AD_____

Award Number: W81XWH-11-1-0337

TITLE: Discovery of Novel Gene Elements Associated With Prostate Cancer Progression

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

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

REPORT DATE: October 2012

TYPE OF REPORT: Annual

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

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE	**2. REPORT TYPE**	**3. DATES COVERED**
29 October 2012	Annual	30 September 2011 - 29 September 2012

4. TITLE AND SUBTITLE
Discovery of Novel Gene Elements Associated With Prostate Cancer Progression

5a. CONTRACT NUMBER

5b. GRANT NUMBER
W81XWH-11-1-0337

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Arul Chinnaiyan

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

E-Mail: chinnaiyangrants@umich.edu

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

University of Michigan
Ann Arbor, Michigan 48109

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

Long non-coding RNAs (ncRNAs) have gained recent attention as potentially important players in cancer biology. This proposal employs innovative next generation sequencing and ab initio transcriptome assembly methodologies to discover novel ncRNAs and incorporate them into expression signatures that stratify indolent versus aggressive prostate cancer. In this reporting period, we describe the identification and characterization of a prostate cancer-specific ncRNA named PCAT-1. Here, we demonstrated that PCAT-1 is a regulator of cell proliferation and is a target of the Polycomb Repressive Complex 2 (PRC2). We further found that patterns of PCAT-1 and PRC2 expression stratified patient tissues into molecular subtypes distinguished by expression signatures of PCAT-1-repressed target genes. PCAT-1 may potentially serve as a novel biomarker of aggressive disease that can be developed for clinical diagnostic and prognostic utility.

15. SUBJECT TERMS
Long non-coding RNAs, ncRNAs, PCAT-1, next generation sequencing

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			USAMRMC
U	U	U	UU	27	**19b. TELEPHONE NUMBER** *(include area code)*

Table of Contents

PROGRESS REPORT
09/30/11 – 09/29/12

INTRODUCTION:

Noncoding RNAs (ncRNAs) are emerging as key players in human cancer, with the potential to serve as novel markers of disease and to reveal uncharacterized aspects of tumor biology. ncRNAs are often cell-type specific, have biologically important roles (1, 2) and may interact with known cancer genes such as EZH2 (3). Indeed, several well-described examples, such as HOTAIR (3, 4) and ANRIL (5, 6), indicate that ncRNAs may have essential roles in cancer biology, typically facilitating epigenetic gene repression *via* chromatin modifying complexes (7, 8). Moreover, ncRNA expression may be correlated with specific clinical phenotype that can be predictive of patient outcomes and thus may have utility in diagnostic tests (4, 9). The characterization of RNA species, their functions, and their clinical applicability is therefore a major area of biological and clinical importance.

Recently, we reported the discovery of cancer-associated ncRNA transcripts (PCATs) that were identified from a cohort of 102 prostate tissues and cells lines. One ncRNA, PCAT-1, was characterized as a prostate-specific regulator of cell proliferation and we showed that it is a target of the Polycomb Repressive Complex 2 (PRC2). Further, patterns of PCAT-1 and PRC2 expression stratified patient tissues into molecular subtypes distinguished by expression signatures of PCAT-1-repressed target genes (12). The findings from this study were published in *Nature Biotechnology*; a copy of the paper is included with this report.

In this project we employ transcriptome sequencing (RNA-seq) on a panel of normal, cancerous and metastatic prostate tissues to discover novel differentially expressed ncRNA transcripts; validate and characterize top-ranking candidates and examine their functional and clinical role in prostate cancer development and progression.

BODY:

To employ next generation sequencing to comprehensively annotate expressed regions in the prostate cancer transcriptome:

We applied high-throughput transcriptome sequencing (RNA-Seq) on a large cohort of >100 prostate cancer samples in order to define the RNA landscape in this disease. We found that nearly 20% of polyadenylated RNA species represent unannotated transcripts, including 121 intergenic, unannotated lncRNAs whose aberrant expression in prostate cancer ranked highly as some of the most differentially-expressed RNAs in this disease (Figure 1). These 121 unannotated transcripts were ranked and named as Prostate Cancer Associated Transcripts (PCATs) as defined by their fold change in localized tumor relative to benign tissue. We validated multiple unannotated transcripts; qPCR for four transcripts (PCAT-114, PCAT-14, PCAT-43, PCAT-1) on two independent cohorts of prostate tissues confirmed predicted cancer-specific expression patterns (data for PCAT-1 is shown in Figure 1B).

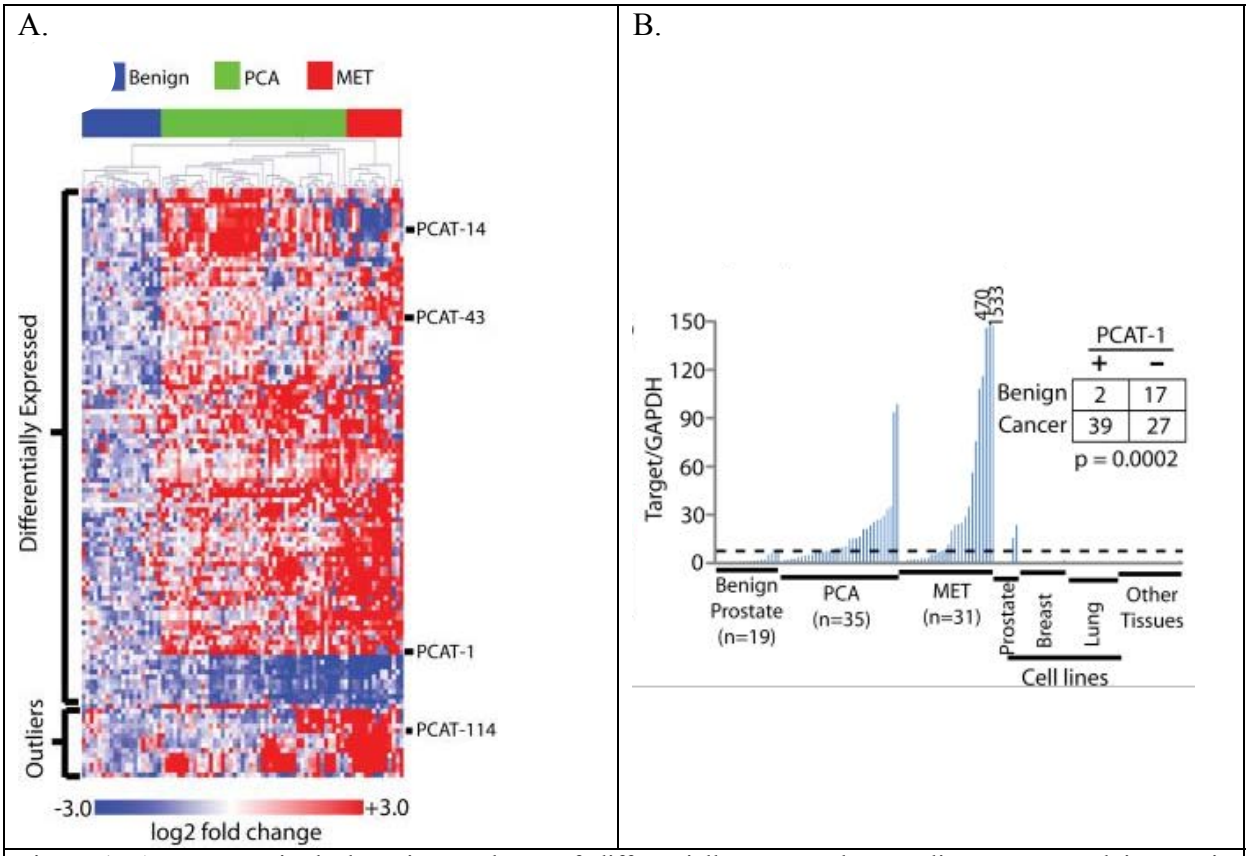

Figure 1: A. Unsupervised clustering analyses of differentially-expressed or outlier unannotated intergenic transcripts clusters benign samples, localized tumors, and metastatic cancers. Expression is plotted as log2 fold change relative to the median of the benign samples. The four transcripts detailed in this study are indicated on the side. B. PCAT-1. qPCR analysis was performed by normalizing to GAPDH and the median expression of the benign samples. *Nat Biotechnol*. 2011;29(8):742-9.

PCAT-1 was strikingly upregulated in a subset of metastatic and high-grade localized (Gleason score ≥7) cancers. To examine the functional role of PCAT-1 in prostate cancer, we stably overexpressed full length PCAT-1 or controls in RWPE benign immortalized prostate cells. We observed a modest but consistent increase in cell proliferation when PCAT-1 was overexpressed at physiological levels (Figure 2, top panel). Next, we designed siRNA oligos to PCAT-1 and performed knockdown experiments in LNCaP cells that express higher levels of PCAT-1. Supporting our overexpression data, knockdown of PCAT-1 with three independent siRNA oligos resulted in a 25% - 50% decrease in cell proliferation in LNCaP cells (Figure 2, bottom panel).

We hypothesized that PCAT-1 may have coordinated expression with the oncoprotein EZH2, a core PRC2 protein that is also upregulated in solid tumors and contributes to a metastatic phenotype (10, 11). Surprisingly, we found that PCAT-1 and EZH2 expression were nearly mutually exclusive (12). This suggests that outlier PCAT-1 and EZH2 expression may define two subsets of high-grade disease. These findings represent the first comprehensive study of lincRNAs in prostate cancer, provide a computational framework for large-scale RNA-Seq

analyses, and describe PCAT-1 as a novel prostate cancer ncRNA functionally implicated in disease progression.

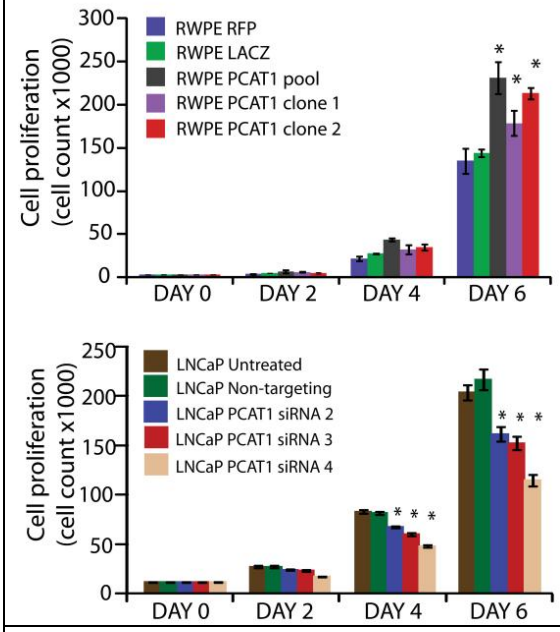

Figure 2: Cell proliferation assays for RWPE benign immortalized prostate cells (Top panel) stably infected with PCAT-1 lentivirus or RFP and LacZ control lentiviruses and in LNCaP using PCAT-1 siRNAs (Bottom panel). An asterisk (*) indicates $p \leq 0.02$ by a two-tailed Students t-test. *Nat Biotechnol.* 2011;29(8):742-9.

We continue to examine other potentially important ncRNAs in prostate cancer. We recently identified a long ncRNA spanning ~250kb in a chromosome 2 gene desert that we have designated "Second Chromosome Locus Associated with Prostate (SCHLAP1)". SCHLAP1 is highly over-expressed in a subset of prostate cancer (about 20%) where it appears to have an effect on invasiveness. We are in the process of elucidating the functional and oncogenic role of SCHAP1.

KEY RESEARCH ACCOMPLISHMENTS:

- We employed transcriptome sequencing on a cohort of 102 prostate tissues and cell lines and found 121 unannotated intergenic ncRNAs as Prostate Cancer Associated Transcripts (PCATs).
- We observed that several PCATs, including PCAT-1, was upregulated markedly in a set of aggressive prostate cancers and
- We found that PCAT-1 upregulation marked a set of aggressive prostate cancers and stratified against EZH2, a prostate cancer oncogene also upregulated in a set of aggressive cancers.
- In vitro, PCAT-1 expression was necessary for cancer cell proliferation and overexpression of PCAT-1 was sufficient to increase benign prostate cell proliferation.

REPORTABLE OUTCOMES:

- Prensner JR, et al. Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression. Nat Biotechnol. 2011;29(8):742-9.

CONCLUSION:

Clinically, ncRNAs are proving to be powerful biomarkers for cancer. ncRNA transcripts can be detected non-invasively in prostate cancer patient fluids such as serum or urine. Previous studies have uncovered PCA3 (prostate cancer antigen-3), a prostate-specific ncRNA upregulated in prostate cancer, as a biomarker for prostate cancer diagnosis (13). Clinical-grade assays for urinary detection of PCA3 are currently available for physicians, and research has shown that adding PCA3 to the standard serum PSA test improves prostate cancer prediction. Future studies exploring the biology of ncRNAs and their role in cancer could lead to the development of additional prostate cancer biomarkers, such as PCAT-1, for the clinical detection and stratification of aggressive prostate cancers.

REFERENCES:

1. Huarte M, et al. A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response. Cell. 2010;142:409–419.
2. Orom UA, et al. Long Noncoding RNAs with Enhancer-like Function in Human Cells. Cell. 2010;143:46–58.
3. Rinn JL, et al. Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. Cell. 2007;129:1311–1323.
4. Gupta RA, et al. Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. Nature. 2010;464:1071–1076.
5. Pasmant E, et al. Characterization of a germ-line deletion, including the entire INK4/ARF locus, in a melanoma-neural system tumor family: identification of ANRIL, an antisense noncoding RNA whose expression coclusters with ARF. Cancer Res. 2007;67:3963–3969.
6. Yap KL, et al. Molecular interplay of the noncoding RNA ANRIL and methylated histone H3 lysine 27 by polycomb CBX7 in transcriptional silencing of INK4a. Mol Cell. 2010;38:662–674.
7. Tsai MC, et al. Long noncoding RNA as modular scaffold of histone modification complexes. Science. 2010;329:689–693.
8. Kotake Y, et al. Long non-coding RNA ANRIL is required for the PRC2 recruitment to and silencing of p15(INK4B) tumor suppressor gene. Oncogene. 2011;30(16):1956-62.
9. de Kok JB, et al. DD3(PCA3), a very sensitive and specific marker to detect prostate tumors. Cancer Res. 2002;62:2695–2698.
10. Kleer CG, et al. EZH2 is a marker of aggressive breast cancer and promotes neoplastic transformation of breast epithelial cells. Proc Natl Acad Sci U S A. 2003;100:11606–11611.
11. Varambally S, et al. The polycomb group protein EZH2 is involved in progression of prostate cancer. Nature. 2002;419:624–629.
12. Prensner JR, et al. Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression. Nat Biotechnol. 2011;29(8):742-9.
13. Prensner JR, Chinnaiyan AM. The emergence of lncRNAs in cancer biology. Cancer Discov. 2011 Oct;1(5):391-407.

Published in final edited form as:
Nat Biotechnol. ; 29(8): 742–749. doi:10.1038/nbt.1914.

Transcriptome Sequencing Identifies PCAT-1, a Novel lincRNA Implicated in Prostate Cancer Progression

John R. Prensner[1,8], **Matthew K. Iyer**[1,8], **O. Alejandro Balbin**[1], **Saravana M. Dhanasekaran**[1,2], **Qi Cao**[1], **J. Chad Brenner**[1], **Bharathi Laxman**[3], **Irfan Asangani**[1], **Catherine Grasso**[1], **Hal D. Kominsky**[1], **Xuhong Cao**[1], **Xiaojun Jing**[1], **Xiaoju Wang**[1], **Javed Siddiqui**[1], **John T. Wei**[4], **Daniel Robinson**[1], **Hari K. Iyer**[5], **Nallasivam Palanisamy**[1,2,6], **Christopher A. Maher**[1,2], and **Arul M. Chinnaiyan**[1,2,4,6,7]

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

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

[3]Department of Medicine, University of Chicago, Chicago, Illinois 60637

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

[5]Department of Statistics, Colorado State University, Fort Collins, Colorado 80523

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

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

Abstract

High-throughput sequencing of polyA+ RNA (RNA-Seq) in human cancer shows remarkable potential to identify both novel markers of disease and uncharacterized aspects of tumor biology, particularly non-coding RNA (ncRNA) species. We employed RNA-Seq on a cohort of 102 prostate tissues and cells lines and performed *ab initio* transcriptome assembly to discover unannotated ncRNAs. We nominated 121 such Prostate Cancer Associated Transcripts (PCATs) with cancer-specific expression patterns. Among these, we characterized *PCAT-1* as a novel prostate-specific regulator of cell proliferation and target of the Polycomb Repressive Complex 2 (PRC2). We further found that high *PCAT-1* and PRC2 expression stratified patient tissues into molecular subtypes distinguished by expression signatures of *PCAT-1*-repressed target genes. Taken together, the findings presented herein identify *PCAT-1* as a novel transcriptional repressor

Please address requests to: Arul M. Chinnaiyan, M.D., Ph.D., Michigan Center for Translational Pathology, University of Michigan Medical School, 1400 E. Medical Center Dr, 5316 Cancer Center, Ann Arbor, Michigan 48109-0602 Phone: (734) 615-4062 Fax: (734) 615-4498 (arul@umich.edu).

[8]These authors contributed equally

Author Contributions

M.K.I., J.R.P. and A.M.C. designed the project and directed experimental studies. M.K.I., O.A.B., C.G., and C.A.M. developed computational platforms and performed sequencing data analysis. M.K.I, O.A.B., and H.I. performed statistical analyses. J.R.P., S.M.D., J.C.B., Q.C., N.P., H.D.K., B.L., X.W., and D.R. performed experimental studies. J.S. and J.T.W. coordinated biospecimens. M.K.I., J.R.P. and A.M.C. interpreted data and wrote the manuscript.

Data Deposition Data from RNA-Seq experiments are deposited at the NCBI Gene Expression Omnibus as GSE25183. *PCAT-1* and *PCAT-14* nucleotide sequences are deposited at GenBank as HQ605084 and HQ605085, respectively.

Disclosures and Competing Financial Interests The University of Michigan has filed for a patent on the detection of gene fusions in prostate cancer, on which A.M.C. is a co-inventor. The diagnostic field of use for ETS gene fusions has been licensed to GenProbe Inc. The University of Michigan has a sponsored research agreement with GenProbe which is unrelated to this study. GenProbe has had no role in the design or experimentation of this study, nor has it participated in the writing of the manuscript.

implicated in subset of prostate cancer patients. These findings establish the utility of RNA-Seq to identify disease-associated ncRNAs that may improve the stratification of cancer subtypes.

Keywords

prostate cancer; transcriptome; next generation sequencing; non-coding RNA; EZH2

Introduction

Recently, next generation transcriptome sequencing (RNA-Seq) has provided a method to delineate the entire set of transcriptional aberrations in a disease, including novel transcripts and non-coding RNAs (ncRNAs) not measured by conventional analyses[1-5]. To facilitate interpretation of sequence read data, existing computational methods typically process individual samples using either short read gapped alignment followed by *ab initio* reconstruction[2, 3], or *de novo* assembly of read sequences followed by sequence alignment[4, 5]. These methods provide a powerful framework to uncover uncharacterized RNA species, including antisense transcripts, short RNAs <250 bps, or long ncRNAs (lincRNAs) >250 bps.

While still largely unexplored, ncRNAs, particularly lincRNAs, have emerged as a new aspect of biology, with evidence suggesting that they are frequently cell-type specific, contribute important functions to numerous systems[6, 7], and may interact with known cancer genes such as *EZH2*[8]. Indeed, several well-described examples, such as *HOTAIR*[8, 9] and *ANRIL*[10, 11], indicate that ncRNAs may be essential actors in cancer biology, typically facilitating epigenetic gene repression via chromatin modifying complexes[12, 13]. Moreover, ncRNA expression may confer clinical information about patient outcomes and have utility as diagnostic tests[9, 14]. The characterization of RNA species, their functions, and their clinical applicability is therefore a major area of biological and clinical importance.

Here, we describe a comprehensive analysis of lincRNAs in 102 prostate cancer tissue samples and cell lines by RNA-Seq. We employ *ab initio* computational approaches to delineate the annotated and unannotated transcripts in this disease, and we find 121 ncRNAs, termed Prostate Cancer Associated Transcripts (PCATs), whose expression patterns distinguish benign, localized cancer, and metastatic cancer samples. Notably, we discover *PCAT-1*, a novel prostate cancer ncRNA alternately demonstrating either repression by PRC2 or an active role in promoting cell proliferation through transcriptional regulation of target genes. Our findings describe the first comprehensive study of lincRNAs in prostate cancer, provide a computational framework for large-scale RNA-Seq analyses, and describe *PCAT-1* as a novel prostate cancer ncRNA functionally implicated in disease progression.

Results

RNA-Seq analysis of the prostate cancer transcriptome

Over two decades of research has generated a genetic model of prostate cancer based on numerous neoplastic events, such as loss of the *PTEN*[15] tumor suppressor gene and gain of oncogenic ETS transcription factor gene fusions[16-18] in large subsets of prostate cancer patients. We hypothesized that prostate cancer similarly harbored disease-associated ncRNAs in molecular subtypes.

To pursue this hypothesis, we employed transcriptome sequencing on a cohort of 102 prostate tissues and cell lines (20 benign adjacent prostates (benign), 47 localized tumors

(PCA), and 14 metastatic tumors (MET) and 21 prostate cell lines). From a total of 1.723 billion sequence fragments from 201 lanes of sequencing (108 paired-end, 93 single read on the Illumina Genome Analyzer and Genome Analyzer II), we performed short read gapped alignment[19] and recovered 1.41 billion mapped reads, with a median of 14.7 million mapped reads per sample (**Supplementary Table 1** for sample information). We used the Cufflinks *ab initio* assembly approach[3] to produce, for each sample, the most probable set of putative transcripts that served as the RNA templates for the sequence fragments in that sample (**Fig. 1a** and **Supplementary Figs. 1** and **2**).

As expected from a large tumor tissue cohort, individual transcript assemblies may exhibit sources of "noise", such as artifacts of the sequence alignment process, unspliced intronic pre-mRNA, and genomic DNA contamination. To exclude these from our analyses, we trained a decision tree to classify transcripts as "expressed" versus "background" on the basis of transcript length, number of exons, recurrence in multiple samples, and other structural characteristics (**Fig. 1b left** and **Supplementary Methods**). The classifier demonstrated a sensitivity of 70.8% and specificity of 88.3% when trained using transcripts that overlapped genes in the AceView database[20], including 11.7% of unannotated transcripts that were classified as "expressed" (**Fig. 1b right**). We then clustered the "expressed" transcripts into a consensus transcriptome and applied additional heuristic filters to further refine the assembly (**Supplementary Methods**). The final *ab initio* transcriptome assembly yielded 35,415 distinct transcriptional loci (**Supplementary Table 2** and **Supplementary Methods**).

Discovery of prostate cancer non-coding RNAs

We compared the assembled prostate cancer transcriptome to the UCSC, Ensembl, Refseq, Vega, and ENCODE gene databases to identify and categorize transcripts (**Fig. 1c**). While the majority of the transcripts (77.3%) corresponded to annotated protein coding genes (72.1%) and non-coding RNAs (5.2%), a significant percentage (19.8%) lacked any overlap and were designated "unannotated" (**Fig. 2a**). These included partially intronic antisense (2.44%), totally intronic (12.1%), and intergenic transcripts (5.25%), consistent with previous reports of unannotated transcription[21, 22, 23]. Due to the added complexity of characterizing antisense or partially intronic transcripts without strand-specific RNA-Seq libraries, we focused on totally intronic and intergenic transcripts.

Global characterization of novel intronic and intergenic transcripts demonstrated that they were more highly expressed (**Fig. 2b**), had greater overlap with expressed sequence tags (ESTs) (**Supplementary Fig. 3**), and displayed a clear but subtle increase in conservation over randomly permuted controls (novel intergenic transcripts $p = 2.7 \times 10^{-4} \pm 0.0002$ for $0.4 < \omega < 0.8$; novel intronic transcripts $p = 2.6 \times 10^{-5} \pm 0.0017$ for $0 < \omega < 0.4$, Fisher's exact test, **Fig. 2c**). By contrast, unannotated transcripts scored lower than protein-coding genes for these metrics, which corroborates data in previous reports[2, 24]. Interestingly, a small subset of novel intronic transcripts showed a profound degree of conservation (**Fig. 2c**, insert). Finally, analysis of coding potential revealed that only 5 of 6,144 transcripts harbored a high quality open reading frame (ORF), indicating that the vast majority of these transcripts represent ncRNAs (**Supplementary Fig. 4**).

To determine whether our unannotated transcripts were supported by histone modifications defining active transcriptional units, we used published prostate cancer ChIP-Seq data for two prostate cell lines[25], VCaP and LNCaP (**Supplementary Table 3**). After filtering our dataset for transcribed repetitive elements known to display alternative patterns of histone modifications[26], we observed a strong enrichment for histone modifications characterizing transcriptional start sites (TSSs) and active transcription, including H3K4me2, H3K4me3, Acetyl-H3 and RNA polymerase II (**Fig. 2d-g**) but not H3K4me1, which characterizes

enhancer regions[27] (**Supplementary Figs. 5** and **6**). Interestingly, intergenic ncRNAs showed greater enrichment compared to intronic ncRNAs in these analyses (**Fig. 2d-g**).

To elucidate global changes in transcript abundance in prostate cancer, we performed a differential expression analysis for all transcripts. We found 836 genes differentially-expressed between benign samples and localized tumors (FDR < 0.01), with annotated protein-coding and ncRNA genes constituting 82.8% and 7.4% of differentially-expressed genes, respectively, including known prostate cancer biomarkers such *AMACR*[28], *HPN*[29], and *PCA3*[14] (**Fig. 2h, Supplementary Fig. 2** and **Supplementary Table 4**). Finally, 9.8% of differentially-expressed genes corresponded to unannotated ncRNAs, including 3.2% within gene introns and 6.6% in intergenic regions.

Characterization of Prostate Cancer Associated Transcripts

As ncRNAs may contribute to human disease[6-9], we identified aberrantly expressed uncharacterized ncRNAs in prostate cancer. We found a total of 1,859 unannotated lincRNAs throughout the human genome. Overall, these intergenic RNAs resided approximately half-way between two protein coding genes (**Supplementary Fig. 7**), and over one-third (34.1%) were ≥10kb from the nearest protein-coding gene, which is consistent with previous reports[30] and supports the independence of intergenic ncRNAs genes. For example, visualizing the Chr15q arm using the Circos program (http://mkweb.bcgsc.ca/circos) illustrated genomic positions of eighty-nine novel intergenic transcripts, including one differentially-expressed gene centromeric to *TLE3* (**Supplementary Fig. 8**).

A focused analysis of the 1,859 unannotated intergenic RNAs yielded 106 that were differentially expressed in localized tumors (FDR < 0.05, **Fig. 3a**). A cancer outlier expression analysis (**Supplementary Methods**) similarly nominated numerous unannotated ncRNA outliers (**Fig. 3b**) as well as known prostate cancer outliers, such as *ERG*[18], *ETV1*[17, 18], *SPINK1*[31] and *CRISP3*[32]. Merging these results produced a set of 121 unannotated transcripts that accurately discriminated benign, localized tumor, and metastatic prostate samples by unsupervised clustering (**Fig. 3a**). Indeed, clustering analyses using novel ncRNA outliers also suggested disease subtypes (**Supplementary Fig. 9**). These 121 unannotated transcripts were ranked and named as Prostate Cancer Associated Transcripts (PCATs) according to their fold change in localized tumor versus benign tissue (**Supplementary Tables 5** and **6**).

Validation of novel ncRNAs

To gain confidence in our transcript nominations, we validated multiple unannotated transcripts *in vitro* by reverse transcription PCR (RT-PCR) and quantitative real-time PCR (qPCR) (**Supplementary Fig. 10**). qPCR for four transcripts (*PCAT-114*, *PCAT-14*, *PCAT-43*, *PCAT-1*) on two independent cohorts of prostate tissues confirmed predicted cancer-specific expression patterns (**Fig. 3c-f** and **Supplementary Fig. 11**). Interestingly, all four are prostate-specific, with minimal expression seen by qPCR in breast (n=14) or lung cancer (n=16) cell lines or in 19 normal tissue types (**Supplementary Table 8**). This is further supported by expression analysis of these transcripts in our RNA-Seq compendium of 13 tumor types, representing 325 samples (**Supplementary Fig. 12**). This tissue specificity was not necessarily due to regulation by androgen receptor signaling, as only *PCAT-14* expression was induced when androgen responsive VCaP and LNCaP cells were treated with the synthetic androgen R1881, consistent with previous data from this locus[17] (**Supplementary Fig. 13**). *PCAT-1* and *PCAT-14* also showed cancer-specific upregulation when tested on a panel of matched tumor-normal samples (**Supplementary Fig. 14**).

Of note, *PCAT-114*, which ranks as the #5 best outlier, just ahead of *ERG* (**Fig. 3b** and **Supplementary Table 7**), appears as part of a large, >500 kb locus of expression in a gene desert in Chr2q31. We termed this region Second Chromosome Locus Associated with Prostate-1 (SChLAP1) (**Supplementary Fig. 15**). Careful analysis of the SChLAP1 locus revealed both discrete transcripts and intronic transcription, highlighting this region as an intriguing aspect of the prostate cancer transcriptome.

PCAT-1, a novel prostate cancer lincRNA

To explore several transcripts more closely, we performed 5' and 3' rapid amplification of cDNA ends (RACE) for *PCAT-1* and *PCAT-14*. Interestingly, the *PCAT-14* locus contained components of viral ORFs from the HERV-K endogenous retrovirus family (**Supplementary Fig. 16**), whereas *PCAT-1* incorporates portions of a *mariner* family transposase[33, 34], an Alu, and a viral long terminal repeat (LTR) promoter region (**Fig. 4a** and **Supplementary Fig. 17**). While *PCAT-14* was upregulated in localized prostate cancer but largely absent in metastases (**Fig. 3c**), *PCAT-1* was strikingly upregulated in a subset of metastatic and high-grade localized (Gleason score ≥7) cancers (**Fig. 3f** and **Supplementary Fig. 11**). Because of this notable profile, we hypothesized that *PCAT-1* may have coordinated expression with the oncoprotein *EZH2*, a core PRC2 protein that is upregulated in solid tumors and contributes to a metastatic phenotype[35, 36]. Surprisingly, we found that *PCAT-1* and *EZH2* expression were nearly mutually exclusive (**Fig. 4b**), with only one patient showing outlier expression of both. This suggests that outlier *PCAT-1* and *EZH2* expression may define two subsets of high-grade disease.

PCAT-1 is located in the chromosome 8q24 gene desert approximately 725 kb upstream of the *c-MYC* oncogene. To confirm that *PCAT-1* is a non-coding gene, we cloned the full-length *PCAT-1* transcript and performed *in vitro* translational assays, which were negative as expected (**Supplementary Fig. 18**). Next, since Chr8q24 is known to harbor prostate cancer-associated single nucleotide polymorphisms (SNPs) and to exhibit frequent chromosomal amplification[37-42], we evaluated whether the relationship between *EZH2* and *PCAT-1* was specific or generalized. To address this, we measured expression levels of *c-MYC* and *NCOA2*, two proposed targets of Chr8q amplification[39, 42], by qPCR. Neither *c-MYC* nor *NCOA2* levels showed striking expression relationships to *PCAT-1*, *EZH2*, or each other (**Supplementary Fig. 19**). Likewise, *PCAT-1* outlier expression was not dependent on Chr8q24 amplification, as highly expressing localized tumors often did not have 8q24 amplification and high copy number gain of 8q24 was not sufficient to upregulate *PCAT-1* (**Supplementary Figs. 20** and **21**).

PCAT-1 Function and Regulation

Despite reports showing that upregulation of the ncRNA *HOTAIR* participates in PRC2 function in breast cancer[9], we do not observe strong expression of this ncRNA in prostate (**Supplementary Fig. 22**), suggesting that other ncRNAs may be important in this cancer. To determine the mechanism for the expression profiles of *PCAT-1* and *EZH2*, we inhibited *EZH2* activity in VCaP cells, which express low-to-moderate levels of *PCAT-1*. Knockdown of *EZH2* by shRNA or pharmacologic inhibition of *EZH2* with the inhibitor 3-deazaneplanocin A (DZNep) caused a dramatic upregulation in *PCAT-1* expression levels (**Fig. 4c,d**), as did treatment of VCaP cells with the demethylating agent 5'deoxyazacytidine, the histone deacetylase inhibitor SAHA, or both (**Fig. 4e**). Chromatin immunoprecipitation (ChIP) assays also demonstrated that SUZ12, a core PRC2 protein, directly binds the *PCAT-1* promoter approximately 1kb upstream of the TSS (**Fig. 4f**). Interestingly, RNA immunoprecipitation (RIP) similarly showed binding of *PCAT-1* to SUZ12 protein in VCaP cells (**Supplementary Fig. 23a**). RIP assays followed by RNase A, RNase H, or DNase I treatment either abolished, partially preserved, or totally preserved this

interaction, respectively (**Supplementary Fig. 23b**). This suggests that *PCAT-1* exists primarily as a single-stranded RNA and secondarily as a RNA/DNA hybrid.

To explore the functional role of *PCAT-1* in prostate cancer, we stably overexpressed full length *PCAT-1* or controls in RWPE benign immortalized prostate cells. We observed a modest but consistent increase in cell proliferation when *PCAT-1* was overexpressed at physiological levels (**Fig. 5a** and **Supplementary Fig. 24**). Next, we designed siRNA oligos to *PCAT-1* and performed knockdown experiments in LNCaP cells, which express higher levels of *PCAT-1* without PRC2-mediated repression (**Supplementary Fig. 25**). Supporting our overexpression data, knockdown of *PCAT-1* with three independent siRNA oligos resulted in a 25% - 50% decrease in cell proliferation in LNCaP cells (**Fig. 5b**), but not control DU145 cells lacking *PCAT-1* expression (**Supplementary Fig. 26**) or VCaP cells, in which *PCAT-1* is expressed but repressed by PRC2 (**Supplementary Fig. 27**).

Gene expression profiling of LNCaP knockdown samples on cDNA microarrays indicated that *PCAT-1* modulates the transcriptional regulation of 370 genes (255 upregulated, 115 downregulated; FDR ≤ 0.01) (**Supplementary Fig. 28** and **Supplementary Table 9**). Gene ontology analysis of the upregulated genes showed preferential enrichment for cellular processes such as mitosis and cell cycle, whereas the downregulated genes had no concepts showing statistical significance (**Fig. 5c** and **Supplementary Table 10**). These results suggest that *PCAT-1*'s function is predominantly repressive in nature, similar to other lincRNAs. We next validated expression changes in three key *PCAT-1* target genes (*BRCA2*, *CENPE* and *CENPF*) whose expression is upregulated upon *PCAT-1* knockdown (**Fig. 5a**) in LNCaP and VCaP cells, the latter of which appear less sensitive to *PCAT-1* knockdown likely due to lower overall expression levels of this transcript.

PCAT-1 signatures in prostate cancer

Because of the regulation of *PCAT-1* by PRC2 in VCaP cells, we hypothesized that knockdown of *EZH2* would also downregulate *PCAT-1* targets as a secondary phenomenon due to the subsequent upregulation of *PCAT-1*. Simultaneous knockdown of *PCAT-1* and *EZH2* would thus abrogate expression changes in *PCAT-1* target genes. Performing this experiment in VCaP cells demonstrated that *PCAT-1* target genes were indeed downregulated by *EZH2* knockdown, and that this change was either partially or completely reversed using siRNA oligos to *PCAT-1* (**Fig. 6a**), lending support to the role of *PCAT-1* as a transcriptional repressor. Taken together, these results suggest that *PCAT-1* biology may exhibit two distinct modalities: one in which PRC2 represses *PCAT-1* and a second in which active *PCAT-1* promotes cell proliferation. *PCAT-1* and PRC2 may therefore characterize distinct subsets of prostate cancer.

To examine our clinical cohort, we used qPCR to measure expression of *BRCA2*, *CENPE*, and *CENPF* in our tissue samples. Consistent with our model, we found that *PCAT-1*-expressing samples tended to have low expression of *PCAT-1* target genes (**Fig. 6b**). Moreover, comparing *EZH2*-outlier and *PCAT-1*-outlier patients (see **Fig. 4b**), we found that two distinct patient phenotypes emerged: those with high *EZH2* tended to have high levels of *PCAT-1* target genes; and those with high *PCAT-1* expression displayed the opposite expression pattern (**Fig. 6c**). Network analysis of the top 20 upregulated genes following *PCAT-1* knockdown with the HefaLMP tool[43] further suggested that these genes form a coordinated network (**Fig. 6d**), corroborating our previous observations. Taken together, these results provide initial data into the composition and function of the prostate cancer ncRNA transcriptome.

Discussion

This study represents the largest RNA-Seq analysis to date and the first to comprehensively analyze a common epithelial cancer from a large cohort of human tissue samples. As such, our study has adapted existing computational tools intended for small-scale use[3] and developed new methods in order to distill large numbers of transcriptome datasets into a single consensus transcriptome assembly that reflects a coherent biological picture.

Among the numerous uncharacterized ncRNA species detected by our study, we have focused on 121 prostate cancer-associated PCATs, which we believe represent a set of uncharacterized ncRNAs that may have important biological functions in this disease. In this regard, these data contribute to a growing body of literature supporting the importance of unannotated ncRNA species in cellular biology and oncogenesis[6-12], and broadly our study confirms the utility of RNA-Seq in defining functionally-important elements of the genome[2-4].

Of particular interest is our discovery of the prostate-specific ncRNA gene *PCAT-1*, which is markedly overexpressed in a subset of prostate cancers, particularly metastases, and may contribute to cell proliferation in these tumors. It is also notable that *PCAT-1* resides in the 8q24 “gene desert” locus, in the vicinity of well-studied prostate cancer risk SNPs and the *c-MYC* oncogene, suggesting that this locus—and its frequent amplification in cancer—may be linked to additional aspects of cancer biology. In addition, the interplay between PRC2 and *PCAT-1* further suggests that this ncRNA may have an important role in prostate cancer progression (**Fig. 6e**). Other ncRNAs identified by this analysis may similarly contribute to prostate cancer as well. Furthermore, recent pre-clinical efforts to detect prostate cancer non-invasively through the collection of patient urine samples have shown promise for several urine-based prostate cancer biomarkers, including the ncRNA *PCA3*[44, 45]. While additional studies are needed, our identification of ncRNA biomarkers for prostate cancer suggests that urine-based assays for these ncRNAs may also warrant investigation, particularly for those that may stratify patient molecular subtypes.

Taken together, our findings support an important role for tissue-specific ncRNAs in prostate cancer and suggest that cancer-specific functions of these ncRNAs may help to “drive” tumorigenesis. We further speculate that specific ncRNA signatures may occur universally in all disease states and applying these methodologies to other diseases may reveal key aspects of disease biology and clinically important biomarkers.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Kalpana Ramnarayanan and Roger Morey for technical assistance with next generation sequencing. We thank Robert J. Lonigro, Shanker Kaylana-Sundaram, Terrence Barrette, and Mike Quist for help with sequencing data analysis, and Rohit Mehra, Bo Han, and Khalid Suleman for prostate tissue specimens. We thank Cole Trapnell and Geo Pertea for assistance with computational analyses. We thank Scott Tomlins, Yi-Mi Wu, Sameek Roychowdhury and members of the Chinnaiyan lab for advice and discussions. We thank Rameen Beroukhim for guidance.

This work was supported in part by the NIH Prostate Specialized Program of Research Excellence grant P50CA69568, the Early Detection Research Network grant UO1 CA111275 (to A.M.C), the US National Institutes of Health R01CA132874-01A1 (to A.M.C.), the Department of Defense grant PC100171 and W81XWH-11-1-0337 (to A.M.C.) and the National Center for Functional Genomics supported by the Department of Defense (to A.M.C.). A.M.C. is supported by the Doris Duke Charitable Foundation Clinical Scientist Award, a Burroughs Welcome Foundation Award in Clinical Translational Research and the Prostate Cancer Foundation. A.M.C. is an American

Cancer Society Research Professor. C.A.M. was supported by the American Association of Cancer Research Amgen Fellowship in Clinical/Translational Research, the Canary Foundation and American Cancer Society Early Detection Postdoctoral Fellowship, and a Prostate Cancer Foundation Young Investigator Award. Q.C. was supported by a Department of Defense Postdoctoral Fellowship grant PC094725. J.R.P was supported by the NIH Cancer Biology Training Grant CA009676-18 and the Department of Defense Predoctoral Fellowship PC094290. M.K.I was supported by the Department of Defense Predoctoral Fellowship W81XWH-11-1-0136. J.R.P and M.K.I are Fellows of the University of Michigan Medical Scientist Training Program.

Online Methods

Cell lines, treatments, and tissues

All prostate cell lines were obtained from the American Type Culture Collection (Manassas, VA), except for PrEC (benign non-immortalized prostate epithelial cells) and PrSMC (prostate smooth muscle cells), which were obtained from Lonza (Basel, Switzerland). Cell lines were maintained using standard media and conditions.

For androgen treatment experiments, LNCaP and VCaP cells were grown in androgen-depleted media for 48 hours and subsequently treated with 5nM methyltrienolone (R1881, NEN Life Science Products) or an equivalent volume of ethanol for 48 hours before harvesting the cells. For drug treatments, VCaP cells were treated with 20uM 5'deoxyazacytidine (Sigma), 500 nM HDAC inhibitor suberoylanilide hydroxamic acid (SAHA) (Biovision Inc.), or both 5'deoxyazacytidine and SAHA. 5'deoxyazacytidine treatments were performed for 6 days with media and drug re-applied every 48 hours. SAHA treatments were performed for 48 hours. DMSO treatments were performed for 6 days. For DZNep treatments, DZNep was dissolved in DMSO and VCAP cells were treated with either 0.1uM of DZNep or vehicle control; RNA was harvested at 72 hours and 144 hours.

Prostate tissues were obtained from the radical prostatectomy series and Rapid Autopsy Program at the University of Michigan tissue core as part of the University of Michigan Prostate Cancer Specialized Program Of Research Excellence (S.P.O.R.E.). All tissue samples were collected with informed consent under an Institutional Review Board (IRB) approved protocol at the University of Michigan.

RNA isolation; cDNA synthesis; and PCR experiments

Total RNA was isolated using Trizol and an RNeasy Kit (Invitrogen) with DNase I digestion according to the manufacturer's instructions. RNA integrity was verified on an Agilent Bioanalyzer 2100 (Agilent Technologies, Palo Alto, CA). cDNA was synthesized from total RNA using Superscript III (Invitrogen) and random primers (Invitrogen). Quantitative Real-time PCR (qPCR) was performed using Power SYBR Green Mastermix (Applied Biosystems, Foster City, CA) on an Applied Biosystems 7900HT Real-Time PCR System. Reverse-transcription PCR (RT-PCR) was performed with Platinum Taq High Fidelity polymerase (Invitrogen). All oligonucleotide primers are listed in **Supplementary Table 12**. For PCR product sequencing, PCR products were resolved on a 1.5% agarose gel, and either sequenced directly or extracted using a Gel Extraction kit (Qiagen) and cloned into pcr4-TOPO vectors (Invitrogen). PCR products were bidirectionally sequenced at the University of Michigan Sequencing Core.

RNA-ligase-mediated rapid amplification of cDNA ends (RACE)

5' and 3' RACE was performed using the GeneRacer RLM-RACE kit (Invitrogen) according to the manufacturer's instructions. RACE PCR products were obtained using

Platinum Taq High Fidelity polymerase (Invitrogen), the supplied GeneRacer primers, and appropriate gene-specific primers indicated in **Supplementary Table 12.**

RNA-Seq library preparation

2μg total RNA was selected for polyA+ RNA using Sera-Mag oligo(dT) beads (Thermo Scientific), and paired-end next-generation sequencing libraries were prepared as previously described[46] using Illumina-supplied universal adaptor oligos and PCR primers (Illumina). Samples were sequenced in a single lane on an Illumina Genome Analyzer I or Genome Analyzer II flowcell using previously described protocols. 36-45mer paired-end reads were according to the protocol provided by Illumina.

Overexpression studies

PCAT-1 full length transcript was cloned into the pLenti6 vector (Invitrogen) along with RFP and LacZ controls. After confirmation of the insert sequence, lentiviruses were generated at the University of Michigan Vector Core and transfected into the benign immortalized prostate cell line RWPE. RWPE cells stably expressing *PCAT-1*, RFP or LacZ were generated by selection with blasticidin (Invitrogen), and 10,000 cells were plated into 12-well plates. Cells were harvested and counted at day 2, day 4, and day 6 post-plating with a Coulter counter.

siRNA knockdown studies

Cells were plated and transfected with 20uM experimental siRNA oligos or non-targeting controls twice, at 12 hours and 36 hours post-plating. Knockdowns were performed with Oligofectamine in OptiMEM media. Knockdown efficiency was determined by qPCR. siRNA sequences (in sense format) for *PCAT-1* knockdown were as follows: siRNA 1 UUAAAGAGAUCCACAGUUAUU; siRNA 2 GCAGAAACACCAAUGGAUAUU; siRNA 3 AUACAUAAGACCAUGGAAAU; siRNA 4 GAACCUAACUGGACUUUAAUU. For *EZH2* siRNA, the following sequence was used: GAGGUUCAGACGAGCUGAUUU.

shRNA knockdown and western blotting

Cells were seeded at 50-60% confluency, incubated overnight, and transfected with EZH2 or non-targeting shRNA lentiviral constructs as described in for 48 hours. GFP+ cells were drug-selected using 1 ug/mL puromycin. RNA and protein were harvested for PCR and Western blotting according to standard protocols. For Western blotting, PVDF membranes (GE Healthcare) were incubated overnight at 4C with either *EZH2* mouse monoclonal (1:1000, BD Biosciences, no. 612666), or *B-Actin* (Abcam, ab8226) for equal loading.

Gene expression profiling

Agilent Whole Human Genome Oligo Microarray (Santa Clara, CA) was used for cDNA profiling of PCAT-1 siRNA knockdown samples or non-targeting control according to standard protocols. All samples were run in technical triplicates against non-targeting control siRNA. Expression array data was processed using the SAM method[47] with an FDR ≤ 0.01. Up- and down-regulated probes were separated and analyzed using the DAVID bioinformatics platform[48].

Chromatin immunoprecipitation

ChIP assays were preformed as previously described[25], where 4 – 7 μg of the following antibodies were used: IgG (Millipore, PP64), SUZ12 (Cell Signaling, #3737), and SUZ12 (Abcam, ab12073). ChIP-PCR reactions were performed in triplicate with SYBRGreen using 1:150th of the ChIP product per reaction.

In vitro translation

Full length *PCAT-1*, Halo-tagged *ERG* or *GUS* positive control were cloned into the PCR2.1 entry vector (Invitrogen) and *in vitro* translational assays were performed using the TnT Quick Coupled Transcription/Translation System (Promega) with 1mM methionine and Transcend Biotin-Lysyl-tRNA (Promega) according to the manufacturer's instructions.

Bioinformatic analyses

Sequencing reads were aligned with TopHat[19], and *ab initio* assembly was performed with Cufflinks[3]. Transcriptome libraries were merged and statistical classifiers were developed and employed to filter low confidence transcripts. Nominated transcripts were compared to UCSC, RefSeq, Vega, Ensembl, and ENCODE database, and coding potential was determined with the txCdsPredict program from UCSC. Transcript conservation was determined with the SiPhy package. Differential expression analysis was performed using SAM methodology, and outlier analysis using a modified COPA method. See the **Supplementary Methods** for details on the bioinformatics methods used.

Statistical analyses for experimental studies

All data are presented as means ± S.E.M. All experimental assays were performed in duplicate or triplicate. Statistical analyses shown in figures represent Fisher's exact tests or two-tailed Student t-tests, as indicated. For details regarding the statistical methods employed during RNA-Seq and ChIP-Seq data analysis, see **Supplementary Methods**.

References

1. Metzker ML. Sequencing technologies - the next generation. Nat Rev Genet. 2010; 11:31–46. [PubMed: 19997069]
2. Guttman M, et al. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. Nat Biotechnol. 2010; 28:503–510. [PubMed: 20436462]
3. Trapnell C, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 2010; 28:511–515. [PubMed: 20436464]
4. Robertson G, et al. De novo assembly and analysis of RNA-seq data. Nat Methods. 2010; 7:909–912. [PubMed: 20935650]
5. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 2008; 18:821–829. [PubMed: 18349386]
6. Huarte M, et al. A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response. Cell. 2010; 142:409–419. [PubMed: 20673990]
7. Orom UA, et al. Long Noncoding RNAs with Enhancer-like Function in Human Cells. Cell. 2010; 143:46–58. [PubMed: 20887892]
8. Rinn JL, et al. Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. Cell. 2007; 129:1311–1323. [PubMed: 17604720]
9. Gupta RA, et al. Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. Nature. 2010; 464:1071–1076. [PubMed: 20393566]

10. Pasmant E, et al. Characterization of a germ-line deletion, including the entire INK4/ARF locus, in a melanoma-neural system tumor family: identification of ANRIL, an antisense noncoding RNA whose expression coclusters with ARF. Cancer Res. 2007; 67:3963–3969. [PubMed: 17440112]
11. Yap KL, et al. Molecular interplay of the noncoding RNA ANRIL and methylated histone H3 lysine 27 by polycomb CBX7 in transcriptional silencing of INK4a. Mol Cell. 2010; 38:662–674. [PubMed: 20541999]
12. Tsai MC, et al. Long noncoding RNA as modular scaffold of histone modification complexes. Science. 2010; 329:689–693. [PubMed: 20616235]
13. Kotake Y, et al. Long non-coding RNA ANRIL is required for the PRC2 recruitment to and silencing of p15(INK4B) tumor suppressor gene. Oncogene. 2010
14. de Kok JB, et al. DD3(PCA3), a very sensitive and specific marker to detect prostate tumors. Cancer Res. 2002; 62:2695–2698. [PubMed: 11980670]
15. Li J, et al. PTEN, a putative protein tyrosine phosphatase gene mutated in human brain, breast, and prostate cancer. Science. 1997; 275:1943–1947. [PubMed: 9072974]
16. Prensner JR, Chinnaiyan AM. Oncogenic gene fusions in epithelial carcinomas. Curr Opin Genet Dev. 2009; 19:82–91. [PubMed: 19233641]
17. Tomlins SA, et al. Distinct classes of chromosomal rearrangements create oncogenic ETS gene fusions in prostate cancer. Nature. 2007; 448:595–599. [PubMed: 17671502]
18. Tomlins SA, et al. Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. Science. 2005; 310:644–648. [PubMed: 16254181]
19. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009; 25:1105–1111. [PubMed: 19289445]
20. Thierry-Mieg D, Thierry-Mieg J. AceView: a comprehensive cDNA-supported gene and transcripts annotation. Genome Biol. 2006; 7(Suppl 1):S12, 11–14. [PubMed: 16925834]
21. Birney E, et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature. 2007; 447:799–816. [PubMed: 17571346]
22. Carninci P, et al. The transcriptional landscape of the mammalian genome. Science. 2005; 309:1559–1563. [PubMed: 16141072]
23. He Y, Vogelstein B, Velculescu VE, Papadopoulos N, Kinzler KW. The antisense transcriptomes of human cells. Science. 2008; 322:1855–1857. [PubMed: 19056939]
24. Guttman M, et al. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature. 2009; 458:223–227. [PubMed: 19182780]
25. Yu J, et al. An integrated network of androgen receptor, polycomb, and TMPRSS2-ERG gene fusions in prostate cancer progression. Cancer Cell. 2010; 17:443–454. [PubMed: 20478527]
26. Day DS, Luquette LJ, Park PJ, Kharchenko PV. Estimating enrichment of repetitive elements from high-throughput sequence data. Genome Biol. 2010; 11:R69. [PubMed: 20584328]
27. Kim TK, et al. Widespread transcription at neuronal activity-regulated enhancers. Nature. 2010; 465:182–187. [PubMed: 20393465]
28. Rubin MA, et al. alpha-Methylacyl coenzyme A racemase as a tissue biomarker for prostate cancer. JAMA. 2002; 287:1662–1670. [PubMed: 11926890]
29. Dhanasekaran SM, et al. Delineation of prognostic biomarkers in prostate cancer. Nature. 2001; 412:822–826. [PubMed: 11518967]
30. van Bakel H, Nislow C, Blencowe BJ, Hughes TR. Most “dark matter” transcripts are associated with known genes. PLoS Biol. 2010; 8:e1000371. [PubMed: 20502517]
31. Tomlins SA, et al. The role of SPINK1 in ETS rearrangement-negative prostate cancers. Cancer Cell. 2008; 13:519–528. [PubMed: 18538735]
32. Bjartell AS, et al. Association of cysteine-rich secretory protein 3 and beta-microseminoprotein with outcome after radical prostatectomy. Clin Cancer Res. 2007; 13:4130–4138. [PubMed: 17634540]
33. Oosumi T, Belknap WR, Garlick B. Mariner transposons in humans. Nature. 1995; 378:672. [PubMed: 7501013]
34. Robertson HM, Zumpano KL, Lohe AR, Hartl DL. Reconstructing the ancient mariners of humans. Nat Genet. 1996; 12:360–361. [PubMed: 8630486]

35. Kleer CG, et al. EZH2 is a marker of aggressive breast cancer and promotes neoplastic transformation of breast epithelial cells. Proc Natl Acad Sci U S A. 2003; 100:11606–11611. [PubMed: 14500907]

36. Varambally S, et al. The polycomb group protein EZH2 is involved in progression of prostate cancer. Nature. 2002; 419:624–629. [PubMed: 12374981]

37. Ahmadiyeh N, et al. 8q24 prostate, breast, and colon cancer risk loci show tissue-specific long-range interaction with MYC. Proc Natl Acad Sci U S A. 2010; 107:9742–9746. [PubMed: 20453196]

38. Al Olama AA, et al. Multiple loci on 8q24 associated with prostate cancer susceptibility. Nat Genet. 2009; 41:1058–1060. [PubMed: 19767752]

39. Beroukhim R, et al. The landscape of somatic copy-number alteration across human cancers. Nature. 2010; 463:899–905. [PubMed: 20164920]

40. Gudmundsson J, et al. Genome-wide association study identifies a second prostate cancer susceptibility variant at 8q24. Nat Genet. 2007; 39:631–637. [PubMed: 17401366]

41. Sotelo J, et al. Long-range enhancers on 8q24 regulate c-Myc. Proc Natl Acad Sci U S A. 2010; 107:3001–3005. [PubMed: 20133699]

42. Taylor BS, et al. Integrative genomic profiling of human prostate cancer. Cancer Cell. 2010; 18:11–22. [PubMed: 20579941]

43. Huttenhower C, et al. Exploring the human genome with functional maps. Genome Res. 2009; 19:1093–1106. [PubMed: 19246570]

44. Laxman B, et al. A first-generation multiplex biomarker analysis of urine for the early detection of prostate cancer. Cancer Res. 2008; 68:645–649. [PubMed: 18245462]

45. Hessels D, et al. DD3(PCA3)-based molecular urine analysis for the diagnosis of prostate cancer. Eur Urol. 2003; 44:8–15. discussion 15-16. [PubMed: 12814669]

46. Maher CA, et al. Chimeric transcript discovery by paired-end transcriptome sequencing. Proc Natl Acad Sci U S A. 2009; 106:12353–12358. [PubMed: 19592507]

47. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A. 2001; 98:5116–5121. [PubMed: 11309499]

48. Dennis G Jr. et al. DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol. 2003; 4:P3. [PubMed: 12734009]

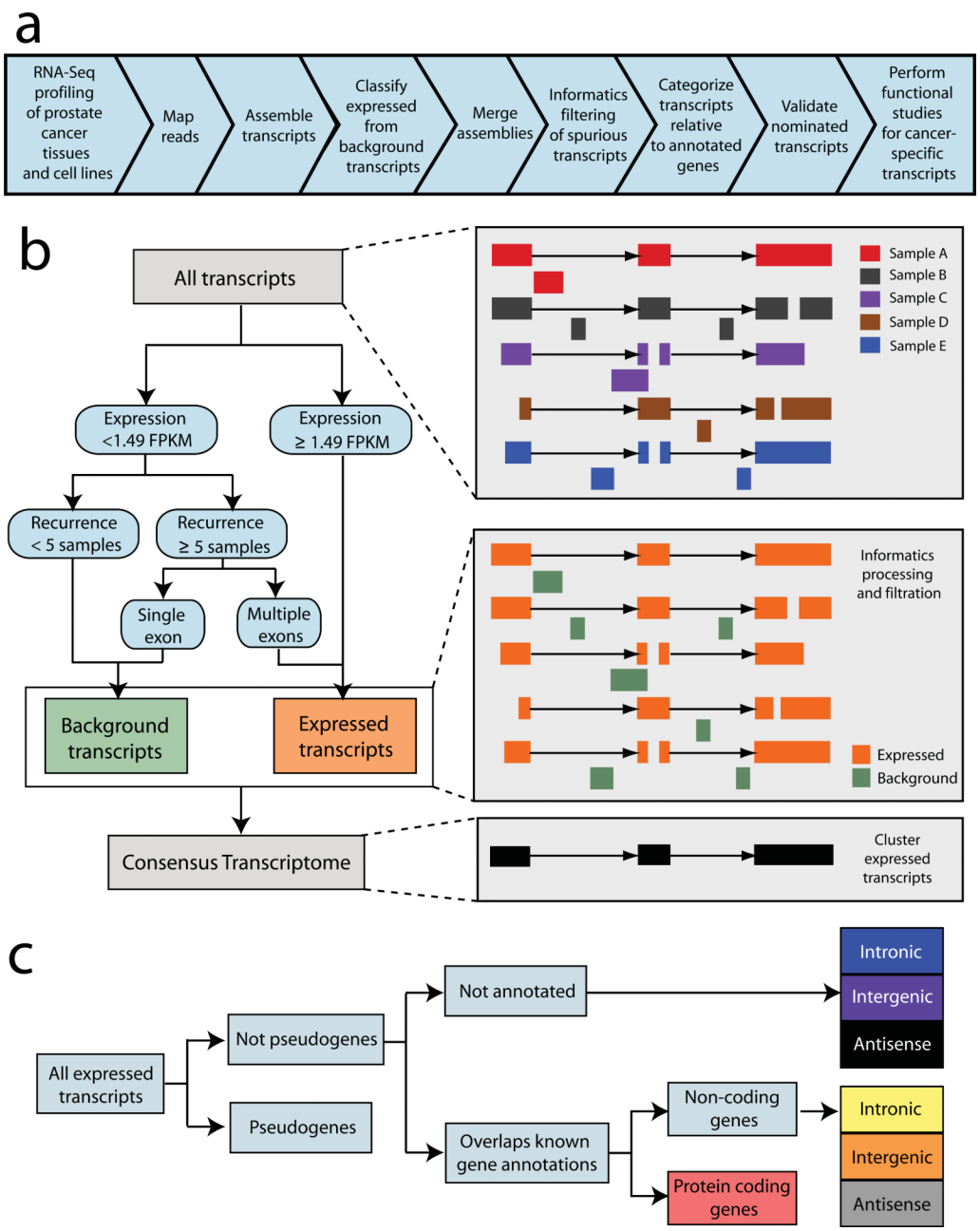

Figure 1. Analysis of transcriptome data for the detection of unannotated transcripts

(**a**) A schematic overview of the methodology employed in this study. (**b**) A graphical representation showing the bioinformatics filtration model used to merge individual transcriptome libraries into a single consensus transcriptome. The merged consensus transcriptome was generated by compiling all individual transcriptome libraries and using a decision tree classifier in order to define high confidence “expressed” transcripts and low confidence “background” transcripts, which were discarded. The example decision tree on the left was produced from transcripts on chromosome 1. The graphics on the right provide a fictional example demonstrating the informatics filtration pipelin e . (**c**) Following informatic processing and filtration of the sequencing data, transcripts were categorized in order to identify unannotated ncRNAs. Transcribed pseudogenes were isolated, and the remaining transcripts were categorized based on overlap with an aggregated set of known

gene annotations into annotated protein coding, non-coding, and unannotated. Both annotated and unannotated ncRNA transcripts were then separated into intronic, intergenic, and antisense categories based on their relationship to protein coding genes.

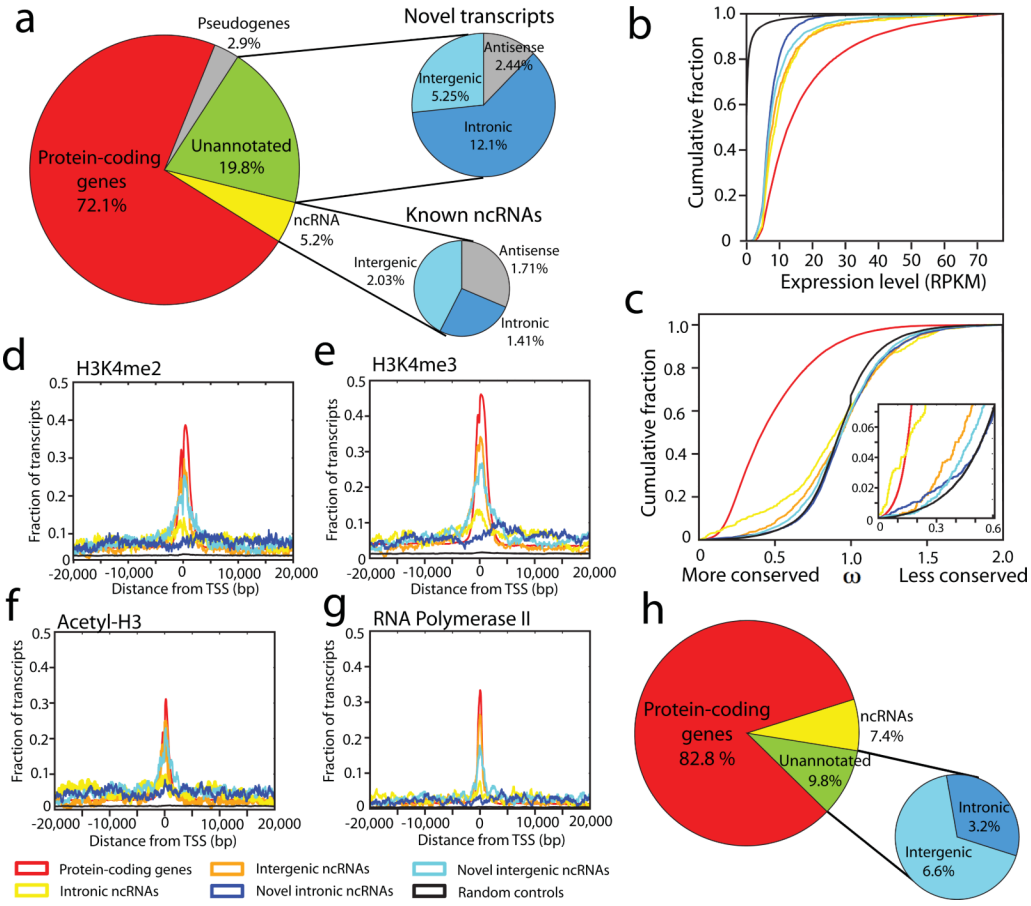

Figure 2. Prostate cancer transcriptome sequencing reveals dysregulation of novel transcripts

(**a**) A global overview of transcription in prostate cancer. The left pie chart displays transcript distribution in prostate cancer. The upper and lower right pie charts display unannotated or annotated ncRNAs, respectively categorized as sense transcripts (intergenic and intronic) and antisense transcripts. (**b**) A line graph showing that unannotated transcripts are more highly expressed (RPKM) than control regions. Negative control intervals were generated by randomly permuting the genomic positions of the transcripts. (**c**) Conservation analysis comparing unannotated transcripts to known genes and intronic controls shows a subtle degree of purifying selection among unannotated transcripts. The insert on the right shows an enlarged view. (**d-g**) Intersection plots displaying the fraction of unannotated transcripts enriched for H3K4me2 (**d**), H3K4me3 (**e**), Acetyl-H3 (**f**) or RNA polymerase II (**g**) at their transcriptional start site (TSS) using ChIP-Seq and RNA-Seq data for the VCaP prostate cancer cell line. The legend for these plots (**b-g**) is shared and located below (**f**) and (**g**). (**h**) A pie chart displaying the distribution of differentially expressed transcripts in prostate cancer (FDR < 0.01).

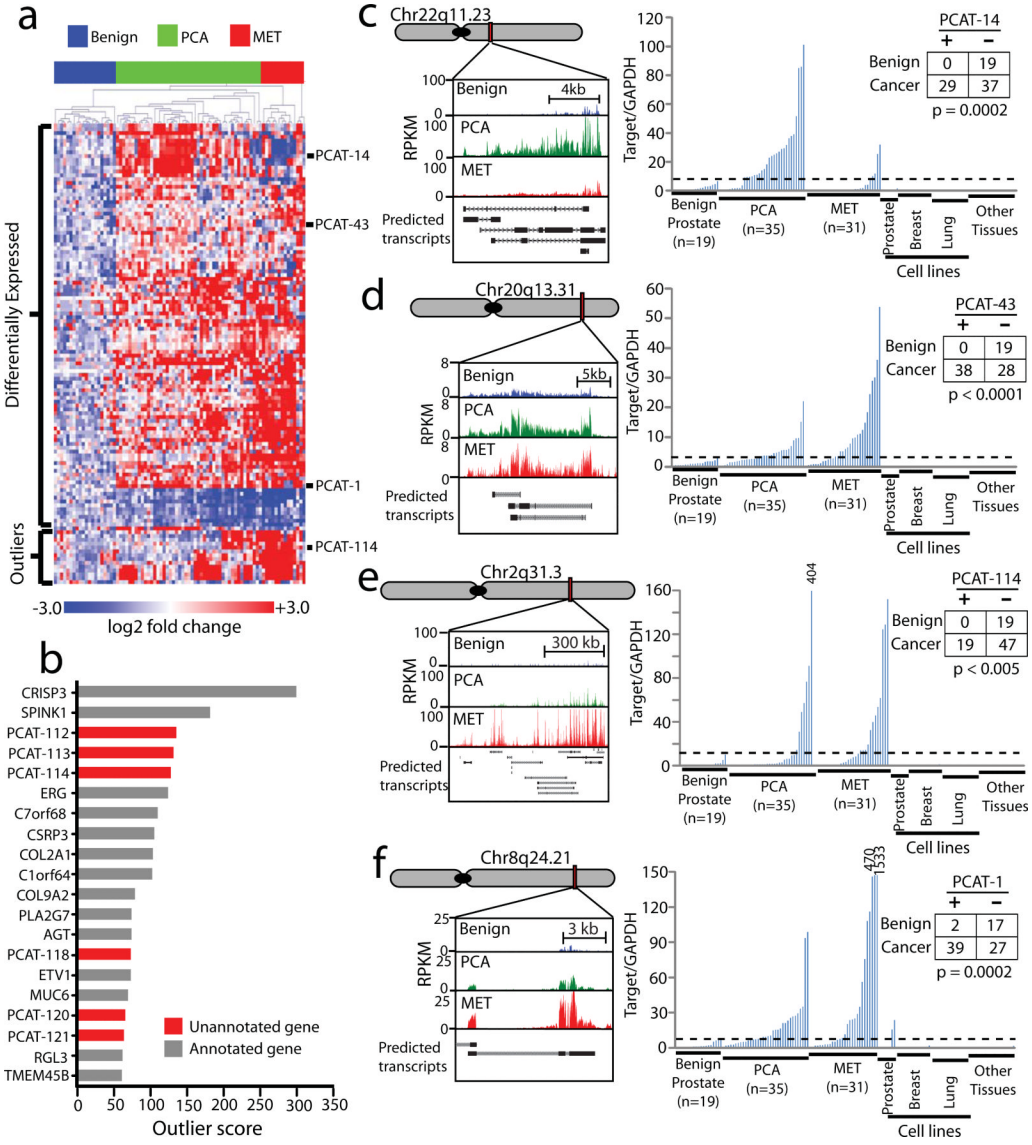

Figure 3. Unannotated intergenic transcripts differentiate prostate cancer and benign prostate samples

(**a**) Unsupervised clustering analyses of differentially-expressed or outlier unannotated intergenic transcripts clusters benign samples, localized tumors, and metastatic cancers. Expression is plotted as log2 fold change relative to the median of the benign samples. The four transcripts detailed in this study are indicated on the side. (**b**) Cancer outlier expression analysis for the prostate cancer transcriptome ranks unannotated transcripts prominently. (**c-f**) qPCR on an independent cohort of prostate and non-prostate samples (Benign (n=19), PCA (n=35), MET (n=31), prostate cell lines (n=7), breast cell lines (n=14), lung cell lines (n=16), other normal samples (n=19), see **Supplementary Table 8**) measures expression levels of four nominated ncRNAs—*PCAT-1, PCAT-43, PCAT-114*, and *PCAT-14*—upregulated in prostate cancer. Inset tables on the right quantify "positive" and "negative" expressing samples using the cut-off value (shown as a black dotted line). Statistical significance was determined using a Fisher's exact test. (**c**) *PCAT-14*. (**d**) *PCAT-43*. (**e**) *PCAT-114* (SChLAP1). (**f**) *PCAT-1*. qPCR analysis was performed by normalizing to *GAPDH* and the median expression of the benign samples.

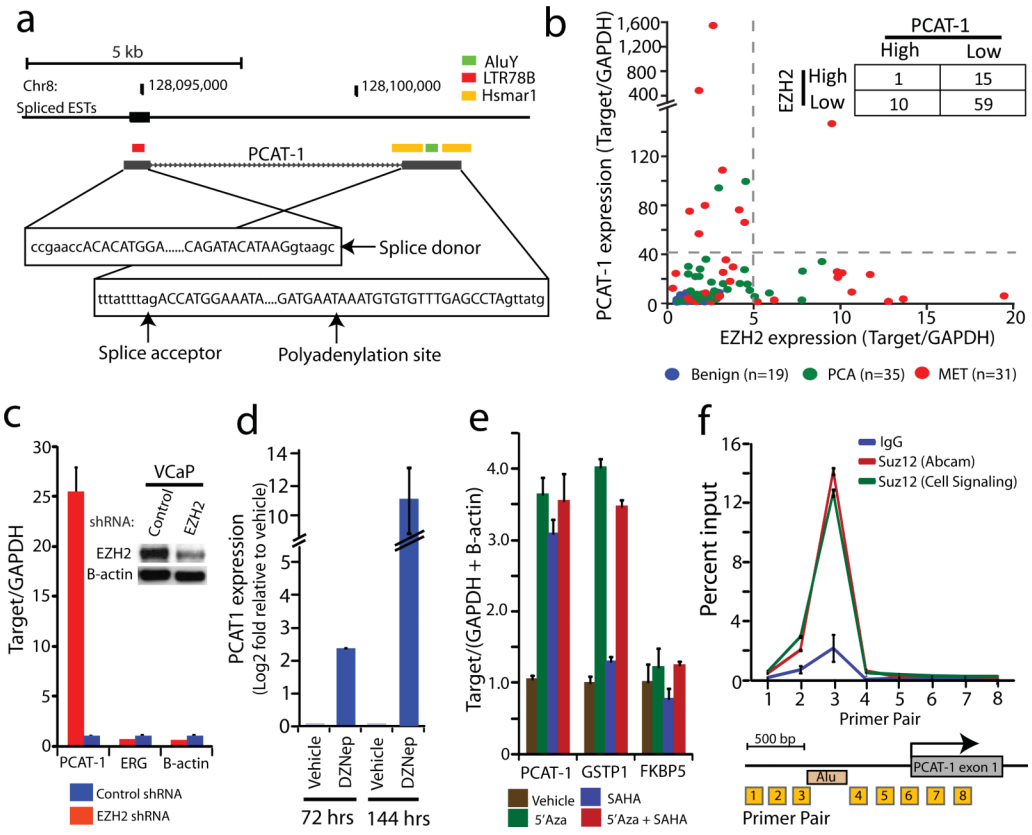

Figure 4. *PCAT-1* is a marker of aggressive cancer and a PRC2-repressed ncRNA

(**a**) The genomic location of *PCAT-1* determined by 5' and 3' RACE, with DNA sequence features indicated by the colored boxes (**b**) qPCR for *PCAT-1* (Y-axis) and *EZH2* (X-axis) on a cohort of benign (n=19), localized tumor (n=35) and metastatic cancer (n=31) samples. The inset table quantifies patient subsets demarcated by the gray dotted lines. (**c**) Knockdown of *EZH2* in VCaP resulted in upregulation of *PCAT-1*. Data were normalized to *GAPDH* and represented as fold change. *ERG* and *B-Actin* serve as negative controls. The inset Western blot indicates *EZH2* knockdown. (**d**) Treatment of VCaP cells with 0.1 μM of the *EZH2* inhibitor DZNep or vehicle control (DMSO) shows increased expression of *PCAT-1* transcript following *EZH2* inhibition. (**e**) *PCAT-1* expression is increased upon treatment of VCaP cells with the demethylating agent 5'Azacytidine, the histone deacetylase inhibitor SAHA, or a combination of both. qPCR data were normalized to the average of (*GAPDH+B-Actin)* and represented as fold change. *GSTP1* and *FKBP5* are positive and negative controls, respectively. (**f**) ChIP assays for *SUZ12* demonstrated direct binding of *SUZ12* to the *PCAT-1* promoter. Primer locations are indicated (boxed numbers) in the *PCAT-1* schematic.

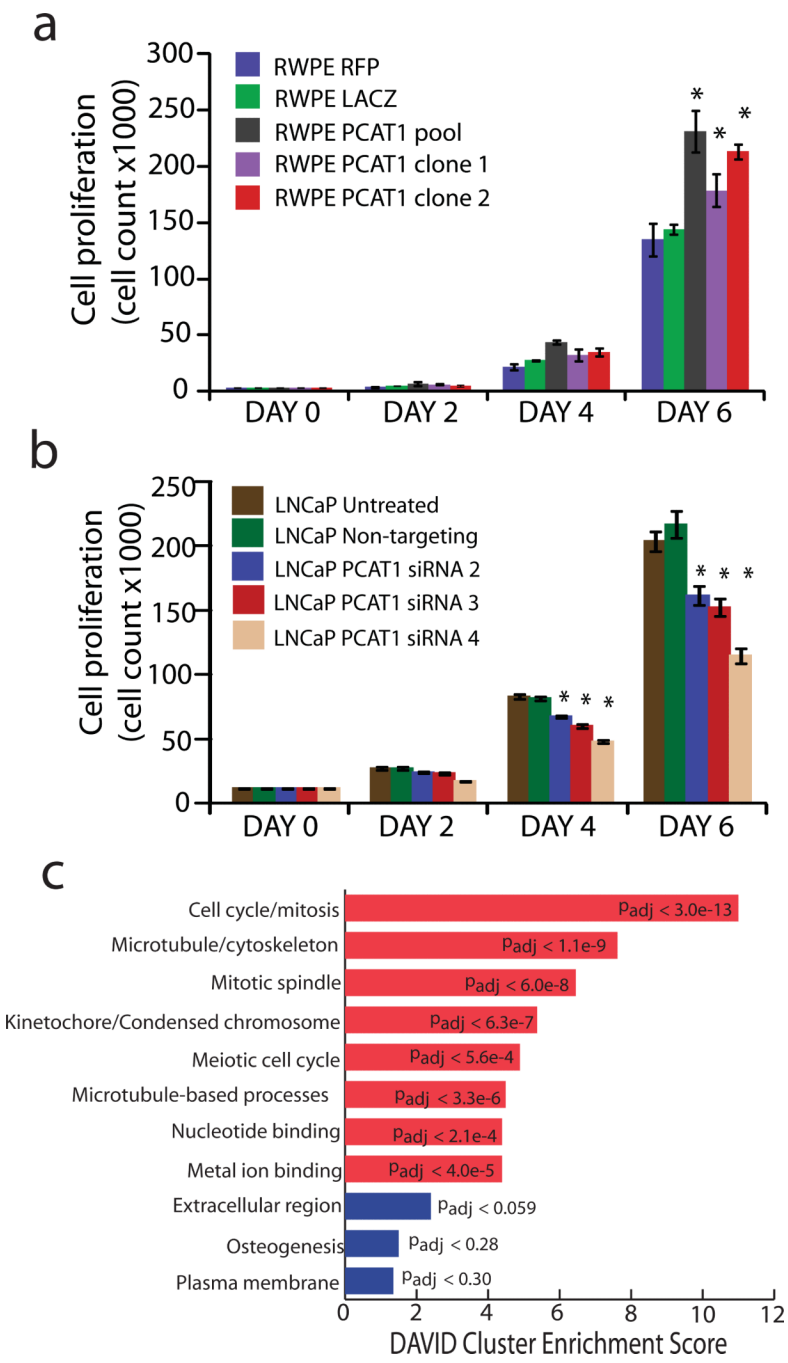

Figure 5. *PCAT-1* promotes cell proliferation

(**a**) Cell proliferation assays for RWPE benign immortalized prostate cells stably infected with *PCAT-1* lentivirus or RFP and LacZ control lentiviruses. An asterisk (*) indicates $p \leq 0.02$ by a two-tailed Students t-test. (**b**) Cell proliferation assays in LNCaP using *PCAT-1* siRNAs. An asterisk (*) indicates $p \leq 0.005$ by a two-tailed Students t-test. (**c**) Gene ontology analysis of *PCAT-1* knockdown microarray data using the DAVID program. Blue bars represent the top hits for upregulated genes. Red bars represent the top hits for downregulated genes. All error bars in this figure are mean ± S.E.M.

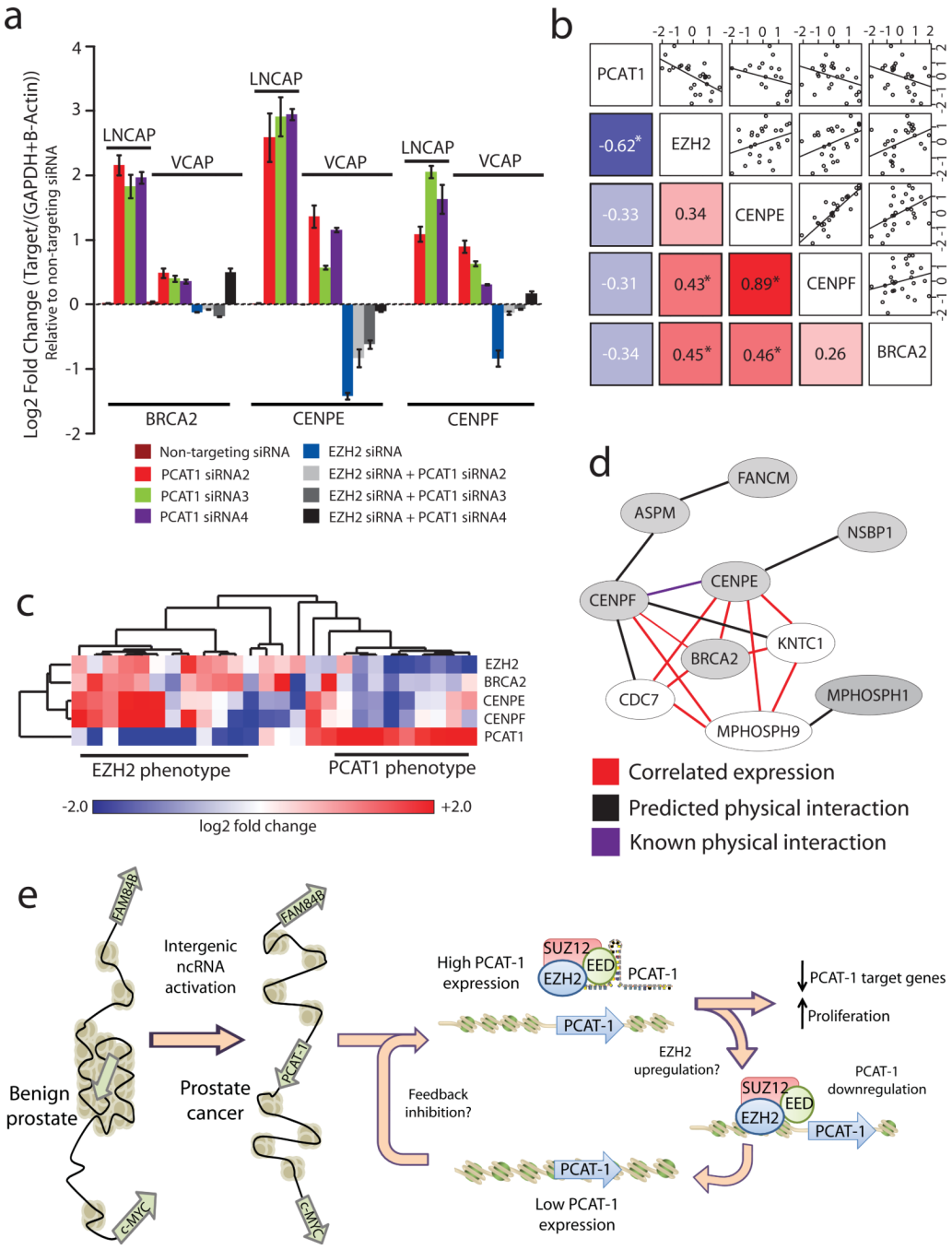

Figure 6. Prostate cancer tissues recapitulate *PCAT-1* signaling

(**a**) qPCR expression of three *PCAT-1* target genes after *PCAT-1* knockdown in VCaP and LNCaP cells, as well as following *EZH2* knockdown or dual *EZH2* and *PCAT-1* knockdown in VCaP cells. qPCR data were normalized to the average of (*GAPDH+B-Actin)* and represented as fold change. Error bars represent mean ± S.E.M. (**b**) Standardized log2-transformed qPCR expression of a set of tumors and metastases with outlier expression of either *PCAT-1* or *EZH2*. The shaded squares in the lower left show Spearman correlation values between the indicated genes (* indicates $p < 0.05$). Blue and red indicate negative or positive correlation, respectively. The upper squares show the scatter plot matrix and fitted trendlines for the same comparisons. (**c**) A heatmap of *PCAT-1* target genes (*BRCA2,*

CENPF, *CENPE*) in *EZH2*-outlier and *PCAT-1*-outlier patient samples (see **Fig. 4b**). Expression was determined by qPCR and normalized as in (**b**). (**d**) A predicted network generated by the HefaLMP program for 7 of 20 top upregulated genes following *PCAT-1* knockdown in LNCaP cells. Gray nodes are genes found following *PCAT-1* knockdown. Red edges indicate co-expressed genes; black edges indicate predicted protein-protein interactions; and purple edges indicate verified protein-protein interactions. **(e)** A proposed schematic representing *PCAT-1* upregulation, function, and relationship to PRC2.