic and/or histopathologic classifiers based on autoantibody profiles.
Contact Information at funding agency: Nancy Sung, 919-991-5100

W81XWH-08-1-0031 (PI: Chinnaiyan) 04/15/08 – 07/14/11 10%
Department of Defense $121,746/yr
Characterization of SPINK1 in Prostate Cancer
Goals: study and define the role of SPINK1 in TMPRSS2-ETS negative prostate cancers and also explore the utility of SPINK1 as a prostate cancer biomarker.
Specific Aims: 1): Determine the role of SPINK1 in prostate cancer cell lines; 2) Explore the mechanism of SPINK1 overexpression in a subset of prostate cancers; 3) Determine the utility of SPINK1 for the non-invasive detection of prostate cancer in urine biospecimens.
Contact Information at funding agency: Grants Officer: Cheryl Lowery, 301-619-7150

PI: Chinnaiyan 01/01/09 – 12/31/13 10%
Doris Duke Foundation $275,000/yr **10% Breast cancer**

Distinguished Clinical Scientist Award for Excellence in "Bench to Bedside" Research
Specific Aims: 1) Develop and employ high-throughput fluorescence in situ hybridization (FISH) in order to interrogate solid tumors for recurrent chromosomal aberrations including gene fusions and translocations; 2) Employ bioinformatics and associated analytical tools to elucidate recurrent gene fusions in common solid tumors;. 3) Employ next generation whole transcriptome and paired-end sequencing of common solid tumors to identify recurrent gene fusions and integrated non-human sequences that may represent pathogens.
Contact Information at funding agency: Grants Officer: Adrienne Fischer, Doris Duke Charitable Foundation, 650 5th Avenue, Fl 19, NY, NY

R01CA132874-01 (PI: Chinnaiyan) 03/01/09– 11/31/13 10%
NIH $166,000/yr
Molecular Sub-typing of Prostate Cancer Based on Recurrent Gene Fusions
Specific Aims: 1) discovery and nomination of novel molecular sub-types of prostate cancer, 2) characterize associations of molecular sub-types of prostate cancer with clinical outcome and/or aggressiveness of disease in a radical prostatectomy cohort, 3) characterize associations of molecular sub-types of prostate cancer with clinical outcome and/or aggressiveness of disease using prostate needle biopsy samples.
Contact Information at funding agency: Grants Management Specialist: Catherine Blount, Email: blountc@mail.nih.gov Phone: 301-496-3179

(PI: Kumar) 01/01/09 – 12/31/09 2.5%
Lustgarten Foundation $100,000/yr
Discovery of Recurrent Gene Fusions in Pancreatic Cancer using High-throughput Sequencing
Goal: carry out a survey of pancreatic cancer transcriptome to identify recurrent gene fusions using high-throughput sequencing.
Specific Aims:
Contact Information at funding agency: LSASSO@cablevision.com, Lustgarden Foundation, 1111 Stewart Avenue, Bethpage, NY, 11714
Role: Co-Investigator

W81XWH-09-2-0014 (PI: Wicha) 03/01/09 – 04/24/10 4%
Department of Defense $443,618/yr **4% breast cancer**
National Functional Genomics Center
Goals: to develop a comprehensive approach to genetics, proteomics and bioinformatics that can help elucidate the mechanisms driving tumorigenesis. This research investigates the notion that cancer stem cells are the key cell component driving tumorigenesis, metastasis and treatment resistance.
Specific Aims: 1) To isolate and achieve molecular characterization of cancer stem cells from human breast, prostate, colon, pancreas, head and neck, brain, ovarian and melanomas. 2) To better define pathways that regulate cancer we will utilize the integrative oncogenomics approaches including HIMAP to elucidate the interacting pathways regulating cancer stem cells. 3) To identify novel genes regulating cancer stem cells we propose to utilize a high throughput siRNA approach to screen for genes which play a functional role in stem cell self-renewal.
Contact Information at funding agency: Dr. Anne Westbrook, e-mail vivian.westbrook@tatrc.org
Role: Co-Investigator

PENDING
American Association for Cancer Research (Dream team leader: Gray) 10/01/09 – 9/30/12 7.5%

Stand Up To Cancer Dream Team Translational Cancer Research $362,190/yr
Personalizing treatment of triple negative, metastatic breast cancer
Goals: developing targeted molecular therapies for breast cancer treatment; test the efficacy of individualized treatment of drug-resistant, triplenegative.
Specific Aims: 1) Compare omic features of 100 drug resistant, TNBCs with those of untreated primary tumors to identify omic features associated with metastasis and/or drug resistance. 2.) Developed improved preclinical biological models of drug resistant, triple-negative breast cancer to facilitate identification of therapeutic approaches that will be effective against TNBC. 3) Identify omic features of metastatic, drug resistant TNBC subsets associated with response to approved and experimental therapeutic agents using novel computational and experimental approaches. 4) Develop and compare computational methods for selection of drugs/combinations for individualized treatment of TNBC patients based on the omic characteristics of their tumors. 5) Conduct an omic-marker-guided clinical trial of therapies predicted to be effective against TNBC subsets. 6) Develop a comprehensive public/patient education and awareness campaign to introduce the consumer community to the new "personalized medicine" concept.
Role: Dream Team Principal (Chinnaiyan)

OVERLAP

Once the pending proposal is award, effort will be reduced on the Burroughs Wellcome Fund project.

APPENDICES:

PDFs of the following journal articles:

1. Rhodes DR, Ateeq B, Cao Q, Tomlins SA, Mehra R, Laxman B, Kalyana-Sundaram S, Lonigro RJ, Helgeson BE, Bhojani MS, Rehemulla A, Kleer CG, Hayes DF, Lucas PC, Varambally S, **Chinnaiyan AM**. AGTR1 overexpression defines a subset of breast cancer and confers sensitivity to losatran, an AGTR1 antagonist. *Proc Natl Acad Sci USA* 2009: 106(25): 10284-10289. PMID: 19487683/PMCID: PMC 2689309

   News Stories from "AGTR1 overexpression defines a subset of breast cancer and confers sensitivity to losatran, an AGTR1 antagonist"

2. Maher CA, Palanismay 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 pair-end transcriptome sequencing. *Proc Natl Acad Sci USA* 2009: 106(30): 12353-12358. PMID: 19592507/PMCID: PMC2708976.

3. Prensner JR, **Chinnaiyan AM**. Oncogenic gene fusions in epithelial carcinomas. *Curr Opinion Genet Dev* 2009 Feb; 19(1):82-91. PMID 19233641/PMCID:PMC2676581

4. 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: 458(7234): 97-101. PMID 19136943/PMCID: PMC2725402

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.

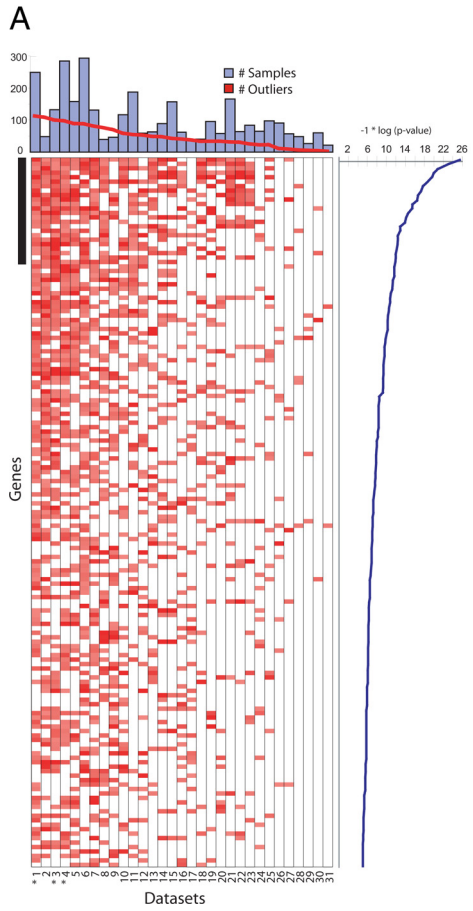
This article is a PNAS Direct Submission.

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.



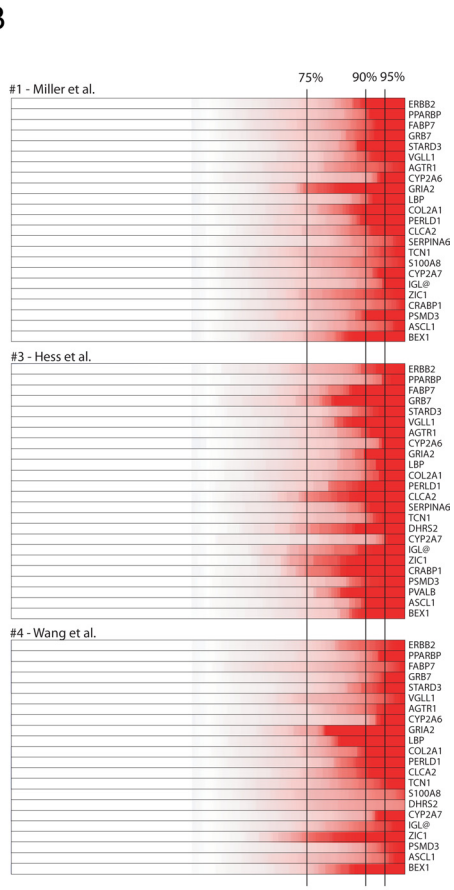


Fig. 1. MetaCOPA analysis of breast cancer gene expression data. (*A*) MetaCOPA map. Each column in the map represents a breast cancer gene expression dataset. The numbers at the base of the map correspond to dataset details (Table S1). Each row indicates a gene. A red cell indicates that the gene was 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 ($P < 1E\text{-}5$). 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 ($P < 1E\text{-}5$) (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%; $P = 3.6E\text{-}26$), 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%; $P = 2.0E\text{-}18$). 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′

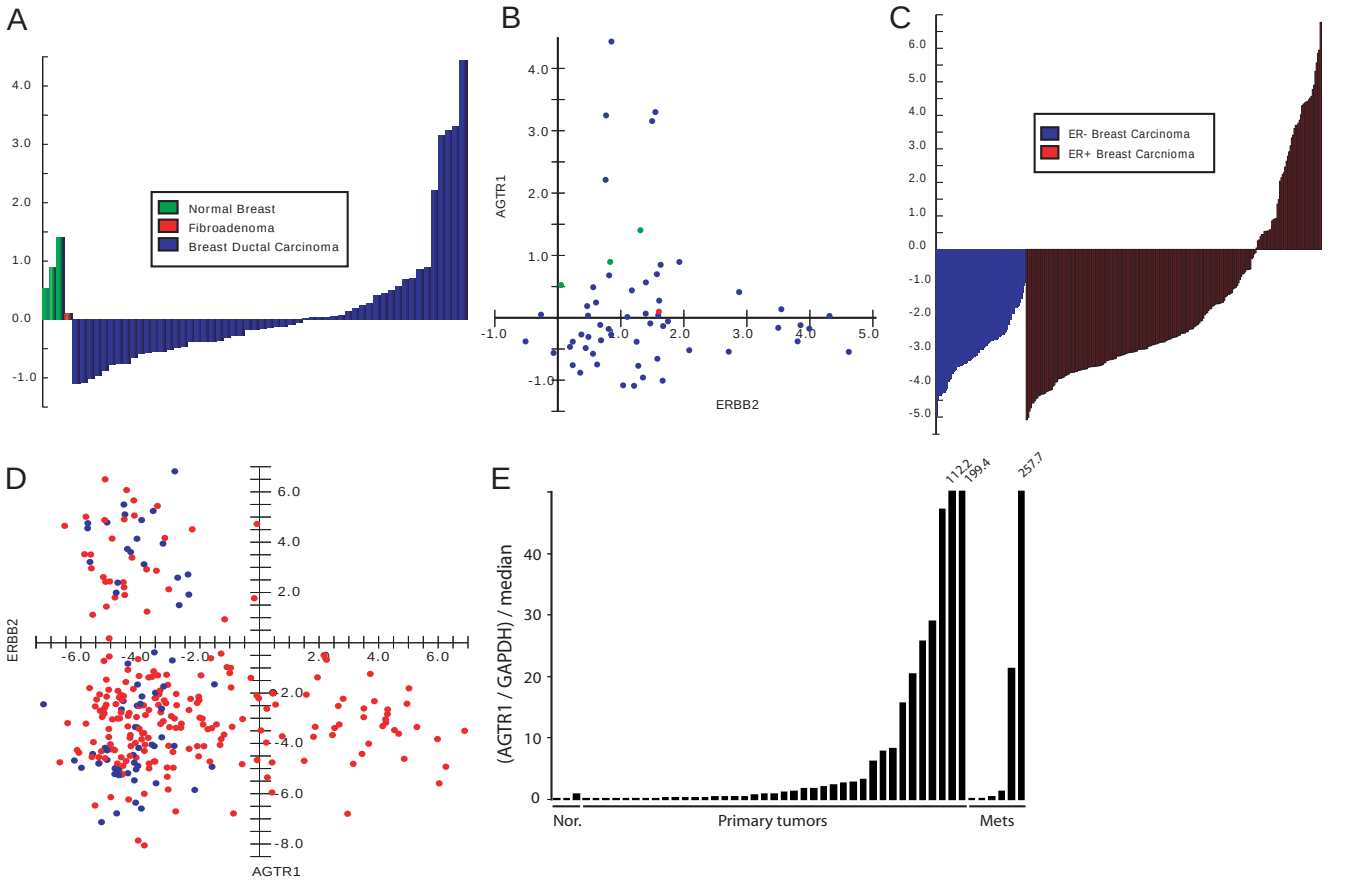


Fig. 2. AGTR1 outlier expression in breast cancer. (*A*) AGTR1 expression profile in the Perou et al. (29) cDNA microarray dataset ($n = 55$). (*B*) In the same dataset, AGTR1 expression vs. ERBB2 expression. (*C*) AGTR1 expression profile in the van de Vijver et al. (30) oligonucleotide dataset, segregated 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 μM) 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

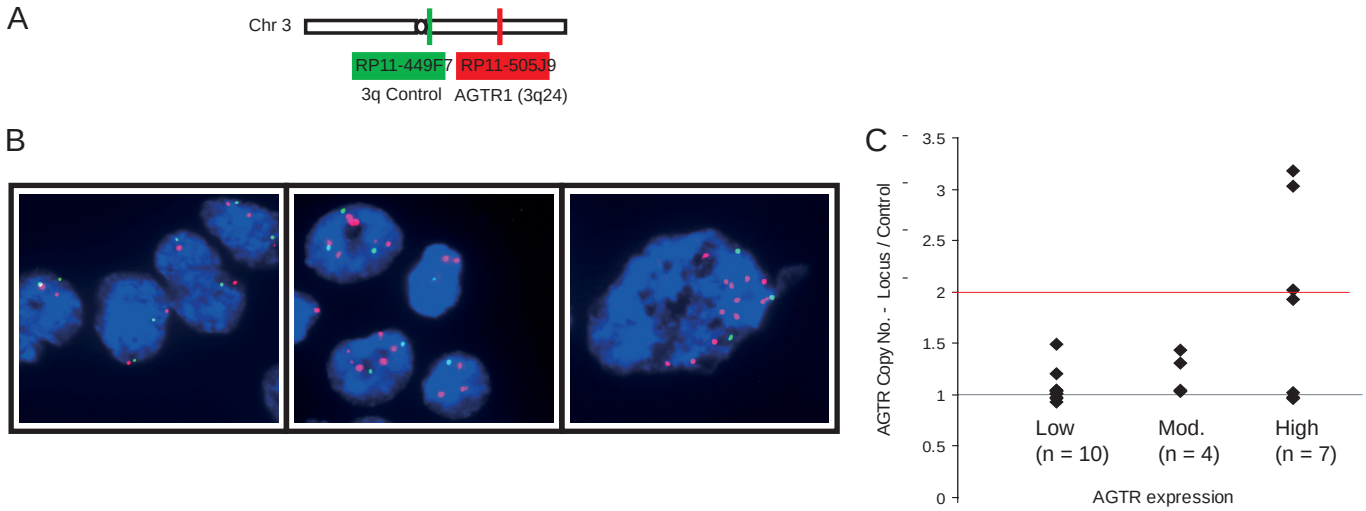


Fig. 3. Copy number analysis of the AGTR1 locus. (*A*) A schematic of probes used for FISH analysis. (*B*) Representative image from FISH analysis. *Left* is taken from a representative negative case. *Middle* and *Right* are images from a representative case with definitive copy number gain of AGTR1. 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 μM AT and losartan. In each of the 4 AGTR1-overexpressing cell lines, we observed an increase in invasion upon stimulation with 1 μM 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 μM 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,

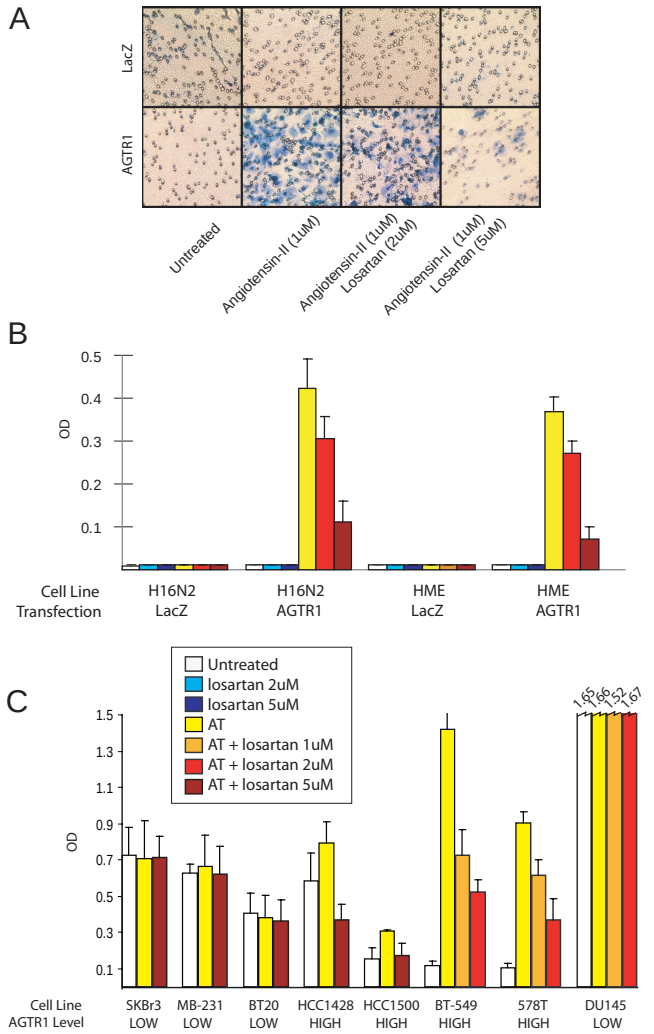


Fig. 4. AGTR1 overexpression and analysis of angiotensin II (AT) and losartan effects on cell invasion. (*A*) Matrigel invasion assays of H16N2 cells infected with adenovirus expressing AGTR1 or LacZ. Cells were cultured in serum-free media and were pretreated with and without AT and losartan. Similar 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 μM 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 μM 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*.

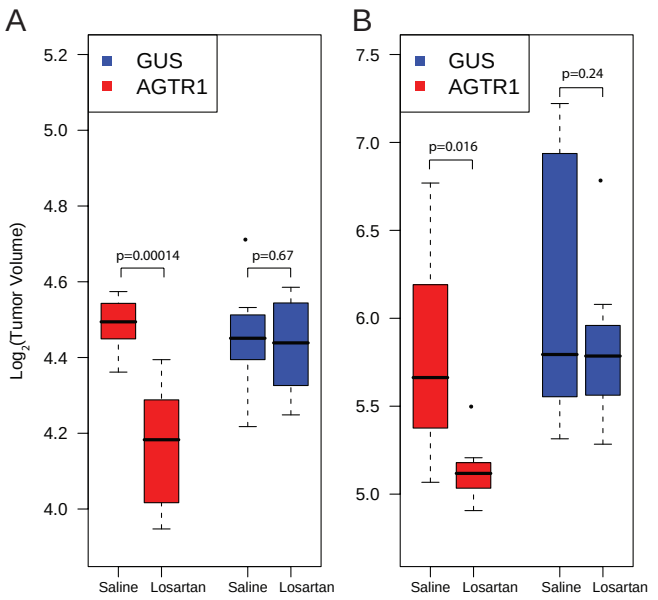


Fig. 5. Effect of losartan treatment on AGTR1- or GUS-overexpressing MCF7 cell xenografts. Female BALB/C nu/nu mice were implanted with 2.5×10^6 stable MCF7 cells overexpressing AGTR1 or GUS resuspended in 100 μL of saline with 20% Matrigel into the mammary fat pad of anesthetized mice. Mice from both 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.

University of Michigan
Comprehensive Cancer Center

News stories from AGTR1 paper

J

Project	# of Articles	Print	Online	Blogs	B'cast	Newswires
AGTR1	12	0	11	0	1	0

Project: AGTR1

Type	Date	Headline Publication / *Journalist*	Audience	Pub. Value
Online	6/3/2009	Blood Pressure Meds May Fight Cancer Genes (Arul Chinnaiyan)		
		Health.com *n/a*	n/a	n/a
Online	6/2/2009	Research Finds New Crop of Breast Cancer Genes (Arul Chinnaiyan)		
		Atlanta Journal And Constitution *n/a*	39,808	$5,727.90
Online	6/2/2009	Breast Cancer Gene Can Be Blocked By Blood Pressure ...(Arul Chinnaiyan)		
		Science Daily *n/a*	n/a	n/a
Online	6/2/2009	Research Finds 'New Crop' of Breast Cancer Genes (Arul Chinnaiyan)		
		Health Day *n/a*	n/a	n/a
Online	6/2/2009	Research Finds New Crop of Breast Cancer Genes (Arul Chinnaiyan)		
		Health.com *n/a*	n/a	n/a
Online	6/2/2009	Heart drug may block breast cancer gene (Arul Chinnaiyan)		
		CNBC *n/a*	n/a	n/a
B'cast	6/1/2009 5:40 PM	Breast Cancer Gene		
		WDIV Detroit *n/a*	n/a	n/a
Online	6/1/2009	Hitting Where It Hurts: Exploiting Cancer Cell 'Addiction' ... (Daniel Rhodes)		
		Science Daily *n/a*	n/a	n/a
Online	6/1/2009	Research Finds 'New Crop' of Breast Cancer Genes (Arul Chinnaiyan)		
		MSN.com *n/a*	n/a	n/a
Online	6/1/2009	Research Finds 'New Crop' of Breast Cancer Genes (Arul Chinnaiyan)		
		U.S. News & World Report *n/a*	n/a	n/a
Online	6/1/2009	BP Drug Blocks Newly Found Breast Cancer Gene (Daniel Rhodes)		
		WebMD *n/a*	n/a	n/a
Online	6/1/2009	Research Finds 'New Crop' of Breast Cancer Genes (Arul Chinnaiyan)		
		Forbes *n/a*	n/a	n/a

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 (2×50 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 $\approx$<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.

Freely available online through the PNAS open access option.

[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.

CELL BIOLOGY

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-

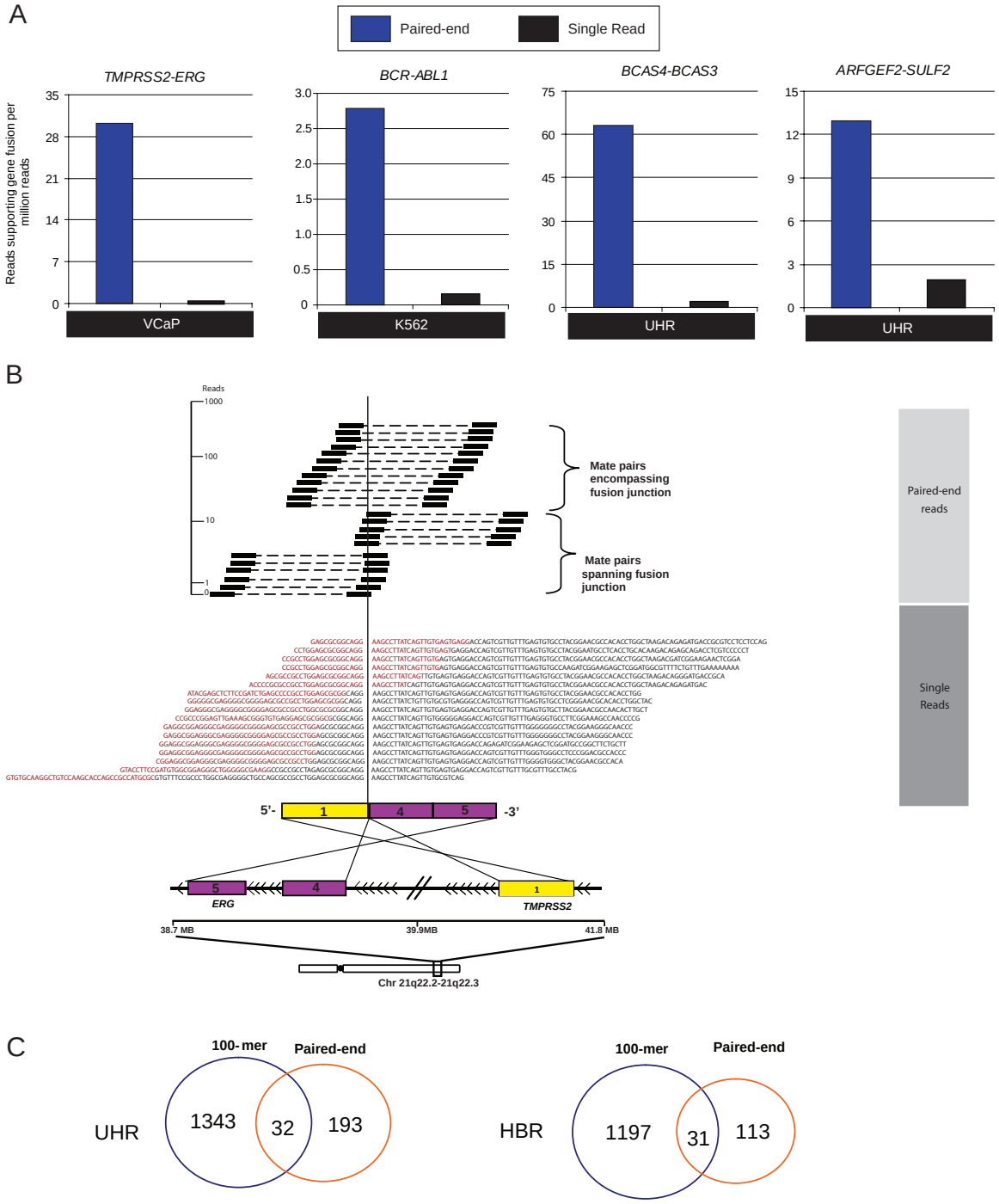


Fig. 1. Dynamic range and sensitivity of the paired-end transcriptome analysis relative to single read approaches. (*A*) Comparison of paired-end (blue) and long single transcriptome reads (black) supporting known gene fusions *TMPRSS2-ERG*, *BCR-ABL1*, *BCAS4-BCAS3*, and *ARFGEF2-SULF2*. (*B*) Schematic 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-

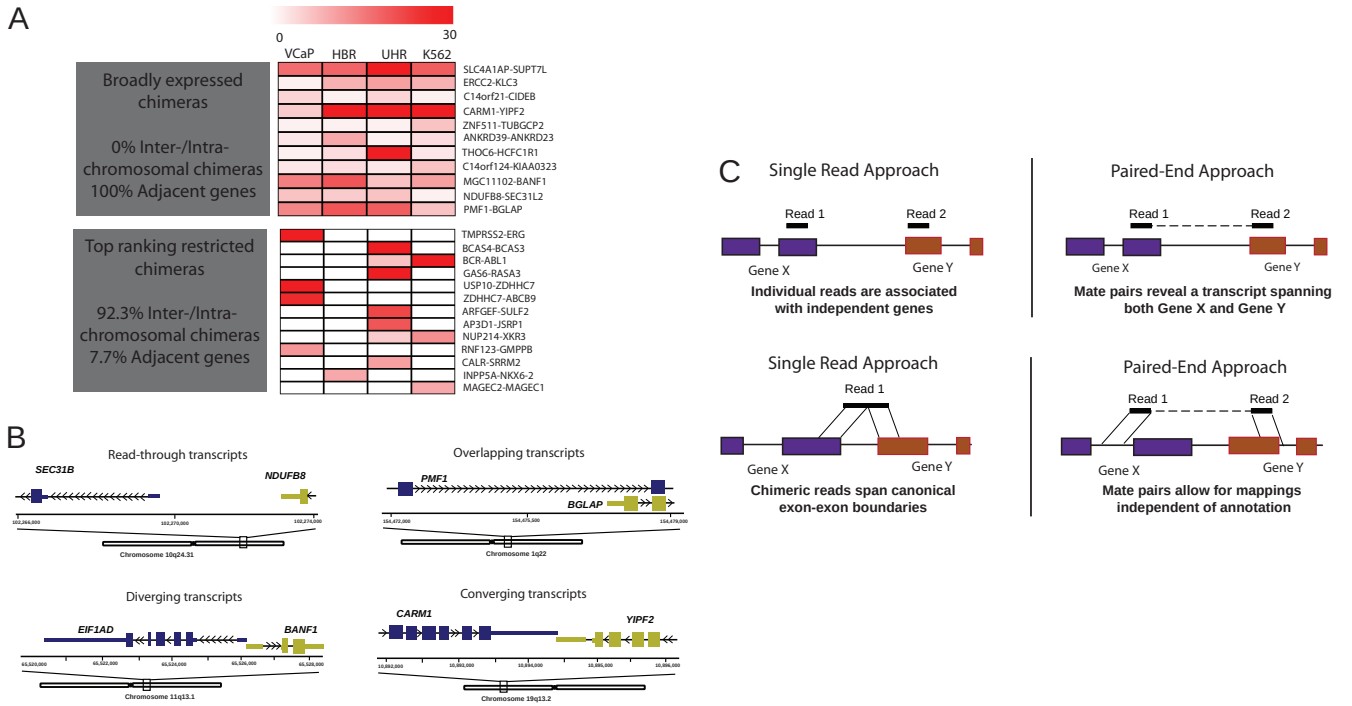


Fig. 2. RNA based chimeras. (*A*) Heatmaps showing the normalized number of reads supporting each read-through chimera across samples ranging from 0 (white) to 30 (red). (*Upper*) The heatmap highlights broadly expressed chimeras in UHR, HBR, VCaP, and K562. (*Lower*) The heatmap highlights the expression 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 ap-

proaches 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.

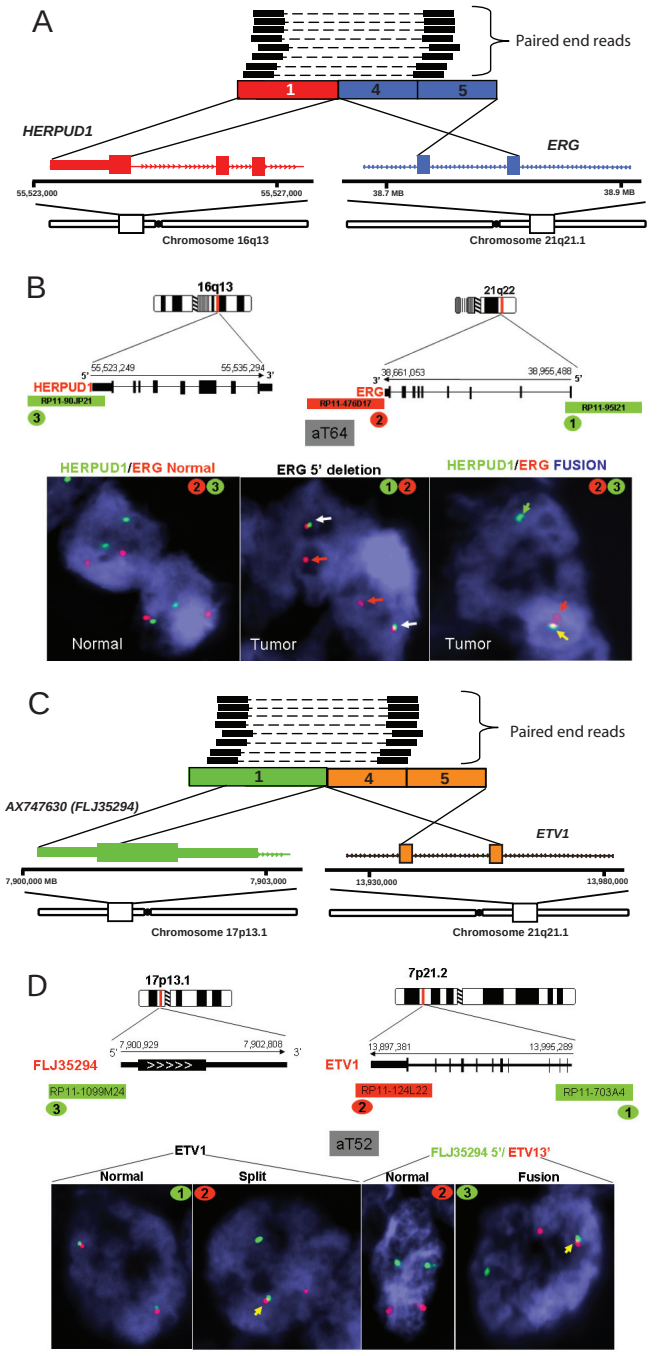


Fig. 3. Discovery of previously undescribed ETS gene fusions in localized prostate cancer. (*A*) Schematic representation of the interchromosomal gene fusion between exon 1 of *HERPUD1* (red), residing on chromosome 16, with exon 4 of *ERG* (blue), located on chromosome 21. (*B*) Schematic representation 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*).

CELL BIOLOGY

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.

Published in final edited form as:
Curr Opin Genet Dev. 2009 February ; 19(1): 82–91. doi:10.1016/j.gde.2008.11.008.

Oncogenic Gene Fusions in Epithelial Carcinomas

John R. Prensner[1] and **Arul M. Chinnaiyan**[1,2,3,4,5]

1*Michigan Center for Translational Pathology, University of Michigan Medical School, 1400 East Medical Center Drive, 5316 CCGC, Ann Arbor, MI 48109, USA*

2*Department of Pathology, University of Michigan Medical School, 1400 East Medical Center Drive, 5316 CCGC, Ann Arbor, MI 48109, USA*

3*Department of Urology, University of Michigan Medical School, 1400 East Medical Center Drive, 5316 CCGC, Ann Arbor, MI 48109, USA*

4*Howard Hughes Medical Institute, University of Michigan Medical School, 1400 East Medical Center Drive, 5316 CCGC, Ann Arbor, MI 48109, USA*

5*The Comprehensive Cancer Center, University of Michigan Medical School, 1400 East Medical Center Drive, 5316 CCGC, Ann Arbor, MI 48109, USA*

Summary of Recent Advances

New discoveries regarding recurrent chromosomal aberrations in epithelial tumors have challenged the view that gene fusions play a minor role in these cancers. It is now known that recurrent fusions characterize significant subsets of prostate, breast, lung and renal-cell carcinomas, among others. This work has generated new insights into the molecular subtypes of tumors and highlighted important advances in bioinformatics, sequencing and microarray technology as tools for gene fusion discovery. Given the ubiquity of tyrosine kinases and transcription factors in gene fusions, further interest in the potential “druggability” of gene fusions with targeted therapeutics has also flourished. Nevertheless, the majority of chromosomal abnormalities in epithelial cancers remain uncharacterized, underscoring the limitations of our knowledge of carcinogenesis and the requirement for further research.

Introduction

The intrigue of chromosomal aberrations in human cancers dates back over 90 years, when early theories about the molecular and genetic origins of cancer were first being discussed. Since then, the genetic basis of cancer has been well established to include certain fundamental tumorigenic processes that accrue within cancer cells: most prominently, chromosomal aberrations, nucleotide substitutions, epigenetic changes and post-transcriptional dysregulation of gene expression [1].

Correspondence to: Arul M. Chinnaiyan, arul@umich.edu.

Disclosure: The University of Michigan has filed for a patent on the recurrent gene fusions in prostate cancer and A.M.C. is named as a co-inventor. The technology has been licensed to Gen-Probe Inc. to be developed as a molecular diagnostic. A.M.C. is a consultant to Gen-Probe.

With more than 50,000 chromosomal alterations annotated in more than 11,500 publications [2], particular interest has focused on the tumorigenic potential of gene fusions. Historically these fusions have been mainly associated with hematological and mesenchymal malignancies. Despite over 440 known gene fusions in benign tumors and cancer [3], only ~15% of these and 10% of known recurrent breakpoint aberrations (RBAs) are found in epithelial tumors, of which only 35% have been characterized [4••]. By contrast, ~90% of known oncogenes are associated with somatic mutations [5]. The subsequent discovery of the TMPRSS2-Ets fusions in prostate cancer by our group [6••], and recurrent fusions lung cancer by others [7•,8•], has fueled investigations of the role of gene fusions in epithelial carcinomas. These findings suggest that numerous, undiscovered gene fusions may be lurking within the cancer genome. Here, we summarize the current state of gene fusions in epithelial cancers, highlighting the technologies that enabled these discoveries.

Historical Perspectives: Gene Fusions

Despite the current swell of interest in gene fusions, the seminal discovery in this field remains Nowell and Hungerford's identification, in 1960, of the BCR-ABL balanced translocation of the long arm of chromosome 22 to the short arm of chromosome 9 [9]. Resulting in constitutive activation of the Abl tyrosine kinase domains, the Bcr-Abl fusion protein is the driving force in chronic myelogenous leukemia (CML) [10,11]. By establishing a causal link between a specific chromosomal lesion and a specific malignancy, BCR-ABL also pioneered cancer therapy: the tyrosine kinase inhibitor, imatinib (Gleevec), was introduced as the first widely used targeted therapeutic [12].

Similar discoveries led to the characterization of causative fusions in a host of other hematological malignancies, including Burkitt's lymphoma, T Cell lymphomas and acute promyleocytic leukemia (AML), which harbors the retinoic acid-sensitive t(15;17) fusion of the transcription factor RARα to PML [13]. Moreover, gene fusions play important roles in many soft tissue tumors, where over 40 known gene fusions have been characterized [14].

Gene Fusions in Epithelial Cancers

As with hematological malignancies, gene fusions in epithelial cancers can be broadly classified into two main groups: the tyrosine kinase (TK) fusions and the transcription factor (TF) fusions. Together, they account for 50% of the genes found in gene fusions (Table 1) [14]. While the two may functionally overlap *in vivo*—TKs can lead to TF phosphorylation, and TFs can influence the expression of TK genes—this distinction is a useful to envision the two major architectural frameworks for fusion proteins.

Tyrosine Kinase Fusions

With BCR-ABL as the presiding paradigm, chromosomal aberrations that activate TKs, especially receptor TKs (RTKs), have long been a focus in cancer biology. Upon extracellular ligand-binding, RTKs activate intracellular signaling pathways by dimerization of the receptor subunits and autophosphorylation of the tyrosine residues [15]. Once initiated, TK activity can lead to numerous cellular responses including increased proliferation, growth, gene expression, and suppression of apoptotic pathways, among others (Figure 1).

Given their widespread functionality in cellular growth and proliferation pathways, it is logical that TKs are prominent 3' partners in oncogenic gene fusions. The 5' partners for such fusions, however, comprise a more variegated group. This is perhaps most readily illustrated by the numerous RET and NTRK1 TK fusions of papillary thyroid cancer. These were linked to at least seventeen total 5' fusion partners that commonly confer dimerization capability through leucine zipper or coiled-coil domains (Table 1) [16]. Interestingly, fusions of RET and NTRK1,

which account for ~50% of papillary thyroid cancers, tend to segregate both from each other [17] and from mutations in the cytosolic kinase BRAF, which is mutated in as much as 40% of thyroid cancers [18].

Recently, several reports have described gene fusions in non-small-cell lung cancers (NSCLC) [7•]. With a prevalence of approximately 5% [7•,19], the EML4-ALK fusion, which has been linked to cellular transformation [20], increased cellular growth and decreased apoptosis [21], defines a subset of NSCLCs, segregating from mutations in EGFR and appearing more commonly in non-smokers [7•]. Rivoka et al. further used phosphoproteomics to conduct a large-scale survey of oncogenic kinases to identify novel gene fusions TFG-ALK and CD74-ROS1 in patients with NSCLC [8•]. In another context, ROS1 has also previously been implicated in rare GOPC-ROS1 fusions in glioblastoma [22].

Transcription Factor Fusions

The story of TF fusions in epithelial cancers spans both the rare oncologic curiosities and the ubiquitous oncologic diseases. As with TK fusions, TFs often form multiple fusion genes by involving many different 5' partners. The MiTF gene family of TFs, for example, define a subset of pediatric papillary renal-cell carcinomas with eight known 5' partners [23,24]. Interestingly, TFE3 and TFEB, two functionally-redundant MiTF factors implicated in these fusions [25,26], contribute to activation of MET RTK signaling, illustrating the interaction between kinases and TFs [27].

Unlike TKs, however, disruption of TF function by gene fusions can cause a dominant-negative effect on the cell. Indeed, dysregulation of TFE3 and TREB leads to a loss of MAD2B-controlled mitotic checkpoint regulation and disruption of tissue-specific development [28, 29]. Moreover, PAX8-PPARγ fusions, found in ~50% of follicular thyroid cancer (FTC) [30] and the follicular variant of papillary thyroid cancer (FVPTC) β1], disrupt PPARγ activation, leading to dysregulated cell-cycle transitions, decreased apoptosis, and cellular transformation [32,33]. Surprisingly, the PAX8-PPARγ fusion, which is overexpressed in fusion-positive tumors, is associated with less aggressive tumor features and a better clinical outcome [34, 35].

Elsewhere, the clinical outcome associated with cancers harboring gene fusions is less sanguine. Rare but poorly differentiated pediatric carcinomas of midline structures, such as those in the head, neck and thorax, possess a distinctive t(15;19) BRD4-NUT fusion [36,37]. Likewise, in secretory breast cancer, a rare form of ductal carcinoma, the recurrent ETV6-NTRK3 fusion has been implicated in increased cellular viability and aberrant cell-cycle progression [38]. Moreover, mucoepidermoid carcinoma and pleiomorphic adenoma, the most common malignant and benign tumors of the salivary glands, respectively, both manifest prominent tumorigenic gene fusions [39-41].

Prostate Cancer

In 2005, our group described recurrent fusions between the Ets family TFs, ERG and ETV1, and the androgen-regulated transmembrane serine protease, TMPRSS2, in prostate cancer [6••]. Subsequently, multiple other 5' fusion partners have been described for ERG and ETV1, as well as other members of the Ets family (Table 2) [6••,42•]. The first major solid cancer to reveal such findings, roughly 60% of prostate cancers harbor a known fusion, of which 80-90% are TMPRSS2-ERG fusions [6••,43-45]. Because ERG and TMPRSS2 reside on the same region of chromosome 21, two mechanisms—an intrachromosomal deletion and an inversion —are implicated in their creation, though ultimately TMPRSS2 contributes only untranslated sequences to the final mRNA transcript [6••,46].

Following this discovery, TMPRSS2-Ets fusions have emerged as a major factor in prostate tumorigenesis, contributing to cellular invasiveness *in vitro* [42•,43,44]. While these fusions are common in pre-malignant prostate lesions [47-49], they are insufficient for the initiation of carcinogenesis in mouse models [46]. Given these data, we and others have posited that ERG cooperates with other early genomic alterations in prostate cancer, such as loss of the tumor suppressor PTEN, to induce an invasive phenotype [46].

Clinically, TMPRSS2-ERG fusions have also been correlated with a poorer prognosis and an increased risk of disease recurrence [50-54], although some discordant results have been found [55]. In this regard, analyzing the clinical impact of these fusions is complicated by the multifocal nature of prostate cancer [56], and recent reports show that the status of TMPRSS2-Ets fusions may be inconsistent in up to 70% of multifocal tumors [57,58]. Given the heterogenous nature of many epithelial cancers, the detection and analysis of gene fusions in other major carcinomas may be impeded by similar complications of multifocality and clonal heterogeneity.

Advances in Gene Fusion Discovery

Our lab has developed several new methodologies for the identification and analysis of gene fusion candidates. In combination with mainstay wet-lab techniques such as fluorescence *in-situ* hybridization (FISH), our research incorporates computational and bioinformatic approaches to gene fusion biology, including cancer outlier profile analysis (COPA) [6••], the microarray compendium Oncomine [59•] and Molecular Concepts Mapping (MCM) analysis [60] (Box 1).

Box 1: Bioinformatic Gene Fusion Analysis

Discovery of fusions by gene expression microarrays often depends on the upregulation of the chimeric transcript or 3' functional end, which can be detected as an outlier. To analyze such data, our lab has developed three core bioinformatic tools.

COPA

Cancer Outlier Profile Analysis (COPA) highlights differential expression of genes screened with microarrays [6••]. By median-centering microarray data, COPA enhances the visibility of outlier genes, which may be candidate gene fusions.

Oncomine

With over 18,000 microarray experiments across 35 tumor types, Oncomine is a compendium used to corroborate expression data across multiple datasets, thereby decreasing the problem of false positives in any given microarray [59•]. Oncomine also visualizes expression data with features such as interactome analysis and Molecular Concepts analysis [59•].

MCM

The Molecular Concepts Map (MCM) nominates potential interactions between biological phenomena within cancer cells [60]. By combining data from Oncomine [59•] and the Connectivity Map [61••], MCM predicts mechanistic pathways, molecular characteristics, and interaction networks for candidate gene fusions.

Recently, other groups have developed new methods to analyze transcriptome and gene expression data. Lamb et al. have devised a bioinformatic tool to predict and nominate interactions between small molecule compounds and human tumors based upon microarray

expression analysis [61••]. This Connectivity Map offers a new paradigm for tumor-specific therapeutics.

To facilitate fusion discovery, Hahn et al. designed an algorithmic approach to query mRNA and expressed sequence tag (EST) databases for incongruous transcript sequences [62], and they nominated 20 putative recurrent fusion genes. While their approach has limitations—for example, they identified only 6 of 22 known Bcr-Abl fusion mRNAs—their findings offer intriguing insight into methods for identifying fusion genes.

Next Generation Sequencing

Recently, high-throughput "massively-parallel" sequencing platforms, including Roche/454, Applied Biosystems/SOLiD and Illumina/Solexa, have provided researchers with tantalizing new tools to study gene fusions. The depth of coverage offered by these platforms permits genome-wide and transcriptome-wide sequencing on a scale not previously feasible. Already, studies analyzing human transcriptomes [63•] and chromosomal breakpoints [64] demonstrate the utility of such modalities. The use of paired-end sequencing, which combines fragmented sample gDNA flanked with known reference sequences, is also a promising method for fusion discovery [65].

Next-generation sequencing platforms, however, also present challenges. Adaptor ligation steps may increase numbers of false positive fusion reads. Genome fragmentation into 30 to 300 bp segments (as compared to 900 bp for capillary sequencing) makes sequence re-assembly more challenging. The sheer volume of sequencing data makes bioinformatic and computational analysis difficult.

To this end, our lab has developed bioinformatic methods to categorize putative gene fusions and eliminate false positive reads. Combining longer ~300 bp reads from Roche/454 with 30-40 bp reads from Illumina/Solexa yields more specific results than either technology alone, allowing the identification of novel gene fusions in prostate cancer (Maher CA et al., in submission). As these technologies become more common, it is likely that many more gene fusions will be identified in this manner. Nevertheless, finding clinically significant, recurrent gene fusions remains challenging, and thus better paradigms may be required to combine these technologies with standard wet-lab techniques in fusion discovery.

Challenges and Future Directions

Significant obstacles still hinder genome and transcriptome analysis. Epithelial cancers, unlike many hematological cancers, frequently display highly aberrant karyotypes that are difficult to characterize cytogenetically. Clonal heterogeneity is common is epithelial cancers, with up to 80% of carcinomas harboring unrelated clones [66,67]. Finally, with the explosion of microarray data in the past decade, databases have been flooded with potential genomic, epigenetic, and transcriptomic aberrations in cancer. Isolating seminal events in tumorigenesis from such volumes is challenging, as false positives remain problematic.

Moving forward, it may be argued that the focus on fusions involving kinases and transcription factors is too narrow. It may be possible that "non-traditional" gene fusions involving protein-folding chaperones and cellular localization proteins, among others, are prominent in certain epithelial cancers. Such bias may partly resolve as computational tools, sequencing technologies and array-based assays become more powerful and precise. With the increased ability to interrogate the genome, putative gene fusions may be detected in a less biased manner. Clinically, this may result in the discovery of "non-traditional" gene fusions that—like BCR-ABL—serve as candidates for targeted therapy (Figure 2). Moreover, fusion transcripts may contribute to novel, non-invasive diagnostics if shed in the urine or detectable in blood serum.

Already, non-invasive clinical tests for TMPRSS2-ERG transcripts are under investigation [68].

Conversely, the clinical picture generated by fusions in epithelial cancers is unclear. Indeed, some fusions, such as PAX8-PPARγ, counter-intuitively seem to characterize less aggressive disease. Yet, data in prostate cancer indicates that fusions may, in fact, define clinically important cancer subtypes. TMPRSS2-ERG fusions generated by intrachromosomal deletions, for example, tend to correspond with worse prognoses than those created by inversions [69]. Additionally, some fusion-based carcinomas are more prominent in pediatric populations, including renal-cell, thyroid, and aggressive midline carcinomas. As research progresses, such epidemiological and demographical data may allow for more specific applications of gene fusion-based targeted therapy.

Conclusions

Long considered a phenomenon of hematological and mesenchymal cancers, gene fusions are now emerging as an important component in epithelial carcinogenesis. With epithelial cancers accounting for 90% of all malignancies and 80% of cancer-related deaths [4••,14], new discoveries, particularly in breast, prostate, lung, and renal-cell carcinomas, show that recurrent gene fusions are widespread across epithelial cancers. Although much work is still needed, new technologies in sequencing, microarrays and bioinformatics hold promise for gene fusion discovery and facilitate the characterization of recurrent gene fusions in major epithelial cancers.

Acknowledgements

We thank Jill Granger for her critical reading of this manuscript. We thank the Howard Hughes Medical Institute, the Burroughs Welcome Foundation, the Early Detection Research Network, the Prostate Cancer Foundation, the U.S. Department of Defense, the National Institutes of Health, and the Specialized Program in Research Excellence (SPORE) program. J.R. Prensner is a fellow in the University of Michigan Medical Scientist Training Program.

References and Recommended Reading

Papers of particular interested, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Hanahan D, Weinberg RA. The hallmarks of cancer. Cell 2000;100:57–70. [PubMed: 10647931]

2. Mitelman F, J B, Mertens F. Mitelman Database of Chromosome Aberrations in Cancer [online]. 2008Edited by

3. Heim S, Mitelman F. Molecular screening for new fusion genes in cancer. Nat Genet 2008;40:685–686. [PubMed: 18509307]

••4. Mitelman F, Mertens F, Johansson B. Prevalence estimates of recurrent balanced cytogenetic aberrations and gene fusions in unselected patients with neoplastic disorders. Genes Chromosomes Cancer 2005;43:350–366.366 [PubMed: 15880352]. The most exhaustive study of chromosomal aberrations across cancer types The authors catalogue and analyze an immense amount of data to distill the most accurate account of chromosomal abnormalities in cancer available.

5. Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, Rahman N, Stratton MR. A census of human cancer genes. Nat Rev Cancer 2004;4:177–183. [PubMed: 14993899]

••6. Tomlins SA, Rhodes DR, Perner S, Dhanasekaran SM, Mehra R, Sun XW, Varambally S, Cao X, Tchinda J, Kuefer R, et al. Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. Science 2005;310:644–648.648 [PubMed: 16254181]. Using novel bioinformatic techniques, the authors describe the first example of a highly prevalent oncogenic gene fusion in a

highly prevalent and lethal epithelial cancer The authors use gene expression arrays to analyze outliers and identify gene fusions This paper reinvigorated research efforts to find fusion events in carcinomas.

•7. Soda M, Choi YL, Enomoto M, Takada S, Yamashita Y, Ishikawa S, Fujiwara S, Watanabe H, Kurashina K, Hatanaka H, et al. Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer. Nature 2007;448:561–566.566 [PubMed: 17625570]. This paper describes the first recurrent gene fusion in non-small-cell lung cancer It identifies a transforming tyrosine kinase fusion that may serve as a promising candidate for targeted therapy.

•8. Rikova K, Guo A, Zeng Q, Possemato A, Yu J, Haack H, Nardone J, Lee K, Reeves C, Li Y, et al. Global survey of phosphotyrosine signaling identifies oncogenic kinases in lung cancer. Cell 2007;131:1190–1203.1203 [PubMed: 18083107]. Employing a phosphoproteomic approach, the authors study phosphotyrosine signaling in over 150 non-small-cell lung cancers and 40 similar cell lines The comprehensive findings give an informatic picture of tyrosine kinase signaling in lung cancer, including both point mutations and nove gene fusions.

9. Nowell PC, Hungerford DA. Chromosome studies on normal and leukemic human leukocytes. J Natl Cancer Inst 1960;25:85–109. [PubMed: 14427847]

10. Lugo TG, Pendergast AM, Muller AJ, Witte ON. Tyrosine kinase activity and transformation potency of bcr-abl oncogene products. Science 1990;247:1079–1082. [PubMed: 2408149]

11. Daley GQ, Van Etten RA, Baltimore D. Induction of chronic myelogenous leukemia in mice by the P210bcr/abl gene of the Philadelphia chromosome. Science 1990;247:824–830. [PubMed: 2406902]

12. Druker BJ, Talpaz M, Resta DJ, Peng B, Buchdunger E, Ford JM, Lydon NB, Kantarjian H, Capdeville R, Ohno-Jones S, et al. Efficacy and safety of a specific inhibitor of the BCR-ABL tyrosine kinase in chronic myeloid leukemia. N Engl J Med 2001;344:1031–1037. [PubMed: 11287972]

13. Goddard AD, Borrow J, Freemont PS, Solomon E. Characterization of a zinc finger gene disrupted by the t(15;17) in acute promyelocytic leukemia. Science 1991;254:1371–1374. [PubMed: 1720570]

14. Mitelman F, Johansson B, Mertens F. Fusion genes and rearranged genes as a linear function of chromosome aberrations in cancer. Nat Genet 2004;36:331–334. [PubMed: 15054488]

15. Blume-Jensen P, Hunter T. Oncogenic kinase signalling. Nature 2001;411:355–365. [PubMed: 11357143]

16. Nikiforov YE. Thyroid carcinoma: molecular pathways and therapeutic targets. Mod Pathol 2008;21 (Suppl 2):S37–43. [PubMed: 18437172]

17. Bongarzone I, Vigneri P, Mariani L, Collini P, Pilotti S, Pierotti MA. RET/NTRK1 rearrangements in thyroid gland tumors of the papillary carcinoma family: correlation with clinicopathological features. Clin Cancer Res 1998;4:223–228. [PubMed: 9516975]

18. Soares P, Trovisco V, Rocha AS, Lima J, Castro P, Preto A, Maximo V, Botelho T, Seruca R, Sobrinho-Simoes M. BRAF mutations and RET/PTC rearrangements are alternative events in the etiopathogenesis of PTC. Oncogene 2003;22:4578–4580. [PubMed: 12881714]

19. Perner S, Wagner PL, Demichelis F, Mehra R, Lafargue CJ, Moss BJ, Arbogast S, Soltermann A, Weder W, Giordano TJ, et al. EML4-ALK fusion lung cancer: a rare acquired event. Neoplasia 2008;10:298–302. [PubMed: 18320074]

20. Choi YL, Takeuchi K, Soda M, Inamura K, Togashi Y, Hatano S, Enomoto M, Hamada T, Haruta H, Watanabe H, et al. Identification of novel isoforms of the EML4-ALK transforming gene in non-small cell lung cancer. Cancer Res 2008;68:4971–4976. [PubMed: 18593892]

21. Koivunen JP, Mermel C, Zejnullahu K, Murphy C, Lifshits E, Holmes AJ, Choi HG, Kim J, Chiang D, Thomas R, et al. EML4-ALK fusion gene and efficacy of an ALK kinase inhibitor in lung cancer. Clin Cancer Res 2008;14:4275–4283. [PubMed: 18594010]

22. Charest A, Lane K, McMahon K, Park J, Preisinger E, Conroy H, Housman D. Fusion of FIG to the receptor tyrosine kinase ROS in a glioblastoma with an interstitial del(6)(q21q21). Genes Chromosomes Cancer 2003;37:58–71. [PubMed: 12661006]

23. Medendorp K, van Groningen JJ, Schepens M, Vreede L, Thijssen J, Schoenmakers EF, van den Hurk WH, Geurts van Kessel A, Kuiper RP. Molecular mechanisms underlying the MiT translocation subgroup of renal cell carcinomas. Cytogenet Genome Res 2007;118:157–165. [PubMed: 18000366]

24. Argani P, Ladanyi M. Translocation carcinomas of the kidney. Clin Lab Med 2005;25:363–378. [PubMed: 15848741]

25. Huan C, Kelly ML, Steele R, Shapira I, Gottesman SR, Roman CA. Transcription factors TFE3 and TFEB are critical for CD40 ligand expression and thymus-dependent humoral immunity. Nat Immunol 2006;7:1082–1091. [PubMed: 16936731]
26. Steingrimsson E, Tessarollo L, Pathak B, Hou L, Arnheiter H, Copeland NG, Jenkins NA. Mitf and Tfe3, two members of the Mitf-Tfe family of bHLH-Zip transcription factors, have important but functionally redundant roles in osteoclast development. Proc Natl Acad Sci U S A 2002;99:4477–4482. [PubMed: 11930005]
27. Tsuda M, Davis IJ, Argani P, Shukla N, McGill GG, Nagai M, Saito T, Lae M, Fisher DE, Ladanyi M. TFE3 fusions activate MET signaling by transcriptional up-regulation, defining another class of tumors as candidates for therapeutic MET inhibition. Cancer Res 2007;67:919–929. [PubMed: 17283122]
28. Mathur M, Samuels HH. Role of PSF-TFE3 oncoprotein in the development of papillary renal cell carcinomas. Oncogene 2007;26:277–283. [PubMed: 16832349]
29. Weterman MA, van Groningen JJ, Tertoolen L, van Kessel AG. Impairment of MAD2B-PRCC interaction in mitotic checkpoint defective t(X;1)-positive renal cell carcinomas. Proc Natl Acad Sci U S A 2001;98:13808–13813. [PubMed: 11717438]
30. Kroll TG, Sarraf P, Pecciarini L, Chen CJ, Mueller E, Spiegelman BM, Fletcher JA. PAX8-PPARgamma1 fusion oncogene in human thyroid carcinoma [corrected]. Science 2000;289:1357–1360. [PubMed: 10958784]
31. Castro P, Rebocho AP, Soares RJ, Magalhaes J, Roque L, Trovisco V, Vieira de Castro I, Cardoso-de-Oliveira M, Fonseca E, Soares P, et al. PAX8-PPARgamma rearrangement is frequently detected in the follicular variant of papillary thyroid carcinoma. J Clin Endocrinol Metab 2006;91:213–220. [PubMed: 16219715]
32. Gregory Powell J, Wang X, Allard BL, Sahin M, Wang XL, Hay ID, Hiddinga HJ, Deshpande SS, Kroll TG, Grebe SK, et al. The PAX8/PPARgamma fusion oncoprotein transforms immortalized human thyrocytes through a mechanism probably involving wild-type PPARgamma inhibition. Oncogene 2004;23:3634–3641. [PubMed: 15077183]
33. Au AY, McBride C, Wilhelm KG Jr, Koenig RJ, Speller B, Cheung L, Messina M, Wentworth J, Tasevski V, Learoyd D, et al. PAX8-peroxisome proliferator-activated receptor gamma (PPARgamma) disrupts normal PAX8 or PPARgamma transcriptional function and stimulates follicular thyroid cell growth. Endocrinology 2006;147:367–376. [PubMed: 16179407]
34. Sahin M, Allard BL, Yates M, Powell JG, Wang XL, Hay ID, Zhao Y, Goellner JR, Sebo TJ, Grebe SK, et al. PPARgamma staining as a surrogate for PAX8/PPARgamma fusion oncogene expression in follicular neoplasms: clinicopathological correlation and histopathological diagnostic value. J Clin Endocrinol Metab 2005;90:463–468. [PubMed: 15483076]
35. Marques AR, Espadinha C, Frias MJ, Roque L, Catarino AL, Sobrinho LG, Leite V. Underexpression of peroxisome proliferator-activated receptor (PPAR)gamma in PAX8/PPARgamma-negative thyroid tumours. Br J Cancer 2004;91:732–738. [PubMed: 15238980]
36. French CA, Miyoshi I, Kubonishi I, Grier HE, Perez-Atayde AR, Fletcher JA. BRD4-NUT fusion oncogene: a novel mechanism in aggressive carcinoma. Cancer Res 2003;63:304–307. [PubMed: 12543779]
37. French CA, Miyoshi I, Aster JC, Kubonishi I, Kroll TG, Dal Cin P, Vargas SO, Perez-Atayde AR, Fletcher JA. BRD4 bromodomain gene rearrangement in aggressive carcinoma with translocation t(15;19). Am J Pathol 2001;159:1987–1992. [PubMed: 11733348]
38. Tognon C, Knezevich SR, Huntsman D, Roskelley CD, Melnyk N, Mathers JA, Becker L, Carneiro F, MacPherson N, Horsman D, et al. Expression of the ETV6-NTRK3 gene fusion as a primary event in human secretory breast carcinoma. Cancer Cell 2002;2:367–376. [PubMed: 12450792]
39. Tonon G, Modi S, Wu L, Kubo A, Coxon AB, Komiya T, O'Neil K, Stover K, El-Naggar A, Griffin JD, et al. t(11;19)(q21;p13) translocation in mucoepidermoid carcinoma creates a novel fusion product that disrupts a Notch signaling pathway. Nat Genet 2003;33:208–213. [PubMed: 12539049]
40. Stenman G. Fusion oncogenes and tumor type specificity--insights from salivary gland tumors. Semin Cancer Biol 2005;15:224–235. [PubMed: 15826837]

41. Kas K, Voz ML, Roijer E, Astrom AK, Meyen E, Stenman G, Van de Ven WJ. Promoter swapping between the genes for a novel zinc finger protein and beta-catenin in pleiomorphic adenomas with t (3;8)(p21;q12) translocations. Nat Genet 1997;15:170–174. [PubMed: 9020842]

•42. Tomlins SA, Laxman B, Dhanasekaran SM, Helgeson BE, Cao X, Morris DS, Menon A, Jing X, Cao Q, Han B, et al. Distinct classes of chromosomal rearrangements create oncogenic ETS gene fusions in prostate cancer. Nature 2007;448:595–599.599 [PubMed: 17671502]. By characterizing multiple types of gene fusions in prostate cancer, the authors show that gene fusions can describe specific molecular subtypes, like those in hematological cancers, in epithelial cancers.

43. Perner S, Demichelis F, Beroukhim R, Schmidt FH, Mosquera JM, Setlur S, Tchinda J, Tomlins SA, Hofer MD, Pienta KG, et al. TMPRSS2:ERG fusion-associated deletions provide insight into the heterogeneity of prostate cancer. Cancer Res 2006;66:8337–8341. [PubMed: 16951139]

44. Soller MJ, Isaksson M, Elfving P, Soller W, Lundgren R, Panagopoulos I. Confirmation of the high frequency of the TMPRSS2/ERG fusion gene in prostate cancer. Genes Chromosomes Cancer 2006;45:717–719. [PubMed: 16575875]

45. Yoshimoto M, Joshua AM, Chilton-Macneill S, Bayani J, Selvarajah S, Evans AJ, Zielenska M, Squire JA. Three-color FISH analysis of TMPRSS2/ERG fusions in prostate cancer indicates that genomic microdeletion of chromosome 21 is associated with rearrangement. Neoplasia 2006;8:465–469. [PubMed: 16820092]

46. Tomlins SA, Laxman B, Varambally S, Cao X, Yu J, Helgeson BE, Cao Q, Prensner JR, Rubin MA, Shah RB, et al. Role of the TMPRSS2-ERG gene fusion in prostate cancer. Neoplasia 2008;10:177–188. [PubMed: 18283340]

47. Perner S, Mosquera JM, Demichelis F, Hofer MD, Paris PL, Simko J, Collins C, Bismar TA, Chinnaiyan AM, De Marzo AM, et al. TMPRSS2-ERG fusion prostate cancer: an early molecular event associated with invasion. Am J Surg Pathol 2007;31:882–888. [PubMed: 17527075]

48. Cerveira N, Ribeiro FR, Peixoto A, Costa V, Henrique R, Jeronimo C, Teixeira MR. TMPRSS2-ERG gene fusion causing ERG overexpression precedes chromosome copy number changes in prostate carcinomas and paired HGPIN lesions. Neoplasia 2006;8:826–832. [PubMed: 17032499]

49. Mosquera JM, Perner S, Genega EM, Sanda M, Hofer MD, Mertz KD, Paris PL, Simko J, Bismar TA, Ayala G, et al. Characterization of TMPRSS2-ERG fusion high-grade prostatic intraepithelial neoplasia and potential clinical implications. Clin Cancer Res 2008;14:3380–3385. [PubMed: 18519767]

50. Rajput AB, Miller MA, De Luca A, Boyd N, Leung S, Hurtado-Coll A, Fazli L, Jones EC, Palmer JB, Gleave ME, et al. Frequency of the TMPRSS2:ERG gene fusion is increased in moderate to poorly differentiated prostate cancers. J Clin Pathol 2007;60:1238–1243. [PubMed: 17259299]

51. Yoshimoto M, Joshua AM, Cunha IW, Coudry RA, Fonseca FP, Ludkovski O, Zielenska M, Soares FA, Squire JA. Absence of TMPRSS2:ERG fusions and PTEN losses in prostate cancer is associated with a favorable outcome. Mod Pathol. 2008

52. Setlur SR, Mertz KD, Hoshida Y, Demichelis F, Lupien M, Perner S, Sboner A, Pawitan Y, Andren O, Johnson LA, et al. Estrogen-dependent signaling in a molecularly distinct subclass of aggressive prostate cancer. J Natl Cancer Inst 2008;100:815–825. [PubMed: 18505969]

53. Demichelis F, Fall K, Perner S, Andren O, Schmidt F, Setlur SR, Hoshida Y, Mosquera JM, Pawitan Y, Lee C, et al. TMPRSS2:ERG gene fusion associated with lethal prostate cancer in a watchful waiting cohort. Oncogene 2007;26:4596–4599. [PubMed: 17237811]

54. Attard G, Clark J, Ambroisine L, Fisher G, Kovacs G, Flohr P, Berney D, Foster CS, Fletcher A, Gerald WL, et al. Duplication of the fusion of TMPRSS2 to ERG sequences identifies fatal human prostate cancer. Oncogene. 2007

55. Saramaki OR, Harjula AE, Martikainen PM, Vessella RL, Tammela TL, Visakorpi T. TMPRSS2:ERG fusion identifies a subgroup of prostate cancers with a favorable prognosis. Clin Cancer Res 2008;14:3395–3400. [PubMed: 18519769]

56. Arora R, Koch MO, Eble JN, Ulbright TM, Li L, Cheng L. Heterogeneity of Gleason grade in multifocal adenocarcinoma of the prostate. Cancer 2004;100:2362–2366. [PubMed: 15160339]

57. Furusato B, Gao CL, Ravindranath L, Chen Y, Cullen J, McLeod DG, Dobi A, Srivastava S, Petrovics G, Sesterhenn IA. Mapping of TMPRSS2-ERG fusions in the context of multi-focal prostate cancer. Mod Pathol. 2007

58. Mehra R, Han B, Tomlins SA, Wang L, Menon A, Wasco MJ, Shen R, Montie JE, Chinnaiyan AM, Shah RB. Heterogeneity of TMPRSS2 gene rearrangements in multifocal prostate adenocarcinoma: molecular evidence for an independent group of diseases. Cancer Res 2007;67:7991–7995. [PubMed: 17804708]

•59. Rhodes DR, Kalyana-Sundaram S, Mahavisno V, Varambally R, Yu J, Briggs BB, Barrette TR, Anstet MJ, Kincead-Beal C, Kulkarni P, et al. Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles. Neoplasia 2007;9:166–180.180 [PubMed: 17356713]. An update of the original Oncomine paper, the authors illustrate the utility of bioinformatics in organizing and analyzing microarray data By grouping microarray datasets into a compendium, the authors developed a bioinformatic tool to interrogate the underlying molecular events in carcinogenesis as well as visualizing molecular interactions Since its launch, Oncomine has become a standard bioinformatic tool.

60. Rhodes DR, Kalyana-Sundaram S, Tomlins SA, Mahavisno V, Kasper N, Varambally R, Barrette TR, Ghosh D, Varambally S, Chinnaiyan AM. Molecular concepts analysis links tumors, pathways, mechanisms, and drugs. Neoplasia 2007;9:443–454. [PubMed: 17534450]

••61. Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, Lerner J, Brunet JP, Subramanian A, Ross KN, et al. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. Science 2006;313:1929–1935.1935 [PubMed: 17008526]. Using small molecule libraries and gene expression data, the authors created a bioinformatic tool to use gene expression profiles to match the effects and applicability of specific small molecule inhibitors on specific tumors By coupling therapeutics with tumor gene expression profiles, this paper provides a powerful model for nominating future cancer therapies.

62. Hahn Y, Bera TK, Gehlhaus K, Kirsch IR, Pastan IH, Lee B. Finding fusion genes resulting from chromosome rearrangement by analyzing the expressed sequence databases. Proc Natl Acad Sci U S A 2004;101:13257–13261. [PubMed: 15326299]

•63. Sultan M, Schulz MH, Richard H, Magen A, Klingenhoff A, Scherf M, Seifert M, Borodina T, Soldatov A, Parkhomchuk D, et al. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. Science 2008;321:956–960.960 [PubMed: 18599741]. Using next-generation sequencing platforms, the authors survey the human transcriptome in human embryonic kidney cells and B cells As the most comprehensive description of the human transcriptome, they find large numbers of transcript-level events not previously reported This study demonstrates the potential of next-generation sequencing in characterizing the transcriptomes of human cancers.

64. Chen W, Kalscheuer V, Tzschach A, Menzel C, Ullmann R, Schulz MH, Erdogan F, Li N, Kijas Z, Arkesteijn G, et al. Mapping translocation breakpoints by next-generation sequencing. Genome Res 2008;18:1143–1149. [PubMed: 18326688]

65. Bashir A, Volik S, Collins C, Bafna V, Raphael BJ. Evaluation of paired-end sequencing strategies for detection of genome rearrangements in cancer. PLoS Comput Biol 2008;4:e1000051. [PubMed: 18404202]

66. Gasparini P, Sozzi G, Pierotti MA. The role of chromosomal alterations in human cancer development. J Cell Biochem 2007;102:320–331. [PubMed: 17722107]

67. Gorunova L, Hoglund M, Andren-Sandberg A, Dawiskiba S, Jin Y, Mitelman F, Johansson B. Cytogenetic analysis of pancreatic carcinomas: intratumor heterogeneity and nonrandom pattern of chromosome aberrations. Genes Chromosomes Cancer 1998;23:81–99. [PubMed: 9739011]

68. Hessels D, Smit FP, Verhaegh GW, Witjes JA, Cornel EB, Schalken JA. Detection of TMPRSS2-ERG fusion transcripts and prostate cancer antigen 3 in urinary sediments may improve diagnosis of prostate cancer. Clin Cancer Res 2007;13:5103–5108. [PubMed: 17785564]

69. Attard G, Clark J, Ambroisine L, Fisher G, Kovacs G, Flohr P, Berney D, Foster CS, Fletcher A, Gerald WL, et al. Duplication of the fusion of TMPRSS2 to ERG sequences identifies fatal human prostate cancer. Oncogene 2008;27:253–263. [PubMed: 17637754]

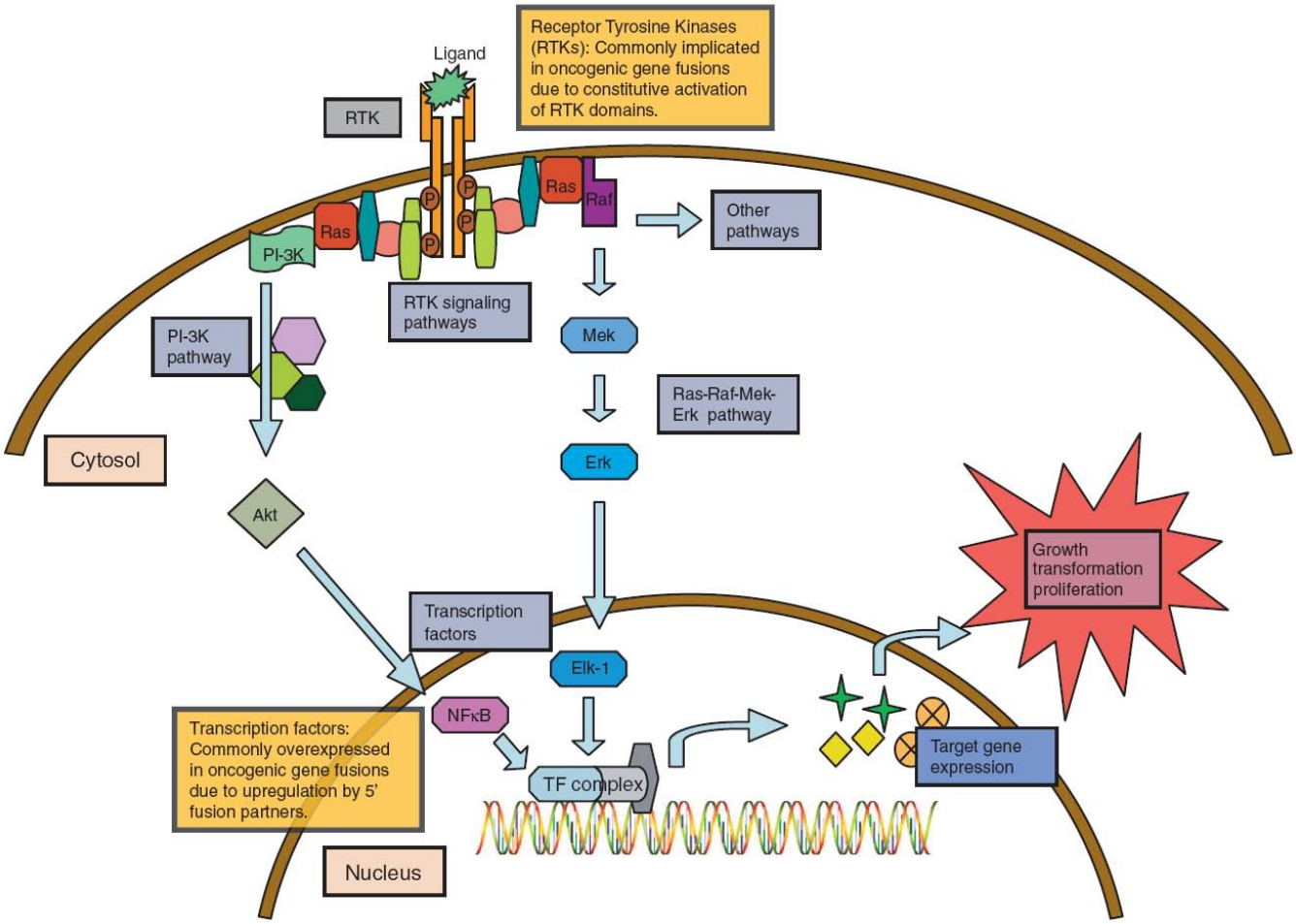


Figure 1. Biochemical Pathways in Gene Fusions

Biochemical effects of gene fusions cluster around tyrosine kinase (TK) signaling pathways, which alter the activity of intracellular proteins, and transcription factor (TF) activity, which control gene expression at the DNA level. Here we outline the examples of the Ras and PI-3K pathways, which are commonly involved downstream of TK activation and are frequently implicated in the oncogenic effects of gene fusions. PI-3K works via increased activity of the master regulator Akt, which controls many cellular processes including the nuclear TF NfκB. Likewise, the Ras-Raf-Mek-Erk pathway promotes activation of TFs, including Elk-1, which is a target of Erk. These signaling pathways and gene expression signatures result in the phenotypic qualities, such as invasiveness and increased proliferation, observed in cancers.

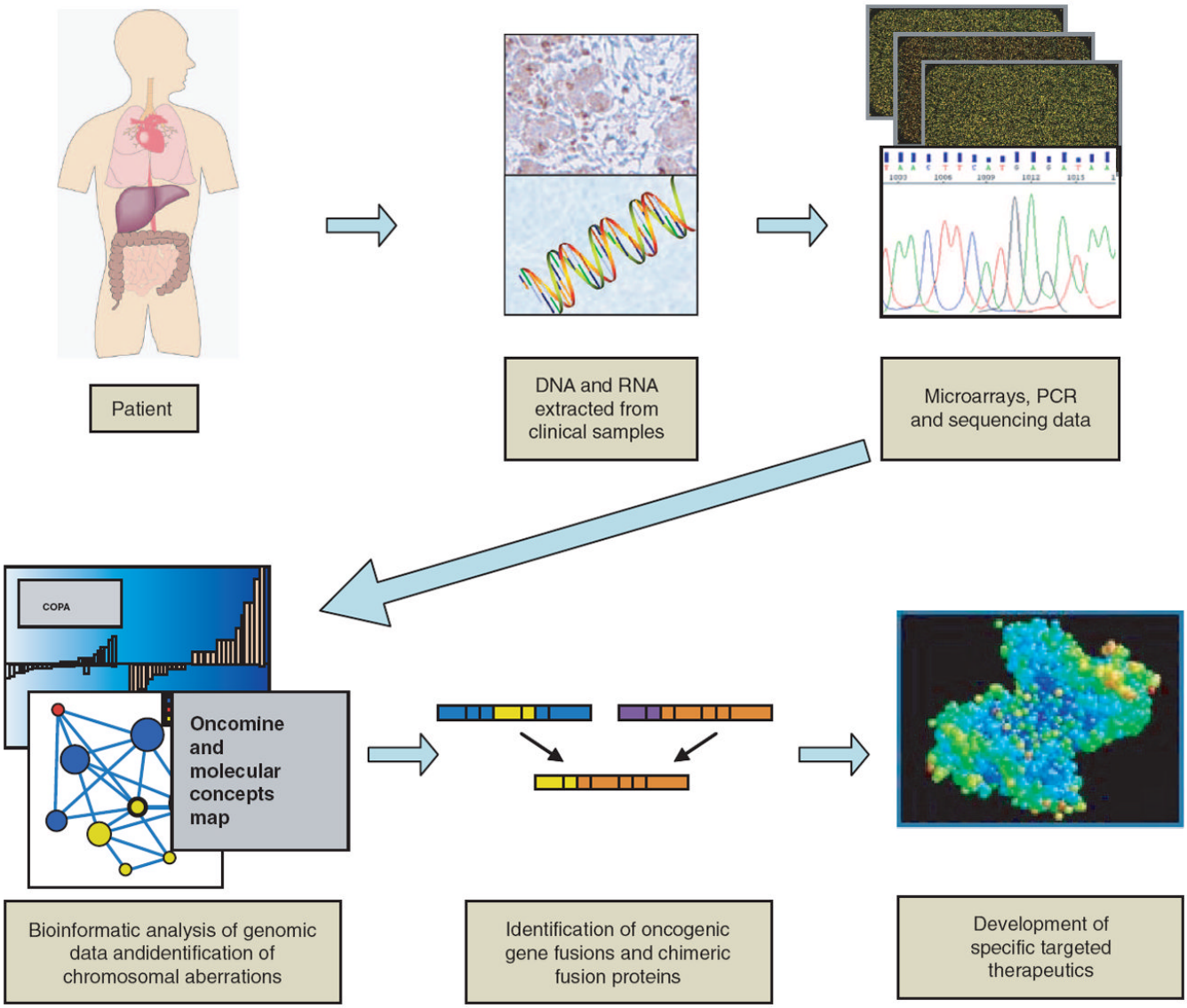


Figure 2. Gene Fusion Discovery and Targeted Therapy

Bioinformatic, sequencing and microarray methods are powerful tools for identifying potential gene fusions in epithelial cancers. By determining the genomic and transcriptomic events in human cancers, clinical management of the disease may be impacted, and gene fusions, such as the TMPRSS2-Ets fusions in prostate cancer, may serve as prominent therapeutic targets. If targeted therapeutics are successfully developed for critical oncogenes, clinical management of cancer may one day be determined based upon genetic evaluation of patient tumors.

Table 1

Gene Fusions in Epithelial Cancers

Gene fusions characterize subsets of several different epithelial carcinomas, including thyroid, prostate, lung, and breast cancer. Gene fusions are broadly classified into two groups: those that contain tyrosine kinases (TKs), which activate intracellular signaling pathways, and those that contain transcription factors (TFs), which control cellular gene expression. Together, TKs and TFs account for 50% of the genes involved in gene fusions. Though most fusions occur at low prevalence rates, some, such as TMPRSS2-ERG in prostate cancer and RET rearrangements in papillary thyroid cancer, among others, are predominant genomic lesions in the disease. Cytogenetically, fusions can be formed by inversions (inv) on a single chromosome, translocations between two genomic loci (t), or intrachromosomal deletions (del). With the exception of pleiomorphic adenomas, this table includes fusions confirmed in human cancer samples. Fusions observed only in tumor-derived cell lines are not included.

Gene Fusions in Carcinomas				
Tyrosine Kinase Fusions				
Papillary Thyroid Carcinoma*	**5' Partner**	**3' Partner**	**Prevalence**	**References**
inv(10)(q11.2;q21)	HRH4	RET	30-80%	Grieco et al. Cell 1990
t(10;17)(q11.2;q23)	Riα	RET	5%	Bongarzone et al. Mol Cell Biol 1993
inv(10)(q11;q22)	NCOA4	RET	15-70%	Bongarzone et al. Cancer Res 1994; Santoro et al. Oncogene 1994
inv(10)(q11;q22)	RFG	RET	<1%	Bongarzone et al. Cancer Res 1994; Santoro et al. Oncogene 1994
t(10;14)(q11.2;q32)	GOLGA5	RET	<1%	Klugbauer et al. Cancer Res 1998
t(7;10)(q32-34;q11.2)	TRIM24	RET	<1%	Klugbauer and Rabes. Oncogene 1999
t(1;10)(p13;q11.2)	TRIM33	RET	<1%	Klugbauer and Rabes. Oncogene 1999
t(10;12)(q11.2;p13.3)	ERC1	RET	<1%	Nakata et al. Genes, Chromosomes Cancer 1999; Liu et al. Thyroid 2005
t(10;14)(q11.2;q22.1)	KTN1	RET	<1%	Salassidis 2000
t(10;18)(q11.2;q21-22)	RFG9	RET	<1%	Klugbauer et al. Cancer Res 2000
t(8;10)(p21-22;q11.2)	PCM1	RET	<1%	Corvi et al. Oncogene 2000
t(6;10)(p21;q11.2)	TRIM27	RET	<1%	Saenko et al. Mutat Res 2003
t(10;14)(q32.12;q11.2)	GOLGA5	RET	<1%	Rabes et al. Clin Cancer Res 2000
t(8;10)(p11.21;q11.2)	HOOK3	RET	<1%	Ciampi et al. Endocr Relat Cancer 2007
inv(1)(q21;q22)	TPM3	NTRK1	In total, 7-12% of	Greco et al. Oncogene 1992

Gene Fusions in Carcinomas				
Tyrosine Kinase Fusions				
Papillary Thyroid Carcinoma*	**5' Partner**	**3' Partner**	**Prevalence**	**References**
inv(1)(q21;q25)	TPM3	TPR	papillary thyroid cancers	Greco et al. Oncogene 1992
inv(1)(q21;q25)	TPR	NTRK1		Greco et al. Genes, Chromosomes Cancer 1997
t(1,3)(q21-22;q11)	TFG	NTRK1		Greco et al. Mol Cell Biol 1995
t(7;7)(q21-22;q34)	AKAP9	BRAF	<1%	Ciampi et al. J Clin Invest 2005
Secretory Breast Cancer				
t(12;15)(p13;q25)	ETV6	NTRK3	>90%	Tognon et al. Cancer Cell 2002
Non-small cell Lung Cancer				
inv(2)(p23;p21) or t(2;2)(p23;p21)	EML4	ALK	2.7 - 6.7%	Soda et al. Nature 2007; Perner et al. Neoplasia 2008
t(6;13)(q22;)	CD74	ROS1	<1%	Rikova et al. Cell 2007
t(2;3)(p23;q12.2)	TFG	ALK	<1%	Rikova et al. Cell 2007
Glioblastoma				
del(6)(q21;q21)	GOPC	ROS1	not reported	Charest et al. PNAS 2003
Transcription Factor Fusions				
Prostate Cancer	**5' Partner**	**3' Partner**	**Prevalence**	**References**
inv(21)(q22.2;q22.3) or del(21)(q22.2;q22.3)	TMPRSS2	ERG	~50%	Tomlins et al. Science 2005
t(1;21)(q32;q22.2)	SLC45A3	ERG	<1%	Han et al. Cancer Res 2008
t(7;21)(p21.2;q22.3)	TMPRSS2	ETV1	5-10%	Tomlins et al. Science 2005
t(7;22)(p21.2;q11.23)	HERV_K_22q11.2 3	ETV1	<1%	Tomlins et al. Nature 2007
t(7;15)(p21.3;q21)	C15orf21	ETV1	1%	Tomlins et al. Nature 2007
t(7;7)(p21.2;p15)	HNRPA2B1	ETV1	1%	Tomlins et al. Nature 2007
t(1;7)(q32;p21.2)	SLC45A3	ETV1	2%	Tomlins et al. Nature 2007
t(2;7)(q36.1p21.2)	ACSL3	ETV1	<1%	Attard et al. Br J Cancer 2008
t(7;14)(p21.2;q13.3-q21.1)	Not Known	ETV1	<1%	Attard et al. Br J Cancer 2008
t(7;17)(p21.2;p13.1)	FLJ35294	ETV1	<1%	Han et al. Cancer Res 2008
t(17;21)(q21;q22.3)	TMPRSS2	ETV4	<5%	Tomlins et al. Cancer Res 2008

Gene Fusions in Carcinomas				
Tyrosine Kinase Fusions				
Papillary Thyroid Carcinoma*	**5' Partner**	**3' Partner**	**Prevalence**	**References**
t(17;19)(q21;q13)	KLK2	ETV4	<1%	Hermans et al. Cancer Res 2008
inv(17;17)(q22;q25)	CANT1	ETV4	<1%	Hermans et al. Cancer Res 2008
t(17;17)(q21;q21)	DDX5	ETV4	<1%	Han et al. Cancer Research 2008
t(3;21)(q27;q22.3)	TMPRSS2	ETV5	<5%	Helgeson et al. Cancer Res 2008
t(1;3)(q32;q27)	SLC45A3	ETV5	<1%	Helgeson et al. Cancer Res 2008
Renal-cell Carcinoma				
t(X;1)(p11;q21)	PRCC	TFE3	In total, 10-15% of all renal tumors	Weterman et al. PNAS 1996; Sidhar et al. Hum Mol Genet 1996
t(X;17)(p11;q25)	ASPSCR1	TFE3		Argani Am J Pathol 2001
t(6;11)(p21.1;q13)	Alpha	TFEB		Davis et al. PNAS 2003
t(X;1)(p11;p34)	SFPQ	TFE3		Clark et al. Oncogene 1997
inv(X)(p11;q12)	NonO	TFE3		Clark et al. Oncogene 1997
t(X;17)(p11.2;q23)	CLTC	TFE3		Argani et al. Oncogene 2003
t(X;17)(p11.2;q25.3)	RCC17	TFE3		Heimann et al. Cancer Res 2001
Salivary Gland Tumors				
Pleiomorphic				
Adenoma				
t(3;8)(p21;q12)	CTNNB1	PLAG1	In total, ~40% of all pleiomorphic adenomas	Kas et al. Nat Genet 1997
t(5;8)(p13;q12)	LIFR	PLAG1		Voz et al. Oncogene 1998
t(8;8)(q12;q11.2)	TCEA1	PLAG1		Atrom et al. Cancer Res 1999
t(8;8)(q12;q11.2)	CHCHD7	PLAG1		Asp et al. Genes Chromosomes Cancer 2006
t(3;13)(p14.2;q13-15)	HMGA2	FHIT	<1%	Geurts et al. Cancer Res 1997
t(9;12)(p12-22;q13-15) or ins(9;12)	HMGA2	NFIB	8-12%	Geurts et al. Oncogene 1998
Mucoepidermoid Carcinoma				
t(11;19)(q21-22;p13)	CRC1	MAML2	30 - 75%	Nordkvist et al. Cancer Genet Cytogen 1994; Tonon et al. Nat Genet 2003

Gene Fusions in Carcinomas				
Tyrosine Kinase Fusions				
Papillary Thyroid Carcinoma*	**5' Partner**	**3' Partner**	**Prevalence**	**References**
t(11;19)(q21-22;p13.11)	CRTC3	MAML2	<1%	Fehr et al. Genes Chromosomes Cancer 2008
Dominant Negative Fusions				
Aggressive Midline Carcinoma				
t(15;19)(q13;p13.1)	BRD4	NUT	~66%	French et al. Cancer Res 2003; French et al. Am J Pathol 2001
t(9;15)(q34;q13)	BRD3	NUT	~10%	French et al. Oncogene 2008
Follicular Thyroid Carcinoma				
t(2,3)(q13;p25)	PAX8	PPARg	25-50%	Kroll et al. Science 2000

* prevalence of RET translocations depends on age and radiation exposure

Table 2

5’ Binding Partners in Ets Fusions

Ets fusions in prostate cancer exhibit a variety of 5’ binding partners that drive overexpression of the Ets transcription factors. Accounting for approximately 90% of these, TMPRSS2-ERG is the most common of the known fusions, followed by TMPRSS2-ETV1. Other fusions feature prostate-specific genes (KLK2, C15orf21, CANT1, SLC45A3), endogenous retroviral elements (HERV_K_22q11.23, FLJ35294), a fatty-acid chain ligase (ACSL3), a DEAD box helicase (DDX5) and a housekeeping gene (HNRPA2B1). With the exception of HNRPA2B1-ETV1, C15ORF21-ETV1, and DDX5-ETV4, all of the 5’ partners display androgen-responsive upregulation.

5’ Fusion Partners in Prostate Cancer		
5’ Binding Partner	**Description**	**References**
TMPRSS2	Androgen-regulated transmembrane serine protease. Fuses with ERG, ETV1, ETV4, and ETV5.	Tomlins et al. Science 2005 Helgeson et al. Cancer Res 2008; Han et al. Cancer Res 2008
HERV_K_22q11.23	An endogenous retroviral element. Fuses with ETV1.	Tomlins et al. Nature 2007
C15orf21	A prostate-specific and androgen-repressed gene. Fuses with ETV1.	Tomlins et al. Nature 2007
HNRPA2B1	A prominent housekeeping gene. Fuses with ETV1.	Tomlins et al. Nature 2007
ACSL3	An isozyme of the long-chain fatty-acid coenzyme A ligase family. Fuses with ETV1	Attard et al. Br J Cancer 2008
FLJ35294	An endogenous retroviral element (HERVK_17p13.1). Fuses with ETV1	Han et al. Cancer Res 2008
DDX5	Putative RNA helicase with a DEAD box polypeptide. Fuses with ETV4.	Han et al. Cancer Res 2008
KLK2	Prostate-specific, androgen-regulated gene. Fuses with ETV4.	Hermans et al. Cancer Res 2008
CANT1	Prostate-specific, androgen-regulated gene. Fuses with ETV4.	Hermans et al. Cancer Res 2008
SLC45A3	Prostate-specific androgen-induced gene. Fuses with ERG, ETV1 and ETV5.	Tomlins at al. Nature 2007; Helgeson et al. Cancer Res 2008; Han et al. Cancer Res 2008

Published in final edited form as:
Nature. 2009 March 5; 458(7234): 97–101. doi:10.1038/nature07638.

Transcriptome Sequencing to Detect Gene Fusions in Cancer

Christopher A. Maher[1,3,†], **Chandan Kumar-Sinha**[1,3,†], **Xuhong Cao**[1,2], **Shanker Kalyana-Sundaram**[1,3], **Bo Han**[1,3], **Xiaojun Jing**[1,3], **Lee Sam**[1,3], **Terrence Barrette**[1,3], **Nallasivam Palanisamy**[1,3], and **Arul M. Chinnaiyan**[1,2,3,4,5,#]

[1]Michigan Center for Translational Pathology, University of Michigan Medical School, Ann Arbor, MI, 48109

[2]Howard Hughes Medical Institute, University of Michigan Medical School, Ann Arbor, MI, 48109

[3]Department of Pathology, University of Michigan Medical School, Ann Arbor, MI, 48109

[4]Department of Urology, University of Michigan Medical School, Ann Arbor, MI, 48109

[5]Comprehensive Cancer Center, University of Michigan Medical School, Ann Arbor, MI, 48109

Abstract

Recurrent gene fusions, typically associated with hematological malignancies and rare bone and soft tissue tumors[1], have been recently described in common solid tumors[2–9]. Here we employ an integrative analysis of high-throughput long and short read transcriptome sequencing of cancer cells to discover novel gene fusions. As a proof of concept we successfully utilized integrative transcriptome sequencing to "re-discover" the *BCR-ABL1* 10 gene fusion in a chronic myelogenous leukemia cell line and the *TMPRSS2-ERG* 2,[3] gene fusion in a prostate cancer cell line and tissues. Additionally, we nominated, and experimentally validated, novel gene fusions resulting in chimeric transcripts in cancer cell lines and tumors. Taken together, this study establishes a robust pipeline for the discovery of novel gene chimeras using high throughput sequencing, opening up an important class of cancer-related mutations for comprehensive characterization.

Keywords

Transcriptome sequencing; Prostate cancer; Bioinformatics; Gene fusions

Characterization of specific genomic aberrations in cancers has led to the identification of several successful therapeutic targets, such as BCR-ABL1, PDGFR, ERBB2, and EGFR etc[11–14], therefore a major goal in cancer research is to identify causal genetic aberrations. Gene fusions resulting from chromosomal rearrangements in cancer are believed to define the most prevalent category of 'cancer genes[15]. Typically, an aberrant juxtaposition of two genes, may encode a fusion protein (e.g., *BCR-ABL1*), or the regulatory elements of one gene may drive the aberrant expression of an oncogene (e.g.,*TMPRSS2-ERG*). While gene fusions have been widely described in rare hematological malignancies and sarcomas[1], the recent discovery of recurrent gene fusions in prostate[2,4] and lung cancers[5–9] points to their role in common solid tumors as well. Considering their prevalence and common characteristics across cancer

Address correspondence and requests for reprints to: Arul M. Chinnaiyan, M.D., Ph.D., Investigator, Howard Hughes Medical Institute, Department of Pathology and Urology, University of Michigan Medical School, 1400 E. Medical Center Drive, 5316 UMCCC, Ann Arbor, MI-48109, Phone: 734-615-4062, Fax: 734-615-4498, E-mail: arul@umich.edu .
† *These authors contributed equally to the work.*

Author Information The gene fusion chimeras have been deposited in GenBank under the accession numbers FJ423742-FJ423755. Correspondence and requests for materials and reprints should be addressed to A.M.C. (arul@umich.edu).

types, gene fusions may be regarded as a distinct class of 'mutations', with a causal role in carcinogenesis, and being strictly confined to cancer cells, they represent ideal diagnostic markers and rational therapeutic targets.

As a proof of concept we carried out whole transcriptome sequencing of the chronic myelogenous leukemia cell line, K562, harboring the classical gene fusion, *BCR-ABL1* 16. Using the Illumina Genome Analyzer, we generated 66.9 million reads of 36 nucleotides in length and screened them for the presence of reads showing partial alignment to exon boundaries from two different genes. While this approach was able to detect *BCR-ABL1*, it was one among a set of 111 other chimeras (with at least 2 reads). Thus, in a *de novo* discovery mode, it would be difficult to pin-point the *BCR-ABL1* fusion in the background of the other putative chimeras. However, when we used the known fusion junction of *BCR-ABL1* (Genbank No. M30829) as the reference sequence, we detected 19 chimeric reads (Supplementary Fig. 1). Thus, we considered an integrative approach for chimera detection, utilizing short read sequencing technology for obtaining deep sequence data and long read technology (Roche 454 sequencing platform) to provide reference sequences for mapping candidate fusion genes.

An important concern in transcriptome sequencing was whether we could detect chimeric transcripts in the background of highly abundant house-keeping genes (i.e., would cDNA normalization be required). To address this, we compared sequences from normalized and non-normalized cDNA libraries of the prostate cancer cell line VCaP, which harbors the gene fusion *TMPRSS2-ERG* (Supplementary Table 1). Overall, the normalized library showed an approximately 3.6-fold reduction in the total number of chimeras nominated. Furthermore, while we expected the normalized library would enrich for the *TMPRSS2-ERG* gene fusion, it failed to reveal any *TMPRSS2-ERG* chimeras suggesting that we would not benefit from normalization in our analyses.

To assess the feasibility of using massively parallel transcriptome sequencing to identify novel gene fusions, we generated non-normalized cDNA libraries from the prostate cancer cell lines VCaP and LNCaP, and a benign immortalized prostate cell line RWPE. As a first step, using the Roche 454 platform, we generated 551,912 VCaP, 244,984 LNCaP, and 826,624 RWPE transcriptome sequence reads, averaging 229.4 nucleotides. These were categorized as completely aligning, partially aligning, or nonmapping to the human reference database (Fig. 1a). Sequence reads that showed partial alignments to two genes (Supplementary Methods) were nominated as first pass candidate chimeras. This yielded 428 VCaP, 247 LNCaP, and 83 RWPE candidates. Admittedly, many of these chimeric sequences could be a result of *trans*-splicing[17] or co-transcription of adjacent genes coupled with intergenic splicing[18], or simply, an artifact of the sequencing protocol. Surprisingly, among the 428 VCaP candidates, only one read spanned the *TMPRSS2-ERG* fusion junction using the long read sequencing platform (Supplementary Table 2).

Next, using the Illumina Genome Analyzer we obtained over 50 million short transcriptome sequence reads from VCaP, LNCaP and RWPE cDNA libraries (Supplementary Table 3). Focusing initially on VCaP cells, we identified the *TMPRSS2-ERG* fusion as one among 57 candidates, many of them likely false positives. To overcome the problem of false positives, lack of depth in long reads, and difficulty in mapping partially aligning short reads, we considered integrating the long and short read sequence data. Following this strategy we found the single long read chimeric sequence spanning *TMPRSS2-ERG* junction from VCaP transcriptome sequence, buttressed by 21 short reads (Fig. 1b), was one of only eight chimeras nominated, overall. Thus, using the integrative approach the total number of false candidates was reduced and the proportion of experimentally validated candidates increased dramatically (Supplementary Fig. 2). Extending the integrative analysis to LNCaP and RWPE sequences provided a total of fifteen chimeric transcripts, of which ten could be experimentally confirmed

(Supplementary Table 4). To ensure that the integration strategy filtered out only false positives and not valid chimeras, we tested a panel of 16 long read chimera candidates that were eliminated upon integration and found that none of them confirmed a fusion transcript by qRT-PCR (Supplementary Fig. 3).

In order to systematically leverage the collective coverage provided by the two sequencing platforms, and to prioritize the candidates, we formulated a scoring function obtained by multiplying the number of chimeric reads derived from either method (Supplementary Table 4). Further, we categorized these chimeras as intra- or inter-chromosomal, based on their location on the same or different chromosomes, respectively. The latter represent *bona fide* gene fusions as do intra-chromosomal chimeras aligning to non-adjacent transcripts; intra-chromosomal chimeras between neighboring genes are classified as (read-throughs). Remarkably, *TMPRSS2-ERG* was our top ranking gene fusion sequence, second only to a read-through chimera *ZNF577-ZNF649*.

In addition to*TMPRSS2-ERG* we identified several new gene fusions in VCaP. One such fusion was between exon 1 of *USP10*, with exon 3 of *ZDHHC7*, both genes located on chromosome 16, approximately 200 kb apart, in opposite orientation (Fig. 2a, Supplementary Discussion). Furthermore, two separate fusions involving the gene *HJURP* on chromosome 2 were identified. A fusion between exon 2 of *EIF4E2* with exon 8 of *HJURP* generated the fusion transcript *EIF4E2-HJURP* and a fusion between exon 9 of *HJURP* with exon 25 of *INPP4A* yielded *HJURP-INPP4A* (Fig. 2b, Supplementary Fig. 4).

Interestingly, based on whole transcriptome sequencing, the highest ranked LNCaP gene fusion was between exon 11 of *MIPOL1* on chromosome 14 with the last exon of *DGKB* on chromosome 7; confirmed by qRT-PCR and FISH (Fig. 3, Supplementary Fig. 5). We recently demonstrated that over-expression of *ETV1*, a member of the oncogenic ETS transcription factor family, plays a role in tumor progression in LNCaP cells[3]. The mechanism of *ETV1* over-expression was attributed to a cryptic insertion of approximately 280 Kb encompassing the *ETV1* gene into an intronic region of *MIPOL1*. Thus, while our previous study suggested that *ETV1* was rearranged without evidence of an *ETV1* fusion transcript, here we show the generation of a surrogate fusion of *MIPOL1* to *DGKB*, which appears to be indicative of an *ETV1* chromosomal aberration.

In addition to gene fusions, we also identified several transcript chimeras between neighboring genes, referred to as read-through events. Overall, the read-through events appear to be more broadly expressed across both malignant and benign samples whereas the gene fusions were cancer cell specific (Supplementary Fig. 6, Supplementary Discussion).

Next, we attempted to extend this methodology to tumor samples that represent the malignant cells often admixed with benign epithelia, stromal, lymphocytic, and vascular cells. Transcriptome sequencing of two *TMPRSS2-ERG* gene fusion positive metastatic prostate cancer tissues, VCaP-Met (from which the VCaP cell line is derived) and Met 3, and one *ERG* negative metastatic prostate tissue, Met 4. Interestingly, in addition to the *TMPRSS2-ERG* fusion sequences detected in both VCaP-Met and Met 3 tissues, three novel gene fusions were identified (Supplementary Fig. 7a). One chimeric transcript from Met 3 involves exon 9 of *STRN4* with exon 2 of *GPSN2* (Supplementary Fig. 7b). *GPSN2* belongs to the steroid 5-alpha reductase family, the enzyme that converts testosterone to dihydrotestosterone (DHT), the key hormone that mediates androgen response in prostate tissues. DHT is known to be highly expressed in prostate cancer, and is a therapeutic target[19]. DHT, like its synthetic analog R1881, has been shown to induce *TMPRSS2-ERG* expression as well as PSA[2]. Additionally, we found exon 10 of *RC3H2* fused to exon 20 of *RGS3* in the VCaP-Met (and VCaP cells)

(Supplementary Fig. 7c). Another novel gene fusion was between exon 1 of *LMAN2* and exon 2 of *AP3S1* (Supplementary Fig. 7d).

Interestingly, one read-through chimera, *SLC45A3-ELK4*, between the fourth exon of *SLC45A3* with exon 2 of *ELK4*, a member of the ETS transcription factor family, was identified in metastatic prostate cancer, Met 4, and the LNCaP cell line suggesting recurrence (Fig. 4a, upper panel). Taqman qRT-PCR assay for this fusion carried out in a panel of cell lines revealed high level of expression in LNCaP cells and much lower levels in other prostate cancer cell lines including 22Rv1, VCaP, and MDA-PCA-2B. Benign prostate epithelial cells, PREC and RWPE and non-prostate cell lines including breast, melanoma, lung, CML, and pancreatic cancer cell lines were negative for this fusion (Fig. 4a, **middle panel**). *SLC45A3* has been earlier reported to be fused to *ETV1* in a prostate cancer sample[3], and notably, it is a prostate specific, androgen responsive gene. Interestingly, the fusion transcript *SLC45A3-ELK4* was also found to be induced by the synthetic androgen R1881 (Fig. 4a, **middle panel, inset**). Further, we interrogated a panel of prostate tissues for this fusion, and found it expressed in seven out of twenty metastatic prostate cancer tissues examined (Fig. 4a, **lower panel**). Interestingly, six of those seven positive cases have been identified as negative for ETS genes *ERG, ETV1, ETV4*, and *ETV5* in our previous work, based on a FISH screen[20]. One *TMPRSS2-ETV1* positive metastatic prostate cancer sample was also found to be positive for *SLC45A3-ELK4* (similar to LNCaP, which is also *ETV1* positive[3]). Unlike the previous ETS gene fusions identified, *SLC45A3-ELK4* is a read-through event between adjacent genes and does not harbor detectable alterations at the DNA level by FISH (Supplementary Figure 8), array CGH (data not shown) or high-density SNP arrays (Supplementary Figure 9).As LNCaP and Met 4 harbor genomic aberrations of *ETV1*, and express high levels of the *SLC45A3-ELK4* chimeric transcript, this suggests that *ETV1* and *ELK4* may cooperate to drive prostate carcinogenesis in those tumors. To our knowledge, *SLC45A3-ELK4* may represent the first description of a recurrent RNA chimeric transcript specific to cancer that does not have a detectable DNA aberration. Overall, *SLC45A3-ELK4* appears to be the only recurrent chimeric transcript identified in our transcriptome sequencing study, as other gene fusions tested in a panel of prostate cancer samples, appear to be restricted to the sample in which they were identified (at least in the limited number of samples we analyzed) and thus may represent rare or private mutations (Supplementary Fig. 10).

Next we tested if the novel gene fusions identified in this study represent acquired somatic mutations or simply, germline variations. Based on qPCR (Supplementary Fig. 11) and FISH (Supplementary Fig. 12–Supplementary Fig. 13) assessment of a representative set of fusion genes on patient matched germline tissues, we found the chimeras restricted to the cancer tissues. Further, we interrogated the 29 genes involved in our gene fusions in the Database of Genomic Variants (http://projects.tcag.ca/variation/) and found only 8 of them with previously reported copy number variations (CNVs) (Supplementary Table 5), but our matched aCGH data did not reveal any copy number variation in those genes (Supplementary Table 6), suggesting that our samples did not harbor CNVs common to the human population.

Based on the gene fusions we have characterized (Supplementary Table 7), we propose a chimera classification system (Fig. 4b). Inter-chromosomal translocation (Class I) involves fusion between two genes on different chromosomes (for example, *BCR-ABL1*). Inter-chromosomal complex rearrangements (Class II) where two genes from different chromosomes fuse together while a third gene follows along and becomes activated (*MIPOL1-DGKB*). Intra-chromosomal deletion (Class III) results when deletion of a genomic region fuses the flanking genes (*TMPRSS2-ERG*). Intra-chromosomal complex rearrangements (Class IV) involve a breakpoint in one gene fusing with multiple regions (*HJURP-EIF4E2*, and *INPP4-HJURP*) and Read-through chimeras (Class V) include chimeric transcripts between neighboring genes (*ZNF649-ZNF577*).

Overall, transcriptome sequencing was found to be a powerful tool for detecting gene fusions, exemplified by our ability to detect multiple gene fusions in cancer cell lines and tissues. One important limitation is in cases where the proximal partner contributes only the regulatory sequence to the fusion and no transcript sequence (e.g, IgH-Myc in Burkitt's lymphoma). While it has been known that gene fusion events can play a causative role in cancer, the current study has demonstrated that a particular cancer cell line or tissue can harbor multiple gene fusions many of which are likely not recurrent. While it is unclear whether these private gene fusions play a role in malignant transformation, they could potentially cooperate with the driver mutation/gene fusions. Similar to the cataloging of point mutations associated with cancer[21–27], it will be important to catalog and investigate the function of the multiple gene fusions present in a single cancer. The discovery of the chimeric transcript *SLC45A3-ELK4* underscores that a refinement of next generation sequencing technologies and attendant analytical tools may well unravel the full scope of these 'dangerous liaisons' in carcinogenesis.

METHODS SUMMARY

Long read sequencing was conducted using 454 FLX Sequencing whereas short read sequencing was performed on the Illumina Genome Analyzer. Q-PCR for fusion candidates were performed using indicated oligonucleotide primers (Supplementary Table 8). Interphase FISH were performed in cell lines and tissues using bacterial artificial chromosome (BAC) probes (Supplementary Fig. 4a, Supplementary Fig 5a, 5c, 5e, Supplementary Fig 8, Supplementary Fig 7d, Supplementary Fig 12, Supplementary Fig 13, Supplementary Fig 14b, and 14d). Oligonucleotide comparative genomic hybridization (aCGH) was performed using Agilent arrays and copy number analysis was conducted in CGH Analytics. Affymetrix Genome-wide Human SNP Array 6.0 was processed using the Affymetrix Genotyping Console. Prostate tissues were obtained from the radical prostatectomy series at the University of Michigan and from the Rapid Autopsy Program, University of Michigan Specialized Program of Research Excellence (S.P.O.R.E.) in prostate cancer.

METHODS

Samples and cell lines

The benign immortalized prostate cell line RWPE and the prostate cancer cell line LNCaP was obtained from the American Type Culture Collection. Primary benign prostatic epithelial cells (PrEC) were obtained from Cambrex Bio Science. The prostate cancer cell line MDA-PCa 2B was provided by E. Keller. The prostate cancer cell line 22-RV1 was provided by J. Macoska. VCaP was derived from a vertebral metastasis from a patient with hormone-refractory metastatic prostate cancer[28], and was provided by Ken Pienta.

Androgen stimulation experiment was carried out with LNCaP and VCaP cells grown in charcoal-stripped serum containing media for 24 h, before treatment with 1% ethanol or 1 nM of methyltrienolone (R1881, NEN Life Science Products) dissolved in ethanol, for 24 and 48 h. Total RNA was isolated with RNeasy mini kit (Qiagen) according to the manufacturer's instructions.

Prostate tissues were obtained from the radical prostatectomy series at the University of Michigan and from the Rapid Autopsy Program[29], University of Michigan Prostate Cancer Specialized Program of Research Excellence Tissue Core. All samples were collected with informed consent of the patients and prior approval of the institutional review board.

454 FLX Sequencing

PolyA+ RNA was purified from 50μg total RNA using two rounds of selection on oligo-dT containing paramagnetic beads using Dynabeads mRNA Purification Kit (Dynal Biotech, Oslo, Norway), according to the manufacturer's instructions. 200 ng mRNA was fragmented at 82° C in Fragmentation Buffer (40 mM Tris-Acetate, 100 mM Potassium Acetate, 31.5 mM Magnesium Acetate, pH 8.1) for 2 minutes. First strand cDNA library was prepared using Superscript II (Invitrogen) according to standard protocols and directional adaptors were ligated to the cDNA ends for clonal amplification and sequencing on the Genome Sequencer FLX.

The adaptor ligation reaction was carried out in Quick Ligase Buffer (New England Biolabs, Ipswich, MA) containing 1.67 μM of the Adaptor A, 6.67 μM of the Adaptor B and 2000 units of T4 DNA Ligase (New England Biolabs, Ipswich, MA) at 37°C for 2 hours. Adapted library was recovered with 0.05% Sera-Mag30 streptavidin beads (Seradyn Inc, Indianapolis, IN) according to manufacturer's instructions. Finally, the sscDNA library was purified twice with RNAClean (Agencourt, Beverly, MA) as per the manufacturer's directions except the amount of beads was reduced to 1.6X the volume of the sample. The purified sscDNA library was analyzed on an RNA 6000 Pico chip on a 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA) to confirm a size distribution between 450 to 750 nucleotides, and quantified with Quant-iT Ribogreen RNA Assay Kit (Invitrogen Corporation, Carlsbad, CA) on a Synergy HT (Bio-Tek Instruments Inc, Winooski, VT) instrument following the manufacturer's instructions. The library was PCR amplified with 2 μM each of Primer A (5'-GCC TCC CTC GCG CCA-3') and Primer B (5'-GCC TTG CCA GCC CGC-3'), 400 μM dNTPs, 1X Advantage 2 buffer and 1 μl of Advantage 2 polymerase mix (Clontech, Mountain View, CA). The amplification reaction was performed at: 96°C for 4 min; 94°C for 30 sec, 64°C for 30 sec, repeating steps 2 and 3 for a total of 20 cycles, followed by 68°C for 3 minutes. The samples were purified using AMPure beads and diluted to a final working concentration of 200,000 molecules per μl. Emulsion beads for sequencing were generated using Sequencing emPCR Kit II and Kit III and sequencing was carried out using 600,000 beads.

Normalization by Subtraction

mRNA from the prostate cancer cell line VCaP was hybridized with the subtractor cell line LNCaP 1st-strand cDNA immobilised on magnetic beads (Dynabeads, Invitrogen), according to the manufacturers instructions. Transcripts common to both the cells were captured and removed by magnetic separation of bead-bound subtractor cDNA and the subtracted VCaP mRNA left in the supernatant was recovered by precipitation and used for generating sequencing library as described. Efficiency of normalization was assessed by qRT-PCR assay of levels of select transcripts in the sample before and after the subtraction (data not shown).

Illumina Genome Analyzer Sequencing

200ng mRNA was fragmented at 70°C for 5 min in a Fragmentation buffer (Ambion), and converted to first strand cDNA using Superscript III (Invitrogen), followed by second strand cDNA synthesis using E coli DNA pol I (Invitrogen). The double stranded cDNA library was further processed by Illumina Genomic DNA Sample Prep kit, and it 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, and was ligated with Illumina's adaptor oligo mix using T4 DNA ligase. Adaptor ligated library was size selected by separating on a 4% agarose gel and cutting out the library smear at 200bp (+/− 25bp). The library was PCR amplified by Phu polymerase (Stratagene), and purified by Qiaquick PCR purification kit (Qiagen). The library was quantified with Quant-iT Picogreen dsDNA Assay Kit (Invitrogen Corporation, Carlsbad, CA) on a Modulus™ Single Tube Luminometer (Turner

Biosystems, Sunnyvale, CA) following the manufacturer's instructions. 10nM library was used to prepare flowcells with approximately 30,000 clusters per lane.

Sequence datasets

Human genome build 18 (hg18) was used as a reference genome. All UCSC and Refseq transcripts were downloaded from the UCSC genome browser (http://genome.ucsc.edu/)[30]. Sequences of previously identified *TMPRSS2-ERGa* fusion transcript (Genbank accession: DQ204772) and *BCR-ABL1* fusion transcript (Genbank accession: M30829) were used for reference.

Short read chimera discovery

Short reads that do not completely align to the human genome, Refseq genes, mitochondrial, ribosomal, or contaminant sequences are categorized as non-mapping. For many chimeras we expect that there will be a larger portion mapping to a fusion partner (major alignment), and smaller portion aligning to the second partner (minor alignment). Our approach is therefore divided into two phases in which we focus on first identifying the major alignment and then performing a more exhaustive approach for identifying the minor alignment. In the first phase all non-mapping reads are aligned against all exons of Refseq genes using Vmatch, a pattern matching program[31]. Only reads that have an alignment of 12 or more nucleotides to an exon boundary are kept as potential chimeras. In the second phase, the non-mapping portion of the remaining reads are then mapped to all possible exon boundaries using a Perl script that utilizes regular expressions to detect alignments of as few as six nucleotides. Only those short reads that show partial alignment to exon boundaries of two separate genes are categorized as chimeras. It is possible to have a chimera that has 28 nucleotides aligning to gene x and 8 nucleotides that align to gene y and z because the 8-mer does not provide enough sequence resolution to distinguish between gene y and gene z. Therefore we would categorize this as two individual chimeras. If a sequence forms more than five chimeras it is discarded because it is ambiguous. To minimize false positives, we require that a predicted gene fusion event has at least two supporting chimeras.

Long and short read integrated chimera discovery

All 454 reads are aligned against the human Refseq collection using BLAT, a rapid mRNA/DNA alignment tool[32]. Using a Perl script, the BLAT output files were parsed to detect potential chimeric reads. A read is categorized as completely aligning if it shows greater than 90% alignment to a known Refseq transcript. These are then discarded as they almost completely align and therefore are not characteristic of a chimera. From the remaining reads, we want to query for reads having partial alignment, with minimal overlap, to two Refseq transcripts representing putative chimeras. To accomplish this, we iterate the all possible BLAT alignments for a putative chimera, extracting only those partial alignments that have no more than a six nucleotide, or two codon, overlap. This step reduces false positive chimeras introduced by repetitive regions, large gene families, and conserved domains. Additionally, while our approach tolerates overlap between the partial alignments, it filters those having more than ten or more nucleotides between the partial alignments.

The short reads (36 nucleotides) generated from the Illumina platform are parsed by aligning them against the Refseq database and the human genome using Eland, an alignment tool for short reads. Reads that align completely or fail quality control are removed leaving only the "non-mapping" reads; a rich source for chimeras. These non-mapping short reads are subsequently aligned against all putative long read chimeras (obtained as described above) using Vmatch[31], a pattern matching program. A Perl script is used to parse the Vmatch output to extract only those reads that span the fusion boundary by at least three nucleotides on each side. Following this integration, the remaining putative chimeras are categorized as inter- or

intra-chromosomal chimeras based on whether the partial alignments are located on different or the same chromosomes, respectively. Those intra-chromosomal chimeras that have partial alignments to adjacent genes are believed to be the product of co-transcription of adjacent genes coupled with intergenic splicing (CoTIS)[18], alternatively known as read-throughs. The remaining intra-chromosomal and all inter-chromosomal chimeras are considered candidate gene fusions.

One additional source of false positive chimeras could be an unknown transcript that is not in Refseq. Due to its absence in the Refseq database, the corresponding long read would not be able to show a complete alignment, but instead show partial hits. Subsequently, short reads spanning this transcript would naturally validate the artificially produced fusion boundary. Therefore, to remove these candidates, we aligned all of the chimeras against the human genome using BLAT. If the long read had greater than 90% alignment to one genomic location, it is considered a novel transcript rather than a chimeric read. The remaining chimeras are given a score which is calculated by multiplying the long read coverage spanning the fusion boundary against the short read coverage spanning the fusion boundary.

Coverage analysis

Transcript coverage for every gene locus was calculated from the total number of passing filter reads that mapped, via ELAND, to exons. The total count of these reads was multiplied by the read length and divided by the longest transcript isoform of the gene as determined by the sum of all exon lengths as defined in the UCSC knownGene table (Mar. 2006 assembly). Nucleotide coverage was determined by enumerating the total reads, based on ELAND mappings, at every nucleotide position within a non-redundant set of exons from all possible UCSC transcript isoforms.

Array CGH analysis

Oligonucleotide comparative genomic hybridization is a high-resolution method to detect unbalanced copy number changes at whole genome level. Competitive hybridization of differentially labeled tumor and reference DNA to oligonucleotide printed in an array format (Agilent Technologies, USA) and analysis of fluorescent intensity for each probe will detect the copy number changes in the tumor sample relative to normal reference genome. We identified genomic breakpoints at regions with a change in copy number level of at least one copy (log ratio ± 0.5) for gains and losses involving more than one probe representing each genomic interval as detected by the aberration detection method (ADM) in CGH analytics algorithm.

Real Time PCR validation

Quantitative PCR (QPCR) was performed using Power SYBR Green Mastermix (Applied Biosystems, Foster City, CA) on an Applied Biosystems Step One Plus Real Time PCR System as described[3]. All oligonucleotide primers were synthesized by Integrated DNA Technologies (Coralville, IA) and are listed in Table S8. *GAPDH* 33, primer was as described. All assays were performed in duplicate or triplicate and results were plotted as average fold change relative to *GAPDH*.

Quantitative PCR for *SLC45A3-ELK4* was carried out by Taqman assay method using fusion specific primers and Probe #7 of Universal Probe Library (UPL), Human (Roche) as the internal oligonucleotide, according to manufacturer's instructions. *PGK1* was used as housekeeping control gene for UPL based Taqman assay (Roche), as per manufacturer's instructions. HMBS (Applied Biosystems, Taqman assay Hs00609297_m1) was used as housekeeping gene control for Taqman assays according to standard protocols (Applied Biosystems).

Fluorescence in situ hybridization (FISH)

FISH hybridizations were performed on VCaP, LNCaP, and FFPE tumor and normal tissues. BAC clones were selected from UCSC genome browser. Following colony purification midi prep DNA was prepared using QiagenTips-100 (Qiagen, USA). DNA was labeled by nick translation labeling with biotin-16-dUTP and digoxigenin-11-dUTP (Roche, USA). Probe DNA was precipitated and dissolved in hybridization mixture containing 50% formamide, 2XSSC, 10% dextran sulphate, and 1% Denhardts solution. About 200ng of labeled probes was hybridized to normal human chromosomes to confirm the map position of each BAC clone. FISH signals were obtained using anti digoxigenin-fluorescein and alexa fluor594 conjugate for green and red colors respectively. Fluorescence images were captured using a high resolution CCD camera controlled by ISIS image processing software (Metasystems, Germany).

Affymetrix Genome-Wide Human SNP Array 6.0

1 μg each of genomic DNA samples was sent to Affymetrix service centers (Center for Molecular Medicine, Grand Rapid, MI and Vanderbilt Affymetrix Genotyping Core, Nashville, TN) for genomic level analysis of 15 samples on the Genome-Wide Human SNP Array 6.0. Copy number analysis was conducted using the Affymetrix Genotyping Console software and visualizations were generated by the Genotyping Console (GTC) browser.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank Illumina and 454 for technical support, Rohit Mehra and Javed Siddiqui for providing tissue samples, Yusong Gong, Sunita Shankar, Xiaosong Wang, and Anjana Menon for technical assistance, Jindan Yu for help with the Illumina Genome Analyzer, and Robert J. Lonigro for helpful discussions. C.A.M. was supported by an NIH Ruth L. Kirschstein post-doctoral 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. This work was supported in part by the National Institutes of Health (to A.M.C.), Department of Defense (to A.M.C.), and the Early Detection Research Network (to A.M.C.).

References

1. Mitelman F, Johansson B, Mertens F. Fusion genes and rearranged genes as a linear function of chromosome aberrations in cancer. Nature genetics 2004;36(4):331. [PubMed: 15054488]
2. Tomlins SA, et al. Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. Science (New York, N.Y 2005;310(5748):644.
3. Tomlins SA, et al. Distinct classes of chromosomal rearrangements create oncogenic ETS gene fusions in prostate cancer. Nature 2007;448(7153):595. [PubMed: 17671502]
4. Kumar-Sinha C, Tomlins SA, Chinnaiyan AM. Recurrent gene fusions in prostate cancer. Nature reviews 2008;8(7):497.
5. Choi YL, et al. Identification of novel isoforms of the EML4-ALK transforming gene in non-small cell lung cancer. Cancer research 2008;68(13):4971. [PubMed: 18593892]
6. Koivunen JP, et al. EML4-ALK Fusion Gene and Efficacy of an ALK Kinase Inhibitor in Lung Cancer. Clin Cancer Res 2008;14(13):4275. [PubMed: 18594010]
7. Perner S, et al. EML4-ALK fusion lung cancer: a rare acquired event. Neoplasia (New York, N.Y 2008;10(3):298.
8. Rikova K, et al. Global Survey of Phosphotyrosine Signaling Identifies Oncogenic Kinases in Lung Cancer. Cell 2007;131:14.
9. Soda M, et al. Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer. Nature 2007;448(7153):561. [PubMed: 17625570]

10. Rowley JD. Chromosome translocations: dangerous liaisons revisited. Nature reviews 2001;1(3):245.
11. Lynch TJ, et al. Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib. The New England journal of medicine 2004;350(21):2129. [PubMed: 15118073]
12. Slamon DJ, et al. Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2. The New England journal of medicine 2001;344(11):783. [PubMed: 11248153]
13. Demetri GD, et al. Efficacy and safety of imatinib mesylate in advanced gastrointestinal stromal tumors. The New England journal of medicine 2002;347(7):472. [PubMed: 12181401]
14. Druker BJ, et al. Five-year follow-up of patients receiving imatinib for chronic myeloid leukemia. The New England journal of medicine 2006;355(23):2408. [PubMed: 17151364]
15. Futreal PA, et al. A census of human cancer genes. Nature reviews 2004;4(3):177.
16. Shtivelman E, Lifshitz B, Gale RP, Canaani E. Fused transcript of abl and bcr genes in chronic myelogenous leukaemia. Nature 1985;315(6020):550. [PubMed: 2989692]
17. Takahara T, Tasic B, Maniatis T, Akanuma H, Yanagisawa S. Delay in synthesis of the 3' splice site promotes trans-splicing of the preceding 5' splice site. Molecular cell 2005;18(2):245. [PubMed: 15837427]
18. Communi D, Suarez-Huerta N, Dussossoy D, Savi P, Boeynaems JM. Cotranscription and intergenic splicing of human P2Y11 and SSF1 genes. The Journal of biological chemistry 2001;276(19):16561. [PubMed: 11278528]
19. Gleave M, et al. The effects of the dual 5alpha-reductase inhibitor dutasteride on localized prostate cancer--results from a 4-month pre-radical prostatectomy study. Prostate 2006;66(15):1674. [PubMed: 16927304]
20. Han B, et al. A fluorescence in situ hybridization screen for E26 transformation-specific aberrations: identification of DDX5-ETV4 fusion protein in prostate cancer. Cancer research 2008;68(18):7629. [PubMed: 18794152]
21. Barber TD, Vogelstein B, Kinzler KW, Velculescu VE. Somatic mutations of EGFR in colorectal cancers and glioblastomas. The New England journal of medicine 2004;351(27):2883. [PubMed: 15625347]
22. Cheung VG, et al. Integration of cytogenetic landmarks into the draft sequence of the human genome. Nature 2001;409(6822):953. [PubMed: 11237021]
23. Greenman C, et al. Patterns of somatic mutation in human cancer genomes. Nature 2007;446(7132): 153. [PubMed: 17344846]
24. Stephens P, et al. A screen of the complete protein kinase gene family identifies diverse patterns of somatic mutations in human breast cancer. Nature genetics 2005;37(6):590. [PubMed: 15908952]
25. Strausberg RL, Buetow KH, Emmert-Buck MR, Klausner RD. The cancer genome anatomy project: building an annotated gene index. Trends Genet 2000;16(3):103. [PubMed: 10689348]
26. Weir BA, et al. Characterizing the cancer genome in lung adenocarcinoma. Nature 2007;450(7171): 893. [PubMed: 17982442]
27. Wood LD, et al. The genomic landscapes of human breast and colorectal cancers. Science (New York, N.Y 2007;318(5853):1108.
28. Korenchuk S, et al. VCaP, a cell-based model system of human prostate cancer. In vivo (Athens, Greece) 2001;15(2):163.
29. Rubin MA, et al. Rapid ("warm") autopsy study for procurement of metastatic prostate cancer. Clin Cancer Res 2000;6(3):1038. [PubMed: 10741732]
30. Karolchik D, et al. The UCSC Table Browser data retrieval tool. Nucleic Acids Res 2004;32(Database issue):D493. [PubMed: 14681465]
31. Abouelhoda MI, Kurtz S, Ohlebusch E. Replacing suffix trees with enhanced suffix arrays. Journal of Discrete Algorithms 2004;2(1):53.
32. Kent WJ. BLAT--the BLAST-like alignment tool. Genome research 2002;12(4):656. [PubMed: 11932250]
33. Vandesompele J, et al. Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. Genome biology 2002;3(7):34.

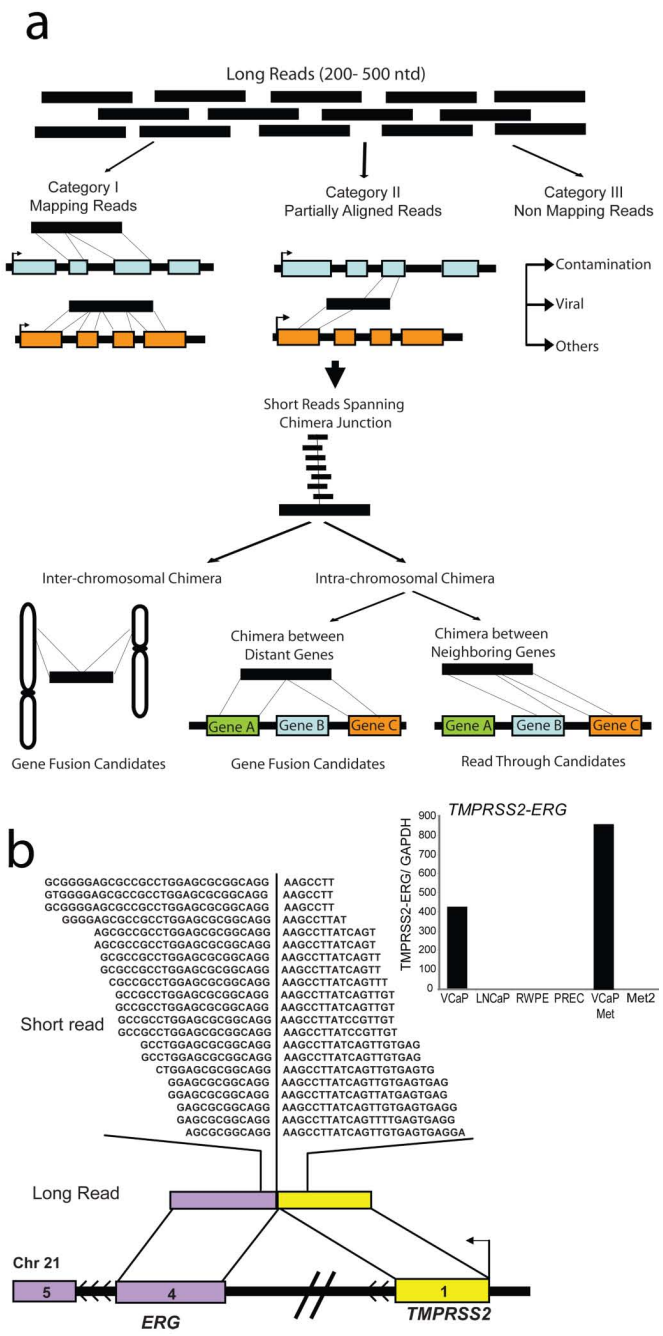


Fig 1. Employing massively parallel sequencing to discover chimeric transcripts in cancer

a, Schema representing our approach to employ transcriptome sequencing to identify chimeric transcripts. 'Long read' sequences compared with the reference database are classified as 'Mapping', 'Partially Aligned', and 'Non-Mapping' reads. Partially aligning reads are considered putative chimeras and are categorized as inter- or intra-chromosomal chimeras. Integration with short read sequence data is utilized for short-listing candidate chimeras and assessing the depth of coverage spanning the fusion junction**b**, "Re-discovery" of*TMPRSS2-ERG* fusion on chromosome 21. Short reads (Illumina) are overlaid on the corresponding long read (454) represented by colored bars. Sequences spanning the fusion junction are indicated by the partition in the short reads. Chromosomal context of the fusion genes is represented by

colored bars punctuated with black lines. Inset displays histogram of qRT-PCR validation of the *TMPRSS2-ERG* transcript.

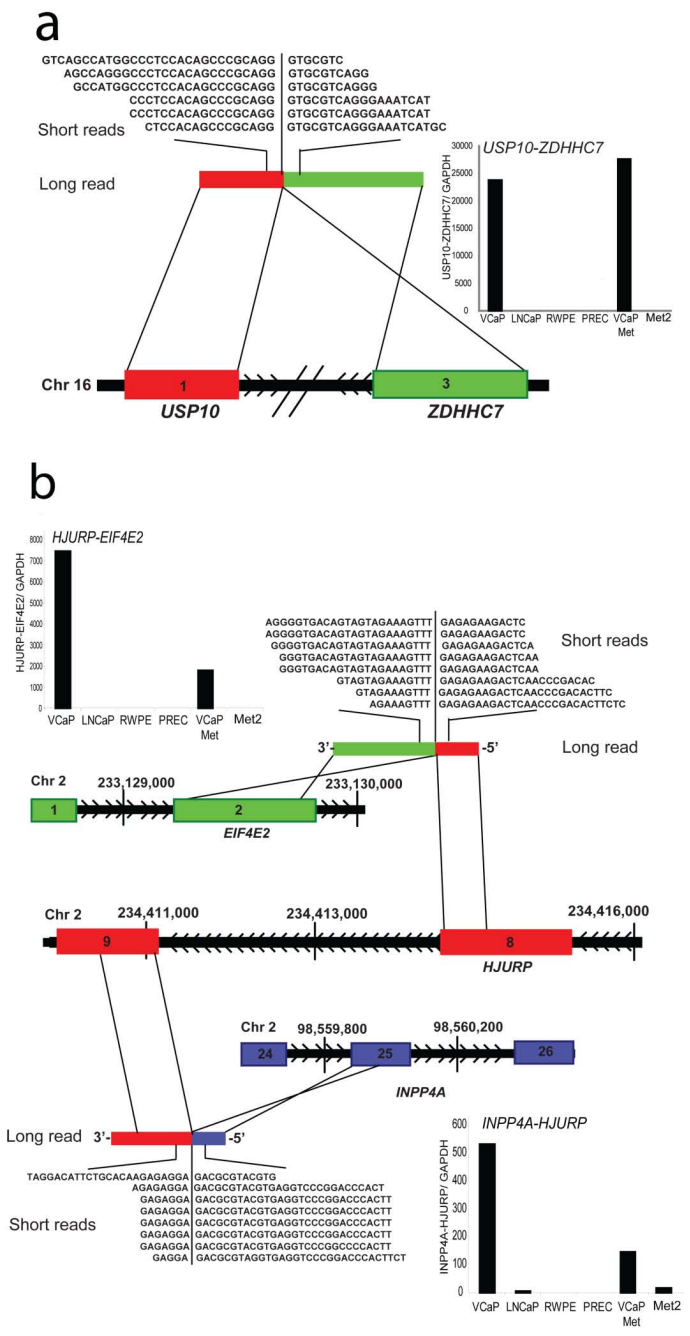


Fig 2. Representative gene fusions characterized in the prostate cancer cell line VCaP

a, Schematic of *USP10-ZDHHC7* fusion on chromosome 16. Exon 1 of *USP10* (red) is fused with exon 3 of *ZDHHC7* (green), located on the same chromosome in opposite orientation. Inset displays histogram of qRT-PCR validation of *USP10-ZDHHC7* transcript. **b**, Schematic of a complex intra-chromosomal rearrangement leading to two gene fusions involving *HJURP* on chromosome 2. Exon 8 of *HJURP* (red) is fused with exon 2 of *EIF4E2* (green) to form *HJURP-EIF4E2*. Exon 25 of *INPP4A* (blue) is fused with exon 9 of *HJURP* (red) to form *INPP4A-HJURP*. Insets display histograms of qRT-PCR validation of *HJURP-EIF4E2* and *INPP4A-HJURP* transcripts.

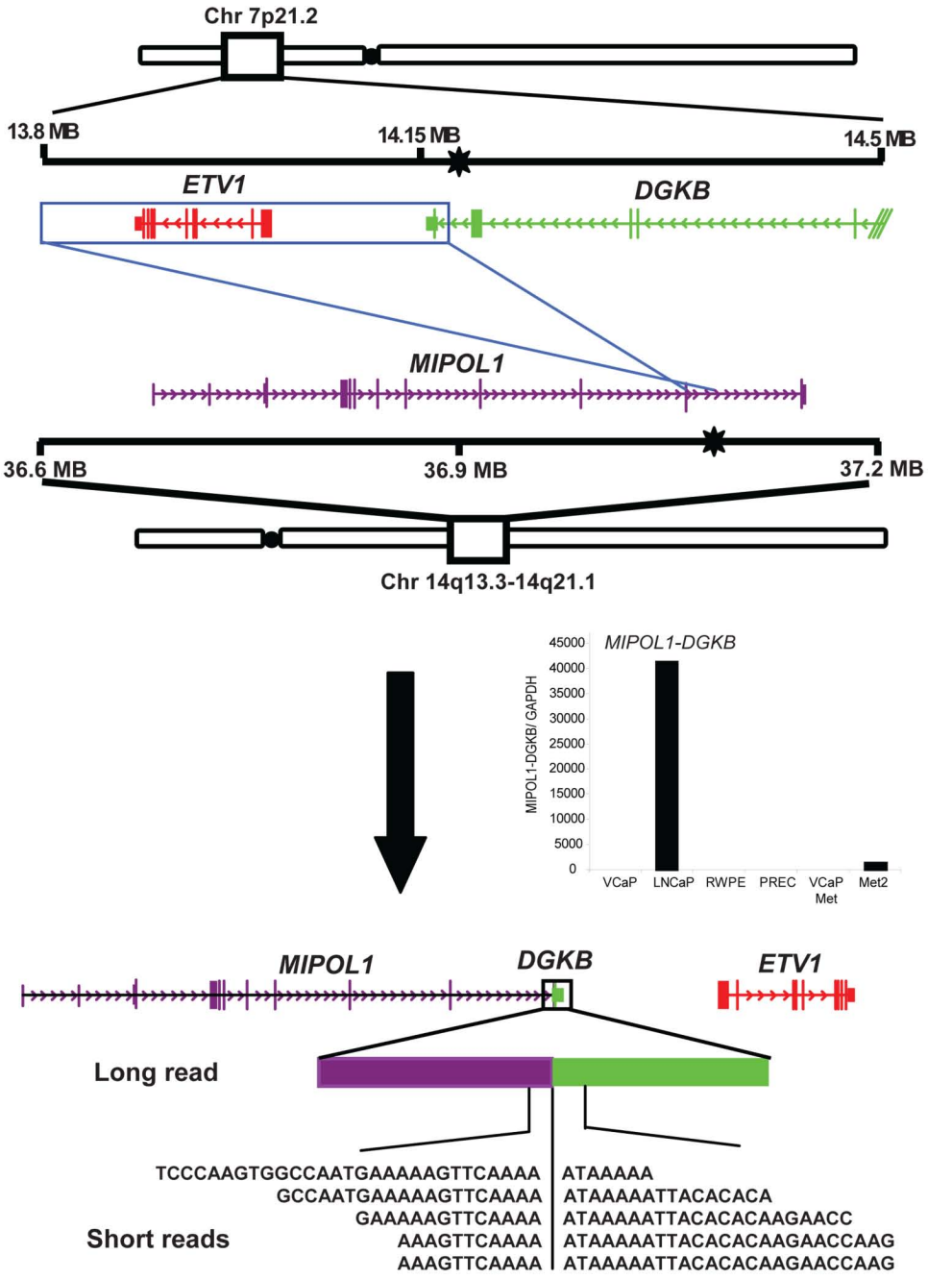


Fig 3. Schematic of *MIPOL1-DGKB* gene fusion in the prostate cancer cell line LNCaP

MIPOL1-DGKB is an inter-chromosomal gene fusion accompanying the cryptic insertion of *ETV1* locus (red) on chromosome 7 into the *MIPOL1*(purple) intron on chromosome 14. Previously determined genomic breakpoints (black stars) are shown in *DGKB* and *MIPOL1*. An insertion event results in the inversion of the 3' end of *DGKB* and *ETV1* into the *MIPOL1* intron between exons 10 and 11. Inset displays histogram of qRT-PCR validation of the *MIPOL1-DGKB* transcript.

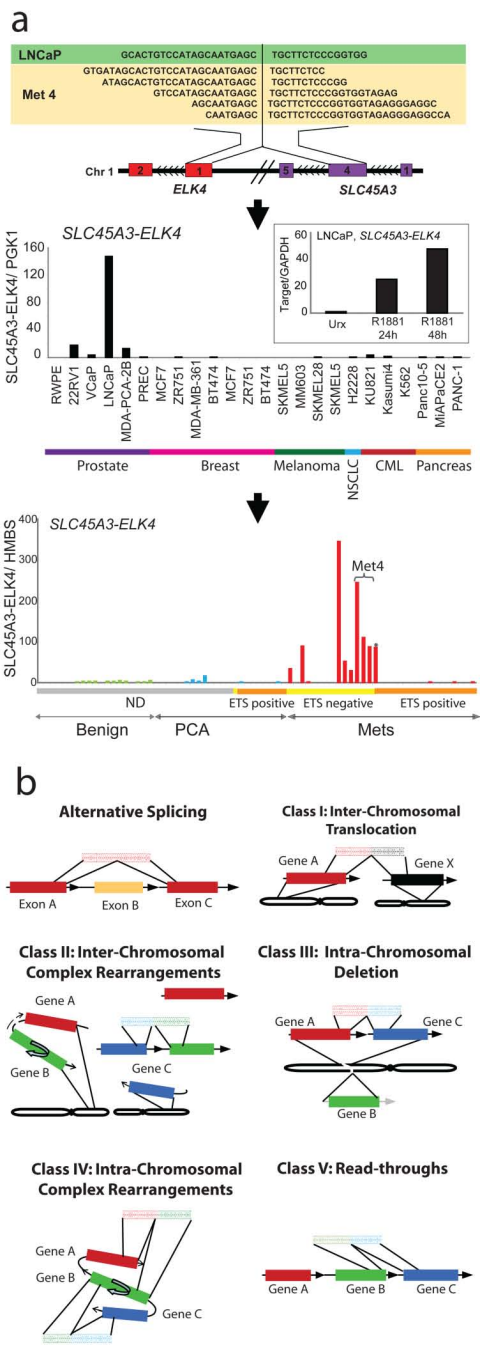


Fig. 4. Discovery of the recurrent *SLC45A3-ELK4* chimera in prostate cancer and a general classification system for chimeric transcripts in cancer

a, Upper panel, schematic of the *SLC45A3-ELK4* chimera located on chromosome 1. Middle panel, qRT-PCR validation of *SLC45A3-ELK4* transcript in a panel of cell lines. Inset, histogram of qRT-PCR assessment of the *SLC45A3-ELK4* transcript in LNCaP cells treated with R1881. Lower panel, histogram of qRT-PCR validation in a panel of prostate tissues-benign adjacent prostate, localized prostate cancer (PCA) and metastatic prostate cancer (Mets). ETS family gene rearrangement status (by FISH) indicated by horizontal colored bars below graph. Grey not determined (ND); yellow, ETS negative; orange, ETS positive. Horizontal bracket indicates three different metastatic tissues from the same patient (Met4).

Asterisk (*) denotes an ETV1 positive sample.**b**, Chimera classification schema (described in the text).