AWARD NUMBER: W81XWH-16-1-0739

TITLE: Developing a PTEN-ERG Signature to Improve Molecular Risk Stratification in Prostate Cancer

PRINCIPAL INVESTIGATOR: Luigi Marchionni

CONTRACTING ORGANIZATION: Johns Hopkins University, Baltimore, MD

REPORT DATE: January 2021

TYPE OF REPORT: Final

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

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE	**2. REPORT TYPE**	**3. DATES COVERED**
January 2021	Final	30Sep2016 – 29Sep2020

4. TITLE AND SUBTITLE

Developing a PTEN-ERG Signature to Improve Molecular Risk Stratification in Prostate Cancer

5a. CONTRACT NUMBER

5b. GRANT NUMBER
W81XWH-16-1-0739

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Luigi Marchionni and Tamara L. Lotan (partnering PI)

E-Mail: marchion@jhu.edu and tlotan1@jhmi.edu

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

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

Johns Hopkins University 1550 Orleans Street CRB2, Rm 1M52 Baltimore, MD 21231

8. PERFORMING ORGANIZATION REPORT NUMBER

Johns Hopkins University Baltimore, MD 21231

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

U.S. Army Medical Research and Development 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

Prostate cancer (PCA) is a clinically and genetically heterogeneous and the development of a molecular classification is critical to distinguish lethal from indolent tumors and minimize overtreatment. Genomic alterations of the PTEN and ERG genes are among the most common in PCA and there is an interest in exploiting these alterations for routine risk assessment. We found that PTEN loss is most strongly associated with PCA death in patients whose tumors do not carry an ERG gene rearrangement, suggesting that ERG absence strengthens PTEN loss association with lethal progression. Despite the widely accessible PTEN/ERG molecular classification, our understanding of their biological interaction along PCA progression remains very limited. Hence, in our study we will perform a comprehensive molecular profiling of well-annotated PCA samples in relation to PTEN and ERG status. Our goals are threefold: 1) to confirm that PTEN/ERG double negative tumors are the most aggressive; 2) to characterize the expression profiles associated with PTEN and ERG alterations; and 3) to determine whether such expression profiles can be used to improve PCA patient stratification into different risk groups.

15. SUBJECT TERMS

Prostate cancer, PTEN, ERG, ETS, MYC, cell cycle, gene expression, Cap Analysis of Gene Expression (CAGE)

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
Unclassified	Unclassified	Unclassified	Unclassified	95	**19b. TELEPHONE NUMBER** *(include area code)*

TABLE OF CONTENTS

1. INTRODUCTION

Prostate cancer (PCA) is a clinically and genetically heterogeneous and the development of a molecular classification is critical to distinguish lethal from indolent tumors and minimize overtreatment. Recent technological advances have enabled extraordinary insights into molecular changes occurring in PCA and the *PTEN* and *ERG* genomic alterations have emerged as the most common in PCA. Furthermore, we have found that PTEN loss is associated with PCA death most strongly in patients carrying *ERG* rearrangements, hence there is an interest in exploiting such alterations for routine risk assessment. Furthermore, despite the fact that PTEN and ERG molecular classification is widely accessible, our understanding of their interaction during disease progression is very limited, and a molecular signature of PTEN/ERG loss in PCA is still lacking.

To address these issues, we have formed a collaborative, multi-disciplinary team – led by a urologic pathologist and computational biologist with expertise in PCA molecular pathology and cancer genomics – to perform a comprehensive molecular assessment of well-annotated prostate cancers in relation to PTEN and ERG status using existing and novel data. Our objectives are threefold: 1) to confirm that the tumors with loss of PTEN and lacking *ERG* rearrangement are among the most aggressive; 2) to characterize the *expression profiles* associated with PTEN and ERG alterations; and 3) to determine whether these *expression profiles* can improve the way we stratify prostate cancer patients into different risk groups.

Findings from our proposed research have the potential for both immediate and long-term clinical and translational research applicability. First, by analyzing several large clinical cohorts from multiple institutions, we will be able to confirm the performance of these biomarkers in patient risk stratification. Second, we will also be able to assess if and how PTEN/ERG *molecular signatures* correlate with lethal disease risk in comparison to currently available prognostic assays. Third, we expect to identify novel molecular alterations responsible for the distinct clinical and biological behavior of tumors based on PTEN and ERG status. Lastly, we will also generate a wealth of information about the biologic drivers of prostate cancer behavior, which shall then be utilized by the entire PCA research community.

2. KEYWORDS

Prostate cancer, PTEN, ERG, ETS, MYC, cell cycle, gene expression, RNA sequencing, Cap Analysis of Gene Expression (CAGE)

3. ACCOMPLISHMENTS

Below are listed tasks, subtasks, and accomplishments for research sites 1 (coordinated by the initiating PI, Dr. Marchionni), and site 2 (coordinated by the partnering PI Dr. Lotan).

SPECIFIC AIM 1 (Dr. Lotan)

Expected tasks and milestones are summarized below.

Specific Aim 1: Validate association of PTEN and ETS status with risk of lethal prostate cancer	**Timeline (Months)**
Major Task 1: Assessing prostatectomy cohorts on multiple tissue microarrays (TMA) for PTEN, ETS, and cell proliferation rate	1-36
Subtask 1: Perform immunostaining for PTEN, ERG and Ki-67 and in situ hybridization on tissue microarrays (TMAs) from JHU and MSKCC cohorts; immunostaining for Ki-67 on HPFS/PHS cohort	1-12
Subtask 2: Score immunostaining and in situ hybridization from Subtask 1 • Digitally scan all slides using Aperio CS slide scanning system in Johns Hopkins OTS Core facility • Segment TMAs and upload to web-based browser, TMAJ (http://tmaj.pathology.jhmi.edu/) • Dr. Lotan, Dr. Gopalan (MSKCC), and pathology fellow supervised by Dr. Lotan perform scoring. Image analysis software (FRiDA on TMAJ) to be used for Ki-67 scoring	13-24

Subtask 3: Analysis of immunostaining and in situ hybridization data from Subtask 2 • Multivariate models to assess association of PTEN/ETS status with metastasis and survival in JHU and MSKCC cohorts • Correlation of PTEN/ETS status with proliferation in JHU, MSKCC and HPFS/PHS cohorts	18-30
Milestone #1: Co-author manuscript on association of PTEN/ETS status with cell cycle gene expression, proliferation rate and risk of metastasis and death in multiple validation cohorts	31-36

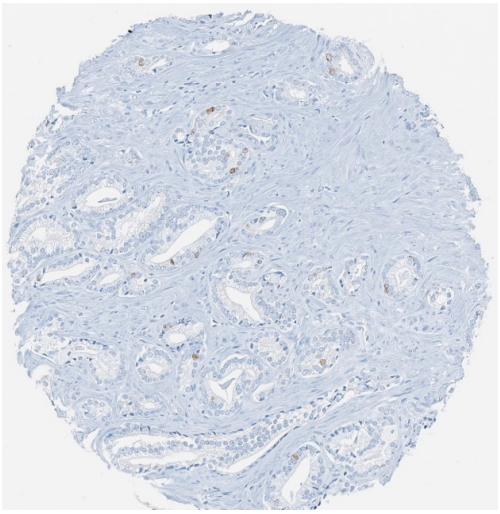


Figure 1: Ki-67 labelling in JHU cohort (200x magnification)

Progress on Major Task 1 – Subtask 1: Major activities for this activities included performing immunostaining and in situ hybridization on the JHU and MSKCC tissue microarray (TMA) cohorts. We have therefore performed and scored PTEN and ERG immunostaining on the JHU and MSKCC TMA cohorts, in addition to ETV1, ETV4 and ETV5 in situ hybridization on the JHU cohorts. Ki-67 immunostaining has been performed as proposed on both cohorts and automated scoring is pending. We also analyzed the PTEN/ERG/ETS data for association with metastasis and death from prostate cancer in the JHU and MSKCC cohorts. We have also correlated PTEN/ERG/ETS status of tumors in these cohorts with gene expression data from the same cohorts.

Progress on Major Task 1 – Subtask 2: The scoring for ETS gene rearrangements (ETV1/4/5) (as well as PTEN and ERG) on the JHU tissue microarrays has been completed and analyzed. Ki-67 immunostaining on those arrays is completed. PTEN and ERG staining as well as Ki-67 staining is completed for these arrays as well. All arrays have been digitally scanned and are viewed on our TMAJ viewer (**Figure 1**). However digital automated scoring of Ki-67 has been challenging since it is difficult to normalize the number of positive detected nuclei (brown) to the negative tumor nuclei (blue). This is because it is difficult to detect all negative tumor nuclei. After examining the possibility of simply normalizing the number of positive tumor cells to the total area of the spot (*i.e.*, density of positive cells per mm^2 of tumor), which is the easiest method, we have decided that this is suboptimal since it can be confounded by the amount of tumor nuclei sampled in the TMA spot (**Figure 1**). In this analysis, we find that lower grade tumors, with fewer tumor nuclei will have inappropriately low Ki-67 density scores. Thus, it is necessary to manually annotate all tumor-containing spots. To pilot this manual annotation algorithm, we used a smaller JHU cohort of ~200 prostate cancer cases where tumor spots could be manually annotated more easily. In this analysis (**Figure 2**, below), we found that median Ki-67 levels (% tumor nuclei staining) were significantly elevated in tumors with PTEN loss compared to those with intact PTEN (p=0.006), regardless of ERG status (p=0.006). Interestingly, cases with PTEN loss that were ERG negative showed increased variability in Ki-67 levels, potentially consistent with our finding of their heterogeneous molecular status (see below). However, we found this to be extraordinarily time-consuming and impractical to perform on 400 spots x 12 TMAs in the MSKCC cohort and 9 TMAs in the full JHU cohort.

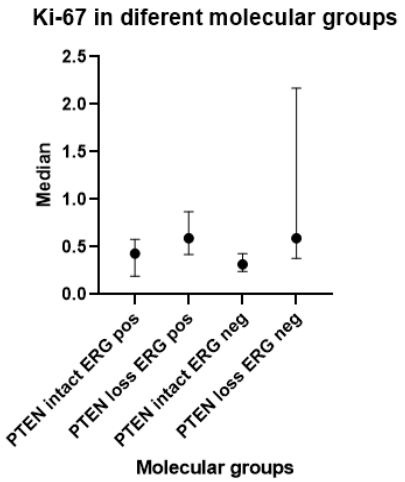


Figure 2: Ki-67 labelling in JHU cohort by PTEN-ERG status. Median Ki-67 is median % tumor nuclei staining.

To automate the tumor annotation process, we piloted a new protocol for dual staining for Ki-67 (brown) and AE1/AE3 keratin (red) with p63 (brown) in the JHU TMAs. This enabled us to train HALO image analysis software to identify the epithelial (glandular) components on each slide and automatically annotate them. Then, using the p63 immunohistochemistry, we can manually exclude benign glands from the analysis. Though still requiring a manual step, this is much more efficient than full manual annotation. We have annotated all JHU TMAs and are now in the process of quantifying the Ki-67 using this algorithm.

Table 1. Ki-67 labeling stratified by PTEN/ERG status in HPFS/PHS cohort

	Ki-67 (% positive tumor nuclei)	P-value	P-value
Any PTEN intact/ERG negative	0.8%	Reference	0.02
Any PTEN intact/ERG positive	0.3%	0.13	0.26
Complete PTEN loss / ERG negative	1.3%	0.02	Reference
Complete PTEN loss / ERG positive	1.1%	0.03	0.68

Progress on Major Task 1 – Subtask 3: For the HPFS/PHS tissue microarrays, we have completed and analyzed Ki-67 immunostaining. Data comparing the percent of tumor cells labeling for Ki-67, stratified by PTEN and ERG status are presented in **Table 1**. Exactly as reported in the JHU cohort described above, there was a statistically significant increase in Ki-67 labelling when comparing tumors that have PTEN loss and those that are PTEN intact, however there was not a significant difference in Ki-67 labeling between tumors with PTEN loss and ERG expression and those with PTEN loss that do not express ERG (p=0.68). Thus, we anticipate we will further validate this finding in the full JHU and MSKCC cohorts as well.

Table 2: Multivariable models of association of PTEN-ERG status with lethal prostate cancer in the MSKCC cohort.

			Univariable		Multivariable*	
	No. Cases	No. Controls	HR (95% CI)	p Value	HR (95% CI)	p Value
PTEN:						
Intact	46	542	Referent	–	Referent	–
Loss	47	156	3.25 (2.16–4.88)	<0.001	1.87 (1.15–3.04)	0.012
ERG:						
Neg	62	384	Referent	–	Referent	–
Pos	30	300	0.64 (0.41–0.99)	0.043	0.64 (0.36–1.11)	0.113
PTEN/ERG:						
PTEN intact/*ERG* neg	35	328	Referent	–	Referent	–
PTEN intact/*ERG* pos	10	200	0.47 (0.23–0.96)	0.037	0.48 (0.18–1.26)	0.136
PTEN loss/*ERG* neg	27	56	3.76 (2.27–6.21)	<0.001	2.31 (1.29–4.14)	0.005
PTEN loss/*ERG* pos	20	100	1.84 (1.06–3.18)	0.030	1.09 (0.56–2.12)	0.809

We have finalized the analysis of PTEN and ERG on the MSKCC cohort and examined the correlation of these molecular alterations with clinical outcomes (lethal prostate cancer) in multivariable models. These results are presented in **Table 2** and **Figure 3** below, and were recently published in *Journal of Urology* (Haney NM, Faisal FA, Lu J, Guedes LB, Reuter VE, Scher HI, Eastham JA, Marchionni L, Joshu C, Gopalan A*, Lotan TL*. PTEN loss with ERG-negative status is associated with lethal disease after radical prostatectomy. *J Urol*. 2020, 203(2):344-350. *Equal Contribution. PMID: 31502941).

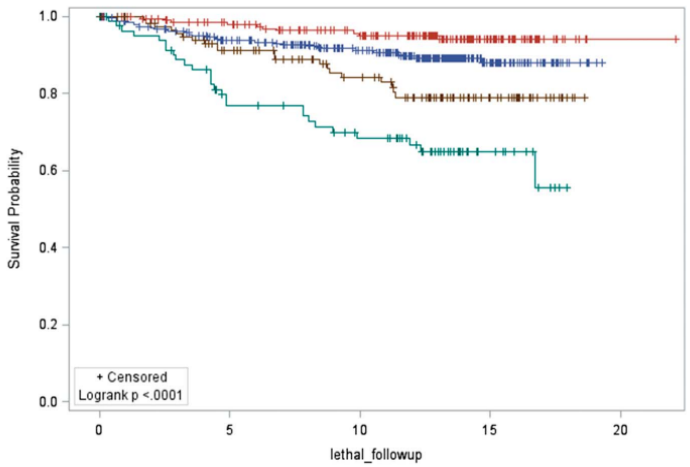


Figure 3: **Kaplan-Meier survival curves of freedom from lethal prostate cancer by PTEN and ERG status**. Blue curve indicates PTEN intact and ERG negative in 363 patients. Red curve indicates PTEN intact and ERG positive in 210 patients. Green curve indicates PTEN loss and ERG negative in 83 patients. Brown curve indicates PTEN loss and ERG positive in 120 patients.

Training and professional development: Nothing to report

Results dissemination to communities of interest: Results from Major Task 1 – Subtask 3 were recently published in *Journal of Urology* (Haney NM, Faisal FA, Lu J, Guedes LB, Reuter VE, Scher HI, Eastham JA, Marchionni L, Joshu C, Gopalan A*, Lotan TL*. PTEN loss with ERG-negative status is associated with lethal disease after radical prostatectomy. *J Urol*. 2020, 203(2):344-350. *Equal Contribution. PMID: 31502941).

SPECIFIC AIM 2 (Dr. Marchionni)

Expected tasks and milestones are summarized below.

Specific Aim 2: Leverage multi-dimensional public domain data to discover genomic features and signaling pathways associated with PTEN loss in ERG-positive and ERG-negative PCa.	**Timeline (Months)**
Major Task 1: Exploratory analysis of genomics datasets	1-6
Subtask 1: Examine gene expression distributions and identify outliers and other potential problems: • Use statistical summaries and visualizations (*e.g.,* principal component analysis, hierarchical clustering) • Apply appropriate transformation to the data if required	1-6
Major Task 2: Classify tumors based on PTEN, ETS, and MKI67 status.	6-24
Subtask 1: Use the EM-algorithm to classify tumors as positive or negative based on the expression levels of PTEN, ETS family members, and MKI67	6-12
Subtask 2: Compare expression-based classification to IHC and in-situ based status from in Specific Aim 1	12-30
Subtask 3: Analysis of PTEN and ETS status in cohorts available from GenomeDX and the public domain • Multivariate models to assess association of PTEN/ETS status based on genes expression dichotomization with metastasis and survival in all cohorts • Correlation of PTEN/ETS status based on genes expression dichotomization with proliferation in all cohorts	12-24
Major Task 3: Comprehensive meta-analysis of differential gene expression programs modulated by PTEN and ETS status in prostate cancer and characterization of their biological and clinical correlates	12-30
Subtask 1: Use generalized linear model to identify genes differentially expressed and differentially modulated by PTEN and ETS in prostate cancer	12-24
Subtask 2: Identification of relevant biological processes and signaling pathways associated with PTEN/ETS molecular signatures in prostate cancer	18-30
Subtask 3: Development and validation of predictive models based on associated with PTEN/ETS molecular signatures in prostate cancer	24-36
Milestone #2: Co-author manuscript on comprehensive meta-analysis of genes and signaling pathways associated with PTEN/ETS status in prostate cancer	24-36
Milestone #3: Co-author manuscript on prognostic values of PTEN/ETS molecular signatures in prostate cancer	24-36

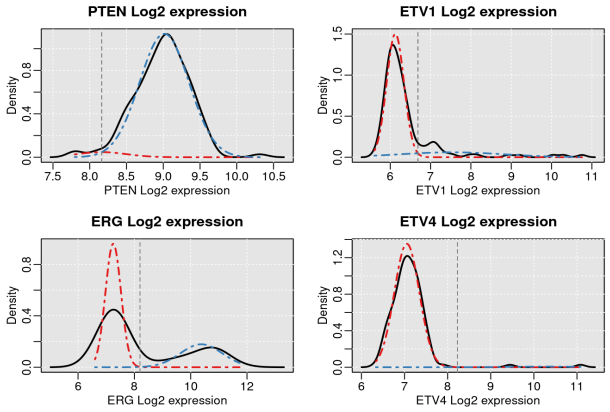


Figure 4: Gene expression distributions for PTEN, ERG, ETV1, and ETV4 in the MSKCC cohort. The underlying distributions from the EM-algorithm are shown in red and blue. ERG and ETV1 expressions are clearly bimodal.

Progress on Major Task 1 – Subtask 1: We have performed exploratory data analysis on all clinically annotated prostate cancer datasets available from the public domain and through the collaboration with GenomeDX. We used statistical summaries and data visualizations techniques (*e.g.*, principal component analysis, hierarchical clustering) to identify outliers and unwanted sources of variation in the data, applying appropriate pre-processing procedures and transformations as required.

Progress on Major Task 2 – Subtask 1: We have used the EM-algorithm to classify tumors as positive or negative based on the expression levels of PTEN, ETS family members, and MKI67. Overall, ERG gene expression proved to be bimodal in all datasets analyzed, with nearly perfect concordance with results from IHC and

CNV status. On the contrary, PTEN classification based on EM-classification of gene expression proved more challenging, with some degree of variation between datasets (an example in **Figure 4** for the MSKCC cohort).

Table 3. Comparison between ERG status and PTEN status based on IHC and EM-algorithm classification. Analyses were performed in the MSKCCC, the HPFS/HPS, and the Natural History cohorts.

	MSKCCC		HPFS/PHS		Natural History	
	ERG	**PTEN**	**ERG**	**PTEN**	**ERG**	**PTEN**
Sensitivity	0.97	0.98	0.47	0.87	0.92	0.01
Specificity	0.84	0.08	0.94	0.24	0.98	1.00
Positive Predictive Value	0.82	0.70	0.88	0.79	0.97	1.00
Negative Predictive Value	0.97	0.67	0.64	0.36	0.95	0.37
Prevalence	0.42	0.69	0.50	0.77	0.41	0.63
Detection Rate	0.41	0.68	0.24	0.67	0.38	0.01
Detection Prevalence	0.50	0.96	0.27	0.84	0.39	0.01
Balanced Accuracy	0.91	0.53	0.71	0.56	0.95	0.51
Overall Accuracy	0.90	0.70	0.71	0.72	0.96	0.38
Kappa	0.89	0.69	0.41	0.12	0.90	0.01

Progress on Major Task 2 – Subtasks 2: We compared results between IHC based assessment of PTEN and ERG expression with classification obtained based on gene expression using the EM algorithm. We performed this analysis on the MSKCCC, the HPFS/HPS, and the Natural History cohorts. For this analysis, IHC status was used as the gold-standard and cross-tabulated with the prediction based on the EM-algorithm classification the ERG and PTEN gene expression levels. Overall, the concordance between IHC and EM-predictions was much higher for ERG status than for PTEN status (**Table 3**). Based on these findings, we decided to develop a more robust, multigene signature for PTEN classification using expression levels.

Progress on Major Task 2 – Subtasks 3: This analysis produced a list of differentially expressed genes associated with ERG and PTEN status. These lists accounted for a core set of genes shared across the different datasets, as well as for genes differentially expressed only in each individual dataset considered. For this reason we therefore decided to focus on genes and pathways identified in a metanalysis in conjunction with the development of a prognostic signature (see below, section **Progress on Major Task 3 – Subtasks 3)**.

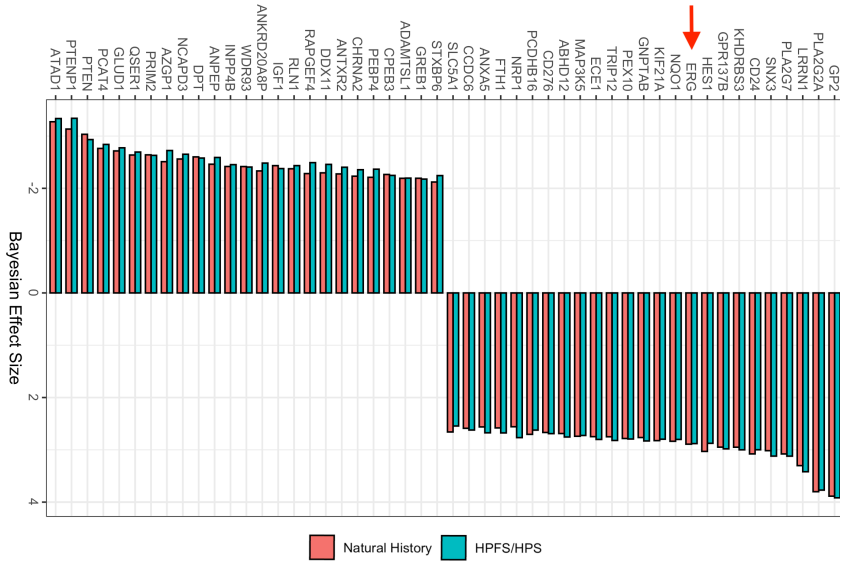


Figure 5. PTEN signature from meta-analysis. PTEN signature obtained by multi-level model for cross-study detection of differential gene expression based on IHC calls on Natural History and HPFS cohorts. Figure shows the effect size of each cohort. ERG is one of the most upregulated genes associated with PTEN loss (red arrow).

Progress on Major Task 3 – Subtask 1: During the first two years of project, we have developed and characterized in depth a consensus signature for PTEN loss using a meta-analytic approach. In the third year of the project, we have investigated the association of this signature with ERG status. This analysis has revealed that the ERG gene itself is among the top upregulated genes in our PTEN loss sigature (**Figure 5**).

Based on this observation, we have therefore hypothesized that our PTEN signature could be heavily influenced by the ERG rearrangement, since this gene encodes a transcription factor. In order to test this hypothesis, we have therefore repeated the meta-analysis by splitting the samples by ERG status and then by fitting two separate Bayesian hierarchical models for differential expression by PTEN status. In the samples with ERG rearrangement, we observed a signature similar to the overall PTEN consensus signature we previously developed in year 2. On the contrary, in the samples without ERG rearrangement, we could not find any significant differences between samples with PTEN loss and PTEN intact.

This finding was surprising, given that PTEN is a powerful tumor suppressor capable of triggering important molecular and functional changes. We speculated that this result could be caused by two reasons: 1) PTEN loss in the absence of ERG rearrangement, does not impact the cell in any significant way; or 2) The absence of ERG

rearrangement generates a high level of heterogeneity that makes it hard to estimate difference between PTEN-null and PTEN intact samples. The first hypothesis, however, is highly unlikely, given the fact that it is well-established that PTEN loss triggers deep changes in cellular metabolism and growth. Therefore, we performed experiments to test if the second hypothesis was true.

In order to test if tumors without ERG rearrangement presented overall higher heterogeneity levels than tumors with it, we stratified the samples based on their PTEN and ERG status. We used the divergence framework available through the R/Bioconductor package 'divergence. Using individual genes (for transcriptomic data) as features of interest, the normal samples were used to estimate baseline profiles and then the divergence was computed for the tumor samples in TCGA and HPFS cohorts. A similar analysis was conducted for the methylation and genomic mutation data from TCGA, using individual CpGs and mutations/copy-number-variation as the features of interest. A random sampling based on the size of the smallest group was extracted from the resulting binary coding to compute the average hamming distances between pairs of samples, this step was performed with 1000 bootstraps.

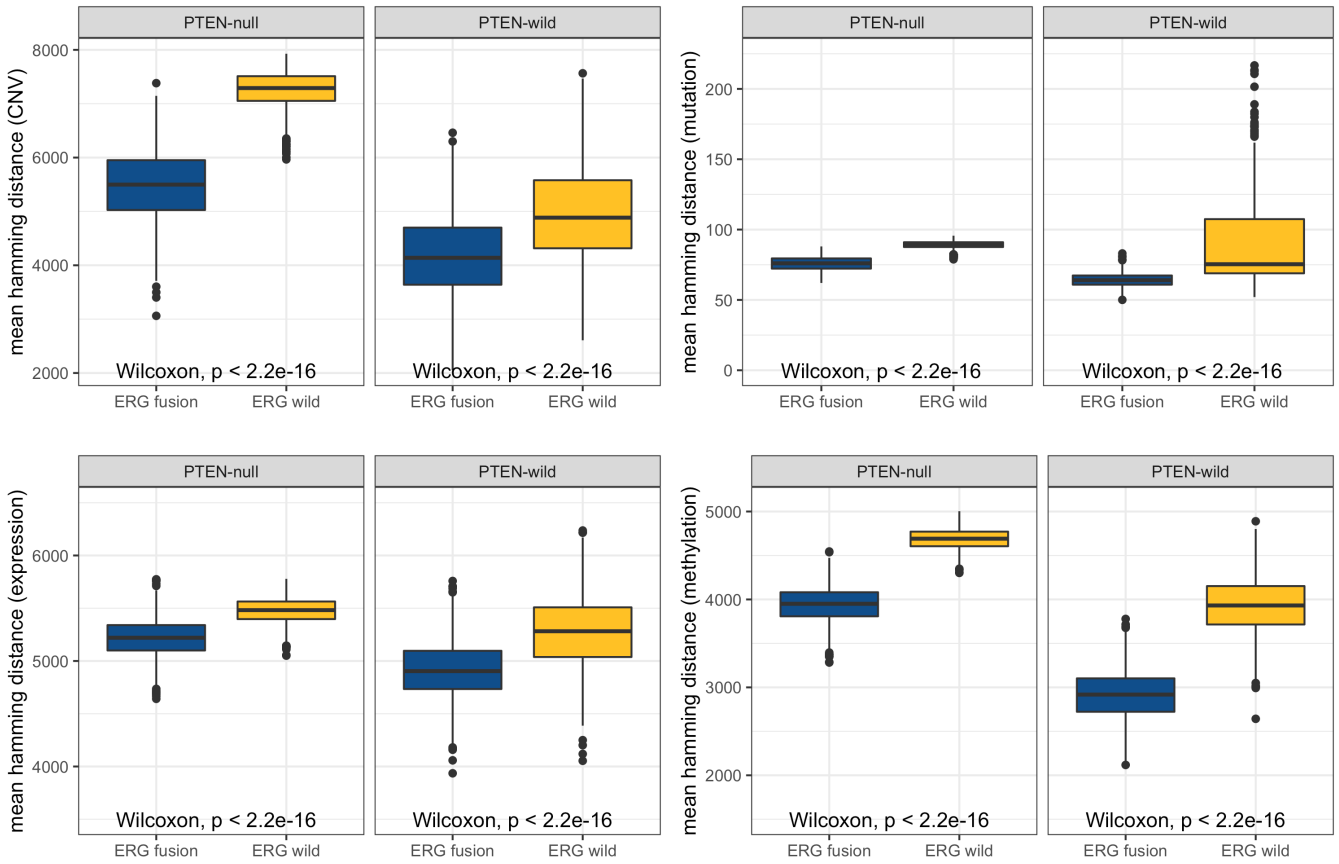


Figure 6. Heterogeneity analysis in ERG positive and negative tumors. Average hamming distance based on 1000 bootstraps intra-samples between each group, showing that samples in absence of ERG rearrangement (ERG wild) presented higher levels of heterogeneity (higher distances) than samples with rearrangement (ERG fusion). **Top left)** Hamming distance based on CNV in TCGA; **Top right)** Hamming distance based on mutation in TCGA; **Bottom left)** Hamming distance based on divergence expression levels in TCGA and **Bottom right)** Hamming distance based on divergence methylation levels in TCGA.

For all molecular data types and for both cohorts, we observed that the intra-group distances between the ERG positive samples (*i.e.*, those with ERG rearrangement) were always significantly higher than between ERG negative tumors, thus confirming our hypothesis (**Figure 6**).

Progress on Major Task 3 – Subtask 2: In our analysis of biological processes and signaling pathways associated with PTEN/ETS molecular signatures in prostate cancer, we saw a strong enrichment in immune related pathways upon PTEN loss (see **Figure 7**). This finding was particularly surprising given that PTEN is itself a key positive regulator of innate immune response. Disruption of PTEN expression has been previously reported to lead to decreased innate immune response. Remarkably, despite the loss of PTEN being associated

with higher expression of the immune checkpoint gene programmed death ligand-1 (PD-L1) in several cancer types this is not true in PCa. So far, current immunotherapeutic interventions, such as PD-1 blockade, in PCa have not been successful. One of the possible reasons is the lack of PD-L1 expression. Therefore, alternative targets must be considered for immunotherapy in PCa. One alternative target is the checkpoint molecule B7-H3 (CD276), whose expression has already been associated with PCa progression and worse prognosis and has been suggested as a target for immunotherapy. CD276 was one of the most concordant up-regulated genes in our signature (**Figure 5**) suggesting that its expression is associated with PTEN loss. The positive enrichment of MHC class II antigen presentation, neutrophil degranulation, vesicle-mediated transport, and FC receptor pathway-related genes suggests that PTEN-null tumors may be immunogenic. This observation has potential implications in the context of precision medicine since immune responsive tumors are more likely to respond to immunotherapies.

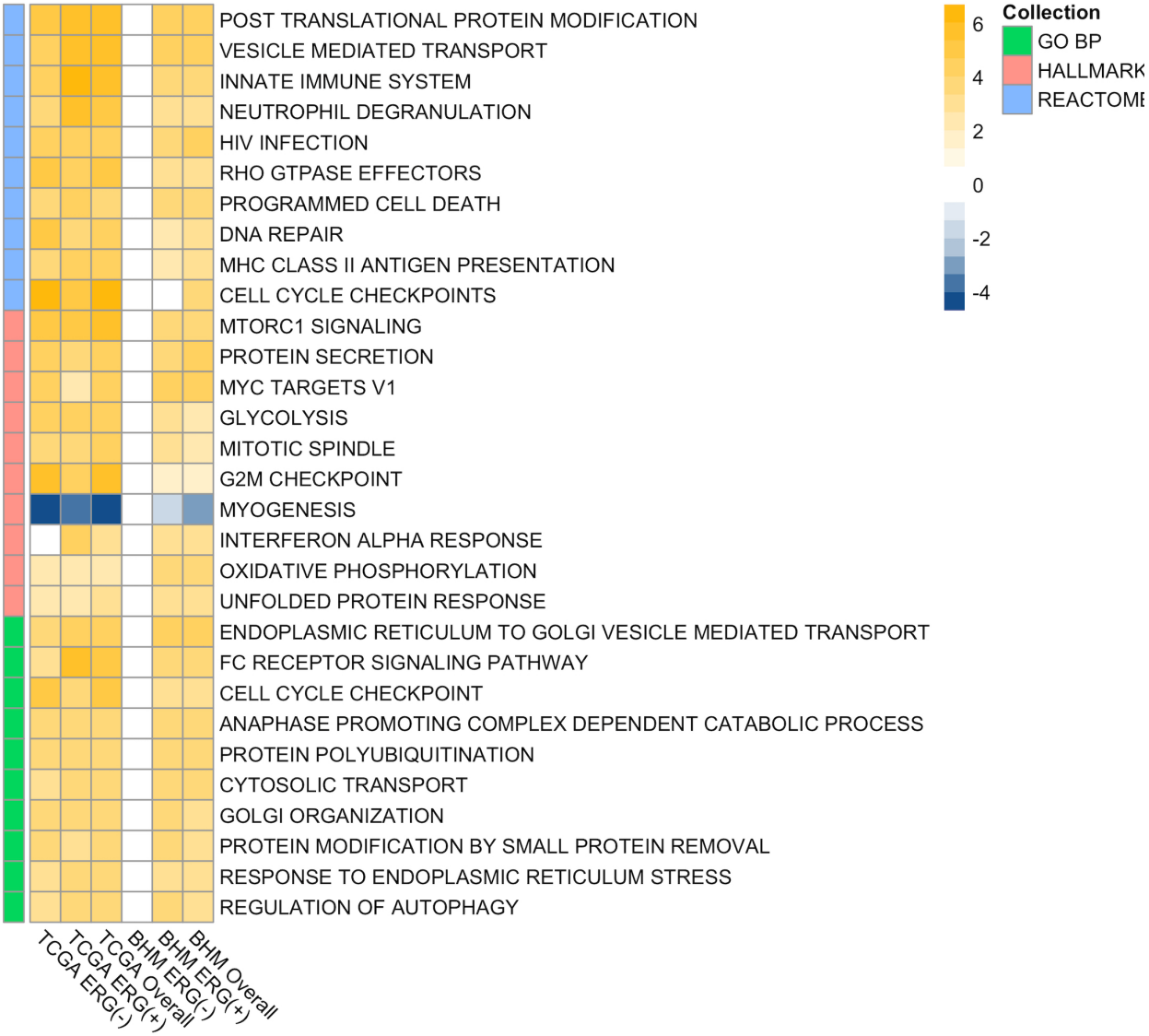


Figure 7. Top enriched gene sets enriched across PTEN-null and PTEN-intact in the TCGA and meta-analysis (BHM) cohorts stratified by ERG status and overall. Heatmap of mean-centered $\log_2$ signed p-values (normalized enriched score multiplied by $\log_{10}$ of p-value) showing the top 10 enriched gene sets of each collection (ranked by signed p-value).

Progress on Major Task 2 and 3 – Subtask 3: During the third year of research, however, we have generated a prognostic gene expression signature for prostate cancer progression using a combination of gene expression data from the public domain, as detailed below. To this end, a total of 674 primary prostate cancer samples (from 3 distinct studies) were used for discovery of the gene signature, while an independent cohort of 248 samples was used for validation and signature performance assessment (see **Table 4**).

Table 4. Collected data sets showing the number of samples and the number of metastasis cases. 3 datasets were used for training and one data set (GSE116918) was used as an independent validation cohort.

GEO accession	GPL	Number of samples (metastasis cases)	Training/Testing
GSE55935	GPL10558	44(8)	Training
GSE51066	GPL5188	85(51)	Training
GSE46691	GPL5188	545(212)	Training
GSE116918	GPL25318	248(22)	**Testing**

First, we have performed a large scale differential expression analysis of gene expression data from different microarray platforms. We have identified 49 up-regulated and 26 down-regulated genes in prostate cancer metastasis cases. We have then further optimized this signature using a "forward search" process reducing the original list to just 14 up-regulated and 17 down-regulated genes. Finally, we have combined the gene expression levels for these genes into a meta-score for use in subsequent analyses, including multivariable Cox proportional hazard model analyses with other clinical and pathological variables (Age, PSA, T-stage ,and Gleason grade).

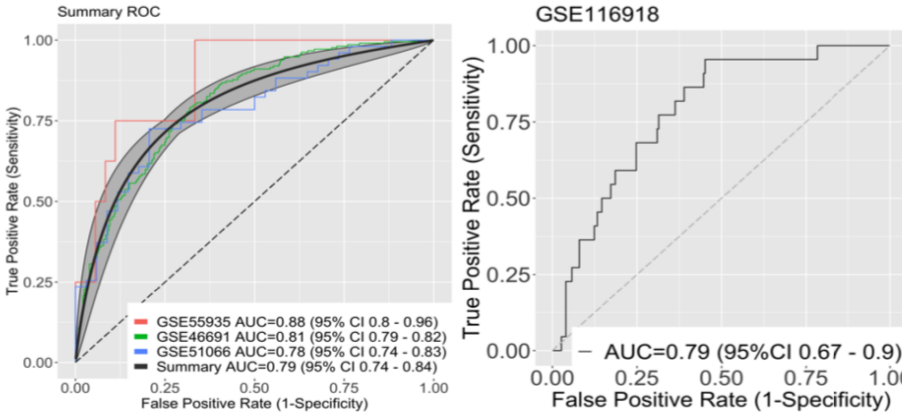


Figure 8. Performance of the prognostic signature. ROC curves in the 3 training data sets with a summary ROC curve of all data sets combined (Left) and ROC curve in the independent testing data set (Right).

To assess the performance of our signature, we measured the area under the receiver operating characteristic curve (AUC). In the training the AUC ranged from 0.78 to 0.88 (**Figure 8**, right panel), while in the testing cohort the AUC was 0.79 (**Figure 8**, left panel), confirming the prognostic value of the signature.We performed Kaplan-Meier analyses in the testing cohort. Patients with higher signature meta-score had worse metastasis-free survival than those with lower score (p-value < 0.0001, see **Figure 9**). Additionally, we also performed survival analyses using individual gene expression levels rather than the signature meta-score. In this analysis, 7 out of 14 up-regulated genes (TMSB10, IQGAP3, CST2, STC2, FOXH1, PTDSS1, HES6) were significantly associated with lower survival, while 8 out of the 17 down-regulated genes (AZGP1, NT5DC1, KCTD14, PTPRN2, UFM1, CCK, KIAA1210, POTEG) were significantly associated with better survival when highly-expressed.

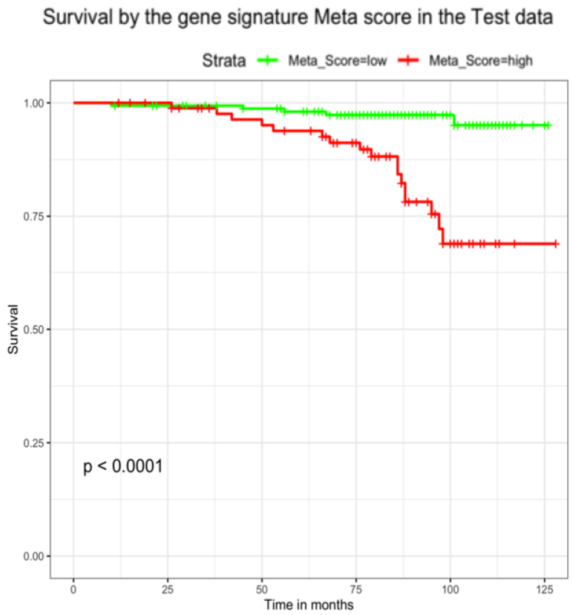


Figure 9. Survival analysis in the testing cohort. Kaplan-Meier curves based on signature meta-score. Patients with low score have a better metastasis-free survival than those with a high score (p-value < 0.0001).

Most importantly, the signature meta-score was the only significant variable in the multivariable Cox regression analysis performed in the testing cohort. The model included

the meta-score together with age, PSA (prostate specific antigen), Gleason grade and T-stage, with a hazard ratio of 5.67 (95% CI : 2.02 - 15.9, see **Figure 10**)

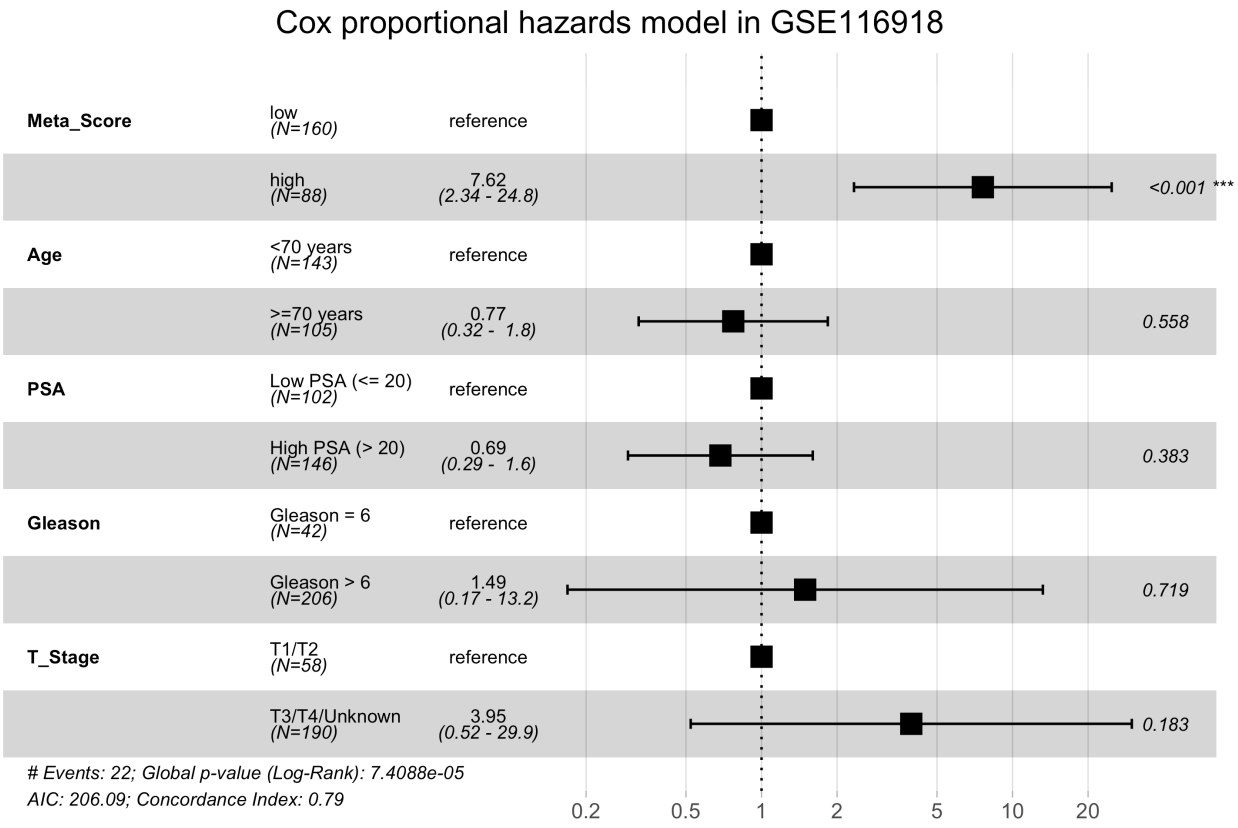


Figure 10. Forest plot for Cox proportional hazards model results in the testing cohort. The signature meta-score is the only significant variable, outperforming other clinical and pathological variables.

Collectively, these analyses show the importance of integrating gene expression data from multiple studies to identify accurate and consistent prognostic signatures. We are currently integrating this signature with PTEN and ERG classification obtained by the EM-algorithm, as previously described.

Training and professional development: Nothing to Report.

Results dissemination to communities of interest: Results from Major Task 1,2,3 were recently published in the following bioRxiv pre-print article: "Transcriptional landscape of PTEN loss in primary prostate cancer", by Eddie Luidy Imada, Diego Fernando Sanchez, Wikum Dinalankara, Thiago Vidotto, Ericka M Ebot, Svitlana Tyekucheva, Gloria Regina Franco, Lorelei Mucci, Massimo Loda, Edward M Schaeffer, Tamara Lotan, Luigi Marchionni. doi: https://doi.org/10.1101/2020.10.08.332049. This article is currently under review in Modern Pathology.

SPECIFIC AIM 3 (Drs. Lotan and Marchionni)

Expected tasks and milestones are summarized below.

Specific Aim 3: Discover and validate gene regulatory and expression signatures associated with PTEN loss on genetically homogeneous ERG-positive and ERG-negative backgrounds.	**Timeline (Months)**
Major Task 1: Select 40 FFPE tumors from Johns Hopkins Surgical Pathology archives (20 ERG-positive and 20 ERG-negative, ETV1-negative). Within each group 10 have heterogeneous PTEN loss, 5 have homogeneous PTEN loss and 5 have intact PTEN by IHC	1-12
Subtask 1: Immunostaining 100 index tumors from Gleason 3+4=7 radical prostatectomies	1-6
Subtask 2: Score staining and select cases	4-8
Subtask 2: Punch blocks and prepare RNA for CAGE	8-12

Major Task 2: Perform CAGE analysis of the tumors resulting from Major Task 1 of Specific Aim3. Technology assessment and troubleshooting in collaboration with Dr. Carninci (RIKEN, Japan)	6-24
Subtask 1: CAGE library preparation, quality assessment, and sequencing • Performed at the Next Generation Sequencing Center (NGSC, Dr. Yegnasubramanian)	6-18
Major Task 3: Bioinformatics analysis of CAGE data generated in Major Task 2 of Specific Aim 3. Technology assessment and troubleshooting in collaboration with Dr. Carninci (RIKEN, Japan)	12-36
Subtask 1: CAGE short reads quality evaluation and alignment to the reference genome • Performed using NGSC computing cluster (Dr. Wheelan)	12-24
Subtask 2: Quantification of expressed genomic regions using CAGE tags • Performed using the School of Public Health (SPH) High Performance Computing Cluster (HPCC)	18-30
Subtask 3: Classification of expressed genomic regions, identification of active enhancers, promoters, and transcript • Performed using SPH HPCC	24-30
Subtask 4: Gene expression regulatory network reconstruction and analysis • Performed using SPH HPCC	24-36
Milestone #4: Co-author manuscript on CAGE analysis of PTEN/ETS status in prostate cancer	30-36

Progress on Major Task 1 – Subtask 1-3 (Dr. Lotan): These activities have been successfully completed.

Progress on Major Task 2 – Subtasks 1 (Dr. Marchionni): During years 1 and 2 of the proposal, we have tested CAGE and nanoCAGE sequencing protocols using high quality RNA obtained from several prostate cancer cell lines. These protocols were optimized for an Illumina mySeq instrument. In year 3 of the proposal, we have focused on optimizing the protocols for RNA samples prepared from tissue specimens. We also worked on developing optimal multiplexing protocols, in order to take advantage of the higher sequencing throughput of the Illumina HiSeq2500 instrument. To this end, we have obtained RNA from 12 tumor samples, prepared the nanoCAGE libraries, and then performed sequencing, as detailed below.

Tumor samples from Major Task 1 were multiplexed and the nanoCAGE protocol was used to the prepare the pooled libraries for sequencing. Before processing the samples on the Illumina HiSeq2500 instrument, we also performed after a successful mini-run on a mySeq instrument. For an unknown reason, however, the sequencer analytical pipeline failed to demultiplex the sequenced samples. We therefore extensively reviewed the experiments and performed an in depth troubleshooting. The quality control analysis in the whole dataset revealed that although the overall sequence quality was good (> 30 Phred Score), there was a high level of duplicated reads (82.6% and 64.2% for R1 and R2, respectively).

We therefore attempted to analyze the sequencing data using an alternative pipeline. Specifically, we tried to process the libraries using the TagDust2 software, which also failed in demultiplexing the libraries. Next, we also aligned the reads to the human genome (hg38) to check if the sequences obtained were originating from the tumor RNA or from the sequencing kit by-products. In this analysis, only about ~6% of the reads aligned uniquely to the human genome, and around ~17% aligned to multiple loci, indicating that most of the sequences obtained were not originating from the human RNA from the tumors. Finally, we tried to align the sequences to the PhiX genome since this DNA was used during the library preparation to increase the library complexity. This analysis revealed that around ~46% of the reads aligned to the PhiX genome, highlighting potential problems during library preparation and/or sequencing (*e.g.*, incorrect primer loading in the Illumina HiSeq2500). Unfortunately, due the COVID-19 pandemic in year 4, the development and troubleshooting of the nanoCAGE libraries had to be halted and this task could not be completed. We, however, were still able to complete our goals with an alternative strategy (see Major Task 3 bellow)

Progress on Major Task 3 – Subtasks 1, 2, 3, and 4 (Dr. Marchionni)

In year 1 and 2 of the project, we have created a comprehensive atlas of gene expression based on recent annotations from the FANTOM consortium based on CAGE-sequencing data (CAGE Associated Transcriptome,

referred as FANTOM-CAT) and the recount2 database. This resource – called FC-R2 – accounts for gene expression summaries for over 109,000 genes across over 70,000 human samples. It encompasses expression information for dozens of thousands lncRNAs genes, including enhancers and promoters. This resource was used as an alternative venue for the study of lncRNAs due to the shortcomings of Major Task 2 – Subtask 1. It enabled us to explore enhancers, promoters and other lncRNAs that have not been explored in this context before.

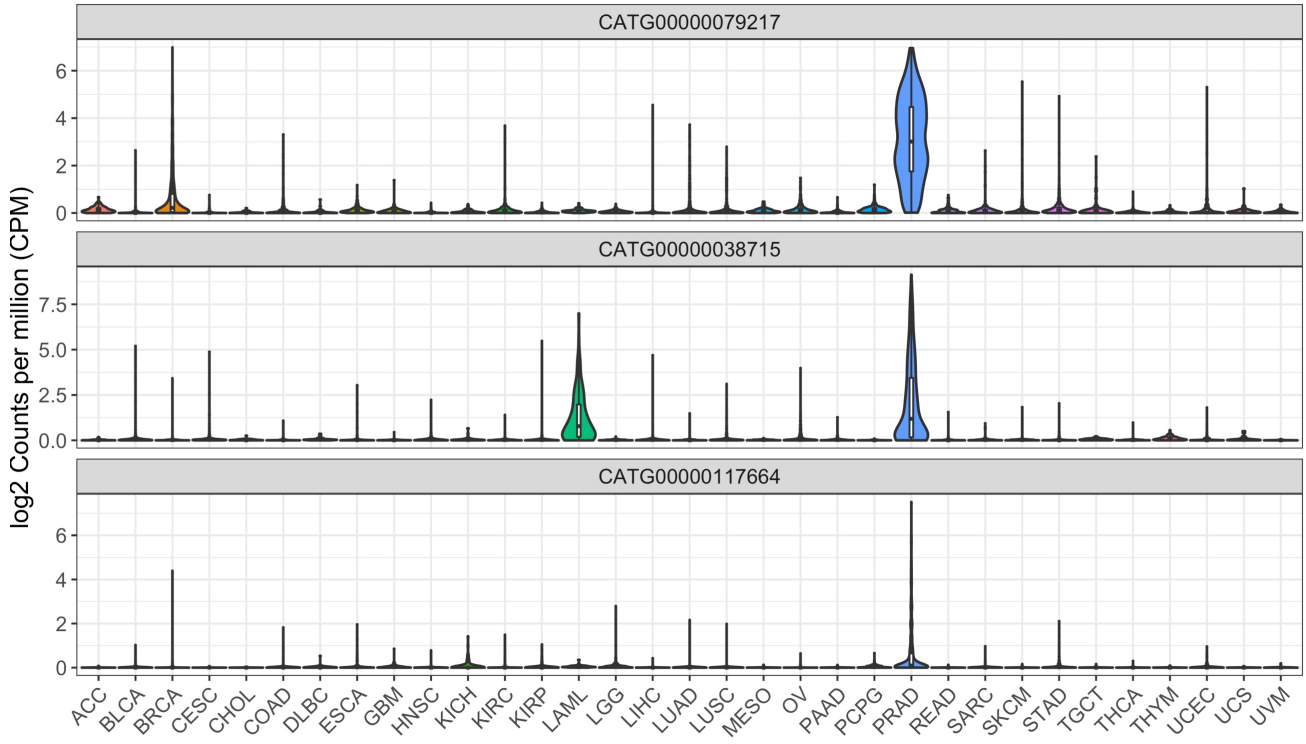


Figure 11 - Expression profiles of novel FANTOM-CAT genes CATG00000038715, CATG00000079217 and CATG00000117664 across 33 cancer types. Violin-plots shows expression (log2 CPM+1) distribution.

In year 3 of the project, we have leveraged the FC-R2 resource and we have performed differential expression analysis between PTEN-null and PTEN-intact samples (see Aim 2 – Major task 3 – subtask 2). In this analysis, we found 264 lnRNAs, including enhancers and promoters, associated with PTEN in PCa, with around half of them not previously reported in association with PCa and were only annotated in the FANTOM-CAT meta-assembly. The FANTOM consortium has recently characterized hundreds of lncRNAs via molecular phenotyping, however none of the lncRNAs associated with PTEN-loss was included in their study, and therefore they still lack an assigned function. In this their study it was shown that the expression levels of genes in the same topological domain are highly correlated only in tissue types in which these genes play a functional role. For this reason, we characterized our novel PTEN associated lncRNAs by analyzing the expression correlation with nearby genes across all cancer types in TCGA.

Among the FANTOM-CAT exclusive genes with the highest fold change in close proximity with coding genes, were CATG00000038715, CATG00000079217 and CATG00000117664. These genes were positioned in the same loci as the genes encoding for CYP4F2, FBXL7, and GPR158, respectively. These lncRNAs genes were mostly expressed in PCa as opposed to other cancer types in TCGA, which might suggest their function are associated with PCa progression (**Figure 11**). All these genes were shown to be highly correlated with their respective "local" coding gene. For example, CATG00000038715 is near CYP4F2 and CYP4F11, which encodes members of the cytochrome P450 enzyme superfamily, and the expression levels of CATG00000038715 and CYP4F2 were found highly correlated almost exclusively in PCa ($R=0.91$, $p < 2.2e-16$) suggesting that CATG00000038715 function might be associated with CYP4F2 in a highly specific manner in PCa (**Figure 12**).

Moreover, all of the coding genes mentioned above (i.e. CYP4F2, FBXL7, and GPR158) are involved in immune response, corroborating with the results of the pathways analysis.

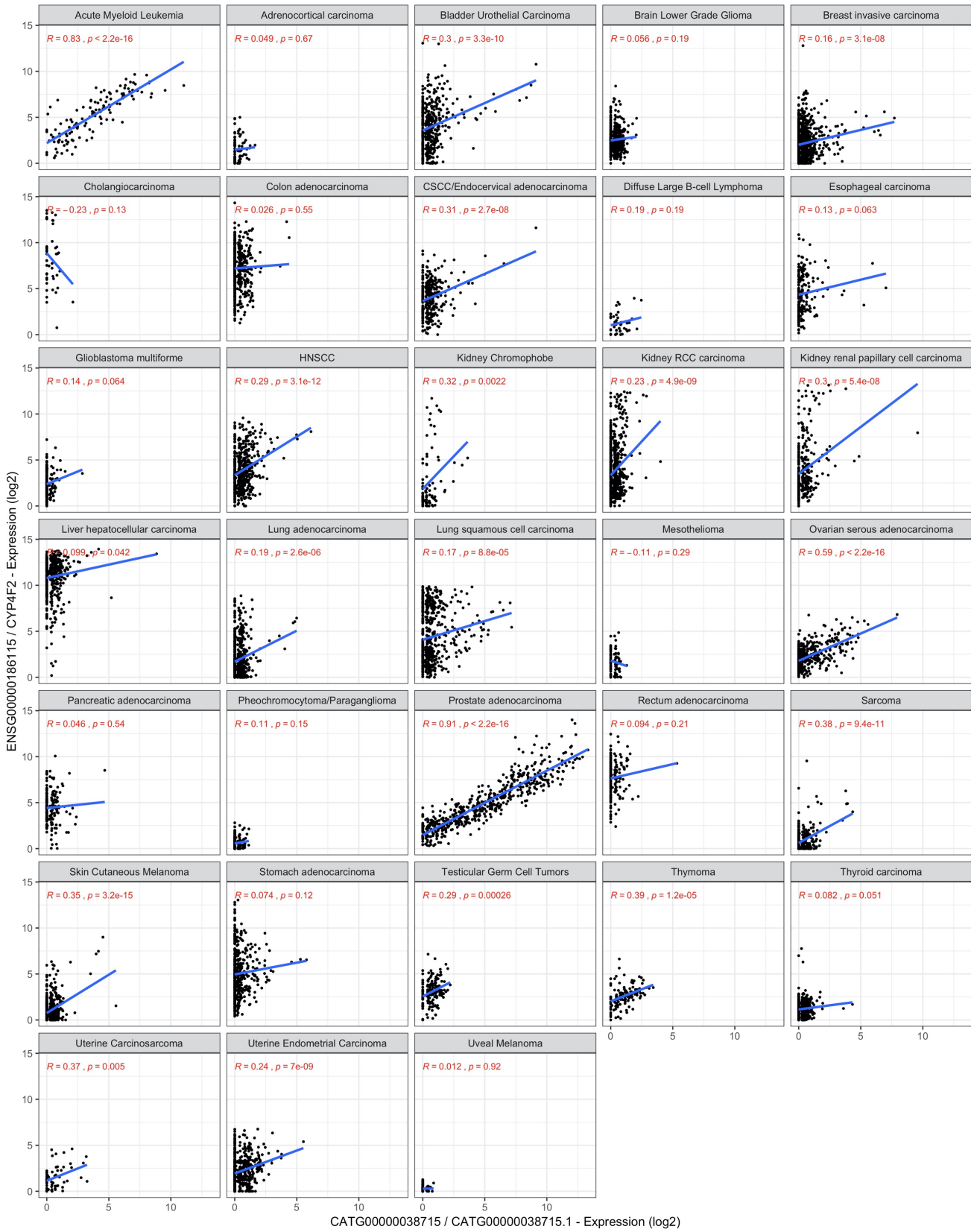


Figure 12 - Person correlation of the unknown gene CATG00000038715 and CYP4F2 across cancer types. CATG00000038715 and CYP4F2 expression are shown to be highly correlated in prostate cancer. Moreover, CATG00000038715 expression is shown to be highly specific to prostate cancer. With exception of leukemia cells, none of the other tumors expressed high levels of CATG00000038715.

Training and professional development: Nothing to Report.

Results dissemination to communities of interest: : Results from Major Task 2 – Subtasks 2, 3, and 4 were recently published in Genome Research: "Recounting the FANTOM Cage Associated Transcriptome", by Eddie-Luidy Imada, Diego Fernando Sanchez, Leonardo Collado-Torres, Christopher Wilks, Tejasvi Matam, Wikum Dinalankara, Aleksey Stupnikov, Francisco Lobo-Pereira, Chi-Wai Yip, Kayoko Yasuzawa, Naoto Kondo, Masayoshi Itoh, Harukazu Suzuki, Takeya Kasukawa, Chung-Chau Hon, Michiel JL de Hoon, Jay W Shin, Piero Carninci, FANTOM consortium, Andrew E Jaffe, Jeffrey T Leek, Alexander Favorov, Gloria R Franco, Ben Langmead, and Luigi Marchionni. doi: https://doi.org/ 10.1101/gr.254656.119

4. IMPACT

Impact on prostate cancer research

We have successfully classified ERG status in all available datasets analyzed. Furthermore, we have successfully reproduced in an independent cohort our previous findings indicating that PTEN loss is associated with a worst prognosis in ERG/ETS-negative patients.

We have successfully applied highly validated IHC and in situ hybridization assays to determine PTEN and ETS status in 2 additional cohorts (MSKCC and JHU) with accompanying gene expression data for future analysis. Association of PTEN with Ki-67 proliferation index has been performed and analyzed for two datasets.

We have developed a consensus molecular signatures of PTEN loss in prostate, showing that PTEN-loss were associated with immune response pathways and biological processes. We have also revealed that ERG negative samples show a higher level of heterogeneity as compared with the ERG positive group, which can be associated with the worst prognosis observed in the former group.

We have generated a comprehensive catalog of expression of coding and non-coding genes using the FANTOM-CAT annotation and the recount2 atlas. Using this resource we have identified hundreds of lncRNAs associated with PTEN and ERG status and investigated potential roles for the top differentially expressed ones.

This project will add significantly to prostate cancer research by further refinement and validation of this prognostic biomarker as we develop expression signatures in the next reporting periods.

Impact on other disciplines: The implementation of the FC-R2 gene expression atlas based on recount2 gene expression summary and the FANTOM-CAT meta-transcriptome will provide a useful resource for studying enhancer and promoter expression in other fields beyond prostate cancer research.

Impact on technology transfer: Nothing to Report.

Impact on society beyond science: Nothing to Report.

5. CHANGES/PROBLEMS

The major change in the research has been the fact that we could not get the CAGE and the nanoCAGE protocols to work properly. For these reason we have developed a bioinformatics pipeline that enables to quantify promoter and enhancer expression from standard RNA-sequencing data. We have then used this pipeline to implement recount2-FANTOM-CAT gene expression atlas. This resource represent a comprehensive compendium of gene expression across the human transcriptome containing over 109,000 genes, greatly expanding the features available for our analyses, by including distinct classes of coding and non-coding genes, such as messenger RNAs, intergenic lncRNAs, and expressed divergent promoters and enhancers. Using this resource we were able to analyze promoter and enhancer expression in PTEN and ERG prostate cancer tumors, ultimately attaining the scientific goals for which the use of CAGE and nanoCAGE were originally proposed.

6. PRODUCTS

As results of the research activities supported on this award the following manuscripts were published:

1. *"Recounting the FANTOM CAGE-Associated Transcriptome."* Eddie Luidy Imada, Diego Fernando Sanchez, Leonardo Collado-Torres, et al. Genome Res. 2020 Jul; 30(7): 1073–1081. doi: 10.1101/gr.254656.119. PMCID: PMC7397872
2. *Functional annotation of human long noncoding RNAs via molecular phenotyping."* Jordan A. Ramilowski, Chi Wai Yip, Saumya Agrawal,et al. Genome Res. 2020 Jul; 30(7): 1060–1072. doi: 10.1101/gr.254219.119 - Correction in: Genome Res. 2020 Sep; 30(9): 13771. PMCID: PMC7397864
3. *"PTEN Loss with ERG Negative Status is Associated with Lethal Disease after Radical Prostatectomy*". Haney NM, Faisal FA, Lu J, et al. J Urol. 2020 Feb;203(2):344-350. doi: 10.1097/JU.0000000000000533. Epub 2019 Sep 10. PMID: 31502941.

As results of the research activities supported on this award the following pre-print were published:

1. *"Transcriptional landscape of PTEN loss in primary prostate cancer"* by Eddie Luidy Imada, Diego Fernando Sanchez, Wikum Dinalankara, et al. Preprint in biorXiv doi: https://doi.org/10.1101/2020.10.08.332049.

As results of the research activities supported on this award the following resources were made available:

1. F2-RC gene expression atlas: http://marchionnilab.org/fcr2.html

7. PARTICIPANTS & OTHER COLLABORATING ORGANIZATIONS

Name:	Luigi Marchionni
Project role:	Initiating Principle Investigator
Researcher Identifier:	0000-0002-7336-8071 (ORCID)
Institution:	Johns Hopkins University
Nearest person month worked:	4 (rounded to 4)
Contribution to Project:	Dr. Marchionni coordinated the project, provided supervision of research activities provided by the fellows, and directly performed the analyses
Funding Support:	NA

Name:	Tamara Lotan
Project role:	Partnering Principle investigator
Researcher Identifier:	0000-0002-0494-9067 (ORCID)
Institution:	Johns Hopkins University
Nearest person month worked:	1
Contribution to Project:	Dr. Lotan coordinated the project, provided supervision of research activities provided by the fellows, and directly performed the analyses
Funding Support:	NA

Name:	Anne Jedlicka
Project role:	Co-investigator
Researcher Identifier:	NA
Institution:	Johns Hopkins University

Nearest person month worked:	1
Contribution to Project:	Dr. Anne Jedlicka coordinated the experiments with CAGE
Funding Support:	NA

Name:	Amanda Dziedzic
Project role:	Research specialist
Researcher Identifier:	NA
Institution:	Johns Hopkins University
Nearest person month worked:	1
Contribution to Project:	Ms. Amanda Dziedzic performed the experiments with CAGE

Name:	Wikum Dinalankara
Project role:	Post-doctoral fellow
Researcher Identifier:	NA
Institution:	Johns Hopkins University
Nearest person month worked:	5 (rounded to 5)
Johns **Contribution to Project:**	Dr. Dinalankara performed bioinformatics and statistical analyses under Dr. Marchionni supervision
Funding Support:	NA

Name:	Eddie Luidy-Imada
Project role:	Post-doctoral fellow
Researcher Identifier:	0000-0001-9527-3703 (ORCID)
Institution:	Johns Hopkins University
Nearest person month worked:	12
Contribution to Project:	Dr. Luidy-Imada performed bioinformatics and statistical analyses under Dr. Marchionni supervision
Funding Support:	NA

Name:	Diego Sanchez Martinez
Project role:	Graduate Student
Researcher Identifier:	NA
Institution:	Johns Hopkins University
Nearest person month worked:	8
Contribution to Project:	Dr. Sanchez Martinez performed bioinformatics and statistical analyses under Dr. Marchionni supervision for developing the prognostic signature.
Funding Support:	NA

Name:	Lotte Mulder
Project role:	Student
Researcher Identifier:	NA
Institution:	Johns Hopkins University
Nearest person month worked:	2
Contribution to Project:	Ms. Mulder performed bioinformatics and statistical analyses under Dr. Marchionni supervision for developing the prognostic signature.
Funding Support:	NA

Name:	Daniella Salles
Project role:	Post-doctoral fellow
Researcher Identifier:	
Institution:	Johns Hopkins University
Nearest person month worked:	1
Contribution to Project:	Dr. Salles performed laboratory analyses under Dr. Lotan supervision
Funding Support:	NA

Name:	Ericka M. Ebot
Project role:	Co-investigator
Researcher Identifier:	NA
Institution:	Harvard T.H. Chan School of Public Health

Nearest person month worked:	1 (rounded to 1)
Contribution to Project:	Dr. Ebot provided analytical support for the PHS/HPHS cohorts
Funding Support:	NA

Name:	Kaushal Asrani
Project role:	Postdoctoral fellow
Researcher Identifier:	NA
Institution:	Johns Hopkins University
Nearest person month worked:	5
Contribution to Project:	Dr. Asrani performed data collection and interpretation supervised by Dr. Lotan
Funding Support:	NA

Name:	Rafael Guerrero-Preston
Project role:	
Researcher Identifier:	NA
Institution:	Johns Hopkins University
Nearest person month worked:	1
Contribution to Project:	
Funding Support:	NA

Name:	Lanlan Ji
Project role:	
Researcher Identifier:	NA
Institution:	Johns Hopkins University
Nearest person month worked:	2
Contribution to Project:	
Funding Support:	NA

Change in active other support

Dr. Marchionni:

- No longer supported by 1U54RR023561-01A1 (Ford)
- No longer supported by KKESH (Eberhart) – this award is completed
- No longer supported by W81XWH-12-PCRP-TIA – this award is completed
- No longer supported by R01CA163594 (Sidransky) – this award is completed
- PC141474 (Tomlins and Schaeffer) – this award is completed
- R21 AI124776-01 (Romerio) – this award is completed
- 1R01CA211695-01A1 (Hurley) – this award is completed
- R01 PA-13-302 (Marchionni) – this award is now active and moved from pending
- R01CA206027 (Sidransky/Hoque) – this award is now active and moved from pending
- R01CA208709 (Sidransky/Hoque) – this award is now active and moved from pending
- W81XWH-16-PCRP-IDA (Lupold) – this award is now active and moved from pending
- U01CA231776 (Marchionni/Tran/Hann) – this award is now active and moved from pending
- R01CA235681 (Hahn) – this award is now active and moved from pending
- W81XWH-19-1-0292 (Lotan) – this award is now active and moved from pending

Dr. Lotan:

- New Award W81XWH-19-1-0292 (Lotan[PI], Title: Epigenomic Landscape of Primary Prostate Cancer in African-American Men, 10% effort)

- New Award W81-XWH-19-1-0781 (Asrani[PI], Title: mTORC1 Regulates MiTF Expression and Lysosomal Biogenesis, 2% effort)
- New Award R01 CA238218 (Pienta [PI], Title: Dissecting the prostate cancer diaspora, 1% effort)
- New Award PC180810 (Luo [PI], Title: Genetic and genomic determinants of homologous recombination repair deficiency as treatment selection markers for lethal prostate cancer, 5% effort)
- New Award PC180375 (Isaacs[PI], Title: Discovery and Functional Analyses of Susceptibility Genes for Lethal Prostate Cancer, 5% effort)
- No longer supported by W81XWH-15-1-0661, R01CA211695, RSG-17-160-01-CSM.

Other organizations were involved
Organization Name: Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA
Organization Name: Memorial Sloan Kettering Cancer Center, New York, NY, USA

Partner's contribution to the project
Collaboration: Dr. Ericka Ebot (Harvard) provided analytical support for the PHS/HPHS cohorts (<1 person/month effort).

Dr. Anu Gopalan (MSKCC) is a pathologist who created the MSKCC TMAs described above and she has participated in Ki-67 scoring and data analysis of these materials after providing them to us (<1 person/month effort).

8. SPECIAL REPORTING REQUIREMENTS

This project (W81XWH-16-1-0739) is a collaborative award with Dr. Tamara Lotan (Partnering PI, award).

9. APPENDICES

Manuscripts are attached below.

PTEN Loss with *ERG* Negative Status is Associated with Lethal Disease after Radical Prostatectomy

Nora M. Haney, Farzana A. Faisal,* Jiayun Lu, Liana B. Guedes, Victor E. Reuter, Howard I. Scher,† James A. Eastham, Luigi Marchionni, Corinne Joshu, Anuradha Gopalan‡ and Tamara L. Lotan‡

From the Departments of Urology (NMH, FAF, TLL), Pathology (LBG, TLL) and Oncology (LM, TLL) and Center for Computational Genomics (LM), Johns Hopkins University School of Medicine and Department of Epidemiology, Johns Hopkins University Bloomberg School of Public Health (JL, CJ), Baltimore, Maryland, and Departments of Pathology (VER, AG), Genitourinary Oncology (HIS) and Urology (JAE), Memorial Sloan Kettering Cancer Center, New York, New York

Abbreviations and Acronyms

AA = African American
ADT = androgen deprivation therapy
AR = androgen receptor
BCR = biochemical recurrence
EA = European American
ERG = ETS-related gene
FISH = fluorescence in situ hybridization
GG = Grade Group
IHC = immunohistochemistry
PCa = prostate cancer
PTEN = phosphatase and tensin homolog
RP = radical prostatectomy
TMA = tissue microarray

Purpose: Few groups have investigated the combined effects of *PTEN* loss and *ERG* expression on the outcomes of metastasis of or death from prostate cancer in surgically treated patients. We examined the association of *PTEN*/*ERG* status with lethal prostate cancer in patients treated with radical prostatectomy.

Materials and Methods: Included in analysis were 791 patients with clinically localized prostate cancer treated with radical prostatectomy at a single institution. Genetically validated immunohistochemistry assays for *PTEN* and *ERG* were performed on tissue microarrays. Multivariable Cox proportional hazard models were used to assess the association of *PTEN*/*ERG* status with lethal prostate cancer (defined as metastasis or prostate cancer specific death), adjusting for patient age, race, pathological grade and stage, and surgical margin status.

Results: Median followup in the cohort was 12.8 years. Of 791 cases 203 (25%) demonstrated *PTEN* loss and 330 of 776 (43%) were *ERG* positive. On multivariable analysis *PTEN* loss (HR 1.9, 95% CI 1.2–3.0, p=0.012) but not *ERG* expression (HR 0.6, 95% CI 0.4–1.1, p=0.11) was associated with an increased risk of lethal prostate cancer. The association of *PTEN* loss with lethal disease only remained among men with *ERG* negative tumors (HR 2.3, 95% CI 1.3–4.1, p=0.005) and not *ERG* positive tumors (HR 1.1, 95% CI 0.6–2.1, p=0.81).

Conclusions: *PTEN* loss is associated with an increased risk of lethal prostate cancer after radical prostatectomy and this risk is most pronounced in the subgroup of patients with *ERG* negative tumors. This work corroborates the use of *PTEN* and *ERG* status for risk stratification in surgically treated patients.

Key words: prostatic neoplasms, prostatectomy, mortality, PTEN phosphohydrolase, oncogene proteins

Accepted for publication September 1, 2019.

Supported by CDMRP (Congressionally Directed Medical Research Programs) Prostate Cancer Research Program Awards W81XWH-12-PCRP-TIA (HIS, TLL, AG, VER) and W81XWH-16-1-0737 (TLL, LM, AG), and NCI (National Cancer Institute) Cancer Center Support Grant 5P30CA006973.

* Correspondence: Department of Urology, Brady Urological Institute, Johns Hopkins University School of Medicine, Baltimore, Maryland 21287 (email: ffaisal1@jhmi.edu).

† Financial interest and/or other relationship with Asterias Biotherapeutics, Ambry Genetics, Konica Minolta, Amgen, ESSA Pharma, Janssen, OncLive Insights, Menarini Silicon, Physicians Education Resource, Sanofi Aventis, WCG Oncology, Epic Sciences, Illumina, Innocrin Pharma and ThermoFisher.

‡ Equal study contribution.

0022-5347/20/2032-0344/0
THE JOURNAL OF UROLOGY®
https://doi.org/10.1097/JU.0000000000000533
Vol. 203, 344-350, February 2020
Printed in U.S.A.

PHOSPHATASE and tensin homolog is a commonly deleted tumor suppressor in PCa. Its loss results in unopposed activity of PI3K and up-regulation of the oncogenic Akt/mTOR signaling pathways.[1] *PTEN* deletion frequently occurs as focal and subclonal events in primary prostate tumors but homogeneous and heterogeneous loss can be reliably detected by genetically validated IHC.[2–4]

PTEN loss is associated with adverse pathological features at RP and an increased risk of BCR after RP.[5–8] Few studies have been done to examine the association of *PTEN* loss with more clinically meaningful outcomes in surgically treated patients, such as metastasis or death.[3,9] Others have been limited to outcomes in conservatively managed cohorts.[10–12]

PTEN loss commonly occurs in tumors with *ERG* gene rearrangements. Fusion of *ERG* (an *ETS* family transcription factor) with the androgen regulated gene *TMPRSS2* is the most common genomic rearrangement found in PCa, occurring in about 50% of patients with PCa who are of European descent.[13] *TMPRSS2-ERG* rearrangements by translocation or deletion in tumor cells subsequently put *ERG* expression under the control of an androgen regulated promoter. While *ERG* rearrangement alone does not predict poor prognosis in surgical cohorts,[14] animal studies suggest that *ERG* rearrangement and *PTEN* loss may work synergistically in tumor progression.[12,15] However, retrospective clinical studies conflict.

Initial studies showed that *PTEN* loss in an *ERG* positive background increased the risk of BCR after surgery[16,17] and yet larger studies have shown that *PTEN* loss predicts BCR regardless of *ERG* status.[2,18] When death from PCa is the primary outcome, *PTEN* loss with *ERG* negative status is associated with worse survival.[3,11,12] Data on a population based, prospective cohort showed that *PTEN* loss was associated with lethal progression after surgery only when *ERG* status was negative.[3] Reid et al performed FISH assays revealing that *PTEN* deletion without *ERG* rearrangement predicted cancer specific death in a conservatively managed cohort.[11] To our knowledge only 1 study has been done to examine *PTEN*/*ERG* status by IHC in a large RP cohort uniformly treated at a single institution.[9] This study showed that while *PTEN* loss predicted metastasis and PCa specific death after RP, *ERG* status did not provide any additional benefit.

Given the paucity of studies and conflicting results, we investigated the combined effects of *PTEN* and *ERG* status on long-term oncologic outcomes in a large, surgically treated cohort from a single institution. Using automated and genetically validated IHC we examined the association of *PTEN* and *ERG* status with lethal PCa after RP.

MATERIALS AND METHODS

Study Population and Tissue Microarray Construction

Institutional Review Board approval was obtained from the 2 participating institutions, namely Memorial Sloan Kettering Cancer Center and the Johns Hopkins Medical Institutions (IRB No. NA 00091198). The cohort consisted of men treated with RP of localized PCa between 1985 and 2003 at Memorial Sloan Kettering Cancer Center.[19] Those who received neoadjuvant or adjuvant ADT, or radiation therapy were excluded from study. Only patients with available slides, blocks and followup information were included in the final cohort, which included 915 RP specimens with a total of 2,745 tumor cores in TMA sets.

H&E slides of the RP specimens were reviewed and slides containing tumor were marked and matched with corresponding paraffin blocks. Tissue cores (0.6 mm) were punched out in triplicate from randomly selected locations in marked tumor areas and mounted in blank recipient blocks using an automated tissue microarrayer (Beecher Instruments, Sun Prairie, Wisconsin). Samples were from the largest tumor focus in most cases. Separate tumor foci were punched only when there were small tumor foci with no dominant nodule. TMAs were tested with validated IHC to determine *PTEN*/*ERG* status.

Clinicopathological and long-term followup information was available on all patients in the final cohort. The primary outcome was lethal PCa, defined as distant metastasis detected on imaging or PCa specific death. This composite definition of lethal PCa was chosen since metastasis-free survival is a strong surrogate for survival in patients with localized PCa.[20]

Immunohistochemistry Assays and Scoring

PTEN and *ERG* IHC were performed in a CLIA (Clinical Laboratory Improvement Amendments) accredited laboratory on the Ventana Discovery Ultra platform (Ventana Medical Systems, Tucson, Arizona) using previously validated protocols.[2–4,21] Briefly, these assays use rabbit antihuman *PTEN* antibody (Clone D4.3 XP, Cell Signaling Technology®) or rabbit antihuman *ERG* antibody (EPR3864). After staining all TMAs were scanned at 20× magnification using an Aperio® device and segmented into TMAJ (http://tmaj.pathology.jhmi.edu/) for scoring.

PTEN and *ERG* protein status was blindly scored by trained urological pathologists (LBG and TLL). *PTEN* was scored as homogeneous *PTEN* loss if all tumor glands sampled in a given case showed cytoplasmic and nuclear *PTEN* loss compared to surrounding internal control benign glands in stroma. *PTEN* was scored as heterogeneous *PTEN* loss if some but not all tumor tissue sampled in a given case showed *PTEN* loss. *PTEN* was scored as intact if all tumor tissue sampled showed *PTEN*. *ERG* was scored as positive if any tumor glands showed nuclear *ERG* expression. *ERG* was scored as negative if no sampled tumor gland showed *ERG* expression.

Statistical Analysis

Clinicopathological characteristics of the *PTEN*/*ERG* status subgroups were compared using the Wilcoxon or Kruskal-Wallis test for continuous variables and the chi-square test for categorical variables. Univariable and multivariable Cox proportional hazard regression models were constructed to estimate the HR and 95% CI of lethal PCa. Patient age and race, pathological grade and stage, and surgical margin status were included in the multivariable model. We used the Kaplan-Meier method to examine the risk of lethal PCa stratified by *PTEN* and *ERG* status.

All tests were 2-sided with $p < 0.05$ considered statistically significant. Analyses were done with SAS®, version 9.4.

RESULTS

Tumor was present on TMA slides in 915 cases, of which 791 (86%) had interpretable staining results with adequate *PTEN* control staining. Of the 791 cases with *PTEN* staining data 776 (93%) had informative staining results for *ERG*. *PTEN* loss was present in 203 of 791 cases (26%), of which 96 (12%) and 107 (14%) showed homogeneous and heterogenous *PTEN* loss, respectively. *ERG* expression was present in 330 of 776 cases (43%). *PTEN* loss was more common among *ERG* positive than *ERG* negative cases (120 of 330 or 36% vs 83 of 446 or 19%, $p < 0.001$). *PTEN* loss and *ERG* expression were more prevalent in EA men than in AA men with *PTEN* loss in 27% of EA vs 9% of AA men ($p = 0.003$) and *ERG* expression in 44% of EA vs 21% of AA men ($p < 0.001$).

On IHC *PTEN* loss was associated with adverse pathological features at RP (table 1). Of the 192 tumors with *PTEN* loss 76 (approximately 38%) were GG 3-5 compared to 108 of 577 GG 3-5 tumors (18%) with intact *PTEN* ($p < 0.001$). Similarly, 98 of 203 *PTEN* loss tumors (48%) were nonorgan confined while 29% with *PTEN* intact demonstrated extraprostatic disease ($p < 0.001$).

Median followup was 12.8 years. Lethal PCa events developed in 92 of the 776 patients (12%) with complete *PTEN* and *ERG* results. On multivariable analysis *PTEN* loss was significantly associated with an increased risk of lethal PCa (HR 1.98, 95% CI 1.15–3.04, $p = 0.012$, table 2). *ERG* expression did not predict lethal PCa on multivariable analysis (HR 0.64, 95% CI 0.36–1.11, $p = 0.113$). Table 2 and the figure show the association of joint categories of *PTEN* loss and *ERG* status with lethal PCa. Compared to cases with *PTEN* intact and *ERG* negative status, *PTEN* loss with *ERG* negative status was the only subgroup significantly associated with an increased risk of lethal progression on univariable analysis (HR 3.76, 95% CI 2.27–6.21, $p < 0.001$) and multivariable analysis (HR 2.31, 95% CI 1.29–4.14, $p = 0.005$, log rank $p < 0.001$). *ERG* positive cases with *PTEN* loss carried a higher risk of lethal disease on univariable analysis (HR 1.84, 95% CI 1.06–3.18, $p = 0.030$).

Table 1. *Clinicopathological characteristics of patients stratified by* PTEN *and* ERG *status*

	PTEN (791 pts)			*ERG* (776 pts)			*PTEN/ERG* (776 pts)				
	Loss	Intact	p Value*	Neg	Pos	p Value*	*PTEN* Intact/*ERG* Neg	*PTEN* Intact/*ERG* Pos	*PTEN* Loss/*ERG* Neg	*PTEN* Loss/*ERG* Pos	p Value*
No. pts	203	588	—	466	330	—	363	210	83	120	—
Median age at RP	62.60	61.32	0.026	62.33	60.36	0.002	62.08	59.55	63.33	61.43	<0.001
No. race (%):			0.016			0.003					0.001
European American	189 (93.1)	512 (87.1)		387 (86.8)	301 (91.2)		309 (85.1)	190 (90.5)	78 (94.0)	111 (92.5)	
African American	5 (2.5)	49 (8.3)		42 (9.4)	11 (3.3)		40 (11.0)	8 (3.8)	2 (2.4)	3 (2.5)	
Other	5 (2.5)	15 (2.6)		9 (2.0)	10 (3.0)		9 (2.5)	5 (2.4)	0	5 (4.2)	
No. Grade Group (%):			<0.001			<0.001					<0.001
1	33 (16.3)	216 (36.7)		119 (26.7)	122 (37.0)		108 (29.8)	100 (47.6)	11 (13.3)	22 (18.3)	
2	83 (40.9)	253 (43.0)		191 (42.8)	141 (42.7)		163 (44.9)	86 (41.0)	28 (33.7)	55 (45.8)	
3	44 (21.7)	56 (9.5)		65 (14.6)	34 (10.3)		46 (12.7)	9 (4.3)	19 (22.9)	25 (20.8)	
4	19 (9.4)	32 (5.4)		39 (8.7)	11 (3.3)		26 (7.2)	5 (2.4)	13 (15.7)	6 (5.0)	
5	13 (6.4)	20 (3.4)		24 (5.4)	9 (2.7)		16 (4.4)	4 (1.9)	8 (9.6)	5 (4.2)	
No. stage (%):			<0.001			0.066					<0.001
T2	105 (51.7)	417 (70.9)		291 (65.3)	219 (66.4)		253 (69.7)	152 (72.4)	38 (45.8)	67 (55.8)	
T3	87 (42.9)	156 (26.5)		135 (30.3)	106 (32.1)		99 (27.3)	55 (26.2)	36 (43.4)	51 (42.5)	
T4	11 (5.4)	15 (2.6)		20 (4.5)	5 (1.5)		11 (3.0)	3 (1.4)	9 (10.8)	2 (1.7)	
No. margin (%):			0.230			0.737					0.695
Neg	122 (60.1)	381 (64.8)		285 (63.9)	207 (62.7)		236 (65.0)	134 (63.8)	49 (59.0)	73 (60.8)	
Pos	81 (39.9)	207 (35.2)		161 (36.1)	123 (37.3)		127 (35.0)	76 (36.2)	34 (41.0)	47 (39.2)	
No. *PTEN* loss (%):		—	—	—	—	—	—	—	—	—	—
Heterogenous	96 (47.3)										
Homogeneous	107 (52.7)										

* Wilcoxon test or Kruskal-Wallis test for continuous variables and chi-square test for categorical variables.

Table 2. *Univariable and multivariable Cox proportional hazard models for lethal prostate cancer*

			Univariable		Multivariable*	
	No. Cases	No. Controls	HR (95% CI)	p Value	HR (95% CI)	p Value
PTEN:						
Intact	46	542	Referent	—	Referent	—
Loss	47	156	3.25 (2.16–4.88)	<0.001	1.87 (1.15–3.04)	0.012
ERG:						
Neg	62	384	Referent	—	Referent	—
Pos	30	300	0.64 (0.41–0.99)	0.043	0.64 (0.36–1.11)	0.113
PTEN/*ERG*:						
PTEN intact/*ERG* neg	35	328	Referent	—	Referent	—
PTEN intact/*ERG* pos	10	200	0.47 (0.23–0.96)	0.037	0.48 (0.18–1.26)	0.136
PTEN loss/*ERG* neg	27	56	3.76 (2.27–6.21)	<0.001	2.31 (1.29–4.14)	0.005
PTEN loss/*ERG* pos	20	100	1.84 (1.06–3.18)	0.030	1.09 (0.56–2.12)	0.809

* Adjusted for age at RP, race, grade group, stage and surgical margin status.

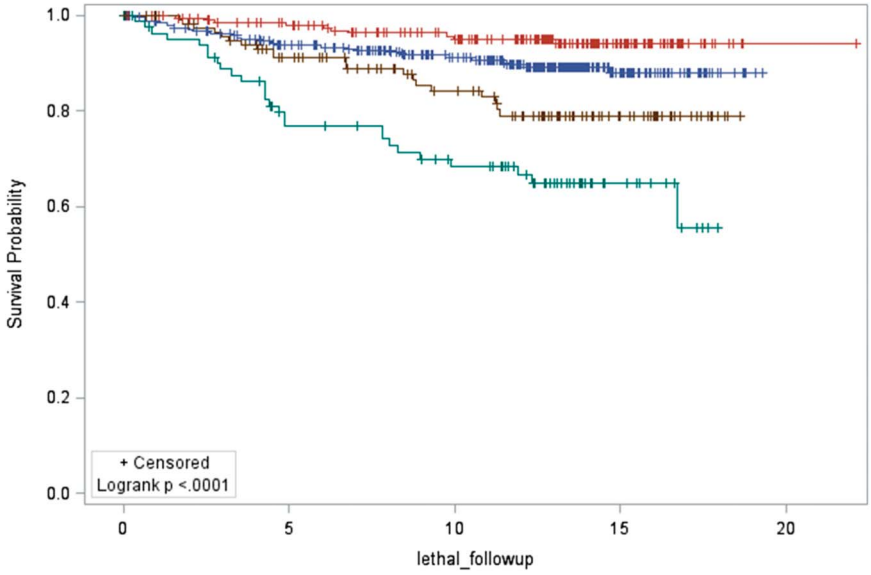
However, this was not significant on multivariable analysis (HR 1.09, 95% CI 0.56–2.12, p = 0.809).

DISCUSSION

Long-term studies demonstrating an association between *PTEN* loss and clinically meaningful outcomes such as metastasis or cancer specific death have been lacking. Moreover, studies of the modifying effect of *ERG* status on *PTEN* loss have conflicted. Using FISH techniques in large cohorts to determine *PTEN*/*ERG* status is time-consuming and technically challenging. Using automated and validated IHC, we found that *PTEN* loss was significantly associated with an approximately twofold increased risk of lethal PCa in a large cohort treated with RP and followed long-term at a single institution. This risk was only significant in the subgroup of patients with *PTEN* loss and with *ERG* negative tumors. Patients with *PTEN* loss but *ERG* positive tumors were not at increased risk for lethal progression. These findings support the clinical usefulness of automated and inexpensive IHC assays for *PTEN* and *ERG* for risk stratification and treatment in post-RP cases.

At this institution we routinely perform *PTEN* and *ERG* IHC testing in GG 1 biopsies. Loss of *PTEN*, particularly when *ERG* is negative, is a relative contraindication to active surveillance. Given these results, we plan to incorporate *PTEN*/*ERG* testing in the RP setting to guide postoperative management.

Our group genetically validated automated IHC for *PTEN* detection to study *PTEN* loss.[2–4] There is high correlation of automated IHC with FISH. Intact *PTEN* immunostaining is 91% specific for the absence of *PTEN* deletion by FISH, and 97% and



Kaplan-Meier survival curves of freedom from lethal PCa by *PTEN* and *ERG* status. Blue curve indicates *PTEN* intact and *ERG* negative in 363 patients. Red curve indicates *PTEN* intact and *ERG* positive in 210 patients. Green curve indicates *PTEN* loss and *ERG* negative in 83 patients. Brown curve indicates *PTEN* loss and *ERG* positive in 120 patients.

65% sensitive for the detection of homozygous and hemizygous deletion by FISH, respectively.[4] The effects of fixation technique and duration, tissue processing type and slide or block age are largely negligible.[22] Interobserver variability is minimal with κ values consistently above 0.9.[3] IHC also provides significant cost and time savings compared to FISH.

Several studies have demonstrated the prognostic role of *PTEN* in predicting upgrading at surveillance biopsy, discontinuation of active surveillance, adverse pathology at RP and BCR after surgery.[5–8,23–25] *PTEN* loss strongly correlates with unfavorable histological features, including intraductal carcinoma, cribriform Gleason pattern 4 and stromogenic PCa.[26]

However, only a few studies have been done to investigate the effect of *PTEN* loss stratified by *ERG* status on metastasis and death outcomes. Reid et al found that *PTEN* deletion without *ERG* rearrangements by FISH increased the risk of cancer specific death in a conservatively managed cohort, although they could not reproduce this finding in a larger cohort.[11] In a surgically treated cohort Ahearn et al used IHC to determine that *PTEN* loss and *ERG* negative tumors were associated with lethal progression.[3] However, data were collected from national population based studies including patients treated for more than 20 years at multiple institutions with self-reporting relied on for followup. Leapman et al analyzed 424 cases treated with RP at a single institution and found that *ERG* status did not add to the c-index of the CAPRA-S (Cancer of the Prostate Risk Assessment Post-Surgical) score and *PTEN*.[9] However, it was not explicitly examined whether cases with *PTEN* loss had worse outcomes when *ERG* was negative compared to those that were *ERG* positive.

It is unclear why some previous studies have shown that *PTEN* loss with *ERG* positive status was associated with the highest risk of BCR after surgery.[16,17,27] Long-term followup has shown that *ERG* negative *PTEN* loss is the subgroup at increased risk for lethal progression.[3,11,12] One important caveat is that due to the frequency of *PTEN* loss among *ERG* positive tumors there are substantially more *ERG* positive tumors with *PTEN* loss than *ERG* negative tumors with *PTEN* loss. Thus, smaller studies were almost certainly underpowered to compare the effects of *PTEN* loss on *ERG* positive and *ERG* negative backgrounds while larger studies revealed no effect of *ERG* status on the association of *PTEN* loss with BCR.

Furthermore, BCR is a different outcome than metastasis or death. Since salvage radiation and ADT are generally introduced after BCR, biomarkers predictive of a response to radiation therapy or ADT may be associated with metastasis and death but not with BCR. Thus, while there may be a lack of interaction between *PTEN* and *ERG* for an association with BCR, this interaction may be seen in cohorts with longer followup (perhaps those in which ADT is introduced early) for an association with metastasis and death. Clearly additional trials are necessary to formally test this hypothetical interaction of *PTEN*/*ERG* with radiation therapy and/or ADT after BCR.

Preclinical studies have been done to examine the influence of *PTEN*/*ERG* status on androgen signaling. *PTEN* loss has been demonstrated to down-regulate AR and AR driven gene transcription.[28,29] Murine models have shown that in the absence of *ERG* expression *PTEN* negative tumors demonstrate diminished AR signatures compared to *PTEN* positive tumors but these signatures are restored to almost normal in the presence of *ERG* expression.[15] Similarly, Blee et al found that tumors in mice with *PTEN* deletions and *TP53* mutations but without *ERG* expression lost AR expression and were resistant to enzalutamide while the same tumors with *ERG* expression maintained AR expression and were sensitive to enzalutamide.[30] They further described the reliance of tumors with *PTEN* and p53 loss (and lacking *ERG* expression) on a separate RB/E2F1 pathway, which could be chemotherapeutically targeted with a CDK4/6 inhibitor such as palbociclib, known for use in breast cancer. It is possible that this androgen independence among *ERG* negative tumors with *PTEN* loss modulates tumor progression and contributes to subsequent metastasis, castrate resistance and PCa specific mortality.

Study limitations include patient selection since only 74 men were nonEA. Therefore, the findings may not be generalizable to AA or other minority men in whom *PTEN* loss and *ERG* rearrangements are significantly less common. Additionally, relevant clinical and pathological information were missing in this patient cohort, including the preoperative prostate specific antigen level and pathological node status. Overall the number of lethal events in our cohort was not high at 92. This raises the potential for overfitting our multivariable model and yet this study remains one of the largest data sets of surgically treated tumors with available *PTEN* and *ERG* status.

As tumors with *PTEN* loss without *ERG* rearrangement were associated with poor prognosis in this cohort, there is the possibility that these 2 subtypes also share specific adverse morphological or histological features.[30] Additionally, other molecular subtypes could be mutually exclusive with *ERG* expression and, thus, contribute to lethal outcomes

in patients with *ERG* negative tumors. Further research can be done to explore the genomic background and molecular underpinnings of the aggressive behavior of *PTEN* loss/*ERG* negative tumors.

Lastly, risk stratification tools, such as the CAPRA-Sm which incorporate clinicopathological parameters after RP, still remain valuable prognostic tools. However, molecular and genomic tests are becoming increasingly available to providers. Additional studies are required to compare *PTEN*/*ERG* IHC tests to commercially available gene panel assays in predictive models. For example, initial studies have suggested that *PTEN* loss performs similarly to the cell cycle proliferation score.[9] However, studies comparing *PTEN* to the OncotypeDx® test as well as to Decipher® are warranted since *PTEN* IHC testing is considerably less expensive than RNA based tests.

CONCLUSIONS

Using a highly validated and automated IHC assay we found that *PTEN* loss was associated with an increased risk of lethal PCa in surgically treated patients. This risk remained significant only in the subgroup of patients with *ERG* negative tumors. This work corroborates the combined use of *PTEN* and *ERG* IHC assays as prognostic tools for risk stratification and treatment management after RP.

REFERENCES

1. Jamaspishvili T, Berman DM, Ross AE et al: Clinical implications of PTEN loss in prostate cancer. Nat Rev Urol 2018; **15:** 222.
2. Lotan TL, Gurel B, Sutcliffe S et al: PTEN protein loss by immunostaining: analytic validation and prognostic indicator for a high risk surgical cohort of prostate cancer patients. Clin Cancer Res 2011; **17:** 6563.
3. Ahearn TU, Pettersson A, Ebot EM et al: A prospective investigation of PTEN loss and ERG expression in lethal prostate cancer. J Natl Cancer Inst 2016; **108.**
4. Lotan TL, Wei W, Ludkovski O et al: Analytic validation of a clinical-grade PTEN immunohistochemistry assay in prostate cancer by comparison with PTEN FISH. Mod Pathol 2016; **29:** 904.
5. Halvorsen OJ, Haukaas SA and Akslen LA: Combined loss of PTEN and p27 expression is associated with tumor cell proliferation by Ki-67 and increased risk of recurrent disease in localized prostate cancer. Clin Cancer Res 2003; **9:** 1474.
6. Bedolla R, Prihoda TJ, Kreisberg JI et al: Determining risk of biochemical recurrence in prostate cancer by immunohistochemical detection of PTEN expression and Akt activation. Clin Cancer Res 2007; **13:** 3860.
7. Chaux A, Peskoe SB, Gonzalez-Roibon N et al: Loss of PTEN expression is associated with increased risk of recurrence after prostatectomy for clinically localized prostate cancer. Mod Pathol 2012; **25:** 1543.
8. Krohn A, Diedler T, Burkhardt L et al: Genomic deletion of PTEN is associated with tumor progression and early PSA recurrence in ERG fusion-positive and fusion-negative prostate cancer. Am J Pathol 2012; **181:** 401.
9. Leapman MS, Nguyen HG, Cowan JE et al: Comparing prognostic utility of a single-marker immunohistochemistry approach with commercial gene expression profiling following radical prostatectomy. Eur Urol 2018; **74:** 668.
10. Cuzick J, Yang ZH, Fisher G et al: Prognostic value of PTEN loss in men with conservatively managed localised prostate cancer. Br J Cancer 2013; **108:** 2582.
11. Reid AH, Attard G, Ambroisine L et al: Molecular characterisation of ERG, ETV1 and PTEN gene loci identifies patients at low and high risk of death from prostate cancer. Br J Cancer 2010; **102:** 678.
12. Bismar TA, Hegazy S, Feng Z et al: Clinical utility of assessing PTEN and ERG protein expression in prostate cancer patients: a proposed method for risk stratification. J Cancer Res Clin Oncol 2018; **144:** 2117.
13. Tomlins SA, Rhodes DR, Perner S et al: Recurrence fusion of TMRPSS2 and ETS transcription factor genes in prostate cancer. Science 2005; **310:** 644.
14. Pettersson A, Graff RE, Bauer SR et al: The TMPRSS2:ERG rearrangement, ERG expression, and prostate cancer outcomes: a cohort study and meta-analysis. Cancer Epidemiol Biomarkers Prev 2012; **21:** 1497.
15. Chen Y, Chi P, Rockowitz S et al: ETS factors reprogram the androgen receptor cistrome and prime prostate tumorigenesis in response to PTEN loss. Nat Med 2013; **19:** 1023.
16. Yoshimoto M, Joshua AM, Cunha IW et al: Absence of TMPRSS2:ERG fusions and PTEN losses in prostate cancer is associated with a favorable outcome. Mod Pathol 2008; **21:** 1451.
17. Leinonen KA, Saramaki OR, Furusato B et al: Loss of PTEN is associated with aggressive behavior in ERG-positive prostate cancer. Cancer Epidemiol Biomarkers Prev 2013; **22:** 2333.
18. Mehra R, Salami SS, Lonigro R et al: Association of ERG/PTEN status with biochemical recurrence after radical prostatectomy for clinically localized prostate cancer. Med Oncol 2018; **35:** 152.
19. Gopalan A, Leversha MA, Satagopan JM et al: TMPRSS2-ERG gene fusion is not associated with outcome in patients treated with prostatectomy. Cancer Res 2009; **69:** 1400.
20. Xie W, Regan MM, Buyse M et al: Metastasis-free survival is a strong surrogate of overall survival in localized prostate cancer. J Clin Oncol 2017; **35:** 3097.
21. Chaux A, Albadine R, Toubaji A et al: Immunohistochemistry for ERG expression as a surrogate for TMPRSS2-ERG fusion detection in prostatic adenocarcinomas. Am J Surg Pathol 2011; **35:** 1014.
22. Guedes LB, Morais CL, Fedor H et al: Effect of preanalytic variables on an automated PTEN immunohistochemistry assay for prostate cancer. Arch Pathol Lab Med 2019; **143:** 338.
23. Lotan TL, Carvalho FL, Peskoe SB et al: PTEN loss is associated with upgrading of prostate cancer from biopsy to radical prostatectomy. Mod Pathol 2015; **28:** 128.
24. Lokman U, Erickson AM, Vasarainen H et al: PTEN loss but not ERG expression in diagnostic biopsies is associated with increased risk of progression and adverse surgical findings in men with prostate cancer on active surveillance. Eur Urol Focus 2018; **4:** 867.
25. Tosoian JJ, Guedes LB, Morais CL et al: PTEN status assessment in the Johns Hopkins active surveillance cohort. Prostate Cancer Prostatic Dis 2019; **22:** 176.
26. Shah RB, Shore KT, Yoon J et al: PTEN loss in prostate adenocarcinoma correlates with specific adverse histologic features (intraductal

carcinoma, cribriform Gleason pattern 4 and stromogenic carcinoma). Prostate 2019; **79:** 1267.

27. Lotan TL, Wei W, Morais CL et al: PTEN loss as determined by clinical-grade immunohistochemistry assay is associated with worse recurrence-free survival in prostate cancer. Eur Urol Focus 2016; **2:** 180.
28. Carver BS, Chapinski C, Wongvipat J et al: Reciprocal feedback regulation of PI3K and androgen receptor signaling in PTEN-deficient prostate cancer. Cancer Cell 2011; **19:** 575.
29. Mulholland DJ, Tran LM, Li Y et al: Cell autonomous role of PTEN in regulating castration-resistant prostate cancer growth. Cancer Cell 2011; **19:** 792.
30. Blee AM, He Y, Yang Y et al: TMPRSS2-ERG controls luminal epithelial lineage and antiandrogen sensitivity in PTEN and TP53-mutated prostate cancer. Clin Cancer Res 2018; **24:** 4551.

EDITORIAL COMMENT

The authors report that loss of *PTEN* expression on IHC using a CLIA certified assay correlated with death from PCa and the prediction was strongest in *ERG* negative tumors. As they note, this finding is somewhat controversial since *PTEN* loss was previously associated with biochemical recurrence in *ERG* positive tumors (reference 27 in article). It is likely that the stratification based on *ERG* status was due to arbitrary differences in other prognostic features such as age and grade between *ERG* positive and *ERG* negative tumors in the cohort, which has been observed previously.[1]

The most important feature of this study is the association of *PTEN* status with hard outcomes, namely metastasis and death from PCa. Similar findings were reported recently using an immunofluorescent based IHC assay.[2] The findings also make biological sense since *PTEN* signaling pathway alterations are common in metastatic PCa, implying that they are selected for during progression.

Developing prognostic biomarkers in PCa has proved challenging, primarily because the Gleason GG is so powerful. Given the repeated association of *PTEN* loss with adverse outcomes, mostly recurrence after surgery but also adverse pathological features and progression in patients on surveillance, this study adds substantially to data arguing that *PTEN* should be used routinely as a tissue based biomarker when assessing PCa biopsies and RPs. It finally might be time for molecular biomarkers that provide prediction of hard outcomes independent of grade and stage, like *PTEN* and *AZGP1*,[3] to be moved into clinical practice.

James D. Brooks
Department of Urology
Stanford University
Stanford, California

REFERENCES

1. Brooks JD, Wei W, Hawley S et al: Evaluation of ERG and SPINK1 by immunohistochemical staining and clinicopathological outcomes in a multi-institutional radical prostatectomy cohort of 1067 patients. PLoS One 2015; **10:** e0132343.
2. Hamid AA, Gray KP, Huang Y, et al: Loss of PTEN expression detected by fluorescence immunohistochemistry predicts lethal prostate cancer in men treated with prostatectomy. Eur Urol Oncol 2019; **2:** 475.
3. Kristensen G, Berg KD, Toft BG et al: Predictive value of AZGP1 following radical prostatectomy for prostate cancer: a cohort study and meta-analysis. J Clin Pathol 2019; **72:** 696.

Resource

Recounting the FANTOM CAGE-Associated Transcriptome

Eddie Luidy Imada,[1,2,11] Diego Fernando Sanchez,[1,11] Leonardo Collado-Torres,[3] Christopher Wilks,[4] Tejasvi Matam,[1] Wikum Dinalankara,[1] Aleksey Stupnikov,[1] Francisco Lobo-Pereira,[5] Chi-Wai Yip,[6] Kayoko Yasuzawa,[6] Naoto Kondo,[6] Masayoshi Itoh,[7] Harukazu Suzuki,[6] Takeya Kasukawa,[6] Chung-Chau Hon,[6] Michiel J.L. de Hoon,[6] Jay W. Shin,[6] Piero Carninci,[6] Andrew E. Jaffe,[3,8,9] Jeffrey T. Leek,[9] Alexander Favorov,[1,10] Gloria R. Franco,[2] Ben Langmead,[4,9,11] and Luigi Marchionni[1,11]

[1]*Department of Oncology, Johns Hopkins University School of Medicine, Baltimore, Maryland 21827, USA;* [2]*Departamento de Bioqúimica e Imunologia, ICB, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, 31270-901, Brazil;* [3]*Lieber Institute for Brain Development, Baltimore, Maryland 21205, USA;* [4]*Department of Computer Science, Johns Hopkins University, Baltimore, Maryland 21218, USA;* [5]*Departamento de Biologia General, ICB, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, 31270-901, Brazil;* [6]*RIKEN Center for Integrative Medical Sciences, Yokohama, 230-0045, Japan;* [7]*RIKEN, Preventive Medicine and Diagnostic Innovation Program, Yokohama, 351-0198, Japan;* [8]*Department of Mental Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland 21205, USA;* [9]*Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland 21205, USA;* [10]*Laboratory of Systems Biology and Computational Genetics, VIGG RAS, 117971 Moscow, Russia*

Long noncoding RNAs (lncRNAs) have emerged as key coordinators of biological and cellular processes. Characterizing lncRNA expression across cells and tissues is key to understanding their role in determining phenotypes, including human diseases. We present here FC-R2, a comprehensive expression atlas across a broadly defined human transcriptome, inclusive of over 109,000 coding and noncoding genes, as described in the FANTOM CAGE-Associated Transcriptome (FANTOM-CAT) study. This atlas greatly extends the gene annotation used in the original *recount2* resource. We demonstrate the utility of the FC-R2 atlas by reproducing key findings from published large studies and by generating new results across normal and diseased human samples. In particular, we (a) identify tissue-specific transcription profiles for distinct classes of coding and noncoding genes, (b) perform differential expression analysis across thirteen cancer types, identifying novel noncoding genes potentially involved in tumor pathogenesis and progression, and (c) confirm the prognostic value for several enhancer lncRNAs expression in cancer. Our resource is instrumental for the systematic molecular characterization of lncRNA by the FANTOM6 Consortium. In conclusion, comprised of over 70,000 samples, the FC-R2 atlas will empower other researchers to investigate functions and biological roles of both known coding genes and novel lncRNAs.

[Supplemental material is available for this article.]

Long noncoding RNAs (lncRNAs) are commonly defined as transcripts longer than 200 nucleotides that are not translated into proteins. This definition is not based on their function, since lncRNAs are involved in distinct molecular processes and biological contexts not yet fully characterized (Batista and Chang 2013). Over the past few years, the importance of lncRNAs has clearly emerged, leading to an increasing focus on decoding the consequences of their modulation and studying their involvement in the regulation of key biological mechanisms during development, normal tissue and cellular homeostasis, and in disease (Esteller 2011; Batista and Chang 2013; Ling et al. 2015).

Given the emerging and previously underestimated importance of noncoding RNAs (ncRNAs), the FANTOM Consortium has initiated the systematic characterization of their biological function. Through the use of Cap Analysis of Gene Expression sequencing (CAGE-seq), combined with RNA-seq data from the public domain, the FANTOM Consortium released a comprehensive atlas of the human transcriptome, encompassing more accurate transcriptional start sites (TSSs) for coding and noncoding genes, including numerous novel long noncoding genes: the FANTOM CAGE-Associated Transcriptome (FANTOM-CAT) (Hon et al. 2017). We hypothesized that these lncRNAs can be measured in many RNA-seq data sets from the public domain and that they have been so far missed by the lack of a comprehensive gene annotation.

Although the systematic analysis of lncRNAs function is being addressed by the FANTOM Consortium in loss-of-function studies, increasing the detection rate of these transcripts

[11]**These authors contributed equally to this work.**
Corresponding author: marchion@jhu.edu
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.254656.119. Freely available online through the *Genome Research* Open Access option.

combining different studies is difficult because of the heterogeneity of analytic methods employed. Current resources that apply uniform analytic methods to create expression summaries from public data do exist but can miss several lncRNAs because of their dependency on a preexisting gene annotation for creating the gene expression summaries (Tatlow and Piccolo 2016; Lachmann et al. 2018). We recently created *recount2* (Collado-Torres et al. 2017b), a collection of uniformly processed human RNA-seq data, wherein we summarized 4.4 trillion reads from over 70,000 human samples from the NCBI Sequence Read Archive (SRA), The Cancer Genome Atlas (TCGA) (The Cancer Genome Atlas Research Network et al. 2013), and the Genotype-Tissue Expression (GTEx) (The GTEx Consortium 2013) projects (Collado-Torres et al. 2017b). Importantly, *recount2* provides annotation-agnostic coverage files that allow requantification using a new annotation without having to reprocess the RNA-seq data.

Given the unique opportunity to access the latest results to the most comprehensive human transcriptome (the *FANTOM-CAT* project) and the *recount2* gene agnostic summaries, we addressed the previously described challenges, building a comprehensive atlas of coding and noncoding gene expression across the human genome: the FANTOM-CAT/*recount2* expression atlas (FC-R2 hereafter). Our resource contains expression profiles for 109,873 putative genes across over 70,000 samples, enabling an unparalleled resource for the analysis of the human coding and noncoding transcriptome.

Results

Building the FANTOM-CAT / *recount2* resource

The *recount2* resource includes a coverage track, in the form of a bigWig file, for each processed sample. We built the FC-R2 expression atlas by extracting expression levels from *recount2* coverage tracks in regions that overlapped unambiguous exon coordinates for the permissive set of FANTOM-CAT transcripts, according to the pipeline shown in Figure 1. Since *recount2*'s coverage tracks do not distinguish between genomic strands, we removed ambiguous segments that presented overlapping exon annotations from both strands (see Methods section and Supplemental Methods). After this disambiguation procedure, the remaining 1,066,515 exonic segments mapped back to 109,869 genes in FANTOM-CAT (out of the 124,047 starting ones included in the permissive set [Hon et al. 2017]). Overall, the FC-R2 expression atlas encompasses 2041 studies with 71,045 RNA-seq samples, providing expression information for 22,116 coding genes and 87,763 noncoding genes, such as enhancers, promoters, and other lncRNAs.

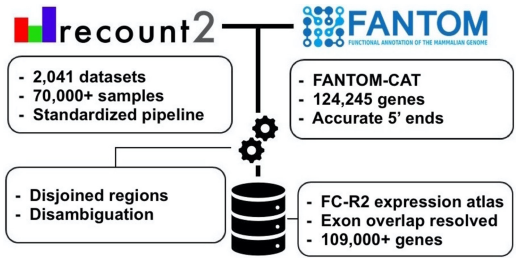


Figure 1. Overview of the FANTOM-CAT/*recount2* resource development. FC-R2 leverages two public resources, the FANTOM-CAT gene models and *recount2*. FC-R2 provides expression information for 109,873 genes, both coding (22,110) and noncoding (87,693). This latter group encompasses enhancers, promoters, and other lncRNAs.

Validating the *FANTOM-CAT / recount2* resource

We first assessed how gene expression estimates in FC-R2 compared to previous gene expression estimates from other projects. Specifically, we considered data from the GTEx Consortium (v6), spanning 9662 samples from 551 individuals and 54 tissues types (The GTEx Consortium 2013). First, we computed the correlation for the GTEx data between gene expression based on the FC-R2 atlas and on the GENCODE (v25) gene model in *recount2*, which has been already shown to be consistent with gene expression estimates from the GTEx project (Collado-Torres et al. 2017b), observing a median correlation ≥0.986 for the 32,922 genes in common. This result supports the notion that our preprocessing steps to disambiguate overlapping exon regions between strands did not significantly alter gene expression quantification.

Next, we assessed whether gene expression specificity, as measured in FC-R2, was maintained across tissue types. To this end, we selected and compared gene expression for known tissue-specific expression patterns, such as keratin 1 (*KRT1*), estrogen receptor 1 (*ESR1*), and neuronal differentiation 1 (*NEUROD1*) (Fig. 2). Overall, all analyzed tissue-specific markers presented nearly identical expression profiles across GTEx tissue types between the alternative gene models considered (see Fig. 2 and Supplemental Fig. S1), confirming the consistency between gene expression quantification in FC-R2 and those based on GENCODE.

We also assessed whether there are genes that are not expressed in any of the normal tissues included in GTEx. Out of 109,869 genes, 681 (0.6%) (see Supplemental Figs. S3, S4) were not expressed in any tissue included in GTEx, and they were overrepresented in the FANTOM-CAT permissive set (χ^2 test, P-value < 2.2×10^{-16}).

Tissue-specific expression of lncRNAs

It has been shown that, although expressed at a lower level, enhancers and promoters are not ubiquitously expressed and are more specific for different cell types than coding genes (Hon et al. 2017). In order to verify this finding, we used GTEx data to assess expression levels and specificity profiles across samples from each of the 54 analyzed tissue types, stratified into four distinct gene categories: coding mRNA, intergenic promoter lncRNA (ip-lncRNA), divergent promoter lncRNA (dp-lncRNA), and enhancer lncRNA (e-lncRNA). Overall, we were able to confirm that these RNA classes are expressed at different levels and that they display distinct specificity patterns across tissues, as shown for primary cell types by Hon et al. (2017), albeit with more variability, likely due to the increased cellular complexity present in tissues. Specifically, coding mRNAs were expressed at higher levels than lncRNAs ($\log_2$ median expression of 6.6 for coding mRNAs, and of 4.1, 3.8, and 3.1 for ip-lncRNA, dp-lncRNA, and e-lncRNA, respectively). In contrast, the expression of enhancers and intergenic promoters was more tissue-specific (median = 0.41 and 0.30, respectively) than that observed for divergent promoters and coding mRNAs (median = 0.13 and 0.09, respectively) (Fig. 3A). Finally, when analyzing the percentage of genes expressed across tissues by category, we observed that coding genes are, in general, more ubiquitous, whereas lncRNAs are more specific, with enhancers showing the lowest percentages of expressed genes (mean ranging from 88.42% to 41.98%) (see Fig. 3B), in agreement

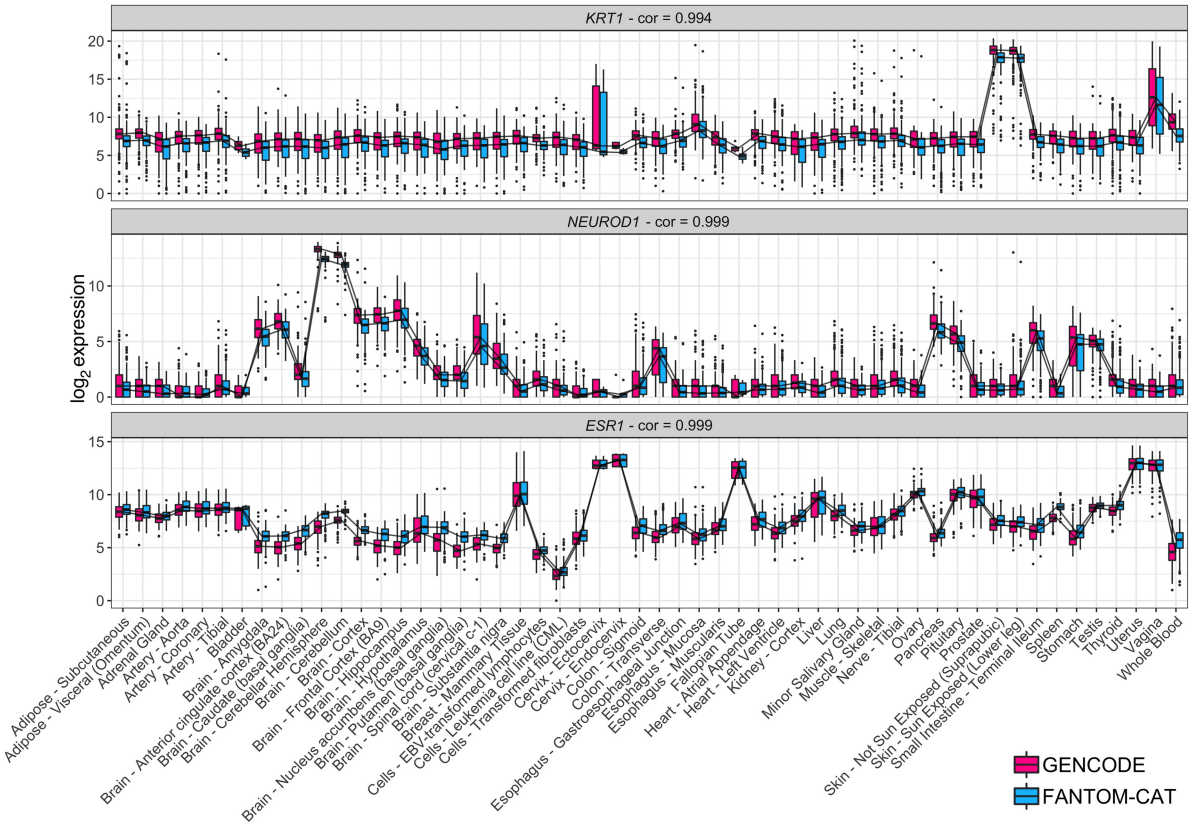


Figure 2. Tissue-specific expression in GTEx. Log_2 expression for three tissue-specific genes (*KRT1*, *NEUROD1*, and *ESR1*) in GTEx data stratified by tissue type using FC-R2- and GENCODE-based quantification. Expression profiles are highly correlated and expressed consistently in the expected tissue types (e.g., *KRT1* is most expressed in skin, *NEUROD1* in brain, and *ESR1* in estrogen-sensitive tissue types like uterus, Fallopian tubes, and breast). Correlations are shown on *top* for each tissue marker. *Center* lines, *upper*/*lower* quartiles, and whiskers represent the median, 25/75 percentiles, and 1.5 interquartile range, respectively. Additional tissue-specific markers are shown in Supplemental Figure S1.

with the notion that enhancer transcription is tissue-specific (Ong and Corces 2011).

Differential expression analysis of coding and noncoding genes in cancer

We analyzed coding and noncoding gene expression in cancer using TCGA data. To this end, we compared cancer to normal samples separately for 13 tumor types, using FC-R2 requantified data. We further identified the differentially expressed genes (DEGs) in common across the distinct cancer types (see Fig. 4). Overall, the number of DEGs varied across cancer types and by gene class, with a higher number of significant coding than noncoding genes (FDR $\leq$ 0.01) (see Table 1). A substantial fraction of these genes was exclusively annotated in the FANTOM-CAT meta-assembly, suggesting that relying on other gene models would result in missing many potential important genes (see Table 1). We then analyzed differential gene expression consensus across the considered cancer types. A total of 41 coding mRNAs were differentially expressed across all of the 13 tumor types after global correction for multiple testing (FDR $\leq 10^{-6}$) (see Supplemental Table S1). For lncRNAs, a total of 28 divergent promoters, four intergenic promoters, and three enhancers were consistently up- or down-regulated across all the 13 tumor types after global correction for multiple testing (FDR $\leq$ 0.1) (see Supplemental Tables S2–S4, respectively).

A usual task performed after differential gene expression analysis is to identify biological processes and pathways associated with the DEGs. To this end, gene set enrichment methods are usually employed; however, this requires detailed gene-to-function annotations, which are mostly lacking for lncRNAs. One possible way to assist prioritizing noncoding transcripts for follow-up functional studies is to identify association with other features along the genome. As an example of this type of analysis, we have assessed the overlap between single-nucleotide polymorphisms (SNPs) associated with cancer in GWAS studies and the list of DEGs we identified. On average, the percentage of DEGs overlapping cancer SNPs ranged from 6.6% in dp-lncRNA to 10.21% in ip-lncRNA across the 13 cancer types (see Supplemental Table S5).

Next, we reviewed the literature to identify functional correlates for these consensus genes. Most of the up-regulated coding genes (Supplemental Table S1) participate in cell cycle regulation, cell division, DNA replication and repair, chromosome segregation, and mitotic spindle checkpoints. Most of the consensus down-regulated mRNAs (Supplemental Table S1) are associated with metabolism and oxidative stress, transcriptional regulation, cell migration and adhesion, and with modulation of DNA damage repair and apoptosis.

Three down-regulated dp-lncRNA genes, *GAS1RR*, *RPL34-DT*, and *RAP2C-AS1*, were reported to be implicated in cancer (Supplemental Table S2). The first one controls epithelial-

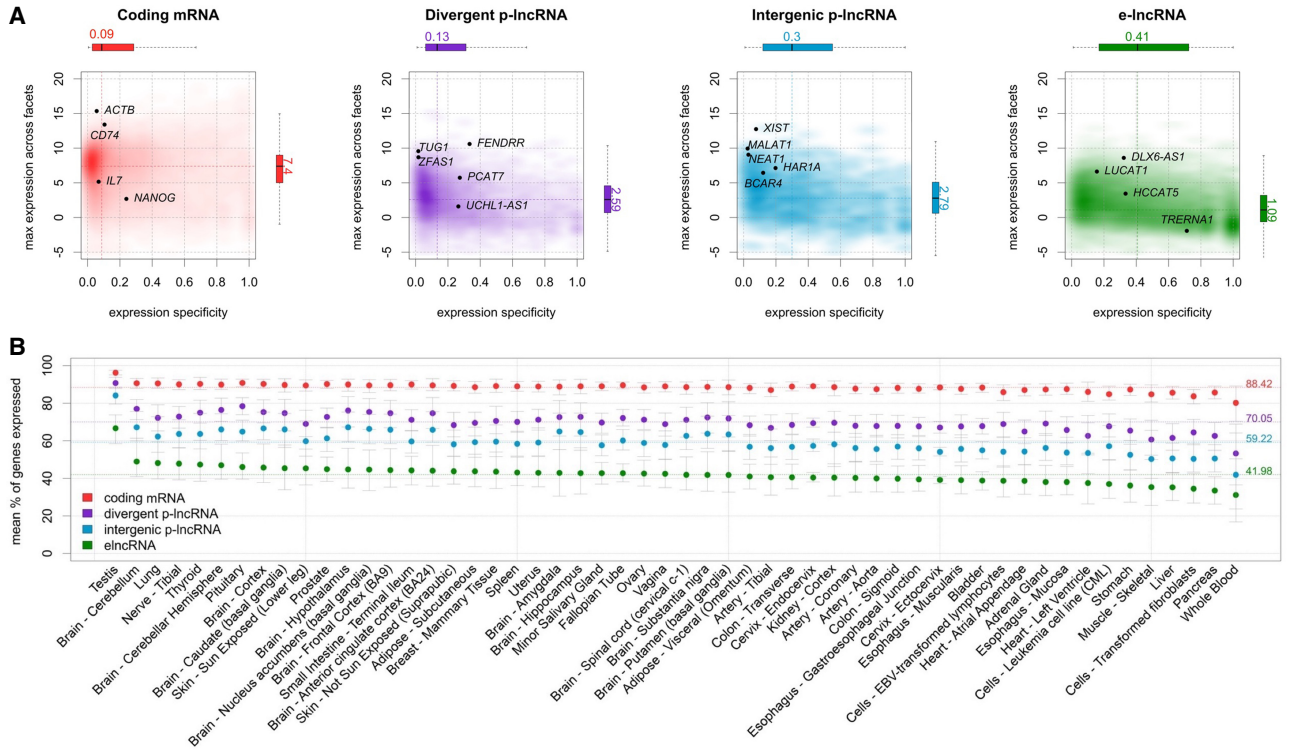


Figure 3. Expression profiles across GTEx tissues. (*A*) Expression level and tissue specificity across four distinct RNA categories. The *y*-axis shows $\log_2$ expression levels representing each gene using its maximum expression in GTEx tissues expressed as transcripts per million (TPM). The *x*-axis shows expression specificity based on entropy computed from median expression of each gene across the GTEx tissue types. Individual genes are highlighted in the figure panels. (*B*) Percentage of genes expressed for each RNA category stratified by GTEx tissue facets. The dots represent the mean among samples within a facet and the error bars represent 99.99% confidence intervals. Dashed lines represent the means among all samples.

mesenchymal transition, the second is associated with tumor size increase, whereas the third is associated with urothelial cancer after kidney cancer transplantation (Zhao et al. 2015b; Shang et al. 2016; Zhou et al. 2016). Among the up-regulated dp-lncRNAs (Supplemental Table S2), *SNHG1* has been implicated in cellular proliferation and migration and invasion of different cancer types, and to be strongly up-regulated in osteosarcoma, nonsmall lung cancer, and gastric cancer (Cao et al. 2013; Sun et al. 2017).

Among the ip-lncRNAs ubiquitously down-regulated (see Supplemental Table S3), *MIR99AHG* has been identified in many different tumor types, including leukemia, breast, vulvar, prostate, and bladder cancer (Emmrich et al. 2014; Sun et al. 2014; Gökmen-Polar et al. 2016; Ni et al. 2016; Li et al. 2017). For instance, in vulvar squamous cell carcinoma, *MIR99AHG* and *MIR31HG* expressions are correlated and associated with tumor differentiation (Ni et al. 2016). Similarly, *MIR99AHG* down-regulation in ER-positive breast cancer is associated with progression, recurrence, and metastasis (Gökmen-Polar et al. 2016). In contrast, increased expression of *SNHG17* (an ip-lncRNA) (see Supplemental Table S3) was associated with short term survival in breast cancer and with tumor size, stage, and lymph node metastasis in colorectal cancer (Zhao et al. 2015a; Ma et al. 2017). In addition, *LINC01311*, another ip-lncRNA (Supplemental Table S3), was found to be up-regulated in liver cancer and metastatic prostate cancer (Zhu et al. 2016). Even though we did not identify any cancer association for common e-lncRNAs, one among those we identified, *LINC02884*, has been previously reported to be up-regulated in late-onset Alzheimer's disease (Humphries et al. 2015). Furthermore, the enhancer lncRNA class also yielded the lowest number of genes in common among all cancer types, reinforcing the concept that enhancers are expressed in a tissue-specific manner (see Fig. 3A and Supplemental Table S4).

Finally, we focused more in depth on prostate cancer (PCa) as a prototypical example, and we were able to confirm previous findings for both coding and noncoding genes (see Supplemental Fig. S2). For coding genes, we confirmed differential expression for known markers of PCa progression and mortality, like *ERG*, *FOXA1*, *RNASEL*, *ARVCF*, and *SLC43A1* (Yu et al. 2010; Lin et al. 2011). Similarly, we also confirmed differential expression for noncoding genes, like *PCA3*, the first clinically approved lncRNA marker for PCa (Bussemakers et al. 1999; de Kok et al. 2002), *PCAT1*, a prostate-specific lncRNA involved in disease progression (Prensner et al. 2011), *MALAT1*, which is associated with PCa poor prognosis (Ren et al. 2013), *CDKN2B-AS1*, an antisense lncRNA up-regulated in PCa that inhibits tumor suppressor genes activity (Kotake et al. 2011; Gutschner and Diederichs 2012), and the *MIR135* host gene, which is associated with castration-resistant PCa (Huang et al. 2015).

Confirming prognostic enhancers

Chen and collaborators have recently surveyed enhancer expression in nearly 9000 patients from TCGA (Chen et al. 2018), using genomic coordinates from the FANTOM5 project (Andersson et al. 2014), identifying 4803 expressed genomic regions with prognostic potential in one or more TCGA tumor types. We therefore

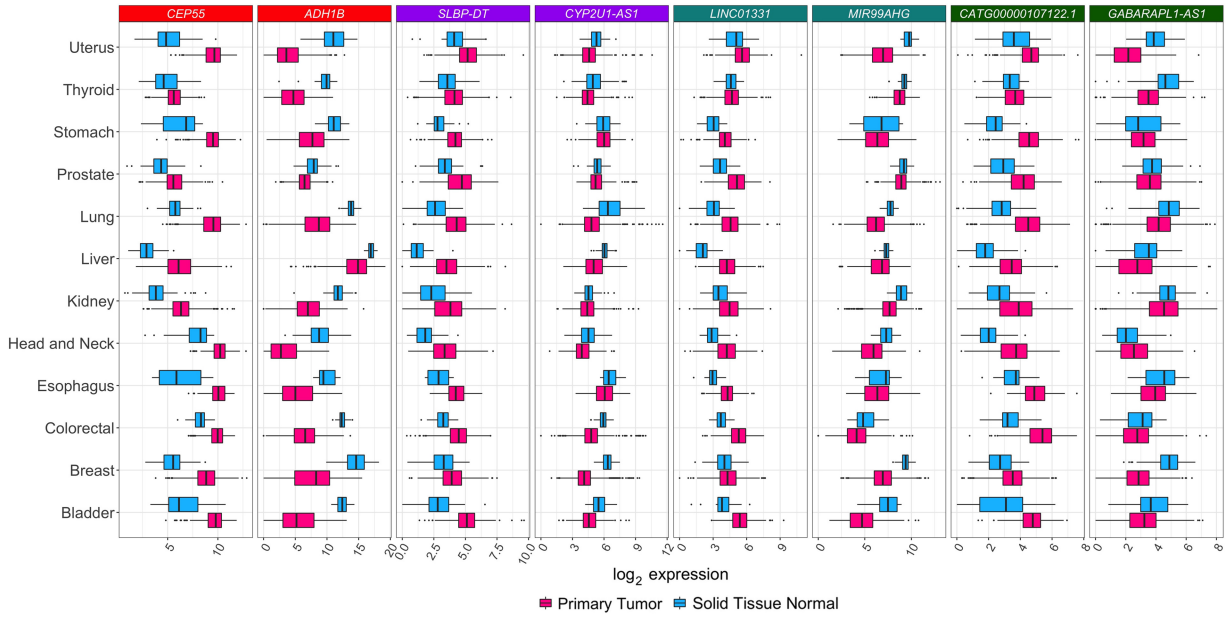


Figure 4. Differential expression for selected transcripts from distinct RNA classes across tumor types. Box plots for selected differentially expressed genes between tumor and normal samples across all 13 tumor types analyzed. For each tissue of origin, the most up-regulated (on the *left*) and down-regulated (on the *right*) gene for each RNA class is shown. *Center* lines, *upper*/*lower* hinges, and the whiskers, respectively, represent the median, the upper and lower quartiles, and 1.5 extensions of the interquartile range. Color coding on the *top* of the figure indicates the RNA classes (red for mRNA, purple for dp-lncRNA, cyan ip-lncRNA, and green for e-lncRNA). These genes were selected after global multiple testing correction across all 13 tumor types (see Supplemental Tables S1–S4).

leveraged the FC-R2 atlas to identify prognostic coding and non-coding genes using both univariate and multivariate Cox proportional hazard models, comparing our results for e-lncRNAs with those reported by Chen and colleagues. To this end, we started by comparing gene annotations and genomic overlap between the studies. This was necessary because Chen and collaborators relied on the enhancer regions reported by Andersson et al. (2014), which is based on the observation of bidirectional transcription. Our resource, on the contrary, relies on the latest updated FANTOM-CAT annotation, which takes into account other features, such as the epigenetic context, when defining RNA categories. Out of the 4803 genomic regions found prognostic by Chen and collaborators (Chen et al. 2018), we could unambiguously map 1218 regions to exons annotated in the FANTOM-CAT gene models for the four RNA categories we considered in our study (corresponding to a total of 1046 unique genes). Overall, despite the mentioned differences in annotation and quantification (see Supplemental Table S6), we were still able to confirm the prognostic value for 466 genes out of the 1046 reported by Chen et al. (2018), including *KLHDC7B-DT* (also known as enhancer 22),

Table 1. Differentially expressed genes in cancer

		dp-lncRNA		e-lncRNA		ip-lncRNA		mRNA	
Cancer type	Total	Up	Down	Up	Down	Up	Down	Up	Down
Bile	7010	200 (60)	313 (90)	186 (89)	203 (99)	47 (12)	84 (17)	2658 (106)	3319 (97)
Bladder	7680	344 (125)	319 (87)	140 (68)	149 (67)	65 (19)	82 (7)	3112 (201)	3469 (61)
Breast	15,290	753 (291)	721 (202)	656 (377)	583 (305)	207 (50)	178 (32)	6109 (296)	6083 (244)
Colorectal	13,685	490 (164)	592 (168)	381 (203)	400 (196)	130 (32)	160 (28)	5538 (371)	5994 (132)
Esophagus	4883	87 (21)	193 (50)	90 (38)	184 (103)	40 (11)	48 (2)	1921 (83)	2320 (77)
Head and neck	10,517	442 (138)	401 (96)	267 (139)	251 (112)	100 (23)	109 (18)	4329 (256)	4618 (53)
Kidney	15,697	734 (238)	820 (281)	535 (299)	486 (209)	203 (45)	200 (48)	6349 (525)	6370 (114)
Liver	10,554	346 (94)	395 (106)	230 (102)	248 (123)	90 (16)	112 (19)	4164 (174)	4969 (95)
Lung	17,143	864 (338)	835 (304)	893 (512)	729 (396)	242 (76)	213 (39)	7523 (532)	5844 (212)
Prostate	13,183	686 (287)	654 (218)	418 (254)	452 (214)	175 (55)	167 (30)	5153 (489)	5478 (128)
Stomach	11,309	528 (213)	518 (164)	462 (291)	436 (240)	144 (51)	129 (22)	4509 (558)	4583 (89)
Thyroid	14,264	752 (284)	804 (318)	527 (295)	594 (332)	161 (39)	174 (47)	5403 (189)	5849 (308)
Uterus	12,906	641 (285)	713 (235)	454 (263)	612 (341)	210 (79)	225 (54)	5135 (335)	4916 (181)
Mean	11,855	528 (195)	560 (178)	403 (225)	410 (211)	140 (39)	145 (28)	4762 (317)	4909 (138)
SD	3650	237 (102)	218 (89)	225 (137)	189 (107)	67 (23)	55 (16)	1557 (167)	1234 (77)

Table summarizes the number of significant DEGs (*FDR* < 0.01) between tumor and normal samples across the 13 cancer types, analyzed for each gene class considered. Counts are for DEGs up- and down-regulated in cancer; values in parentheses are the number of genes exclusively annotated in the FANTOM-CAT gene model. Mean and standard deviation across cancer types are shown at the bottom.

which was highlighted as a promising prognostic marker for kidney cancer (Supplemental Fig. S5).

We then considered the FANTOM-CAT RNA classes across the different tumor types. We were able to identify a variable number of genes significantly associated with overall survival (FDR ≤ 0.05) in univariate Cox proportional hazards models (see Supplemental Tables S7–S10). Among the consensus DEGs identified across all tumor types, 40 out of 41 coding mRNAs, 25 out of 28 dp-lncRNAs, four out of four ip-lncRNAs, and two out of three e-lncRNAs were found to be associated with survival (see Supplemental Tables S11–S14). Kaplan–Meier curves for selected differentially expressed genes for each RNA category are shown in Supplemental Figure S6. Finally, we performed multivariable analysis controlling for relevant clinical and pathological characteristics in each tumor type. Overall, despite a number of genes being associated with such variables, we obtained similar results (see Supplemental Tables S15–S22).

Discussion

The importance of lncRNAs in cell biology and disease has clearly emerged in the past few years, and different classes of lncRNAs have been shown to play crucial roles in cell regulation and homeostasis (Quinn and Chang 2016). For instance, enhancers—a major category of gene regulatory elements, which has been shown to be expressed (Andersson et al. 2014; Arner et al. 2015)—play a prominent role in oncogenic processes (Herz et al. 2014; Sur and Taipale 2016) and other human diseases (Hnisz et al. 2013). Despite their importance, however, there is a scarcity of large-scale data sets investigating enhancers and other lncRNA categories, in part due to the technical difficulty in applying high-throughput techniques such as ChIP-seq and Hi-C over large cohorts, and to the use of gene models that do not account for them in transcriptomics analyses. Furthermore, the large majority of the lncRNAs that are already known—and that have been shown to be associated with some phenotype—are still lacking functional annotation.

To address these needs, the FANTOM Consortium has first constructed the FANTOM-CAT metatranscriptome, a comprehensive atlas of coding and noncoding genes with robust support from CAGE-seq data (Hon et al. 2017); then, it has undertaken a large scale project to systematically target lncRNAs and characterize their function using a multipronged approach (Ramilowski et al. 2020). In a complementary effort, we have leveraged public domain gene expression data from *recount2* (Collado-Torres et al. 2017a,b) to create a comprehensive gene expression compendium across human cells and tissues based on the FANTOM-CAT gene model, with the ultimate goal of facilitating lncRNAs annotation through association studies. To this end, the FC-R2 atlas is already in use in the FANTOM6 project (https://fantom.gsc.riken.jp/6/) to successfully characterize lncRNA expression in human samples (Ramilowski et al. 2020).

In order to validate our resource, we have compared the gene expression summaries based on FANTOM-CAT gene models with previous, well-established gene expression quantifications, demonstrating virtually identical profiles across tissue types overall and for specific tissue markers. We have then confirmed that distinct classes of coding and noncoding genes differ in terms of overall expression level and specificity pattern across cell types and tissues. We also have observed a small subset of genes that were not expressed in the large majority of the samples analyzed in the GTEx project. These genes were mostly classified as small RNAs and enhancers, which was expected given that the RNA-seq libraries included in *recount2* did not target small RNAs, and enhancers are usually expressed at a lower level. We further reveal that this subset of genes not expressed in any normal tissue is also associated with a lower level of support of the corresponding FANTOM-CAT gene models (Hon et al. 2017).

Furthermore, using the FC-R2 atlas, we were also able to identify mRNAs, promoters, enhancers, and other lncRNAs that are differentially expressed in cancer, both confirming previously reported findings and identifying novel cancer genes exclusively annotated in the FANTOM-CAT gene models, which have been therefore missed in prior analyses with TCGA data. Finally, we confirmed the prognostic value for some of the enhancer regions recently reported by Chen and colleagues in the TCGA (Chen et al. 2018) by performing a systematic screening for survival association of both coding and noncoding genes that are quantifiable in the FC-R2 resource. Overall, we identified several genes with potential prognostic value across the analyzed cancer types in TCGA; however, further corroboratory studies in independent patient cohorts are necessary to validate these associations.

Collectively, by confirming findings reported in previous studies, our results demonstrate that the FC-R2 gene expression atlas is a reliable and powerful resource for exploring both the coding and noncoding transcriptome, providing compelling evidence and robust support to the notion that lncRNA gene classes, including enhancers and promoters, despite not being yet fully understood, portend significant biological functions. Our resource, therefore, constitutes a suitable and promising platform for future large scale studies in cancer and other human diseases, which in turn hold the potential to reveal important cues to the understanding of their biological, physiological, and pathological roles, potentially leading to improved diagnostic and therapeutic interventions.

Finally, all results, data, and code from the FC-R2 atlas are available as a public tool. With uniformly processed expression data for over 70,000 samples and 109,873 genes ready to analyze, we want to encourage researchers to dive deeper into the study of ncRNAs, their interaction with coding and noncoding genes, and their influence on normal and disease tissues. We hope this new resource will help pave the way to develop new hypotheses that can be followed to unwind the biological role of the transcriptome as a whole.

Methods

Data and preprocessing

The complete FANTOM-CAT gene catalog (inclusive of robust, intermediate, and permissive sets) was obtained from the FANTOM Consortium within the frame of the FANTOM6 project (Ramilowski et al. 2020). The genes were annotated using official HUGO Gene Nomenclature Committee (HGNC) symbols (https://www.genenames.org) when available. For genes without HGNC symbols, we named them according to HGNC instructions (see Supplemental Table S23). The remaining genes were referred to using the official ID from the Consortium that annotated the gene (Ensembl/FANTOM). This catalog accounts for 124,245 genes supported by CAGE peaks, and it includes those described by Hon et al. (2017). In order to remove ambiguity due to overlapping among exons from distinct genes, the BED files containing the coordinates for all genes and exons were processed with the *GenomicRanges* R/Bioconductor package (Lawrence et al. 2013) to obtain disjoint (nonoverlapping) exon coordinates. To avoid losing strand information from annotation, we processed data using a two-step

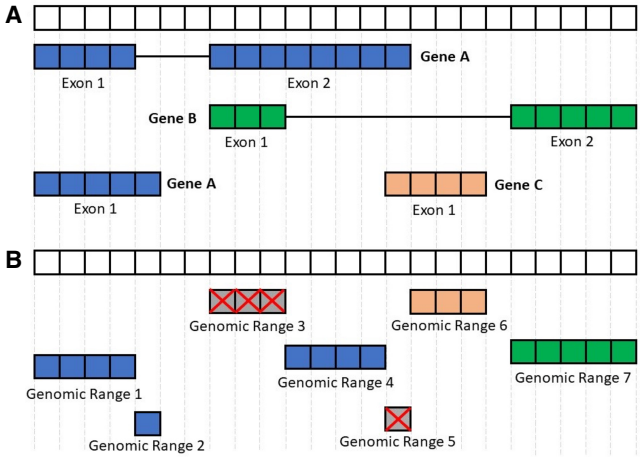


Figure 5. Processing the FANTOM-CAT genomic ranges. This figure summarizes the disjoining and exon disambiguation processes performed before extracting expression information from *recount2* using the FANTOM-CAT gene models. (*A*) Representation of a genomic segment containing three distinct, hypothetical genes: gene A having two isoforms, and genes B and C with one isoform each. Each box can be interpreted as one nucleotide along the genome. Colors indicate the three different genes. (*B*) Representation of disjoint exon ranges from example in panel *A*. Each feature is reduced to a set of nonoverlapping genomic ranges. The disjoint genomic ranges mapping back to two or more distinct genes are removed (crossed gray boxes). After removal of ambiguous ranges, the expression information for the remaining ones is extracted from *recount2* and summarized at the gene level.

approach by first disjoining overlapping segments on the same strand and then across strands (Fig. 5). The genomic ranges (disjoint exon segments) that mapped back to more than one gene were discarded. The expression values for these ranges were then quantified using *recount.bwtool* (Ellis et al. 2018) (code at https://github.com/LieberInstitute/marchionni_projects). The resulting expression quantifications were processed to generate RangedSummarizedExperiment objects compatible with the *recount2* framework (Collado-Torres et al. 2017a,b) (code available from https://github.com/eddieimada/fcr2). Thus, the FC-R2 atlas provides expression information for coding and noncoding genes (including enhancers, divergent promoters, and intergenic lncRNAs) for 9662 samples from the GTEx project, 11,350 samples from TCGA, and over 50,000 samples from the SRA.

Correlation with other studies

To test if the preprocessing steps used for FC-R2 had a major impact on gene expression quantification, we compared our data to the published GTEx expression values obtained from *recount2* (version 2, https://jhubiostatistics.shinyapps.io/recount/). Specifically, we first compared the expression distribution of tissue-specific genes across different tissue types and then computed the Pearson's correlation for each gene in common across the original *recount2* gene expression estimates based on GENCODE and our version based on the FANTOM-CAT transcriptome.

Expression specificity of tissue facets

We analyzed the expression level and specificity of each gene stratified by RNA category (i.e., mRNA, e-lncRNA, dp-lncRNA, ip-lncRNA) using the same approach described by Hon et al. (2017) (see Supplemental Methods). Briefly, overall expression levels for each gene were represented by the maximum transcript per million (TPM) values observed across all samples within each tissue type in GTEx. Gene specificity was based on the empirical entropy computed using the mean expression value across tissue types. The 99.99% confidence intervals for the expression of each category by tissue type were calculated based on TPM values. Genes with a TPM greater than 0.01 were considered to be expressed.

Identification of differentially expressed genes

We analyzed differential gene expression in 13 cancer types, comparing primary tumor with normal samples using TCGA data from the FC-R2 atlas. Gene expression summaries for each cancer type were split by RNA category (coding mRNA, intergenic promoter lncRNA, divergent promoter lncRNA, and enhancer lncRNA) and then analyzed independently. A generalized linear model approach, coupled with empirical Bayes moderation of standard errors (Smyth 2004), was used to identify differentially expressed genes between groups. The model was adjusted for the three most relevant coefficients for data heterogeneity as estimated by surrogate variable analysis (SVA) (Leek and Storey 2007). Correction for multiple testing was performed across RNA category by merging the resulting *P*-values for each cancer type and applying the Benjamini–Hochberg method (Benjamini and Hochberg 1995). Overlapping between DEG and GWAS SNPs was performed using the FANTOM-CAT gene regions coordinates and the SNPs positions obtained from the GWAS catalog (Buniello et al. 2019).

Prognostic analysis

To evaluate the prognostic potential of the genes in FC-R2, we performed both multivariate and univariate Cox proportional hazards regression analysis separately for each RNA class (22,106 mRNAs, 17,404 e-lncRNAs, 6204 dp-lncRNAs, and 1948 ip-lncRNAs) across each of the 13 TCGA cancer types with available survival follow-up information (see Supplemental Methods; Supplemental Table S24). Genes with FDR ≤ 0.05, using the Benjamini–Hochberg correction (Benjamini and Hochberg 1995) within each cancer type and RNA class, were deemed significant prognostic factors. We further analyzed the prognostic value of the consensus differentially expressed genes we identified comparing tumors to normal samples by intersecting the corresponding gene lists with those obtained by Cox proportional regression. Finally, in order to compare our results to previous prognostic analyses, we obtained data on enhancers position and prognostic potential from Chen et al. (2018), performed a liftOver to the hg38 genome assembly to match FC-R2 coordinates, and assessed the overlap between prognostic genes identified in the two studies.

Data access

All data are available from http://marchionnilab.org/fcr2.html. Expression data can be directly accessed through https://jhubiostatistics.shinyapps.io/recount/ and the recount Bioconductor package (v1.9.5 or newer) at https://bioconductor.org/packages/recount as *RangedSummarizedExperiment* objects organized by the Sequence Read Archive (SRA) study ID. The data can be loaded using R-programming language and are ready to be analyzed using Bioconductor packages, or the data can be exported to other formats for use in another environment. All code used in this manuscript is available for reproducibility and transparency at GitHub (https://github.com/eddieimada/fcr2 and https://github.com/LieberInstitute/marchionni_projects). A compressed archive with all scripts used is also available as Supplemental Code.

Competing interest statement

The authors declare no competing interests.

Acknowledgments

This publication was made possible through support from the National Institutes of Health–National Cancer Institute (NIH-NCI) grants P30CA006973 (L.M. and A.F.), 1U01CA231776 (W.D. and L.M.), U01CA196390 (L.M. and A.S.), and R01CA200859 (W.D. and L.M.); and the National Institutes of Health–National Institute of General Medical Sciences (NIH-NIGMS) grants R01GM118568 (C.W. and B.L.) and R21MH109956-01 (L.C.-T. and A.E.J.); and the U.S. Department of Defense (DoD) office of the Congressionally Directed Medical Research Programs (CDMRP) award W81XWH-16-1-0739 (E.L.I. and L.M.); Russian Academic project 0112-2019-0001 and Russian Foundation for Basic Research project 17-00-00208 (A.F.); and Fundação de Amparo a Pesquisa do Estado de Minas Gerais award BDS-00493-16 (E.L.I. and G.R.F.). *recount2* and FC-R2 are hosted on SciServer, a collaborative research environment for large-scale data-driven science. It is being developed at, and administered by, the Institute for Data Intensive Engineering and Science (IDIES) at Johns Hopkins University. SciServer is funded by the National Science Foundation Award ACI-1261715. For more information about SciServer, visit http://www.sciserver.org/.

Author contributions: L.M. conceived the idea; L.M., E.L.I., A.F., and B.L. designed the study; E.L.I., D.F.S., T.M., W.D., A.S., L.C.-T., and L.M. performed the analysis; E.L.I., D.F.S., F.L.-P., G.R.F., and L.M. interpreted the results; L.C.-T., C.W., C-W.Y., K.Y, N.K., M.I., H.S., T.K., C.-C.H., M.J.L.deH., J.W.S., P.C., A.E.J., J.T.L., and B.L. provided the data and tools; E.L.I., D.F.S., L.C.-T., B.L., and L.M. wrote the manuscript; all authors reviewed and approved the manuscript.

References

Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, Chen Y, Zhao X, Schmidl C, Suzuki T, et al. 2014. An atlas of active enhancers across human cell types and tissues. *Nature* **507:** 455–461. doi:10.1038/nature12787

Arner E, Daub CO, Vitting-Seerup K, Andersson R, Lilje B, Drabløs F, Lennartsson A, Rönnerblad M, Hrydziuszko O, Vitezic M, et al. 2015. Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells. *Science* **347:** 1010–1014. doi:10.1126/science.1259418

Batista PJ, Chang HY. 2013. Long noncoding RNAs: cellular address codes in development and disease. *Cell* **152:** 1298–1307. doi:10.1016/j.cell.2013.02.012

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Series B* **57:** 289–300. doi:10.1111/j.2517-6161.1995.tb02031.x

Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, McMahon A, Morales J, Mountjoy E, Sollis E, et al. 2019. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res* **47:** D1005–D1012. doi:10.1093/nar/gky1120

Bussemakers MJ, van Bokhoven A, Verhaegh GW, Smit FP, Karthaus HF, Schalken JA, Debruyne FM, Ru N, Isaacs WB. 1999. *DD3*: a new prostate-specific gene, highly overexpressed in prostate cancer. *Cancer Res* **59:** 5975–5979.

The Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM. 2013. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet* **45:** 1113–1120. doi:10.1038/ng.2764

Cao WJ, Wu HL, He BS, Zhang YS, Zhang ZY. 2013. Analysis of long non-coding RNA expression profiles in gastric cancer. *World J Gastroentero* **19:** 3658–3664. doi:10.3748/wjg.v19.i23.3658

Chen H, Li C, Peng X, Zhou Z, Weinstein JN, Cancer Genome Atlas Research Network, Liang H. 2018. A pan-cancer analysis of enhancer expression in nearly 9000 patient samples. *Cell* **173:** 386–399.e12. doi:10.1016/j.cell.2018.03.027

Collado-Torres L, Nellore A, Jaffe AE. 2017a. recount workflow: accessing over 70,000 human RNA-seq samples with Bioconductor. *F1000Res* **6:** 1558. doi:10.12688/f1000research.12223.1

Collado-Torres L, Nellore A, Kammers K, Ellis SE, Taub MA, Hansen KD, Jaffe AE, Langmead B, Leek JT. 2017b. Reproducible RNA-seq analysis using *recount2*. *Nat Biotechnol* **35:** 319–321. doi:10.1038/nbt.3838

de Kok JB, Verhaegh GW, Roelofs RW, Hessels D, Kiemeney LA, Aalders TW, Swinkels DW, Schalken JA. 2002. $DD3^{PCA3}$, a very sensitive and specific marker to detect prostate tumors. *Cancer Res* **62:** 2695–2698.

Ellis SE, Collado-Torres L, Jaffe A, Leek JT. 2018. Improving the value of public RNA-seq expression data by phenotype prediction. *Nucleic Acids Res* **46:** e54. doi:10.1093/nar/gky102

Emmrich S, Streltsov A, Schmidt F, Thangapandi VR, Reinhardt D, Klusmann JH. 2014. LincRNAs *MONC* and *MIR100HG* act as oncogenes in acute megakaryoblastic leukemia. *Mol Cancer* **13:** 171. doi:10.1186/1476-4598-13-171

Esteller M. 2011. Non-coding RNAs in human disease. *Nat Rev Genet* **12:** 861–874. doi:10.1038/nrg3074

Gokmen-Polar Y, Zavodszky M, Chen X, Gu X, Kodira C, Badve S. 2016. Abstract P2-06-05: LINC00478: a novel tumor suppressor in breast cancer. *Cancer Res* **76:** P2–06–05.

The GTEx Consortium. 2013. The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45:** 580–585. doi:10.1038/ng.2653

Gutschner T, Diederichs S. 2012. The hallmarks of cancer: a long non-coding RNA point of view. *RNA Biol* **9:** 703–719. doi:10.4161/rna.20481

Herz HM, Hu D, Shilatifard A. 2014. Enhancer malfunction in cancer. *Mol cell* **53:** 859–866. doi:10.1016/j.molcel.2014.02.033

Hnisz D, Abraham BJ, Lee TI, Lau A, Saint-André V, Sigova AA, Hoke HA, Young RA. 2013. Super-enhancers in the control of cell identity and disease. *Cell* **155:** 934–947. doi:10.1016/j.cell.2013.09.053

Hon CC, Ramilowski JA, Harshbarger J, Bertin N, Rackham OJL, Gough J, Denisenko E, Schmeier S, Poulsen TM, Severin J, et al. 2017. An atlas of human long non-coding RNAs with accurate 5′ ends. *Nature* **543:** 199–204. doi:10.1038/nature21374

Huang X, Yuan T, Liang M, Du M, Xia S, Dittmar R, Wang D, See W, Costello BA, Quevedo F, et al. 2015. Exosomal miR-1290 and miR-375 as prognostic markers in castration-resistant prostate cancer. *Eur Urol* **67:** 33–41. doi:10.1016/j.eururo.2014.07.035

Humphries CE, Kohli MA, Nathanson L, Whitehead P, Beecham G, Martin E, Mash DC, Pericak-Vance MA, Gilbert J. 2015. Integrated whole transcriptome and DNA methylation analysis identifies gene networks specific to late-onset Alzheimer's disease. *J Alzheimers Dis* **44:** 977–987. doi:10.3233/JAD-141989

Kotake Y, Nakagawa T, Kitagawa K, Suzuki S, Liu N, Kitagawa M, Xiong Y. 2011. Long non-coding RNA *ANRIL* is required for the PRC2 recruitment to and silencing of $p15^{INK4B}$ tumor suppressor gene. *Oncogene* **30:** 1956–1962. doi:10.1038/onc.2010.568

Lachmann A, Torre D, Keenan AB, Jagodnik KM, Lee HJ, Wang L, Silverstein MC, Ma'ayan A. 2018. Massive mining of publicly available RNA-seq data from human and mouse. *Nat Commun* **9:** 1366. doi:10.1038/s41467-018-03751-6

Lawrence M, Huber W, Pages H, Aboyoun P, Carlson M, Gentleman R, Morgan MT, Carey VJ. 2013. Software for computing and annotating genomic ranges. *PLoS Comput Biol* **9:** e1003118. doi:10.1371/journal.pcbi.1003118

Leek JT, Storey JD. 2007. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet* **3:** 1724–1735. doi:10.1371/journal.pgen.0030161

Li S, Li B, Zheng Y, Li M, Shi L, Pu X. 2017. Exploring functions of long non-coding RNAs across multiple cancers through co-expression network. *Sci Rep* **7:** 754. doi:10.1038/s41598-017-00856-8

Lin DW, FitzGerald LM, Fu R, Kwon EM, Zheng SL, Kolb S, Wiklund F, Stattin P, Isaacs WB, Xu J, et al. 2011. Genetic variants in the *LEPR*, *CRY1*, *RNASEL*, *IL4*, and *ARVCF* genes are prognostic markers of prostate cancer-specific mortality. *Cancer Epidem Biomar* **20:** 1928–1936. doi:10.1158/1055-9965.EPI-11-0236

Ling H, Vincent K, Pichler M, Fodde R, Berindan-Neagoe I, Slack FJ, Calin GA. 2015. Junk DNA and the long non-coding RNA twist in cancer genetics. *Oncogene* **34:** 5003–5011. doi:10.1038/onc.2014.456

Ma Z, Gu S, Song M, Yan C, Hui B, Ji H, Wang J, Zhang J, Wang K, Zhao Q. 2017. Long non-coding RNA SNHG17 is an unfavourable prognostic factor and promotes cell proliferation by epigenetically silencing P57 in colorectal cancer. *Mol BioSystems* **13:** 2350–2361. doi:10.1039/C7MB00280G

Ni S, Zhao X, Ouyang L. 2016. Long non-coding RNA expression profile in vulvar squamous cell carcinoma and its clinical significance. *Oncol Rep* **36:** 2571–2578. doi:10.3892/or.2016.5075

Ong CT, Corces VG. 2011. Enhancer function: new insights into the regulation of tissue-specific gene expression. *Nat Rev Genet* **12:** 283–293. doi:10.1038/nrg2957

Prensner JR, Iyer MK, Balbin OA, Dhanasekaran SM, Cao Q, Brenner JC, Laxman B, Asangani IA, Grasso CS, Kominsky HD, et al. 2011. Transcriptome sequencing across a prostate cancer cohort identifies *PCAT-1*, an unannotated lincRNA implicated in disease progression. *Nat Biotechnol* **29:** 742–749. doi:10.1038/nbt.1914

Quinn JJ, Chang HY. 2016. Unique features of long non-coding RNA biogenesis and function. *Nat Rev Genet* **17:** 47–62. doi:10.1038/nrg.2015.10

Ramilowski JA, Yip CW, Agrawal S, Chang JC, Ciani Y, Kulakovskiy IV, Mendez M, Ooi JLC, Ouyang JF, Parkinson N, et al. 2020. Functional annotation of human long noncoding RNAs via molecular phenotyping. *Genome Res* (this issue). doi:10.1101/gr.254219.119

Ren S, Liu Y, Xu W, Sun Y, Lu J, Wang F, Wei M, Shen J, Hou J, Gao X, et al. 2013. Long noncoding RNA MALAT-1 is a new potential therapeutic target for castration resistant prostate cancer. *J Urology* **190:** 2278–2287. doi:10.1016/j.juro.2013.07.001

Shang D, Zheng T, Zhang J, Tian Y, Liu Y. 2016. Profiling of mRNA and long non-coding RNA of urothelial cancer in recipients after renal transplantation. *Tumor Biol* **37:** 12673–12684. doi:10.1007/s13277-016-5148-1

Smyth GK. 2004. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* **3:** Article3. doi:10.2202/1544-6115.1027

Sun D, Layer R, Mueller AC, Cichewicz MA, Negishi M, Paschal BM, Dutta A. 2014. Regulation of several androgen-induced genes through the repression of the miR-99a/let-7c/miR-125b-2 miRNA cluster in prostate cancer cells. *Oncogene* **33:** 1448–1457. doi:10.1038/onc.2013.77

Sun Y, Wei G, Luo H, Wu W, Skogerbø G, Luo J, Chen R. 2017. The long noncoding RNA *SNHG1* promotes tumor growth through regulating transcription of both local and distal genes. *Oncogene* **36:** 6774–6783. doi:10.1038/onc.2017.286

Sur I, Taipale J. 2016. The role of enhancers in cancer. *Nat Rev Cancer* **16:** 483–493. doi:10.1038/nrc.2016.62

Tatlow PJ, Piccolo SR. 2016. A cloud-based workflow to quantify transcript-expression levels in public cancer compendia. *Sci Rep* **6:** 39259. doi:10.1038/srep39259

Yu J, Yu J, Mani RS, Cao Q, Brenner CJ, Cao X, Wang X, Wu L, Li J, Hu M, et al. 2010. An integrated network of androgen receptor, polycomb, and TMPRSS2-ERG gene fusions in prostate cancer progression. *Cancer Cell* **17:** 443–454. doi:10.1016/j.ccr.2010.03.018

Zhao W, Luo J, Jiao S. 2015a. Comprehensive characterization of cancer subtype associated long non-coding RNAs and their clinical implications. *Sci Rep* **4:** 6591. doi:10.1038/srep06591

Zhao J, Liu Y, Zhang W, Zhou Z, Wu J, Cui P, Zhang Y, Huang G. 2015b. Long non-coding RNA Linc00152 is involved in cell cycle arrest, apoptosis, epithelial to mesenchymal transition, cell migration and invasion in gastric cancer. *Cell Cycle* **14:** 3112–3123. doi:10.1080/15384101.2015.1078034

Zhou M, Hou Y, Yang G, Zhang H, Tu G, Du Y-e, Wen S, Xu L, Tang X, Tang S, et al. 2016. LncRNA-Hh strengthen cancer stem cells generation in twist-positive breast cancer via activation of hedgehog signaling pathway. *Stem Cells* **34:** 55–66. doi:10.1002/stem.2219

Zhu S, Li W, Liu J, Chen CH, Liao Q, Xu P, Xu H, Xiao T, Cao Z, Peng J, et al. 2016. Genome-scale deletion screening of human long non-coding RNAs using a paired-guide RNA CRISPR–Cas9 library. *Nat Biotechnol* **34:** 1279–1286. doi:10.1038/nbt.3715

Received July 12, 2019; accepted in revised form February 11, 2020.

Recounting the FANTOM CAGE-Associated Transcriptome

Eddie Luidy Imada, Diego Fernando Sanchez, Leonardo Collado-Torres, et al.

Genome Res. 2020 30: 1073-1081 originally published online February 20, 2020
Access the most recent version at doi:10.1101/gr.254656.119

Resource

Functional annotation of human long noncoding RNAs via molecular phenotyping

Jordan A. Ramilowski,[1,2,47] Chi Wai Yip,[1,2,47] Saumya Agrawal,[1,2] Jen-Chien Chang,[1,2] Yari Ciani,[3] Ivan V. Kulakovskiy,[4,5] Mickaël Mendez,[6] Jasmine Li Ching Ooi,[2] John F. Ouyang,[7] Nick Parkinson,[8] Andreas Petri,[9] Leonie Roos,[10,11] Jessica Severin,[1,2] Kayoko Yasuzawa,[1,2] Imad Abugessaisa,[1,2] Altuna Akalin,[12] Ivan V. Antonov,[13] Erik Arner,[1,2] Alessandro Bonetti,[2] Hidemasa Bono,[14] Beatrice Borsari,[15] Frank Brombacher,[16,17] Christopher J.F. Cameron,[18,23,46] Carlo Vittorio Cannistraci,[19,20] Ryan Cardenas,[21] Melissa Cardon,[1] Howard Chang,[22] Josée Dostie,[23] Luca Ducoli,[24] Alexander Favorov,[25,26] Alexandre Fort,[2] Diego Garrido,[15] Noa Gil,[27] Juliette Gimenez,[28] Reto Guler,[16,17] Lusy Handoko,[2] Jayson Harshbarger,[2] Akira Hasegawa,[1,2] Yuki Hasegawa,[2] Kosuke Hashimoto,[1,2] Norihito Hayatsu,[1] Peter Heutink,[29] Tetsuro Hirose,[30] Eddie L. Imada,[26] Masayoshi Itoh,[2,31] Bogumil Kaczkowski,[1,2] Aditi Kanhere,[21] Emily Kawabata,[2] Hideya Kawaji,[31] Tsugumi Kawashima,[1,2] S. Thomas Kelly,[1] Miki Kojima,[1,2] Naoto Kondo,[2] Haruhiko Koseki,[1] Tsukasa Kouno,[1,2] Anton Kratz,[2] Mariola Kurowska-Stolarska,[32] Andrew Tae Jun Kwon,[1,2] Jeffrey Leek,[26] Andreas Lennartsson,[33] Marina Lizio,[1,2] Fernando López-Redondo,[1,2] Joachim Luginbühl,[1,2] Shiori Maeda,[1] Vsevolod J. Makeev,[25,34] Luigi Marchionni,[26] Yulia A. Medvedeva,[13,34] Aki Minoda,[1,2] Ferenc Müller,[21] Manuel Muñoz-Aguirre,[15] Mitsuyoshi Murata,[1,2] Hiromi Nishiyori,[1,2] Kazuhiro R. Nitta,[1,2] Shuhei Noguchi,[1,2] Yukihiko Noro,[2] Ramil Nurtdinov,[15] Yasushi Okazaki,[1,2] Valerio Orlando,[35] Denis Paquette,[23] Callum J.C. Parr,[1] Owen J.L. Rackham,[7] Patrizia Rizzu,[29] Diego Fernando Sánchez Martinez,[26] Albin Sandelin,[36] Pillay Sanjana,[21] Colin A.M. Semple,[37] Youtaro Shibayama,[1,2] Divya M. Sivaraman,[1,2] Takahiro Suzuki,[1,2] Suzannah C. Szumowski,[2] Michihira Tagami,[1,2] Martin S. Taylor,[37] Chikashi Terao,[1] Malte Thodberg,[36] Supat Thongjuea,[2] Vidisha Tripathi,[38] Igor Ulitsky,[27] Roberto Verardo,[3] Ilya E. Vorontsov,[25] Chinatsu Yamamoto,[2] Robert S. Young,[39] J. Kenneth Baillie,[8] Alistair R.R. Forrest,[1,2,40] Roderic Guigó,[15,41] Michael M. Hoffman,[42] Chung Chau Hon,[1,2] Takeya Kasukawa,[1,2] Sakari Kauppinen,[9] Juha Kere,[33,43] Boris Lenhard,[10,11,44] Claudio Schneider,[3,45] Harukazu Suzuki,[1,2] Ken Yagi,[1,2] Michiel J.L. de Hoon,[1,2] Jay W. Shin,[1,2] and Piero Carninci[1,2]

[47]**These authors contributed equally to this work.**
Corresponding authors: michiel.dehoon@riken.jp, jay.shin@riken.jp, carninci@riken.jp

Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.254219.119. Freely available online through the *Genome Research* Open Access option.

30:1060–1072 Published by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/20; www.genome.org

[1]RIKEN Center for Integrative Medical Sciences, Yokohama, Kanagawa 230-0045, Japan; [2]RIKEN Center for Life Science Technologies, Yokohama, Kanagawa 230-0045, Japan; [3]Laboratorio Nazionale Consorzio Interuniversitario Biotecnologie (CIB), Trieste 34127, Italy; [4]Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, Moscow 119991, Russia; [5]Institute of Protein Research, Russian Academy of Sciences, Pushchino 142290, Russia; [6]Department of Computer Science, University of Toronto, Toronto, Ontario M5S 1A1, Canada; [7]Program in Cardiovascular and Metabolic Disorders, Duke-National University of Singapore Medical School, Singapore 169857, Singapore; [8]Roslin Institute, University of Edinburgh, Edinburgh EH25 9RG, United Kingdom; [9]Center for RNA Medicine, Department of Clinical Medicine, Aalborg University, Copenhagen 9220, Denmark; [10]Institute of Clinical Sciences, Faculty of Medicine, Imperial College London, London W12 0NN, United Kingdom; [11]Computational Regulatory Genomics, MRC London Institute of Medical Sciences, London W12 0NN, United Kingdom; [12]Berlin Institute for Medical Systems Biology, Max Delbrük Center for Molecular Medicine in the Helmholtz Association, Berlin 13125, Germany; [13]Institute of Bioengineering, Research Center of Biotechnology, Russian Academy of Sciences, Moscow 117312, Russia; [14]Graduate School of Integrated Sciences for Life, Hiroshima University, Higashi-Hiroshima City 739-0046, Japan; [15]Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Barcelona, Catalonia 08003, Spain; [16]International Centre for Genetic Engineering and Biotechnology (ICGEB), University of Cape Town, Cape Town 7925, South Africa; [17]Institute of Infectious Diseases and Molecular Medicine (IDM), Department of Pathology, Division of Immunology and South African Medical Research Council (SAMRC) Immunology of Infectious Diseases, Faculty of Health Sciences, University of Cape Town, Cape Town 7925, South Africa; [18]School of Computer Science, McGill University, Montréal, Québec H3G 1Y6, Canada; [19]Biomedical Cybernetics Group, Biotechnology Center (BIOTEC), Center for Molecular and Cellular Bioengineering (CMCB), Center for Systems Biology Dresden (CSBD), Cluster of Excellence Physics of Life (PoL), Department of Physics, Technische Universität Dresden, Dresden 01062, Germany; [20]Center for Complex Network Intelligence (CCNI) at the Tsinghua Laboratory of Brain and Intelligence (THBI), Department of Bioengineering, Tsinghua University, Beijing 100084, China; [21]Institute of Cancer and Genomic Sciences, College of Medical and Dental Sciences, University of Birmingham, Birmingham B15 2TT, United Kingdom; [22]Center for Personal Dynamic Regulome, Stanford University, Stanford, California 94305, USA; [23]Department of Biochemistry, Rosalind and Morris Goodman Cancer Research Center, McGill University, Montréal, Québec H3G 1Y6, Canada; [24]Institute of Pharmaceutical Sciences, Swiss Federal Institute of Technology, Zurich 8093, Switzerland; [25]Department of Computational Systems Biology, Vavilov Institute of General Genetics, Russian Academy of Sciences, Moscow 119991, Russia; [26]Department of Oncology, Johns Hopkins University, Baltimore, Maryland 21287, USA; [27]Department of Biological Regulation, Weizmann Institute of Science, Rehovot 76100, Israel; [28]Epigenetics and Genome Reprogramming Laboratory, IRCCS Fondazione Santa Lucia, Rome 00179, Italy; [29]Genome Biology of Neurodegenerative Diseases, German Center for Neurodegenerative Diseases (DZNE), Tübingen 72076, Germany; [30]Graduate School of Frontier Biosciences, Osaka University, Suita 565-0871, Japan; [31]RIKEN Preventive Medicine and Diagnosis Innovation Program (PMI), Saitama 351-0198, Japan; [32]Institute of Infection, Immunity, and Inflammation, University of Glasgow, Glasgow, Scotland G12 8QQ, United Kingdom; [33]Department of Biosciences and Nutrition, Karolinska Institutet, Huddinge 14157, Sweden; [34]Moscow Institute of Physics and Technology, Dolgoprudny 141701, Russia; [35]Biological and Environmental Sciences and Engineering Division, King Abdullah University of Science and Technology, Thuwal 23955-6900, Kingdom of Saudi Arabia; [36]Department of Biology and BRIC, University of Copenhagen, Denmark, Copenhagen N DK2200, Denmark; [37]MRC Human Genetics Unit, University of Edinburgh, Edinburgh EH4 2XU, United Kingdom; [38]National Centre for Cell Science, Pune, Maharashtra 411007, India; [39]Centre for Global Health Research, Usher Institute, University of Edinburgh, Edinburgh EH8 9AG, United Kingdom; [40]Harry Perkins Institute of Medical Research, QEII Medical Centre and Centre for Medical Research, The University of Western Australia, Nedlands, Perth, Western Australia 6009, Australia; [41]Universitat Pompeu Fabra (UPF), Barcelona, Catalonia 08002, Spain; [42]Princess Margaret Cancer Centre, Toronto, Ontario M5G 1L7, Canada; [43]Stem Cells and Metabolism Research Program, University of Helsinki and Folkhälsan Research Center, 00290 Helsinki, Finland; [44]Sars International Centre for Marine Molecular Biology, University of Bergen, Bergen N-5008, Norway; [45]Department of Medicine and Consorzio Interuniversitario Biotecnologie p.zle Kolbe 1 University of Udine, Udine 33100, Italy; [46]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, Connecticut 06510, USA

Long noncoding RNAs (lncRNAs) constitute the majority of transcripts in the mammalian genomes, and yet, their functions remain largely unknown. As part of the FANTOM6 project, we systematically knocked down the expression of 285 lncRNAs in human dermal fibroblasts and quantified cellular growth, morphological changes, and transcriptomic responses using Capped Analysis of Gene Expression (CAGE). Antisense oligonucleotides targeting the same lncRNAs exhibited global concordance, and the molecular phenotype, measured by CAGE, recapitulated the observed cellular phenotypes while providing additional insights on the affected genes and pathways. Here, we disseminate the largest-to-date lncRNA knockdown data set with molecular phenotyping (over 1000 CAGE deep-sequencing libraries) for further exploration and highlight functional roles for *ZNF213–AS1* and *lnc–KHDC3L–2*.

[Supplemental material is available for this article.]

Over 50,000 loci in the human genome transcribe long noncoding RNAs (lncRNAs) (Iyer et al. 2015; Hon et al. 2017), which are defined as transcripts at least 200 nucleotides (nt) long with low or no protein-coding potential. Although lncRNA genes outnumber protein-coding genes in mammalian genomes, they are comparatively less conserved (Ulitsky 2016), lowly expressed, and more cell-type-specific (Hon et al. 2017). However, the evolutionary conservation of lncRNA promoters (Carninci et al. 2005) and the structural motifs of lncRNAs (Chu et al. 2015; Xue et al. 2016) suggest that lncRNAs are fundamental biological regulators. To date, only a few hundred human lncRNAs have been extensively characterized (de Hoon et al. 2015; Quek et al. 2015; Volders et al. 2015; Ma et al. 2019), revealing their roles in regulating transcription (Engreitz et al. 2016b), translation (Carrieri et al. 2012), and chromatin state (Gupta et al. 2010; Guttman et al. 2011; Guttman and Rinn 2012; Quinn and Chang 2016; Ransohoff et al. 2018).

Our recent FANTOM5 computational analysis showed that 19,175 (out of 27,919) human lncRNA loci are functionally implicated (Hon et al. 2017). Yet, genomic screens are necessary to comprehensively characterize each lncRNA. One common approach of gene knockdown followed by a cellular phenotype assay typically characterizes a small percentage of lncRNAs for a single observable phenotype. For example, a recent large-scale screening using CRISPR interference (CRISPRi) found that ~3.7% of targeted lncRNA loci are essential for cell growth or viability in a cell-type-specific manner (Liu et al. 2017). In addition, CRISPR-Cas9 experiments targeting splice sites identified ~2.1% of lncRNAs that affect growth of K562 (Liu et al. 2018), and a CRISPR activation study revealed ~0.11% lncRNAs to be important for drug resistance in melanoma (Joung et al. 2017). However, many of these studies target the genomic DNA, potentially perturbing the chromatin architecture, or focus on a single cellular assay, possibly missing other relevant functions and underlying molecular pathways.

As a part of the FANTOM6 pilot project, we established an automated high-throughput cell culture platform to suppress 285 lncRNAs expressed in human primary dermal fibroblasts (HDFs) using antisense LNA-modified GapmeR antisense oligonucleotide (ASO) technology (Roux et al. 2017). We then quantified the effect of each knockdown on cell growth and morphology using real-time imaging, followed by Cap Analysis Gene Expression (CAGE) (Murata et al. 2014) deep sequencing to reveal molecular pathways associated with each lncRNA. In contrast to cellular phenotyping, molecular phenotyping provides a detailed assessment of the response to a lncRNA knockdown at the molecular level, allowing biological pathways to be associated to lncRNAs even in the absence of an observable cellular phenotype. All data and analysis results are publicly available (see Data access), and results can be interactively explored using our in-house portal (https://fantom.gsc.riken.jp/zenbu/reports/#FANTOM6).

Results

Selection and ASO-mediated knockdown of lncRNA targets

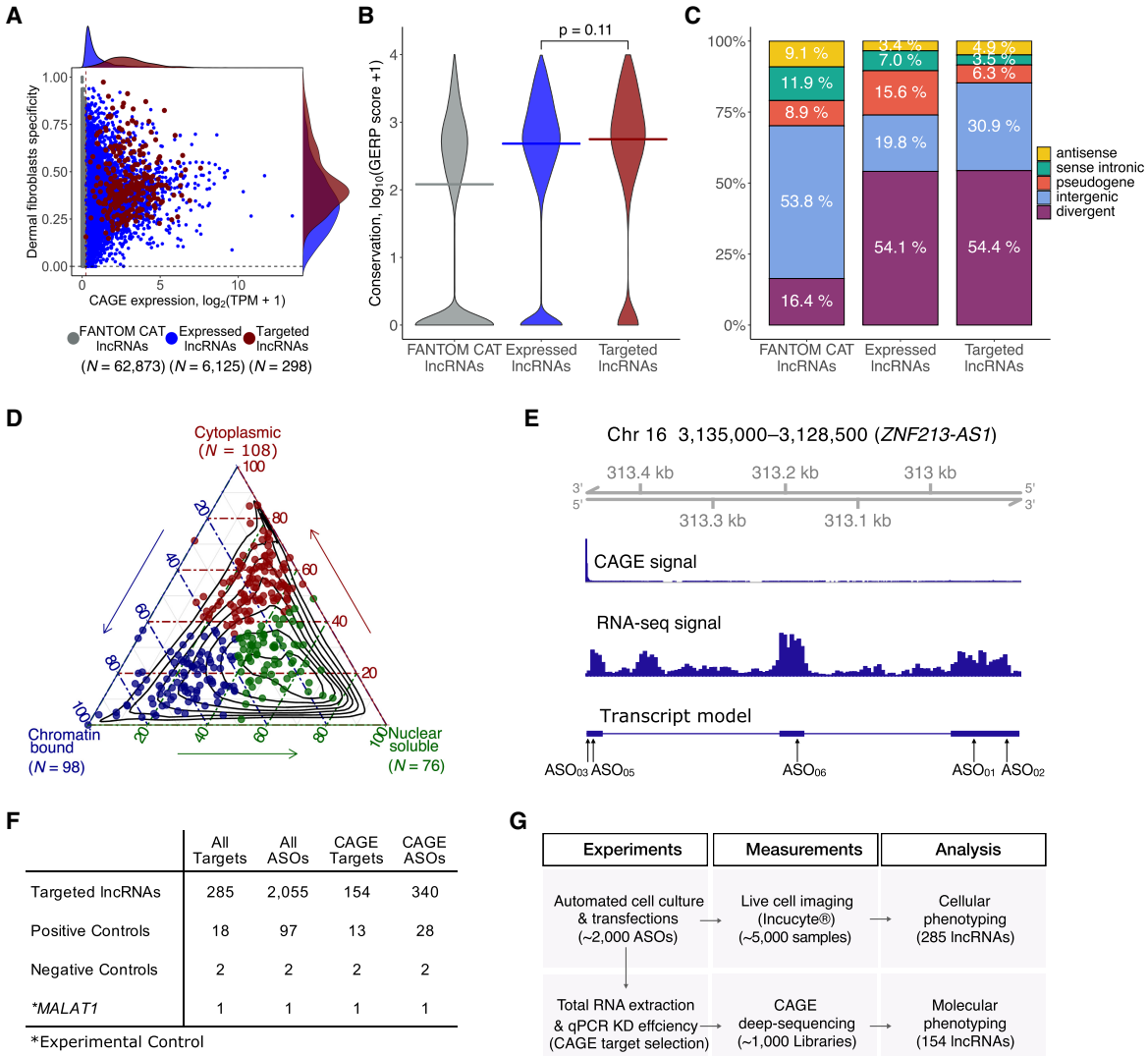
Human dermal fibroblasts are nontransformed primary cells that are commonly used for investigating cellular reprogramming (Takahashi et al. 2007; Ambasudhan et al. 2011), wound healing (Li and Wang 2011), fibrosis (Kendall and Feghali-Bostwick 2014), and cancer (Kalluri 2016). Here, an unbiased selection of lncRNAs expressed in HDFs was performed to choose 285 lncRNAs for functional interrogation (Methods; Supplemental Table S1; Fig. 1A–C). Using RNA-seq profiling of fractionated RNA, we annotated the lncRNA subcellular localization as the chromatin-bound (35%), nucleus-soluble (27%), or cytoplasmic (38%) (Fig. 1D). We then designed a minimum of five non-overlapping antisense oligonucleotides against each lncRNA (Supplemental Methods; Supplemental Table S2; Fig. 1E,F) and transfected them individually using an automated cell culture platform to minimize experimental variability (Fig. 1G). The overall knockdown efficiencies across 2021 ASOs resulted in median value of 45.4%, and we could successfully knockdown 879 out of 2021 (43.5%) ASOs (>40% knockdown efficiency in at least two primer pairs or >60% in one primer pair) (Supplemental Table S2). ASOs targeting exons or introns were equally effective, and knockdown efficiencies were independent of the genomic class, expression level, and subcellular localization of the lncRNA (Supplemental Fig. S1A–D).

A subset of lncRNAs are associated with cell growth and morphology changes

To evaluate the effect of each lncRNA knockdown on cell growth and morphology, we imaged ASO-transfected HDFs in duplicate every 3 h for a total of 48 h (Supplemental Table S3) and estimated their growth rate based on cell confluence measurements (Fig. 2A,B). First, we observed across all ASOs that changes in cell growth and morphological parameters were significantly correlated with knockdown efficiency (Supplemental Fig. S1E). Considering both successful knockdown and significant growth inhibition (Student's two-sided *t*-test FDR $\leq$ 0.05), 246 out of 879 ASOs (~28%) showed cellular phenotype (Fig. 2C; Supplemental Table S3).

To assess globally whether the observed growth inhibition is lncRNA-specific, we used all 194 lncRNAs successfully targeted by at least two ASOs (Supplemental Fig. S2A) and found that ASOs targeting the same lncRNA were significantly more likely to have a concordant growth response than ASOs targeting different lncRNA (empirical $P = 0.00037$) (Supplemental Methods; Supplemental Fig. S2B). However, different ASOs targeting the same lncRNA typically showed different effects on growth, possibly due to variable knockdown efficiencies or differences in targeted lncRNA isoforms, as well as off-target effects. To reliably identify target-specific cellular phenotype, we applied conditional cutoffs based on the number of successful ASOs per each lncRNA (Supplemental Methods; Supplemental Fig. S2C) and identified 15/194 lncRNAs (7.7%) with growth phenotype (adjusted background <5%) (Supplemental Fig. S2D). We validated *A1BG-AS1*, which was previously implicated in cell growth (Bai et al. 2019), *CATG00000089639*, *RP11-195F19.9*, and *ZNF213-AS1* by measuring the MKI67 proliferation protein marker upon knockdown with siRNAs and with selected ASOs (Fig. 2D; Supplemental Fig. S2E).

In addition to cell growth, we also explored changes in cell morphology (Fig. 2E). Using a machine learning-assisted workflow (Methods), each cell was segmented and its morphological features representing various aspects of cell shapes and sizes were quantified (Fig. 2F; Supplemental Table S3; Carpenter et al. 2006). As an example, knockdown of 14/194 lncRNAs (7.2%) affected the spindle-like morphology of fibroblasts, as indicated by a consistent decrease in their observed eccentricity without reducing the cell number, suggesting possible cellular transformation toward epithelial-like states. Collectively, we observed 59/194 lncRNAs (~30%) affecting cell growth and/or morphological parameters (Fig. 2G; Supplemental Table S3).



F

	All Targets	All ASOs	CAGE Targets	CAGE ASOs
Targeted lncRNAs	285	2,055	154	340
Positive Controls	18	97	13	28
Negative Controls	2	2	2	2
MALAT1	1	1	1	1

*Experimental Control

G

Figure 1. Selection of lncRNA targets, their properties, and the study overview. (*A*) CAGE expression levels at $\log_2$TPM (tags per million) and human dermal fibroblasts (HDFs) specificity of lncRNAs in the FANTOM CAT catalog (Hon et al. 2017) ($N = 62,873$; gray), lncRNAs expressed in HDFs ($N = 6125$; blue), and targeted lncRNAs ($N = 285$; red). The dashed vertical line indicates most lowly expressed lncRNA target (~0.2 TPM). (*B*) Gene conservation levels of lncRNAs in the FANTOM CAT catalog (gray), lncRNAs expressed in HDFs (blue), and targeted lncRNAs (red). Crossbars indicate the median. No significant difference is observed when comparing targeted and expressed in HDF lncRNAs (Wilcoxon $P = 0.11$). (*C*) Similar to that in *B* but for genomic classes of lncRNAs. Most of the targeted lncRNAs and those expressed in HDFs are expressed from divergent promoters. (*D*) Subcellular localization (based on relative abundances from RNA-seq fractionation data) for targeted lncRNAs. Chromatin-bound ($N = 98$; blue); nuclear soluble ($N = 76$; green); cytoplasmic ($N = 108$; red). Black contours represent the distribution of all lncRNAs expressed in HDFs. (*E*) Example of *ZNF213-AS1* loci showing transcript model, CAGE, and RNA-seq signal along with targeting ASOs. (*F*) Number of ASOs for target lncRNAs and controls used in the experiment. (*G*) Schematics of the study.

Molecular phenotyping by CAGE recapitulates cellular phenotypes and highlights functions of lncRNAs

Next, we selected 340 ASOs with high knockdown efficiencies (mostly >50%; median 71.4%) and sequenced 970 CAGE libraries to analyze 154 lncRNAs (Fig. 3A; Supplemental Table S4). To assess functional implications by individual ASOs, we performed differential gene expression, Motif Activity Response Analysis (MARA) (The FANTOM Consortium et al. 2009), and Gene Set Enrichment Analysis (GSEA) (Fig. 3B–F; Subramanian et al. 2005), and compared them with cellular phenotype.

We globally observed significant knockdown-mediated transcriptomic changes (which generally correlated with KD efficiency) (Supplemental Fig. S3A), with ~57% of ASOs showing at least 10 differentially expressed genes (FDR ≤ 0.05; abs[$\log_2$FC] > 0.5). For 84 divergent-antisense lncRNAs (targeted by 186 independent ASOs) (Supplemental Methods), we found their partner gene to be generally unchanged (median abs[$\log_2$FC] = ~0.13), with an exception of two significantly down-regulated and three significantly up-regulated genes (FDR ≤ 0.05) (Supplemental Fig. S3B). We have, however, noticed a common response in a large number of ASOs (~30%–35% of all responding ASOs), such as down-regulation of cell-cycle-related pathways, up-regulated stress genes and pathways, or altered cell metabolism and energetics (Supplemental Fig. S3C,D).

When comparing knockdown-mediated molecular and cellular response, we found that transcription factor motifs that

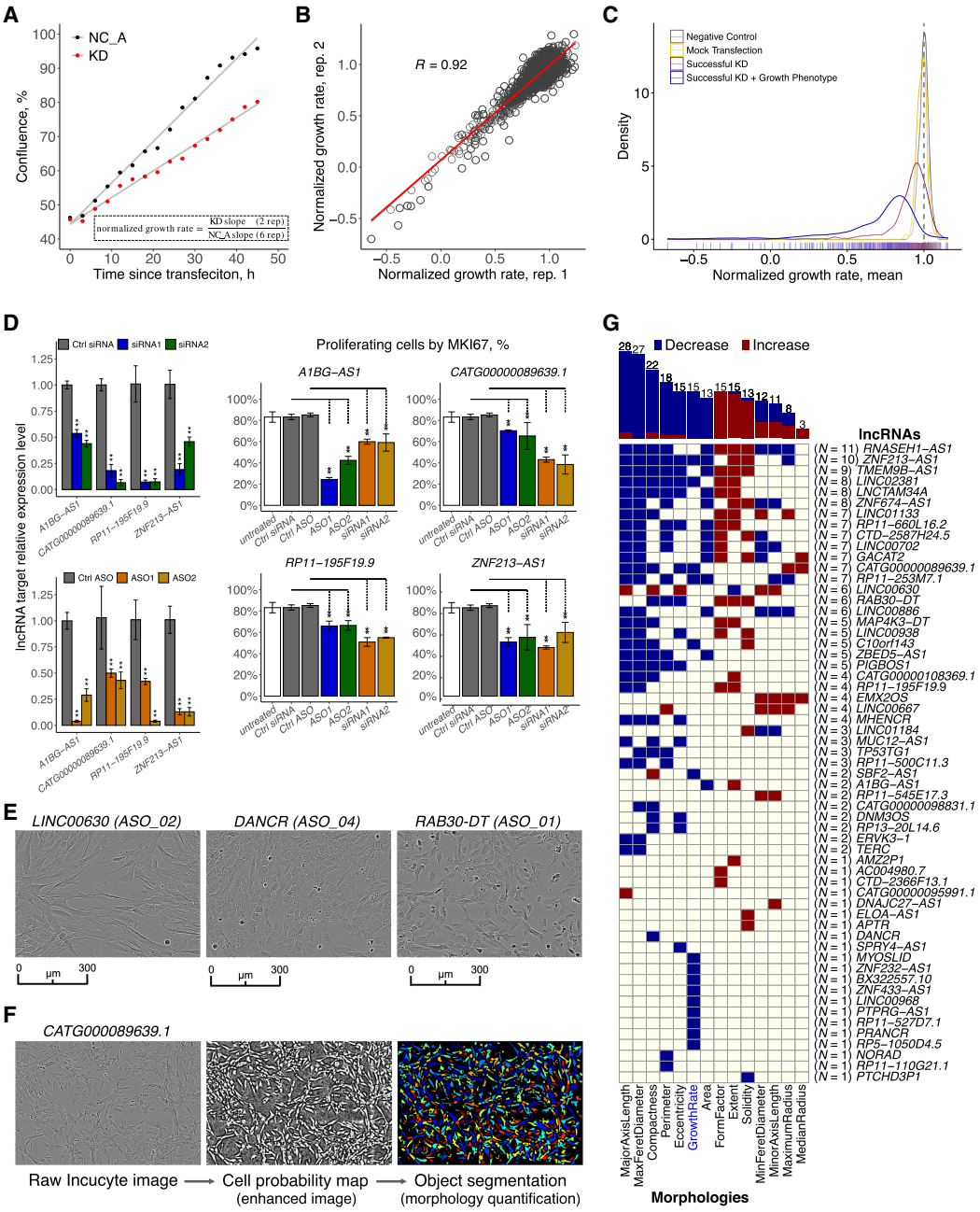


Figure 2. Cell growth and morphology assessment. (*A*) Selected example (*PTPRG1-AS1*) showing the normalized growth rate estimation using a matching NC_A (negative control). (*B*) Correlation of the normalized growth rate for technical duplicates across 2456 Incucyte samples. (*C*) Density distribution of normalized growth rates (technical replicates averaged) 252 ASOs targeting lncRNAs with successful knockdown (KD) and growth phenotype (blue) consistent in two replicates (FDR < 0.05 as compared to matching NC_A; 246 ASOs inhibited growth), 627 ASOs targeting lncRNAs with successful KD (purple), 270 negative control (NC_A) samples (gray), and 90 mock-transfected cells (Lipofectamine only) samples (yellow). (*D*) MKI67 staining (growth inhibition validation) for four selected lncRNA targets after siRNA and ASOs suppression. (*E*) Incucyte cell images of selected distinct cell morphologies changes upon an lncRNA KD. (*F*) An overview of the cell morphology imaging processing pipeline using a novel lncRNA target, *CATG000089639.1*, as an example. (*G*) lncRNAs ($N = 59$) significantly (FDR < 0.05) and consistently (after adjusting for the number of successfully targeting ASOs) affecting cell growth ($N = 15$) and cell morphologies ($N = 44$).

promote cell growth, including TFDP1, E2F1,2,3, and EP300, were positively correlated with the measured cell growth rate, whereas transcription factor motifs known to inhibit growth or induce apoptosis (e.g., PPARG, SREBPF, and STAT2,4,6) were negatively correlated (Fig. 3D; Supplemental Fig. S4A; Supplemental Table S6). Moreover, correlations of growth with GSEA pathways (Fig. 3F; Supplemental Fig. S4B; Supplemental Table S6) or with FANTOM5 coexpression clusters (Supplemental Fig. S4C) showed that cell growth and replication-related pathways were positively correlated with the measured growth rate, whereas those related

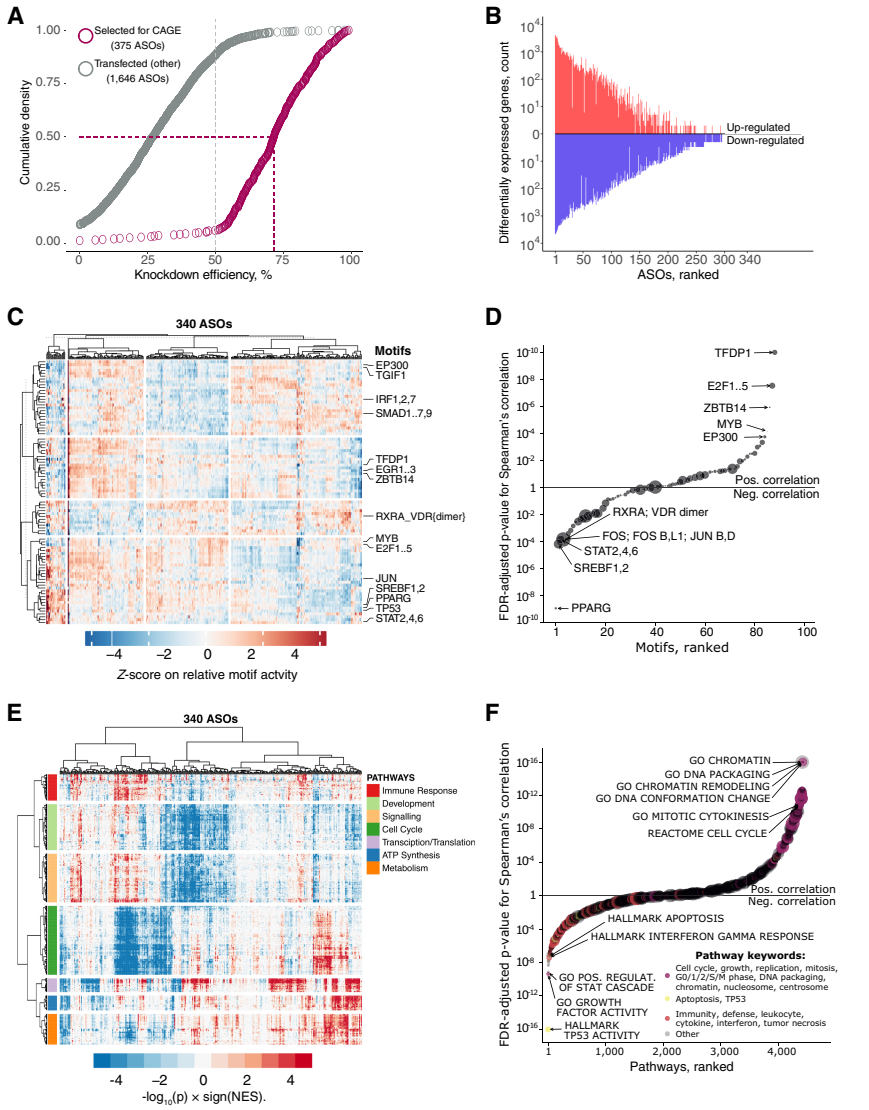


Figure 3. CAGE predicts cellular phenotypes. (*A*) RT-qPCR knockdown efficiency for 2021 ASO-transfected samples (targeted lncRNAs only). Gray dashed line indicates 50% KD efficiency generally required for CAGE selection. Purple dashed lines indicate median KD efficiency (71.5%) for 375 ASOs selected for CAGE sequencing. After quality control, 340 ASOs targeting lncRNAs were included for further analysis. (*B*) Distribution of significantly differentially expressed genes (up-regulated: FDR < 0.05, *Z*-score > 1.645, $\log_2$FC > 0.5; and down-regulated: FDR < 0.05, *Z*-score < −1.645, $\log_2$FC < −0.5) across all 340 ASOs. (*C*) Motif Response Activity Analysis (MARA) across 340 ASOs. Scale indicates *Z*-score of the relative motif activity (the range was set to abs[*Z*-score] = <5 for visualization purposes). (*D*) Correlation between normalized growth rate and motif activities across 340 ASOs targeting lncRNAs with highlighted examples. Motif sizes shown are scaled based on the HDF expression of their associated TFs (range 1 to ~600 TPM). (*E*) Enriched biological pathways across 340 ASOs. Scale indicates GSEA enrichment value calculated as $-\log_{10}$(p) × sign(NES). (*F*) Same as in *D* but for selected GSEA pathways. Pathways sizes are scaled based on the number of associated genes.

to immunity, and cell stress and cell death were negatively correlated. We found that among 53 ASOs implicated in a growth-inhibition pathway based on the CAGE profiles, only 43% of them showed growth inhibition in the real-time imaging. This might suggest better sensitivity of transcriptomic profiling when detecting phenotypes as compared to live cell imaging methods, which are more prone to a delayed cellular response to the knockdown.

Additionally, morphological changes were reflected in the molecular phenotype assessed by CAGE (Supplemental Fig. S4D). Cell radius and axis length were associated with GSEA categories related to actin arrangement and cilia, whereas cell compactness was negatively correlated with apoptosis. The extensive molecular phenotyping analysis also revealed pathways not explicitly associated with cell growth and cell morphology, such as transcription, translation, metabolism, development, and signaling (Fig. 3E).

Next, to globally assess whether individual ASO knockdowns lead to lncRNA-specific effects, we scaled the expression change of each gene across the whole experiment and compared differentially expressed genes (Fig. 3B) of all possible ASO pairs targeting the same lncRNA target versus different lncRNAs (Supplemental Methods; Supplemental Table S5). We found that the concordance of the same target group was significantly greater than that of the different target group (comparing the Jaccard indices across 10,000 permutations) (Supplemental Fig. S5A), suggesting that ASO knockdowns are nonrandom and lead to more lncRNA specific effects than the nontargeting ASO pairs. Further, by requiring at least five common DEGs (FDR ≤ 0.05, abs[$\log_2$FC] > 0.5, abs[*Z*-score] > 1.645) and ASO-pairs significantly above the nontargeting ASO pairs background ($P \leq 0.05$), we identified 16 ASO pairs, targeting 13 lncRNAs, exhibiting reproducible knockdown-mediated molecular responses in human dermal fibroblasts (Supplemental Fig. S5B). Corresponding GSEA pathways and MARA motifs of these 16 ASO pairs are shown in Supplemental Figure S5C.

siRNA validation experiments

To evaluate whether the lncRNA-specific effects can be measured by other knockdown technologies, nine lncRNAs, with relatively mild growth phenotype, were subjected to siRNA knockdown. Measuring transcriptional response, we noted that higher concordance was observed for ASO modality alone (Supplemental Fig. S5D). The observed discrepancies in the transcriptional response between ASO- and siRNA-mediated knockdowns could be contributed by their mode of action and variable activities in different subcellular compartments. Next, a concordant response was found for (5/36) ASO-siRNA pairs targeting three lncRNAs (Supplemental Fig. S5E; Supplemental Table S5), enriched in the cytoplasm (*MAPKAPK5-AS1*), soluble nuclear fraction (*LINC02454*), and in the chromatin-bound fraction (*A1BG-AS1*). Although we cannot completely exclude the technical artifacts of each technology, concordant cellular response

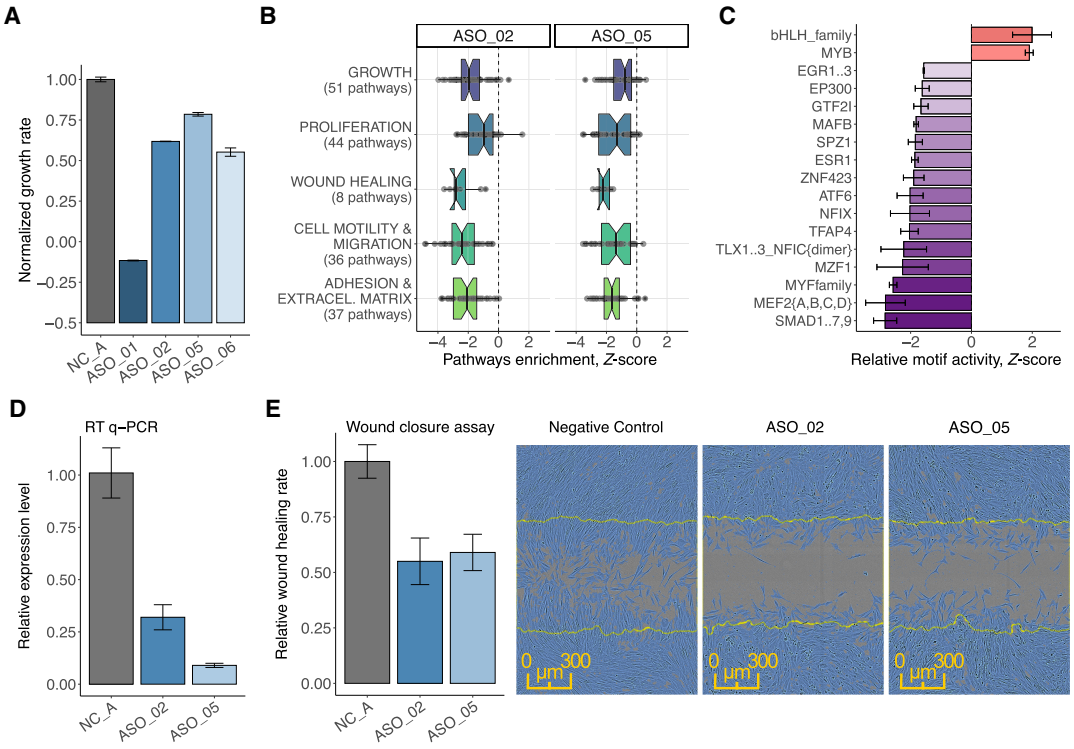


Figure 4. *ZNF213-AS1* regulates cell growth, migration, and proliferation. (*A*) Normalized growth rate across four distinct ASOs (in duplicate) targeting *ZNF213-AS1* as compared to six negative control samples (shown in gray). (*B*) Enrichment of biological pathways associated with growth, proliferation, wound healing, migration, and adhesion for ASO_02 and ASO_05. (*C*) Most consistently down- and up-regulated transcription factor binding motifs including those for transcription factors known to modulate growth, migration, and proliferation such as for example EGR family, EP300, GTF2I. (*D*) Knockdown efficiency measured by RT-qPCR after wound closure assay (72 h posttransfection) showing sustained suppression (65%–90%) of *ZNF213-AS1*. (*E*) Transfected, replated, and mitomycin C (5 µg/mL)-treated HDF cells were scratched and monitored in the Incucyte imaging system. Relative wound closure rate calculated during the 24 h postscratching shows 40%–45% reduction for the two targeting ASOs (ASO_02 [$N=10$] and ASO_05 [$N=13$]) as compared to NC_A transfection controls ($N=33$, shown in gray) and the representative images of wound closure assay 16 h postscratching.

exhibited by using ASOs alone suggests that lncRNAs, in part, are essential regulatory elements in cells. Yet, our study generally warrants a careful assessment of specific findings from different knockdown technologies, including CRISPR-inhibition, and demonstrates a requirement of using multiple replicates in a given target per each modality.

ZNF213–AS1 is associated with cell growth and migration

Extensive molecular and cellular phenotype data for each ASO knockdown can be explored using our portal https://fantom.gsc.riken.jp/zenbu/reports/#FANTOM6. As an example of an lncRNA associated with cell growth and morphology (Fig. 2G), we showcase *ZNF213-AS1* (*RP11-473M20.14*). This lncRNA is highly conserved in placental mammals, moderately expressed (~eight CAGE tags per million) in HDFs, and enriched in the chromatin-bound fraction. Four distinct ASOs (ASO_01, ASO_02, ASO_05, and ASO_06) strongly suppressed expression of *ZNF213-AS1*, whereas expression of the *ZNF213* sense gene was not significantly affected in any of the knockdowns. The four ASOs caused varying degrees of cell growth inhibition (Fig. 4A). ASO_01 and ASO_06 showed a reduction in cell number, as well as an up-regulation of apoptosis and immune and defense pathways in GSEA, suggesting cell death. While cell growth inhibition observed for ASO_02 and ASO_05 was confirmed by MKI67 marker staining (Fig. 2D; Supplemental Table S7), the molecular phenotype revealed suppression of GSEA pathways related to cell growth, as well as to cell proliferation, motility, and extracellular structure organization (Fig. 4B). We also observed consistent down-regulation of motifs related to the observed cellular phenotype, for example, EGR1, EP300, SMAD1…7,9 (Fig. 4C).

As cell motility pathways were affected by the knockdown, we tested whether *ZNF213-AS1* could influence cell migration. Based on the wound-closure assay after transient cell growth inhibition (mitomycin C and serum starvation) (Supplemental Fig. S2F,G), we observed a substantial reduction of wound closure rate (~40% over a 24-h period) in the *ZNF213-AS1*-depleted HDFs (Fig. 4D, E). The reduced wound healing rate should thus mainly reflect reduced cell motility, further confirming affected motility pathways predicted by the molecular phenotype.

As these results indicated a potential role of *ZNF213-AS1* in cell growth and migration, we used FANTOM CAT Recount 2 atlas (Imada et al. 2020), which incorporates The Cancer Genome Atlas (TCGA) data set (Collado-Torres et al. 2017), and found relatively higher expression of *ZNF213-AS1* in acute myeloid leukemia (LAML) and in low-grade gliomas (LGG) as compared to other cancers (Supplemental Fig. S6A). In LAML, the highest expression levels were associated with mostly undifferentiated states, whereas in LGG, elevated expression levels were found in oligodendrogliomas, astrocytomas, and in IDH1 mutated tumors, suggesting that *ZNF213-AS1* is involved in modulating

differentiation and proliferation of tumors (Supplemental Fig. S6B–E). Further, univariate Cox proportional hazard analysis as well as Kaplan-Meier curves for LGG were significant and consistent with our findings (HR = 0.61, BH FDR = 0.0079). The same survival analysis on LAML showed a weak association with poor prognostic outcome, but the results were not significant (Supplemental Fig. S6F,G).

RP11–398K22.12 (*KHDC3L–2*) regulates *KCNQ5* in *cis*

Next, we investigated in detail *RP11-398K22.12* (ENSG00000229852), where the knockdowns by two independent ASOs (ASO_03, ASO_05) successfully reduced the expression of the target lncRNA (67%–82% knockdown efficiency, respectively) and further down-regulated its neighboring genes, *KCNQ5* and its divergent partner novel lncRNA *CATG00000088862.1* (Fig. 5A). Although the two genomic loci occupy Chromosome 6 and are 650 kb away, Hi-C analysis (Supplemental Methods; Supplemental Fig. S7; Supplemental Table S8) showed that they are located within the same topologically associated domain (TAD) and spatially colocalized (Fig. 5B). Moreover, chromatin-enrichment and single molecule RNA-FISH of *RP11-398K22.12* (Fig. 5C; Supplemental Table S9) suggested its highly localized *cis*-regulatory role.

In FANTOM5 (Hon et al. 2017), expression levels of *RP11-398K22.12*, *KCNQ5*, and *CATG00000088862.1* were enriched in brain and nervous system samples, whereas GTEx (The GTEx Consortium 2015) showed their highly specific expression in the brain, particularly in the cerebellum and the cerebellar hemisphere (Fig. 5D). GTEx data also showed that expression of *RP11-398K22.12* was highly correlated with the expression of *KCNQ5* and *CATG00000088862.1* across neuronal tissues (Fig. 5E,F), with the exception of cerebellum and the cerebellar hemisphere, potentially due to relatively lower levels of *KCNQ5* and *CATG00000088862.1*, whereas levels of *RP11-398K22.12* remained relatively higher. Additionally, we found an eQTL SNP (rs14526472) overlapping with *RP11-398K22.12* and regulating expression of *KCNQ5* in brain caudate ($P = 4.2 \times 10^{-6}$; normalized effect size −0.58). All these findings indicate that *RP11-398K22.12* is implicated in the nervous system by maintaining the expression of *KCNQ5* and *CATG000 00088862.1* in a *cis*-acting manner.

Discussion

This study systematically annotates lncRNAs through molecular and cellular phenotyping by selecting 285 lncRNAs from human dermal fibroblasts across a wide spectrum of expression, conservation levels and subcellular localization enrichments. Using ASO technology allowed observed phenotypes to be associated to the lncRNA transcripts, whereas, in contrast, CRISPR-based approaches may synchronically influence the transcription machinery at the site of the divergent promoter or affect regulatory elements of the targeted DNA site. Knockdown efficiencies obtained with ASOs were observed to be independent of lncRNA expression levels, subcellular localization, and of their genomic annotation, allowing us to apply the same knockdown technology to various classes of lncRNAs.

We investigated the *cis*-regulation of nearby divergent promoters, which has been reported as one of the functional roles of lncRNA (Luo et al. 2016). However, in agreement with previous studies (Guttman et al. 2011), we did not observe general patterns in the expression response of divergent promoters (Supplemental Fig. S3B). Recent studies suggest that transcription of lncRNA loci that do not overlap with other transcription units may influence RNA polymerase II occupancy on neighboring promoters and gene bodies (Engreitz et al. 2016a; Cho et al. 2018). Thus, it is plausible that transcription of targeted lncRNA was maintained, despite suppression of mature or nascent transcripts using ASOs. This further suggests that the functional responses described in this study are due to interference of processed transcripts present either in the nucleus, the cytoplasm, or both. Although it is arguable that ASOs may interfere with general transcription by targeting the 5′-end of nascent transcripts and thus releasing RNA polymerase II, followed by exonuclease-mediated decay and transcription termination (aka "torpedo model") (Proudfoot 2016), most of the ASOs were designed across the entire length of the transcript. Since we did not broadly observe dysregulation in nearby genes, interference of transcription or splicing activity is less likely to occur.

We observed a reduction in cell growth for ∼7.7% of our target lncRNA genes, which is in line with previous experiments using CRISPRi-pooled screening, which reported 5.9% (in iPS cells) of lncRNAs exhibiting a cell growth phenotype (Liu et al. 2017). Although these rates are much lower than for protein-coding genes (Sokolova et al. 2017), recurrent observations of cell growth phenotypes (including cell death) strongly suggest that a substantial fraction of lncRNAs play an essential role in cellular physiology and viability. Further, when applying image-based analysis, we found that lncRNAs affect cell morphologies (Fig. 2G), which has not been so far thoroughly explored.

Several lncRNAs such as *MALAT1*, *NEAT1*, and *FIRRE* have been reported to orchestrate transcription, RNA processing, and gene expression (Kopp and Mendell 2018) but are not essential for mouse development or viability. These observations advocate for assays that can comprehensively profile the molecular changes inside perturbed cells. Therefore, in contrast to cell-based assays, functional elucidation via molecular phenotyping provides comprehensive information that cannot be captured by a single phenotypic assay. Herein, the number of overlapping differentially expressed genes between two ASOs of the same lncRNA targets indicated that 10.9% of lncRNAs exert a reproducible regulatory function in HDF.

Although the features of selected lncRNAs are generally similar to those of other lncRNAs expressed in HDFs (Fig. 1B–D), the cell-type-specific nature of lncRNAs and the relatively small sampling size (119 lncRNAs with knockdown transcriptome profiles) used in our study may not fully represent the whole extent of lncRNA in other cell types. However, lncRNA targets that did not exhibit a molecular phenotype may be biologically relevant in other cell types or cell states (Li and Chang 2014; Liu et al. 2017). At the same time, our results showed that particular lncRNAs expressed broadly in other tissues (e.g., in the human brain) were functional in HDFs (such as *RP11-398K22.12*). Although the exact molecular mechanisms of *RP11-398K22*.12 are not yet fully understood, its potential role in HDFs suggests that lncRNAs may be functionally relevant across multiple tissues in spite of the cell-type-specific expression of lncRNAs.

Further, we used siRNA technology to knockdown lncRNA targets as a method for independent validation. When comparing the transcriptomes perturbed by ASOs and siRNAs, concordance was observed only for three out of nine lncRNAs. This discrepancy is likely due to different modes of actions of the two technologies. Whereas ASOs invoke RNase H-mediated cleavage, primarily active

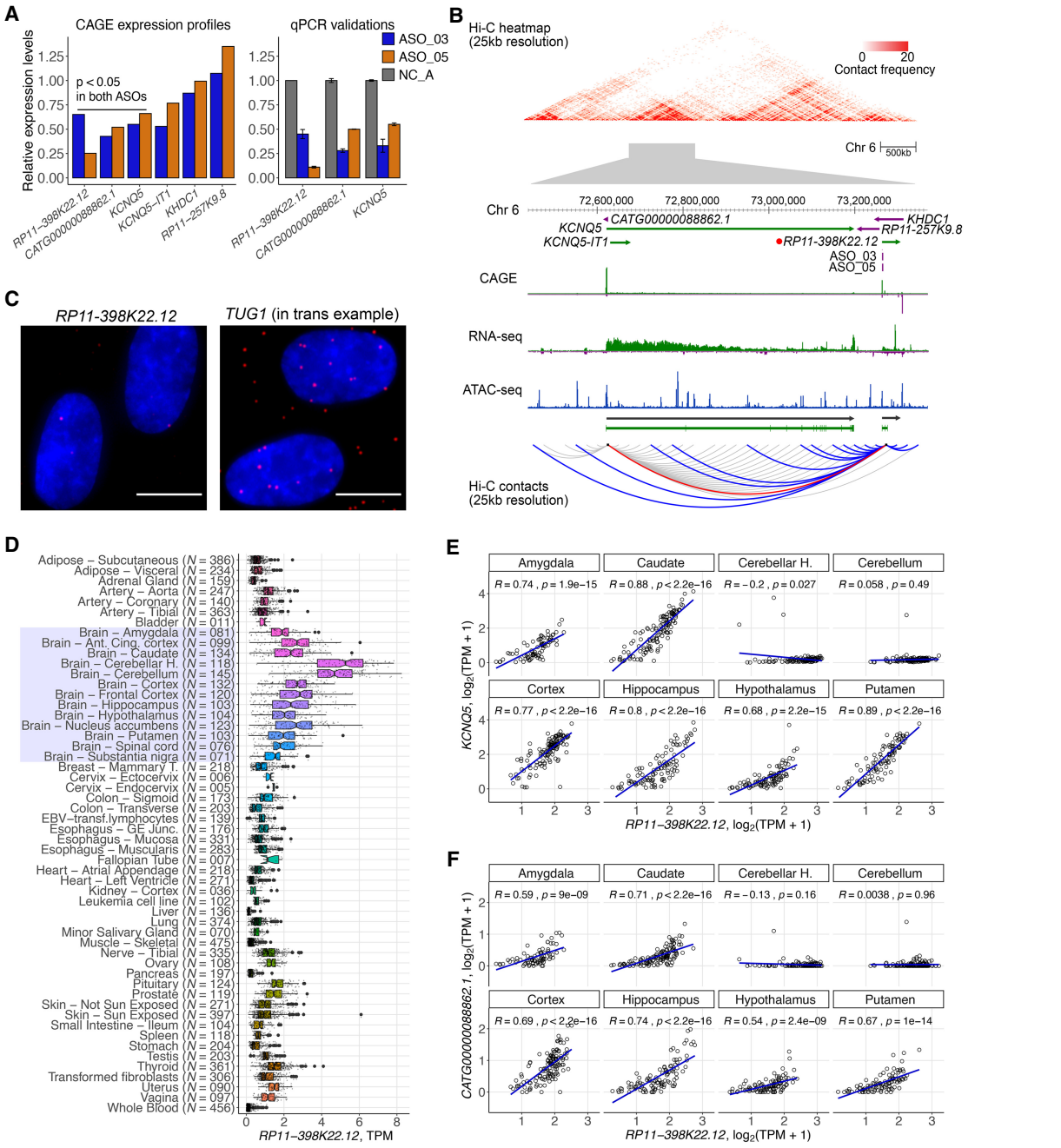


Figure 5. *RP11-398K22.12* down-regulates *KCNQ5* and *CATG00000088862.1* in *cis*. (*A*) Changes in expression levels of detectable genes in the same topologically associated domain (TAD) as *RP11-398K22.12* based on Hi-C analysis. Both *KCNQ5* and *CATG00000088862.1* are down-regulated ($P<0.05$) upon the knockdown of *RP11-398K22.12* by two independent ASOs in CAGE analysis (*left*) as further confirmed with RT-qPCR (*right*). (*B*) (*Top*) Representation of the chromatin conformation in the 4-Mb region proximal to the TAD containing *RP11-398K22.12,* followed by the locus gene annotation, CAGE, RNA-seq, and ATAC-seq data for native HDFs. (*Bottom*) Schematic diagram showing Hi-C predicted contacts of *RP11-398K22.12* (blue) and *KCNQ5* (gray) (25-kb resolution, frequency ≥ 5) in HDF cells. Red line indicates *RP11-398K22.12* and *KCNQ5* contact. (*C*) FISH image for *RP11-398K22.12*, suggesting proximal regulation. *TUG1* FISH image (suggesting *trans* regulation) is included as a comparison; (bar = 10 µm). (*D*) GTEx atlas across 54 tissues ($N=9662$ samples) shows relatively high expression levels of *RP11-398K22.12* in 13 distinct brain regions samples (highlighted). (*E*) Expression correlation for *RP11-398K22.12* and *KCNQ5* in eight out of 13 distinct brain regions, as highlighted in *D*. (*F*) Expression correlation for *RP11-398K22.12* and *CATG00000088862.1* in eight out of 13 distinct brain regions, as highlighted in *D*.

in the nucleus, the siRNAs use the RNA-inducing silencing complex (RISC) mainly active in the cytoplasm. LncRNAs are known to function in specific subcellular compartments (Chen 2016) and their maturity, secondary structures, isoforms, and functions could be vastly different across compartments (Johnsson et al. 2013). Since the majority of functional lncRNAs are reported to be inside the nucleus (Palazzo and Lee 2018; Sun et al. 2018), ASO-mediated knockdowns, which mainly target nuclear RNAs,

are generally more suitable for functional screenings of our lncRNA (62% found in the nuclear compartment). Besides, the dynamics of secondary effects mediated by different levels of knockdown from different technologies are likely to be observed as discordance when considering the whole transcriptome, where this kind of discordance has been reported previously (Stojic et al. 2018). In contrast, in the MKI67 assay, where only a single feature such as growth phenotype is assayed, siRNA knockdown revealed higher reproducibility with ASO knockdown. This suggested that the growth phenotype might be triggered by different specific pathways in ASO- and siRNA-knockdowns.

Previous studies suggest that lncRNAs regulate gene expression in *trans* epigenetically, via direct or indirect interaction with regulators such as DNMT1 (Di Ruscio et al. 2013) or by directly binding to DNA (triplex) (Mondal et al. 2015) or other RNA-binding proteins (Tichon et al. 2016). Analysis of cellular localization by fractionation followed by RNA-seq and in situ hybridization can indicate whether a given lncRNA may act in *trans* by quantifying its abundance in the nuclear soluble fraction as compared to cytoplasm. Although most lncRNAs in the nuclear soluble fraction may affect pathways associated with chromatin modification, additional experiments to globally understand their interaction partners will elucidate the molecular mechanism behind *trans*-acting lncRNAs (Li et al. 2017; Sridhar et al. 2017).

In summary, our study highlights the functional importance of lncRNAs regardless of their expression, localization, and conservation levels. Molecular phenotyping is a powerful and generally more sensitive to knockdown-mediated changes platform to reveal the functional relevance of lncRNAs that cannot be observed based on the cellular phenotypes alone. With additional molecular profiling techniques, such as RNA duplex maps in living cells to decode common structural motifs (Lu et al. 2016), and Oxford Nanopore Technology (ONT) to annotate the full-length variant isoforms of lncRNAs (Hardwick et al. 2019), the structure-to-functional relationship of lncRNAs may be elucidated further in the future.

Methods

Gene models and lncRNA target selections

The gene models used in this study were primarily based on the FANTOM CAGE-associated transcriptome (CAT) at permissive level as defined previously (Hon et al. 2017). From this merged assembly, there were ~2000 lncRNAs robustly expressed in HDFs (TPM ≥ 1). However, we selected lncRNA knockdown targets in an unbiased manner to broadly cover various types of lncRNAs (TPM ≥ 0.2). Briefly, we first identified a list of the lncRNA genes expressed in HDFs, with RNA-seq expression at least 0.5 fragments per kilobase per million and CAGE expression at least 1 tag per million. Then, we manually inspected each lncRNA locus in the ZENBU genome browser for (1) its independence from neighboring genes on the same strand (if any), (2) support from RNA-seq (for exons and splicing junctions) and CAGE data (for TSSs) of its transcript models, and (3) support from histone marks at TSSs for transcription initiation (H3K27ac) and along the gene body for elongation (H3K36me3), from the Roadmap Epigenomics Consortium (Roadmap Epigenomics Consortium et al. 2015). A representative transcript model, which best represents the RNA-seq signal, was manually chosen from each locus for design of antisense oligonucleotides. In total, 285 lncRNA loci were chosen for ASO suppression. Additional controls (*NEAT1*, protein coding genes) (Supplemental Table S1) were added, including *MALAT1* as an experimental control. For details, please refer to the Supplemental Methods.

ASO design

ASOs were designed as RNase H-recruiting locked nucleic acid (LNA) phosphorothioate gapmers with a central DNA gap flanked by 2–4 LNA nucleotides at the 5′ and 3′ ends of the ASOs. For details, please refer to the Supplemental Methods.

Automated cell culturing, ASO transfection, and cell harvesting

Robotic automation (Hamilton) was established to provide a stable environment and accurate procedural timing control for cell culturing and transfection. In brief, trypsin-EDTA detachment, cell number and viability quantification, cell seeding, transfection, and cell harvesting were performed with automation. All transfections were divided into 28 runs on a weekly basis. ASO transfection was performed with duplication. In each run, there were 16 independent transfections with ASO negative control A (NC_A, Exiqon) and 16 wells transfected with an ASO targeting *MALAT-1* (Exiqon).

The HDF cells were seeded in 12-well plates with 80,000 cells in each well 24 h prior to the transfection. A final concentration of 20 nM ASO and 2 µL Lipofectamine RNAiMAX (Thermo Fisher Scientific) were mixed in 200 µL Opti-MEM (Thermo Fisher Scientific). The mixture was incubated at room temperature for 5 min and added to the cells, which were maintained in 1 mL complete medium. The cells were harvested 48 h posttransfection by adding 200 µL RLT buffer from the RNeasy 96 kit (Qiagen) after PBS washing. The harvested lysates were kept at −80°C. RNA was extracted from the lysate for real-time quantitative RT-PCR (Supplemental Methods).

ASO transfection for real-time imaging

The HDF cells were transfected manually in 96-well plates to facilitate high-throughput real-time imaging. The cells were seeded 24 h before transfection at a density of 5200 cells per well. A final concentration of 20 nM ASO and 2 µL Lipofectamine RNAiMAX (Thermo Fisher Scientific) were mixed in 200 µL Opti-MEM (Thermo Fisher Scientific). After incubating at room temperature for 5 min, 18 µL of the transfection mix was added to 90 µL complete medium in each well. The ASOs were divided into 14 runs and transfected in duplicate. Each plate accommodated six wells of NC_A control, two wells of *MALAT1* ASO control, and two wells of mock-transfection (Lipofectamine alone) control.

Phase-contrast images of transfected cells were captured every 3 h for 2 d with three fields per well by the Incucyte live-cell imaging system (Essen Bioscience). The confluence in each field was analyzed by the Incucyte software. The mean confluence of each well was taken along the timeline until the mean confluence of the NC_A control in the same plate reached 90%. The growth rate in each well was calculated as the slope of a linear regression. A normalized growth rate of each replicate was calculated as the growth rate divided by the mean growth rate of the six NC_A controls from the same plate. Negative growth rate was derived when cells shrink and/or detach. As these rates of cell depletion could not be normalized by the rate of growth, negative values were maintained to indicate severe growth inhibition. Student's *t*-test was performed between the growth rate of the duplicated samples and the six NC_A controls, assuming equal variance.

Cell morphology quantification

For each transfection, a representative phase-contrast image at a single time point was exported from the Incucyte time-series. These raw images were first transformed to probability maps of cells by pixel classification using ilastik (1.3.2) (Berg et al. 2019). The trained model was then applied to all images where the predicted probability maps of cells (grayscale, 16 bits tiff format) were subsequently used for morphology quantification in CellProfiler (3.1.5) (Carpenter et al. 2006). For details, please refer to the Supplemental Methods.

MKI67 staining upon lncRNA knockdown

For the selected four lncRNA targets showing >25% growth inhibition, we used two siRNAs and two ASOs with independent sequences. The transfected cells were fixed by adding prechilled 70% ethanol and incubated at −20°C. The cells were washed with FACS buffer (2% FBS in PBS, 0.05% NaN3) twice. FITC-conjugated MKI67 (20Raj1, eBioscience) was applied to the cells and subjected to flow cytometric analysis. Knockdown efficiency by siRNA was determined by real-time quantitative RT-PCR using the same three primer pairs as for ASO knockdown efficiency. For details, please refer to the Supplemental Methods.

Wound closure assay

The HDF cells were transfected with 20 nM ASO as described earlier in 12-well plates. The cells were replated at 24 h posttransfection into a 96-well ImageLock plate (Essen BioScience) at a density of 20,000 cells per well. At 24 h after seeding, cells form a spatially uniform monolayer with 95%–100% cell confluence. The cells were incubated with 5 µg/mL mitomycin C for 2 h to inhibit cell division. Then, medium was refreshed and a uniform scratch was created in each well by the WoundMaker (Essen BioScience). The closure of the wound was monitored by Incucyte live-cell imaging system (Essen Bioscience) every 2 h for 24 h. The RNA was harvested after the assay for real-time quantitative RT-PCR. For details, please refer to the Supplemental Methods.

Cap analysis of gene expression (CAGE)

Four micrograms of purified RNA were used to generate libraries according to the nAnT-iCAGE protocol (Murata et al. 2014). For details, please refer to the Supplemental Methods.

Chromosome conformation capture (Hi-C)

Hi-C libraries were prepared essentially as described previously (Lieberman-Aiden et al. 2009; Fraser et al. 2015a) with minor changes to improve the DNA yield of Hi-C products (Fraser et al. 2015b). For details, please refer to the Supplemental Methods.

Data access

All raw and processed sequencing data generated in this study have been submitted to the DNA Data Bank of Japan (DDBJ; https://www.ddbj.nig.ac.jp/) under accession numbers DRA008311, DRA008312, DRA008436, and DRA008511 or can be accessed through the FANTOM6 project portal https://fantom.gsc.riken.jp/6/datafiles. The analysis results can be downloaded from https://fantom.gsc.riken.jp/6/suppl/Ramilowski_et_al_2020/data/ and interactively explored using our in-house portal https://fantom.gsc.riken.jp/zenbu/reports/#FANTOM6.

Competing interest statement

The authors declare no competing interests.

Acknowledgments

We thank Linda Kostrencic, Hiroto Atsui, Emi Ito, Nobuyuki Takeda, Tsutomu Saito, Teruaki Kitakura, Yumi Hara, Machiko Kashiwagi, and Masaaki Furuno at RIKEN Yokohama for assistance in arranging collaboration agreements, ethics applications, computational infrastructure, and the FANTOM6 meetings. We also thank RIKEN GeNAS for generation and sequencing of the CAGE libraries and subsequent data processing. FANTOM6 was made possible by a Research Grant for RIKEN Center for Life Science Technology, Division of Genomic Technologies (CLST DGT) and RIKEN Center for Integrative Medical Sciences (IMS) from MEXT, Japan. I.V.K. and I.E.V. were supported by Russian Foundation for Basic Research (RFBR) 18-34-20024, B.B. is supported by the fellowship 2017FI_B00722 from the Secretaria d'Universitats i Recerca del Departament d'Empresa i Coneixement (Generalitat de Catalunya) and the European Social Fund (ESF), A. Favorov was supported by National Institutes of Health (NIH) P30 CA006973 and RFBR 17-00-00208, D.G. is supported by a "la Caixa"-Severo Ochoa pre-doctoral fellowship (LCF/BQ/SO15/52260001), E.L.I. and L.M. were supported by NIH National Cancer Institute Grant R01CA200859 and Department of Defense (DOD) award W81XWH-16-1-0739, M.K.-S. was supported by Versus Arthritis UK 20298, A.L. was supported by the Swedish Cancer Society, The Swedish Research Council, the Swedish Childhood Cancer fund, Radiumhemmets forsknigsfonder; V.J.M. was supported by the Russian Academy of Sciences Project 0112-2019-0001; Y.A.M. was supported by Russian Science Foundation (RSF) grant 18-14-00240, A.S. was supported by Novo Nordisk Foundation, Lundbeck Foundation, Danish Cancer Society, Carlsberg Foundation, Independent Research Fund Denmark, A.R.R.F. is currently supported by an Australian National Health and Medical Research Council Fellowship APP1154524, M.M.H. was supported by Natural Sciences and Engineering Research Council of Canada (RGPIN-2015-3948), C.S. was supported by the Interuniversity Consortium for Biotechnology (CIB) from the Italian Ministry of Education, University and Research (MIUR) grant n.974, CMPT177780. J. Luginbühl was supported by Japan Society for the Promotion of Science (JSPS) Postdoctoral Fellowship for Foreign Researchers. C.J.C.P. was supported by RIKEN Special Post-Doctoral Research (SPDR) fellowship.

References

Ambasudhan R, Talantova M, Coleman R, Yuan X, Zhu S, Lipton SA, Ding S. 2011. Direct reprogramming of adult human fibroblasts to functional neurons under defined conditions. *Cell Stem Cell* **9:** 113–118. doi:10.1016/j.stem.2011.07.002

Bai J, Yao B, Wang L, Sun L, Chen T, Liu R, Yin G, Xu Q, Yang W. 2019. lncRNA A1BG-AS1 suppresses proliferation and invasion of hepatocellular carcinoma cells by targeting miR-216a-5p. *J Cell Biochem* **120:** 10310–10322. doi:10.1002/jcb.28315

Berg S, Kutra D, Kroeger T, Straehle CN, Kausler BX, Haubold C, Schiegg M, Ales J, Beier T, Rudy M, et al. 2019. ilastik: interactive machine learning for (bio)image analysis. *Nat Methods* **16:** 1226–1232. doi:10.1038/s41592-019-0582-9

Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, Oyama R, Ravasi T, Lenhard B, Wells C, et al. 2005. The transcriptional landscape of the mammalian genome. *Science* **309:** 1559–1563. doi:10.1126/science.1112014

Carpenter AE, Jones TR, Lamprecht MR, Clarke C, Kang I, Friman O, Guertin DA, Chang J, Lindquist RA, Moffat J, et al. 2006. CellProfiler: image

analysis software for identifying and quantifying cell phenotypes. *Genome Biol* **7:** R100. doi:10.1186/gb-2006-7-10-r100
Carrieri C, Cimatti L, Biagioli M, Beugnet A, Zucchelli S, Fedele S, Pesce E, Ferrer I, Collavin L, Santoro C, et al. 2012. Long non-coding antisense RNA controls *Uchl1* translation through an embedded SINEB2 repeat. *Nature* **491:** 454–457. doi:10.1038/nature11508
Chen L-L. 2016. Linking long noncoding RNA localization and function. *Trends Biochem Sci* **41:** 761–772. doi:10.1016/j.tibs.2016.07.003
Cho SW, Xu J, Sun R, Mumbach MR, Carter AC, Chen YG, Yost KE, Kim J, He J, Nevins SA, et al. 2018. Promoter of lncRNA gene *PVT1* is a tumor-suppressor DNA boundary element. *Cell* **173:** 1398–1412.e22. doi:10.1016/j.cell.2018.03.068
Chu C, Zhang QC, da Rocha ST, Flynn RA, Bharadwaj M, Calabrese JM, Magnuson T, Heard E, Chang HY. 2015. Systematic discovery of Xist RNA binding proteins. *Cell* **161:** 404–416. doi:10.1016/j.cell.2015.03.025
Collado-Torres L, Nellore A, Kammers K, Ellis SE, Taub MA, Hansen KD, Jaffe AE, Langmead B, Leek JT. 2017. Reproducible RNA-seq analysis using *recount2*. *Nat Biotechnol* **35:** 319–321. doi:10.1038/nbt.3838
De Hoon M, Shin JW, Carninci P. 2015. Paradigm shifts in genomics through the FANTOM projects. *Mamm Genome* **26:** 391–402. doi:10.1007/s00335-015-9593-8
Di Ruscio A, Ebralidze AK, Benoukraf T, Amabile G, Goff LA, Terragni J, Figueroa ME, De Figueiredo Pontes LL, Alberich-Jorda M, Zhang P, et al. 2013. DNMT1-interacting RNAs block gene-specific DNA methylation. *Nature* **503:** 371–376. doi:10.1038/nature12598
Engreitz JM, Haines JE, Perez EM, Munson G, Chen J, Kane M, McDonel PE, Guttman M, Lander ES. 2016a. Local regulation of gene expression by lncRNA promoters, transcription and splicing. *Nature* **539:** 452–455. doi:10.1038/nature20149
Engreitz JM, Ollikainen N, Guttman M. 2016b. Long non-coding RNAs: spatial amplifiers that control nuclear structure and gene expression. *Nat Rev Mol Cell Biol* **17:** 756–770. doi:10.1038/nrm.2016.126
The FANTOM Consortium, Suzuki H, Forrest ARR, Van Nimwegen E, Daub CO, Balwierz PJ, Irvine KM, Lassmann T, Ravasi T, Hasegawa Y, et al. 2009. The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line. *Nat Genet* **41:** 553–562. doi:10.1038/ng.375
Fraser J, Ferrai C, Chiariello AM, Schueler M, Rito T, Laudanno G, Barbieri M, Moore BL, Kraemer DCA, Aitken S, et al. 2015a. Hierarchical folding and reorganization of chromosomes are linked to transcriptional changes in cellular differentiation. *Mol Syst Biol* **11:** 852. doi:10.15252/msb.20156492
Fraser J, Williamson I, Bickmore WA, Dostie J. 2015b. An overview of genome organization and how we got there: from FISH to Hi-C. *Microbiol Mol Biol Rev* **79:** 347–372. doi:10.1128/MMBR.00006-15
The GTEx Consortium. 2015. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348:** 648–660. doi:10.1126/science.1262110
Gupta RA, Shah N, Wang KC, Kim J, Horlings HM, Wong DJ, Tsai M-C, Hung T, Argani P, Rinn JL, et al. 2010. Long non-coding RNA *HOTAIR* reprograms chromatin state to promote cancer metastasis. *Nature* **464:** 1071–1076. doi:10.1038/nature08975
Guttman M, Rinn JL. 2012. Modular regulatory principles of large non-coding RNAs. *Nature* **482:** 339–346. doi:10.1038/nature10887
Guttman M, Donaghey J, Carey BW, Garber M, Grenier JK., Munson G, Young G, Lucas AB, Ach R, Bruhn L, et al. 2011. lincRNAs act in the circuitry controlling pluripotency and differentiation. *Nature* **477:** 295–300. doi:10.1038/nature10398
Hardwick SA, Bassett SD, Kaczorowski D, Blackburn J, Barton K, Bartonicek N, Carswell SL, Tilgner HU, Loy C, Halliday G, et al. 2019. Targeted, high-resolution RNA sequencing of non-coding genomic regions associated with neuropsychiatric functions. *Front Genet* **10:** 309. doi:10.3389/fgene.2019.00309
Hon C-C, Ramilowski JA, Harshbarger J, Bertin N, Rackham OJL, Gough J, Denisenko E, Schmeier S, Poulsen TM, Severin J, et al. 2017. An atlas of human long non-coding RNAs with accurate 5′ ends. *Nature* **543:** 199–204. doi:10.1038/nature21374
Imada E-L, Sanchez DF, Collado-Torres L, Wilks C, Matam T, Dinalankara W, Stupnikov A, Lobo-Pereira F, Yip C-W, Yasuzawa K, et al. 2020. Recounting the FANTOM CAGE-Associated Transcriptome. *Genome Res* (this issue). doi:10.1101/gr.254656.119
Iyer MK, Niknafs YS, Malik R, Singhal U, Sahu A, Hosono Y, Barrette TR, Prensner JR, Evans JR, Zhao S, et al. 2015. The landscape of long noncoding RNAs in the human transcriptome. *Nat Genet* **47:** 199–208. doi:10.1038/ng.3192
Johnsson P, Ackley A, Vidarsdottir L, Lui W-O, Corcoran M, Grandér D, Morris KV. 2013. A pseudogene long-noncoding-RNA network regulates *PTEN* transcription and translation in human cells. *Nat Struct Mol Biol* **20:** 440–446. doi:10.1038/nsmb.2516
Joung J, Engreitz JM, Konermann S, Abudayyeh OO, Verdine VK, Aguet F, Gootenberg JS, Sanjana NE, Wright JB, Fulco CP, et al. 2017. Genome-scale activation screen identifies a lncRNA locus regulating a gene neighbourhood. *Nature* **548:** 343–346. doi:10.1038/nature23451
Kalluri R. 2016. The biology and function of fibroblasts in cancer. *Nat Rev Cancer* **16:** 582–598. doi:10.1038/nrc.2016.73
Kendall RT, Feghali-Bostwick CA. 2014. Fibroblasts in fibrosis: novel roles and mediators. *Front Pharmacol* **5:** 123. doi:10.3389/fphar.2014.00123
Kopp F, Mendell JT. 2018. Functional classification and experimental dissection of long noncoding RNAs. *Cell* **172:** 393–407. doi:10.1016/j.cell.2018.01.011
Li L, Chang HY. 2014. Physiological roles of long noncoding RNAs: insight from knockout mice. *Trends Cell Biol* **24:** 594–602. doi:10.1016/j.tcb.2014.06.003
Li B, Wang JH-C. 2011. Fibroblasts and myofibroblasts in wound healing: force generation and measurement. *J Tissue Viability* **20:** 108–120. doi:10.1016/j.jtv.2009.11.004
Li X, Zhou B, Chen L, Gou L-T, Li H, Fu X-D. 2017. GRID-seq reveals the global RNA–chromatin interactome. *Nat Biotechnol* **35:** 940–950. doi:10.1038/nbt.3968
Lieberman-Aiden E, van Berkum NL, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, Williams L, et al. 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326:** 289–293. doi:10.1126/science.1181369
Liu SJ, Horlbeck MA, Cho SW, Birk HS, Malatesta M, He D, Attenello FJ, Villalta JE, Cho MY, Chen Y, et al. 2017. CRISPRi-based genome-scale identification of functional long noncoding RNA loci in human cells. *Science* **355:** eaah7111. doi:10.1126/science.aah7111
Liu Y, Cao Z, Wang Y, Guo Y, Xu P, Yuan P, Liu Z, He Y, Wei W. 2018. Genome-wide screening for functional long noncoding RNAs in human cells by Cas9 targeting of splice sites. *Nat Biotechnol* **36:** 1203–1210. doi:10.1038/nbt.4283
Lu Z, Zhang QC, Lee B, Flynn RA, Smith MA, Robinson JT, Davidovich C, Gooding AR, Goodrich KJ, Mattick JS, et al. 2016. RNA duplex map in living cells reveals higher-order transcriptome structure. *Cell* **165:** 1267–1279. doi:10.1016/j.cell.2016.04.028
Luo S, Lu JY, Liu L, Yin Y, Chen C, Han X, Wu B, Xu R, Liu W, Yan P, et al. 2016. Divergent lncRNAs regulate gene expression and lineage differentiation in pluripotent cells. *Cell Stem Cell* **18:** 637–652. doi:10.1016/j.stem.2016.01.024
Ma L, Cao J, Liu L, Du Q, Li Z, Zou D, Bajic VB, Zhang Z. 2019. LncBook: a curated knowledgebase of human long non-coding RNAs. *Nucleic Acids Res* **47:** D128–D134. doi:10.1093/nar/gky960
Mondal T, Subhash S, Vaid R, Enroth S, Uday S, Reinius B, Mitra S, Mohammed A, James AR, Hoberg E, et al. 2015. *MEG3* long noncoding RNA regulates the TGF-β pathway genes through formation of RNA-DNA triplex structures. *Nat Commun* **6:** 7743. doi:10.1038/ncomms8743
Murata M, Nishiyori-Sueki H, Kojima-Ishiyama M, Carninci P, Hayashizaki Y, Itoh M. 2014. Detecting expressed genes using CAGE. *Methods Mol Biol* **1164:** 67–85. doi:10.1007/978-1-4939-0805-9_7
Palazzo AF, Lee ES. 2018. Sequence determinants for nuclear retention and cytoplasmic export of mRNAs and lncRNAs. *Front Genet* **9:** 440. doi:10.3389/fgene.2018.00440
Proudfoot NJ. 2016. Transcriptional termination in mammals: stopping the RNA polymerase II juggernaut. *Science* **352:** aad9926. doi:10.1126/science.aad9926
Quek XC, Thomson DW, Maag JLV, Bartonicek N, Signal B, Clark MB, Gloss BS, Dinger ME. 2015. lncRNAdb v2.0: expanding the reference database for functional long noncoding RNAs. *Nucleic Acids Res* **43:** D168–D173. doi:10.1093/nar/gku988
Quinn JJ, Chang HY. 2016. Unique features of long non-coding RNA biogenesis and function. *Nat Rev Genet* **17:** 47–62. doi:10.1038/nrg.2015.10
Ransohoff JD, Wei Y, Khavari PA. 2018. The functions and unique features of long intergenic non-coding RNA. *Nat Rev Mol Cell Biol* **19:** 143–157. doi:10.1038/nrm.2017.104
Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, et al. 2015. Integrative analysis of 111 reference human epigenomes. *Nature* **518:** 317–330. doi:10.1038/nature14248
Roux BT, Lindsay MA, Heward JA. 2017. Knockdown of nuclear-located enhancer RNAs and long ncRNAs using locked nucleic acid GapmeRs. *Methods Mol Biol* **1468:** 11–18. doi:10.1007/978-1-4939-4035-6_2
Sokolova M, Turunen M, Mortusewicz O, Kivioja T, Herr P, Vähärautio A, Björklund M, Taipale M, Helleday T, Taipale J. 2017. Genome-wide screen of cell-cycle regulators in normal and tumor cells identifies a differential response to nucleosome depletion. *Cell Cycle* **16:** 189–199. doi:10.1080/15384101.2016.1261765

Sridhar B, Rivas-Astroza M, Nguyen TC, Chen W, Yan Z, Cao X, Hebert L, Zhong S. 2017. Systematic mapping of RNA-chromatin interactions in vivo. *Curr Biol* **27:** 602–609. doi:10.1016/j.cub.2017.01.011

Stojic L, Lun ATL, Mangei J, Mascalchi P, Quarantotti V, Barr AR, Bakal C, Marioni JC, Gergely F, Odom DT. 2018. Specificity of RNAi, LNA and CRISPRi as loss-of-function methods in transcriptional analysis. *Nucleic Acids Res* **46:** 5950–5966. doi:10.1093/nar/gky437

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, et al. 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci* **102:** 15545–15550. doi:10.1073/pnas.0506580102

Sun Q, Hao Q, Prasanth KV. 2018. Nuclear long noncoding RNAs: key regulators of gene expression. *Trends Genet* **34:** 142–157. doi:10.1016/j.tig.2017.11.005

Takahashi K, Tanabe K, Ohnuki M, Narita M, Ichisaka T, Tomoda K, Yamanaka S. 2007. Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell* **131:** 861–872. doi:10.1016/j.cell.2007.11.019

Tichon A, Gil N, Lubelsky Y, Havkin Solomon T, Lemze D, Itzkovitz S, Stern-Ginossar N, Ulitsky I. 2016. A conserved abundant cytoplasmic long noncoding RNA modulates repression by Pumilio proteins in human cells. *Nat Commun* **7:** 12209. doi:10.1038/ncomms12209

Ulitsky I. 2016. Evolution to the rescue: using comparative genomics to understand long non-coding RNAs. *Nature Reviews Genetics* **17:** 601–614. doi:10.1038/nrg.2016.85

Volders P-J, Verheggen K, Menschaert G, Vandepoele K, Martens L, Vandesompele J, Mestdagh P. 2015. An update on LNCipedia: a database for annotated human lncRNA sequences. *Nucleic Acids Res* **43:** D174–D180. doi:10.1093/nar/gku1060

Xue Z, Hennelly S, Doyle B, Gulati AA, Novikova IV, Sanbonmatsu KY, Boyer LA. 2016. A G-rich motif in the lncRNA *Braveheart* interacts with a zinc-finger transcription factor to specify the cardiovascular lineage. *Mol Cell* **64:** 37–50. doi:10.1016/j.molcel.2016.08.010

Received July 12, 2019; accepted in revised form June 24, 2020.

***Genome Research* 28:** 285–294 (2018)

Corrigendum: 3′ UTR lengthening as a novel mechanism in regulating cellular senescence

Meng Chen, Guoliang Lyu, Miao Han, Hongbo Nie, Ting Shen, Wei Chen, Yichi Niu, Yifan Song, Xueping Li, Huan Li, Xinyu Chen, Ziyue Wang, Zheng Xia, Wei Li, Xiao-Li Tian, Chen Ding, Jun Gu, Yufang Zheng, Xinhua Liu, Jinfeng Hu, Gang Wei, Wei Tao, and Ting Ni

The authors would like to correct Figure 3, panel J, in which the rightmost upper image of SA-β-gal stained 293T cells following short hairpin RNA (shRNA)-mediated knockdown of *RRAS2* with sh769 (*RRAS2*-KD-sh769) was inadvertently, and due to a labeling error, taken from the same original source image presented in the middle upper panel, which shows increased SA-β-gal activity following *RRAS2* knockdown by a different shRNA (sh646). This correction does not affect any of the conclusions of the article. The corrected image representative of *RRAS2*-KD-sh769 is provided below, and Figure 3 has been updated in the article online.

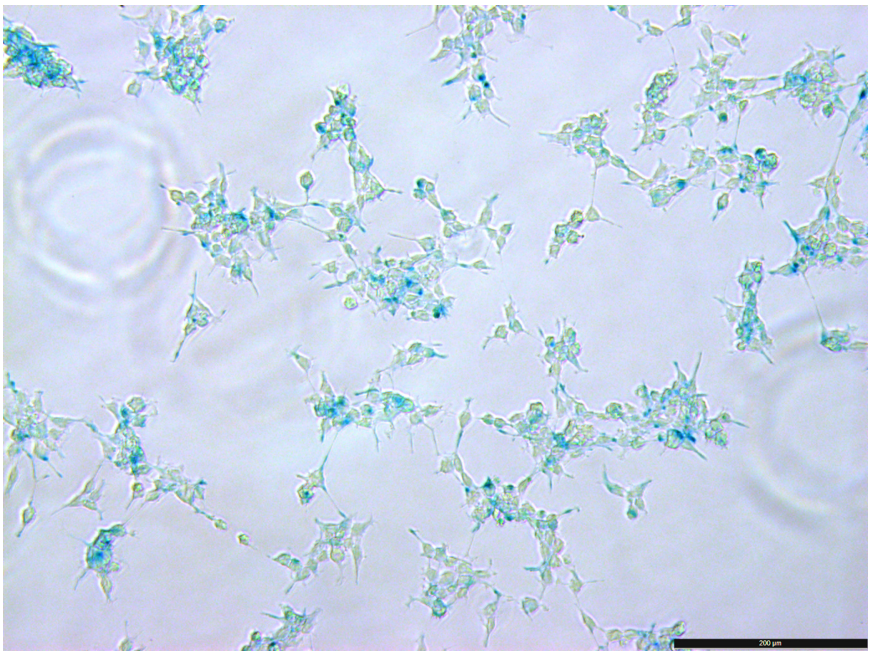


Figure 3. Panel J, rightmost upper image.

The authors thank Ning Yuan Lee for bringing this error to their attention and apologize for any confusion this may have caused.

Additionally, the authors have provided a revised Supplemental Figure S7 file in which the redundant successive Supplemental figure files have been removed. This can be found in the Revised Supplemental Material online.

doi: 10.1101/gr.270165.120

***Genome Research* 30:** 1060–1072 (2020)

Corrigendum: Functional annotation of human long noncoding RNAs via molecular phenotyping

Jordan A. Ramilowski, Chi Wai Yip, Saumya Agrawal, Jen-Chien Chang, Yari Ciani, Ivan V. Kulakovskiy, Mickaël Mendez, Jasmine Li Ching Ooi, John F. Ouyang, Nick Parkinson, Andreas Petri, Leonie Roos, Jessica Severin, Kayoko Yasuzawa, Imad Abugessaisa, Altuna Akalin, Ivan V. Antonov, Erik Arner, Alessandro Bonetti, Hidemasa Bono, Beatrice Borsari, Frank Brombacher, Christopher J.F. Cameron, Carlo Vittorio Cannistraci, Ryan Cardenas, Melissa Cardon, Howard Chang, Josée Dostie, Luca Ducoli, Alexander Favorov, Alexandre Fort, Diego Garrido, Noa Gil, Juliette Gimenez, Reto Guler, Lusy Handoko, Jayson Harshbarger, Akira Hasegawa, Yuki Hasegawa, Kosuke Hashimoto, Norihito Hayatsu, Peter Heutink, Tetsuro Hirose, Eddie L. Imada, Masayoshi Itoh, Bogumil Kaczkowski, Aditi Kanhere, Emily Kawabata, Hideya Kawaji, Tsugumi Kawashima, S. Thomas Kelly, Miki Kojima, Naoto Kondo, Haruhiko Koseki, Tsukasa Kouno, Anton Kratz, Mariola Kurowska-Stolarska, Andrew Tae Jun Kwon, Jeffrey Leek, Andreas Lennartsson, Marina Lizio, Fernando López-Redondo, Joachim Luginbühl, Shiori Maeda, Vsevolod J. Makeev, Luigi Marchionni, Yulia A. Medvedeva, Aki Minoda, Ferenc Müller, Manuel Muñoz-Aguirre, Mitsuyoshi Murata, Hiromi Nishiyori, Kazuhiro R. Nitta, Shuhei Noguchi, Yukihiko Noro, Ramil Nurtdinov, Yasushi Okazaki, Valerio Orlando, Denis Paquette, Callum J.C. Parr, Owen J.L. Rackham, Patrizia Rizzu, Diego Fernando Sánchez Martinez, Albin Sandelin, Pillay Sanjana, Colin A.M. Semple, Youtaro Shibayama, Divya M. Sivaraman, Takahiro Suzuki, Suzannah C. Szumowski, Michihira Tagami, Martin S. Taylor, Chikashi Terao, Malte Thodberg, Supat Thongjuea, Vidisha Tripathi, Igor Ulitsky, Roberto Verardo, Ilya E. Vorontsov, Chinatsu Yamamoto, Robert S. Young, J. Kenneth Baillie, Alistair R.R. Forrest, Roderic Guigó, Michael M. Hoffman, Chung Chau Hon, Takeya Kasukawa, Sakari Kauppinen, Juha Kere, Boris Lenhard, Claudio Schneider, Harukazu Suzuki, Ken Yagi, Michiel J.L. de Hoon, Jay W. Shin, and Piero Carninci

The authors would like to correct the misspelling of an author's name and the inadvertent omission of two affiliations for that author, which are as follows: Christopher J.F. Cameron, Department of Biochemistry, Rosalind and Morris Goodman Cancer Research Center, McGill University, Montréal, Québec H3G 1Y6, Canada and Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, Connecticut 06510, USA.

These updates are reflected in the revised manuscript online.

doi: 10.1101/gr.270330.120

Functional annotation of human long noncoding RNAs via molecular phenotyping

Jordan A. Ramilowski, Chi Wai Yip, Saumya Agrawal, et al.

Genome Res. 2020 30: 1060-1072 originally published online July 27, 2020
Access the most recent version at doi:10.1101/gr.254219.119

Transcriptional landscape of PTEN loss in primary prostate cancer

Eddie Luidy Imada[1,2,3], Diego Fernando Sanchez[2], Wikum Dinalankara[1,2], Thiago Vidotto[4,] Ericka M Ebot[5], Svitlana Tyekucheva[5], Gloria Regina Franco[3], Lorelei Mucci[1], Massimo Loda[1], Edward M Schaeffer[6], Tamara Lotan[2,3], and Luigi Marchionni[1,5,*]

[1] Department of Pathology and Laboratory Medicine, Weill Cornell Medicine, New York, NY, USA

[2] Department of Oncology, Johns Hopkins University School of Medicine, Baltimore, MD, USA

[3] Departamento de Bioquímica e Imunologia, ICB, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brazil

[4] Department of Pathology, Johns Hopkins University School of Medicine, Baltimore, MD, USA

[5] Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA

[6] Department of Urology, Northwestern University, Evanston, IL, US

[*] Correspondence to: lum4003@med.cornell.edu

ABSTRACT

PTEN is the most frequently lost tumor suppressor in primary prostate cancer (PCa) and its loss is associated with aggressive disease. However, the transcriptional changes associated with PTEN loss in PCa have not been described in detail. Here, we applied a meta-analysis approach, leveraging two large PCa cohorts with experimentally validated PTEN and ERG status, to derive a transcriptomic signature of *PTEN* loss, while also accounting for potential confounders due to *ERG* rearrangements. Strikingly, the signature indicates a strong activation of both innate and adaptive immune systems upon *PTEN* loss, as well as an expected activation of cell-cycle genes. Moreover, we made use of our recently developed FC-R2 expression atlas to expand this signature to include many non-coding RNAs recently annotated by the FANTOM consortium. With this resource, we analyzed the TCGA-PRAD cohort, creating a comprehensive transcriptomic landscape of *PTEN* loss in PCa that comprises both the coding and an extensive non-coding counterpart.

Introduction

Previous molecular studies have explored the genomic heterogeneity of prostate adenocarcinomas (PCa) revealing distinct molecular subsets characterized by common genome alterations (1–3). Among these molecular alterations, loss of the tumor suppressor gene phosphatase and tensin homolog (*PTEN*) – which is implicated in the negative-regulation of the PI3K-AKT-mTOR pathway – has been identified as one of the most common genomic drivers of primary PCa (4,5). Since alterations in the PI3K pathway are present in more than 30% of human cancers, the identification of an expression signature associated with *PTEN* loss has been investigated in different tumor contexts, including breast, bladder, lung, and PCa (6,7).

Assessment of *PTEN* status by fluorescence in situ hybridization (FISH) and immunohistochemistry (IHC) in large clinical PCa cohorts has shown a consistent association with adverse pathological features such as high Gleason score, extra-prostatic extension, as well as prognostic value for biochemical recurrence and cancer-related death (4,8). IHC-based assessment of *PTEN* status has been shown to correlate tightly with genomic alterations of the *PTEN* locus and captures not only loss of the gene, but also mutation and epigenetic changes that lead to *PTEN* functional inactivation(4,9,10) and the potential clinical utility of PTEN IHC as a valuable prognostic marker has been demonstrated previously (11–14).

Though PTEN is involved in a myriad of cellular processes spanning cellular proliferation to tumor microenvironment interactions (5), the transcriptional landscape related to *PTEN* expression has not yet been explored in depth, and the role of long non-coding RNAs (lncRNAs) remains elusive (15). These observations, added to the evidence that subtle *PTEN* downregulation can lead to cancer susceptibility (16), demonstrate the important role of *PTEN* in cancer biology but also highlight the need for additional studies.

Similarly, gene rearrangements of the ETS transcription factor, *ERG*, with the androgen-regulated gene Transmembrane Serine Protease 2 (*TMPRSS2*) are present in ~50% of PCa from patients of European descent. *TMPRSS2-ERG* fusion (herein denoted as ERG^{+} for fusion present and ERG^{-} for absence of fusion) has been shown to activate the PI3K-kinase pathway similarly to *PTEN* loss (17), leading to increased proliferation and invasion. Importantly, tumors harboring *TMPRSS2-ERG* rearrangements show an enrichment for *PTEN* loss (17,18). The co-

53 occurrence of these two genomic alterations makes it challenging to dissect the contributions of each to the
54 transcriptomic landscape.

55 The goal of this study was to elucidate the transcriptional landscape of *PTEN* loss in PCa through the analysis
56 of two large and very well clinically-curated cohorts, for which *PTEN* and *ERG* status was assessed by clinical-grade IHC:
57 The Natural History (NH) cohort, in which patients that underwent radical prostatectomy for clinically localized PCa did
58 not receive neoadjuvant therapy or adjuvant hormonal therapy prior to documented distant metastases (19); and the
59 Health Professionals Follow-up Study (HPFS) cohort in which the patients were followed for over 25 years (20). Based
60 on IHC-assessed *PTEN* status for these cohorts, we built a *PTEN*-loss signature highly concordant across the
61 independent datasets, in both presence and absence of *TMPRSS2-ERG* fusion. Overall, this *PTEN*-loss signature was
62 associated with cellular processes associated with aggressive tumor behavior (*e.g.*, increased motility and proliferation)
63 and, surprisingly, with increases in gene sets related to the immune response. In addition, through our recently
64 developed FANTOM-CAT/recount2 (FC-R2) resource (21) and copy-number-variation data, we expanded this signature
65 beyond coding genes and report the non-coding RNA repertory resulting from *PTEN* loss.
66

Methods

Data collection and Immunostaining

All expression data used in this work were gathered from public domain databases. In this work, we made use of three cohorts: FC-R2 TCGA, Natural History (NH), and Health Professionals Follow-up Study (HPFS). Information about each cohort is summarized in Table 1. Information about *PTEN* status by immunohistochemistry for the HPFS cohort was readily available and therefore obtained from the public domain. For NH cohort samples, IHC staining for *PTEN* and *ERG* were performed using a previously validated protocol (22). Last, for TCGA we used the Copy Number Variation (CNV) called by the GISTIC algorithm to define *PTEN* status and expectation-maximization algorithm to define ERG status.

Meta-analysis of NH and HPFS cohorts

We performed a meta-analysis approach using a Bayesian hierarchical multi-level model (BHM) for cross-study detection of differential gene expression implemented in the Bioconductor package XDE (23) on microarray-based cohorts to obtain a *PTEN*-null signature from *PTEN* IHC validated samples. The model was fitted using the delta gp model with empirical starting values and 1000 bootstraps were performed. All remaining parameters were set to default values. This analysis was also performed stratifying the samples by *ERG* status to evaluate the impact of the *ERG* rearrangement in the signature.

Differential expression analysis in the TCGA cohort

A generalized linear model (GLM) approach coupled with empirical Bayes moderation of standard errors and voom precision weigths (24,25) was used to detect differentially expressed genes in the TCGA cohort. The models were adjusted for surrogate variables with the SVA package (26). Adjusted p-values controlling for multiple hypothesis testing were performed using the Benjamini-Hochberg method and genes with false discovery rate (FDR) equal or less than 0.1 were reported (27).

Gene set enrichment analysis (GSEA)

The results from the meta-analysis performed in the NH and HPFS cohort were ranked by the weighted size effect (average of the posterior probability of concordant differential expression multiplied by the Bayesian effect size of

each cohort). The results from the TCGA cohort were ranked by t-statistics. Ranked lists were tested for gene set enrichment. Gene set enrichment analysis (GSEA) was performed using a Monte Carlo adaptive multilevel splitting approach, implemented in the fgsea (28) package. A collection of gene sets (Hallmarks, REACTOME, and GO Biological Processes) were obtained from the Broad Institute MSigDB database. The androgen response gene set was obtained from Scheaffer et al (29). Gene sets with less than 15 and more than 1500 genes were removed from the analysis, except for the GO biological processes whose max size was set to 300 to avoid overly generic gene sets. The enriched pathways were collapsed to maintain only independent ones using the function collapsePathways from fgsea.

Results

Meta-analysis of Natural History and Health Professionals Follow-Up Study cohorts

We sought to obtain a consensus signature of *PTEN* loss that could be reproduced across independent cohorts. We utilized a meta-analysis approach leveraging a multi-level model for cross-study detection of differential gene expression (DGE). We fitted a Bayesian hierarchical model (BHM) for analysis of differential expression across multiple studies that allowed us to aggregate data from two previously described tissue microarray-based cohorts where *PTEN* and *ERG* status was determined by IHC (Table 1 and Figure 1) and we derived a *PTEN*-loss signature (Figure 2). In this analysis, we observed 813 genes for which the differential expression was highly concordant (Bayesian Effect Size (BES) ≥ 1, posterior probability of concordant differential expression (PPCDE) ≥ 0.95) (Table S1).

The consequences of *PTEN* loss on cell cycle regulation and tumor cell invasion has been extensively reported previously (4,30,31). Accordingly, beyond *PTEN* itself, the top DEG genes in our signature reflected this profile (Figure 2 and Table S1). Dermatopontin (*DPT*) (BES = -2.59, PPCDE = 1) and Alanyl membrane aminopeptidase (*ANPEP*) (BES = -2.53, PPCDE = 1) were found down-regulated upon *PTEN* loss. Leucine-Rich Repeat Neuronal 1 (*LRRN1*) was among the genes up-regulated upon *PTEN* loss (BES = 3.36, PPCDE = 1). These and other genes found differentially expressed upon *PTEN* loss have all been shown to be associated with a more aggressive phenotype in several cancer types (5) .

Notably, we found *ERG* among the top upregulated genes in the signature (Figure 2). As expected (18,32,33), *ERG* rearrangement was more common among cases with *PTEN* loss compared to intact *PTEN* in all cohorts (Fisher exact test, $p \leq 0.001$). Given this enrichment, it was not surprising that *ERG* was among the most up-regulated genes in the BHM signature, as well as *PLA2G7*, which has been shown to be among the most highly overexpressed genes in *ERG*-rearranged PCa compared to those lacking *ERG* rearrangements (34). The presence of ERG and ERG-regulated transcripts in the *PTEN*-loss signature suggested that this signature might be confounded by enrichment of *ERG* rearranged tumors among the tumors with PTEN loss.

Since *ERG* rearrangements represent a major driver event in PCa and *PTEN* loss is enriched in *ERG*-rearranged tumors, we next investigated the role of ERG in our PTEN-loss signature. To this end, we repeated the Bayesian hierarchical model for the analysis of differential expression by stratifying the samples by *ERG* status. In the background with *ERG* rearrangement, we observed a similar signature to the previous overall *PTEN*-loss signature, but without the aforementioned *ERG*-associated genes (Supplementary figure S1 and Supplementary table S2). However, in the

absence of *ERG* rearrangement, we could not find any significant differences between samples with or without *PTEN* loss. This was unexpected given that *PTEN* is a powerful tumor suppressor capable of triggering multiple molecular changes.

Extending the PTEN-loss signature

To validate our PTEN loss signatures in an orthogonal cohort, we next examined the TCGA PRAD cohort (35), where *PTEN* status was estimated by genomic copy number (CN) assessment, which was closely aligned with PTEN gene expression (Figure S3). We recently developed a comprehensive expression atlas based on the FANTOM-CAT annotations. This meta-assembly is currently the broadest collection of the human transcriptome (21,36). These gene models include many novel lncRNA categories such as enhancers and promoters, allowing the signature to be further expanded beyond the coding repertoire. We used TCGA expression data from the FC-R2 expression atlas (21) to perform DGE analysis stratified by the *PTEN* status as derived from CN analysis. We also performed the same analysis in a stratified manner as in the HPFS and NH cohorts, using the ERG expression with expectation maximization (EM) algorithm to define ERG status given the bimodal nature of ERG expression in PCa. Interestingly, we were able to detect differential expression between *PTEN*-null and PTEN-intact samples without *ERG* rearrangement in the TCGA cohort, which used high-throughput sequencing as opposed to gene expression microarrays, suggesting that there the lack of signal in the previous analysis can be a reflection of the potential limitations with the later technology.

We observed 521 differentially expressed genes (DEG) when comparing *PTEN*-null and *PTEN*-wild-type samples (FDR $\leq$ 0.01, LogFC $\geq$ 1), of which 257 were coding genes and 264 were non-coding genes (Supplementary Table S3). When stratifying the samples by ERG status, we obtained 435 and 364 DEG in the background with and without ERG rearrangement (Supplementary Table S4 and S5), respectively, with similar proportions of coding and non-coding genes. Using Correspondence-at-the-top (CAT) analysis of the coding genes, we observed a higher concordance than expected by chance between the TCGA PTEN-loss signature and that from the BHM (Figure S4). This confirmed that CN is a reasonable proxy to IHC-staining in TCGA which allowed us to expand this signature beyond coding RNAs.

In this analysis, we were able to detect a variety of lncRNAs that are already known to be involved in PCa development and progression. Notably, several differentially expressed lncRNAs were already reported to be

associated with PCa (37–46) (e.g. *PCA3*, *PCGEM1*, *SCHLAP1*, *KRTAP5-AS1*, *Mir-596*) (Supplementary Table S3-S5). *PCA3* is a prostate-specific lncRNA overexpressed in PCa tissue. Similarly, lncRNA *PCGEM1* expression is increased and highly specific in PCa where it promotes cell growth and it has been associated with high-risk PCa patients (41,42). On the other hand, *KRTAP5-AS1* expression has not been directly associated with PCa.

Also ranked high among lncRNAs differentially expressed were the lncRNAs *SChLAP1* and its uncharacterized antisense neighbor *AC009478.1*. *SchLAP1* is overexpressed in a subset of PCa where it antagonizes the tumor-suppressive function of the SWI/SNF complex and can independently predict poor outcomes (45,46). On the other hand, the role of *AC009478.1* in PCa development is still unknown. Interestingly, *SchLAP1* and *AC009478.1* expression is strongly correlated in the TCGA datasets only in PCa ($R = 0.94$, $p < 2.2e-26$) and bladder cancer ($R = 0.85$, $p < 2.2e-26$) (Figure S5).

Strikingly, a substantial proportion of lncRNAs associated with *PTEN* loss were not yet associated with PCa. Out of the 264 DE non-coding genes, 134 were novel and annotated only in the FANTOM-CAT meta-assembly annotation (Table 2). Among the FANTOM-CAT exclusive genes, those with the highest fold change in close proximity with coding genes were *CATG00000038715*, *CATG00000079217*, and *CATG00000117664* (Figure S6). These genes were mostly expressed in PCa as opposed to other cancer types in the TCGA dataset (Figure 3).

Among the downregulated genes were *CATG00000038715* and *CATG00000079217*. *CATG00000038715* is in close proximity to *CYP4F2* and *CYP4F11*, encoding members of the cytochrome P450 enzyme superfamily. Expression of *CATG00000038715* and *CYP4F2* are highly correlated ($R=0.91$, $p < 2.2e-16$) in PCa, and expression of the former was highly specific for PCa (Figure S7). *CATG00000079217* is in close proximity to the coding gene *FBXL7*, an F-box gene which is a component of the E3 ubiquitin ligase complex. While expression of *FBXL7* and *CATG00000079217* showed only a weak correlation ($R=0.14$, $p < 7.4e-4$), *CATG00000079217* expression was notably higher in PCa and breast cancer than in other cancers, and it was moderately correlated with several PCa biomarkers (e.g. *KLK2*, *KLK3*, *STEAP2*, *PCGEM1*, *SLC45A3*) (41,42,47–51) ($R=0.37-0.57$, $p < 2.2e-16$) in TCGA.

CATG00000117664 was among the most upregulated lncRNA and it is located near *GPR158*, a G protein coupled receptor highly expressed in brain. The expression between *GPR158* were correlated ($R=0.54$, $p < 2.2e-16$), and *CATG00000117664* expression was shown to be highly specific to PCa (52) (Figure S7).

182

183 **PTEN loss induces the innate and adaptive immune system**

184 We performed Gene Set Enrichment Analysis (GSEA) using fgsea (28) and tested both the BHM- and TCGA-generated
185 molecular signatures for enrichment in three collections of the Molecular Signature Database (MSigDB) (53,54):
186 HALLMARKS, REACTOME, and GO Biological Processes (BP). Results were similar in both signatures, with positive
187 enrichment of proliferation and cell cycle-related gene sets (e.g. MYC1 targets, MTORC1 signaling, cell cycle
188 checkpoints, and DNA repair) and both innate and adaptive immune system associated gene sets (e.g. Neutrophil
189 degranulation, MHC antigen presentation, interferon-alpha, and gamma) (Figure 4-5 and Supplementary Table S6-
190 S20). The positive enrichment of MHC antigen presentation, interferon-alpha and -gamma in PTEN-null tumors is
191 consistent with our previous study showing that the absolute density of T-cells is increased in PCa with PTEN loss (55).

192 Since *PTEN*-null tumors are known to have decreased androgen output, which is a strong suppressor of
193 inflammatory immune cells (29,56,57), we hypothesized that this decrease in androgen levels could activate an
194 immune response. We, therefore, performed a GSEA analysis using a collection of androgen-regulated genes from
195 Schaeffer et al. (29) to test if the *PTEN*-null signature was enriched in this gene set. Both the TCGA- and BHM-signature
196 were shown to be positively enriched in genes that were shown to be repressed upon dihydrotestosterone treatment
197 (NES =1.39-155, FDR $\leq$ 0.05) (Figure S8).

198

Discussion

With an estimated prevalence of up to 50%, *PTEN* loss is recognized as one of the major driving events in PCa (58). *PTEN* antagonizes PI3K-AKT/PKB and is a key modulator of the AKT-mTOR signaling pathways which are important in regulating cell growth and proliferation. Accordingly, *PTEN* loss is consistently associated with more aggressive disease features and poor outcomes. Saal and collaborators previously generated a transcriptomic signature of *PTEN* loss in breast cancer (6). While this signature was correlated with worse patient outcomes in breast and other independent cancer datasets, including PCa, the signature unsurprisingly fails to capture key characteristics of PCa such as *ERG*-rearrangement (6,11). Significantly, a transcriptomic signature reflecting the landscape of *PTEN* loss in PCa has not been described to date.

Immunohistochemistry (IHC) assay is a clinically utilized technique to determine the status of the *PTEN* gene, with high sensitivity and specificity for underlying genomic deletions (59) (Figure 1). Therefore, we analyzed transcriptome data from two large PCa cohorts – the Health Professional Follow-up Study (HPFS) and the Natural History (NH) study – for which IHC-based PTEN and ERG status was available (n = 390 and 207, respectively), deriving a *PTEN*-loss gene expression signature specific to PCa (Figure 2 and Supplementary Table S1). Genes that are associated with increased proliferation and invasion in several cancer types, such as *DPT*, *ANPEP* and *LRRN1*, were among the most concordant DEG in this signature. *DPT* has been shown to inhibit cell proliferation through MYC repression and to be down-regulated in both oral and thyroid cancer (60,61). It has also been shown to control cell adhesion and invasiveness, with low expression leading to a worst prognosis (61,62). *ANPEP* is known to play an important role in cell motility, invasion, and metastasis progression (62,63), and lower expression of this gene has been associated with the worst prognosis (64). *LRRN1* is a direct transcriptional target of *MYCN*, and an enhancer of EGFR and IGRF signaling pathway (65). Higher levels of *LRRN1* expression promote tumor cell proliferation, inhibiting cell apoptosis, and play an important role in preserving pluripotency-related proteins through *AKT* phosphorylation (65–67), leading to a poor clinical outcome in gastric and brain cancer.

Notably, *ERG* was shown to be upregulated in our signature, which led us to perform a stratified analysis to avoid capturing signals driven mostly by *ERG* overexpression. Surprisingly, we were not able to detect significant differences by *PTEN* status in the HPFS and NH cohorts, which were quantified by gene expression microarrays, in the ERG^- samples. Conversely, when analyzing the TCGA cohort, we were able to detect significant changes by *PTEN* status

in the *ERG*$^{-}$ samples (Supplementary Tables S3-S5). However, given the known limitations of gene expression microarrays performed on formalin fixed material, such as the limited dynamic range of expression values (68), we believe that the HPFS and NH datasets were limited by the technology employed. Nevertheless, concordance between the BHM- and TCGA- cohorts were similar in both the overall and the *ERG*$^{+}$ background comparison (Supplementary Figure S4).

We observed in the TCGA cohort several lncRNAs that have already been associated with PCa progression were found in our signature. PCA3 acts by a variety of mechanisms such as down-regulation of the oncogene *PRUNE2* and up-regulation of the *PRKD3* gene by acting as a miRNA sponge for *mir-1261* leading to increase proliferation and migration(37,38). Conversely, knockdown of *PCA3* can lead to partial reversion of epithelial-mesenchymal transition (EMT) (39) which can lead to increased cell invasion, motility, and survival (40). Although *KRTAP5-AS1* has not been associated with PCa, it has recently shown that *KRTAP5-AS1* can act as a miRNA sponge for miRNAs, such as *mir-596*, which targets the oncogene *CLDN4* which enhances the invasion capacity of cancer cells and promote EMT (40,43), thereby overexpression of *KRTAP5-AS1* can lead increased levels of *CLDN4* (44). *Mir-596* has also been shown to be overexpressed in response to androgen signaling and associated with anti-androgen therapy resistance (44).

Moreover, many lncRNAs exclusively annotated in the FANTOM-CAT were associated with PTEN-loss and were shown to be expressed mostly in PCa (Figure 3). Since these genes are novel genes without elucidated function, we analyzed potential roles for these genes by looking at coding genes located in the same loci. Among the top DE lncRNAs, genes within proximity to coding genes were *CATG00000038715*, *CATG00000079217*, and *CATG00000117664* (Figure S6) which are positioned in the same loci as *CYP4F2*, *FBXL7*, and *GPR158*, respectively. *CYP4F2* is involved in the process of inactivating and degrading leukotriene B4 (*LTB4*). *LTB4* is a key gene in the inflammatory response that is produced in leukocytes in response to inflammatory mediators and can induce the adhesion and activation of leukocytes on the endothelium.(69). *FBXL7* regulates mitotic arrest by degradation of *AURKA*, which is known to promote inflammatory response and activation of NF-κB (70,71). Likewise, increase expression of *GPR158* is reported to stimulate cell proliferation in PCa cell lines, and it is linked to neuroendocrine differentiation (72).

We consistently observed a strong enrichment in immune response genes and gene sets upon *PTEN* loss (Figure 4 and Supplementary Tables S6-S20). Immune-associated genes (i.e. *GP2* and *PLA2G2A*) were found amongst

the top up-regulated genes in our signature (Figure 2). Positive enrichment of Interferon-alpha- and gamma-response genes (FDR ≤ 0.01) further suggests that a strong immuno-responsive environment, with both innate and adaptive systems activated, is developed in *PTEN*-null tumors (Figure 5). The positive enrichment of MHC class II antigen presentation, neutrophil degranulation, vesicle-mediated transport, and FC receptor pathway-related genes suggests that *PTEN*-null tumors may be immunogenic (Figure 4). This finding was particularly surprising given that *PTEN* is itself a key positive regulator of innate immune response, controlling the import of *IRF3*, which is responsible for IFN production. Accordingly, disruption of PTEN expression has previously been reported to lead to decreased innate immune response (73). Conversely, it has also been hypothesized that the increased genomic instability caused by, or associated with, *PTEN* loss can increase immunogenicity in the tumor micro-enviroment (TME) (74). This finding is of particular interest given that immune-responsive tumors can be good candidates for immunotherapy-based approaches.

Remarkably, despite loss of *PTEN* being associated with higher expression of the immune checkpoint gene programmed death ligand-1 (*PD-L1*) in several cancer types (75,76) this is not true in PCa (77). So far, current immunotherapeutic interventions, such as *PD-1* blockade, in PCa have not been successful. One of the possible reasons is the lack of *PD-L1* expression (77). Therefore, alternative targets must be considered for immunotherapy in PCa. One alternative target is the checkpoint molecule *B7-H3* (*CD276*), whose expression has already been associated with PCa progression and worse prognosis (78) and has been suggested as a target for immunotherapy (79,80). *CD276* was one of the most concordant up-regulated genes in our signature (Figure 2) suggesting that its expression is associated with *PTEN* loss. Interestingly, *B7-H3* expression may be down-regulated by androgens (81).

The effects of androgen on the immune system has already been extensively studied and reviewed (56). Androgens are known to suppress inflammatory immune cells and to impair the development and function of B- and T-cells (57). We, therefore, hypothesized that the decreased levels of androgen in *PTEN*-null TME could lead to an unsuppressed immune system. By testing our signature for enrichment in androgen-related genes (AR) derived from Schaeffer et al. (29), we observed that upon *PTEN*-loss, androgen-sensitive genes that are typically suppressed by DHT are positively enriched, indicating that androgen levels or androgen response in *PTEN*-null tumors may be lower than in their *PTEN*-intact counterparts (Figure S8). This decrease in AR-signaling has been described in *PTEN*-null tumors, in

278 which activation of PI3K pathway inhibits AR activity. (82). Furthermore, AR inhibition activates AKT signaling by
279 inhibiting AKT phosphatase levels further boosting cell proliferation (82), which has also been noted in this study
280 (Figure 3). Finally, in the non-coding repertoire, both *PCA3* and *PCGEM1* are modulated by androgen (83,84) and were
281 down-regulated upon *PTEN* loss which tracks with the observed decreased androgen response in *PTEN*-null tumors
282 (Figure S6 and S8).
283

Conclusion

Altogether, we have generated a highly concordant gene signature for *PTEN* loss in PCa across three independent datasets. We show that this signature was highly enriched in proliferation and cell cycle genes, leading to a more aggressive phenotype upon *PTEN* loss, which is concordant with the literature. Moreover, we have shown that *PTEN* loss is associated with an increase in both innate and adaptive immune response. Although the literature shows that *PTEN* loss usually leads to immuno-suppression, we find evidence that this finding may be reversed in PCa. This observation has potential implications in the context of precision medicine since immune responsive tumors are more likely to respond to immunotherapies. Therefore, *PTEN*-null tumors might benefit more from this approach than *PTEN*-intact tumors. Potentially, *PTEN* status can guide immunotherapy combination with other approaches such as androgen ablation.

Finally, by leveraging the FC-R2 resource, we were able to highlight many lncRNAs that may be associated with PCa progression. Although functional characterization these lncRNAs is beyond the scope of this study, we have shown that these novel lncRNAs are highly specific to PCa and track with several coding mRNAs and lncRNAs already reported to be involved in PCa development and progression, most notably, genes involved in immune response. By providing a PCa-specific signature for PTEN loss, as well as highlighting potential new players, we hope to empower further studies on the mechanisms leading to the development of PCa as well its more aggressive subtypes aiding in the future development of potential biomarkers, drug targets and guide therapies choice.

Acknowledgments

This publication was made possible through support from the National Institutes of Health–National Cancer Institute (NIH-NCI) grants U01CA196390, and R01CA200859; and the U.S. Department of Defense (DoD) office of the Congressionally Directed Medical Research Programs (CDMRP) award W81XWH-16-1-0739 and W81XWH-16-1-0737; and Fundação de Amparo a Pesquisa do Estado de Minas Gerais award BDS-00493-16.

Author contributions statement

L.M. and T.L. conceived the idea; L.M., E.L.I. and T.L. designed the study; E.L.I., D.F.S., W.D., T.L., and L.M. performed the analysis; E.L.I., D.F.S., T.L., T.V., G.R.F., and L.M. interpreted the results; T.L., E.M.E., S.T., L.M., M.L., and E.M.S. provided data and tools; E.L.I., D.F.S., T.L., and L.M. wrote the manuscript; all authors reviewed and approved the manuscript.

References

1. Cancer Genome Atlas Research Network, TCGA. The Molecular Taxonomy of Primary Prostate Cancer. Cell [Internet]. 2015 Nov 5;163(4):1011–25. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26544944

2. Taylor BS, Schultz N, Hieronymus H, Gopalan A, Xiao Y, Carver BS, et al. Integrative genomic profiling of human prostate cancer. Cancer Cell. 2010;18(1):11–22.

3. Baca S, Garraway L. The genomic landscape of prostate cancer. Front Endocrinol (Lausanne) [Internet]. 2012;3:69. Available from: https://www.frontiersin.org/article/10.3389/fendo.2012.00069

4. Jamaspishvili T, Berman DM, Ross AE, Scher HI, De Marzo AM, Squire JA, et al. Clinical implications of PTEN loss in prostate cancer. Nat Rev Urol [Internet]. 2018;15(4):222–34. Available from: http://dx.doi.org/10.1038/nrurol.2018.9

5. Lee Y-R, Chen M, Pandolfi PP. The functions and regulation of the PTEN tumour suppressor: new modes and prospects. Nat Rev Mol Cell Biol [Internet]. 2018;19(9):1–16. Available from: http://dx.doi.org/10.1038/s41580-018-0015-0

6. Saal LH, Johansson P, Holm K, Gruvberger-Saal SK, She Q-BBQ-B, Maurer M, et al. Poor prognosis in carcinoma is associated with a gene expression signature of aberrant PTEN tumor suppressor pathway activity. Proc Natl Acad Sci [Internet]. 2007 May 1;104(18):7564–9. Available from: http://www.pnas.org/content/104/18/7564

7. Ong CW, Maxwell P, Alvi MA, McQuaid S, Waugh D, Mills I, et al. A gene signature associated with PTEN activation defines good prognosis intermediate risk prostate cancer cases. J Pathol Clin Res. 2018;4(2):103–13.

8. Morais CL, Han JS, Gordetsky J, Nagar MS, Anderson AE, Lee S, et al. Utility of PTEN and ERG immunostaining for distinguishing high-grade PIN from intraductal carcinoma of the prostate on needle biopsy. Am J Surg Pathol [Internet]. 2015 Feb;39(2):169–78. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25517949

9. Lotan TL, Wei W, Ludkovski O, Morais CL, Guedes LB, Jamaspishvili T, et al. Analytic validation of a clinical-grade PTEN immunohistochemistry assay in prostate cancer by comparison with PTEN FISH. Nat Genet. 2016;29(8):904–14.

10. Lotan TL, Heumann A, Rico SD, Hicks J, Lecksell K, Koop C, et al. PTEN loss detection in prostate cancer: comparison of PTEN immunohistochemistry and PTEN FISH in a large retrospective prostatectomy cohort. Oncotarget. 2017;8(39):65566–76.

11. Han B, Mehra R, Lonigro RJ, Wang L, Suleman K, Menon A, et al. Fluorescence in situ hybridization study shows association of PTEN deletion with ERG rearrangement during prostate cancer progression. Mod Pathol [Internet]. 2009 Aug 1;22(8):1083–93. Available from: http://www.nature.com/articles/modpathol200969

12. Leapman MS, Nguyen HG, Cowan JE, Xue L, Stohr B, Simko J, et al. Comparing Prognostic Utility of a Single-marker Immunohistochemistry Approach with Commercial Gene Expression Profiling Following Radical Prostatectomy. Eur Urol [Internet]. 2018;74(5):668–75. Available from: http://www.ncbi.nlm.nih.gov/pubmed/30181067

13. Ahearn TU, Pettersson A, Ebot EM, Gerke T, Graff RE, Morais CL, et al. A Prospective Investigation of PTEN Loss and ERG Expression in Lethal Prostate Cancer. J Natl Cancer Inst [Internet]. 2016 Feb;108(2). Available from: http://www.ncbi.nlm.nih.gov/pubmed/26615022

14. Lotan TL, Tomlins SA, Bismar TA, Van der Kwast TH, Grignon D, Egevad L, et al. Report From the International Society of Urological Pathology (ISUP) Consultation Conference on Molecular Pathology of Urogenital Cancers. I. Molecular Biomarkers in Prostate Cancer. Am J Surg Pathol [Internet]. 2020 Jul;44(7):e15–29. Available from: http://www.ncbi.nlm.nih.gov/pubmed/32044806

15. Misawa A, Takayama KI, Inoue S. Long non-coding RNAs and prostate cancer. Cancer Sci. 2017;108(11):2107–14.

16. Alimonti A, Carracedo A, Clohessy JG, Trotman LC, Nardella C, Egia A, et al. Subtle variations in Pten dose determine cancer susceptibility. Nat Genet. 2010;42(5):454–8.

17. King JC, Xu J, Wongvipat J, Hieronymus H, Carver BS, Leung DH, et al. Cooperativity of TMPRSS2-ERG with PI3-kinase pathway activation in prostate oncogenesis. Nat Genet [Internet]. 2009 May;41(5):524–6. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19396167

18. Carver BS, Tran J, Gopalan A, Chen Z, Shaikh S, Carracedo A, et al. Aberrant ERG expression cooperates with loss of PTEN to promote cancer progression in the prostate. Nat Genet [Internet]. 2009 May;41(5):619–24. Available from: http://www.nature.com/doifinder/10.1038/ng.370

19. Ross AE, Johnson MH, Yousefi K, Davicioni E, Netto GJ, Marchionni L, et al. Tissue-based Genomics Augments Post-prostatectomy Risk Stratification in a Natural History Cohort of Intermediate- and High-Risk Men. Eur Urol [Internet]. 2016 Jan;69(1):157–65. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26058959

20. Penney KL, Sinnott JA, Tyekucheva S, Gerke T, Shui IM, Kraft P, et al. Association of prostate cancer risk variants with gene expression in normal and tumor tissue. Cancer Epidemiol Biomarkers Prev [Internet]. 2015 Jan;24(1):255–60. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25371445

21. Imada EL, Sanchez DF, Collado-Torres L, Wilks C, Matam T, Dinalankara W, et al. Recounting the FANTOM CAGE--Associated Transcriptome. Genome Res. 2020;30(7):gr--254656.

22. Lotan TL, Gurel B, Sutcliffe S, Esopi D, Liu W, Xu J, et al. PTEN Protein Loss by Immunostaining: Analytic Validation and Prognostic Indicator for a High Risk Surgical Cohort of Prostate Cancer Patients. Clin Cancer Res [Internet]. 2011 Oct 15;17(20):6563–73. Available from: http://clincancerres.aacrjournals.org/cgi/doi/10.1158/1078-0432.CCR-11-1244

23. Scharpf RB, Tjelmeland H, Parmigiani G, Nobel AB. A Bayesian model for cross-study differential gene expression. J Am Stat Assoc. 2009;104(488):1295–310.

372 24. Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments.
373 Stat Appl Genet Mol Biol [Internet]. 2004;3:Article3. Available from:
374 http://www.ncbi.nlm.nih.gov/pubmed/16646809?dopt=abstract
375 25. Law CW, Chen Y, Shi W, Smyth GK. voom: precision weights unlock linear model analysis tools for RNA-seq read counts.
376 Genome Biol [Internet]. 2014;15(2):R29. Available from: http://genomebiology.biomedcentral.com/articles/10.1186/gb-
377 2014-15-2-r29
378 26. Leek JT, Storey JD. Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS Genet [Internet].
379 2007 Sep;3(9):1724–35. Available from:
380 http://www.plosgenetics.org/article/info%3Adoi%2F10.1371%2Fjournal.pgen.0030161
381 27. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R
382 Stat Soc [Internet]. 1995;Series B,(1):289–300. Available from:
383 http://www.dm.uba.ar/materias/analisis_expl_y_conf_de_datos_de_exp_de_marrays_Mae/2006/1/teoricas/FDR
384 1995.pdf
385 28. Sergushichev AA. An algorithm for fast preranked gene set enrichment analysis using cumulative statistic calculation.
386 BioRxiv [Internet]. 2016;60012. Available from: https://www.biorxiv.org/content/10.1101/060012v1
387 29. Schaeffer EM, Marchionni L, Huang Z, Simons B, Blackman A, Yu W, et al. Androgen-induced programs for prostate
388 epithelial growth and invasion arise in embryogenesis and are reactivated in cancer. Oncogene [Internet]. 2008 Dec
389 4;27(57):7180–91. Available from: http://www.ncbi.nlm.nih.gov/pubmed/18794802
390 30. Leinonen KA, Saramaki OR, Furusato B, Kimura T, Takahashi H, Egawa S, et al. Loss of PTEN Is Associated with Aggressive
391 Behavior in ERG-Positive Prostate Cancer. Cancer Epidemiol Biomarkers Prev [Internet]. 2013 Dec 1;22(12):2333–44.
392 Available from: http://cebp.aacrjournals.org/cgi/doi/10.1158/1055-9965.EPI-13-0333-T
393 31. Yoshimoto M, Ludkovski O, DeGrace D, Williams JL, Evans A, Sircar K, et al. PTEN genomic deletions that characterize
394 aggressive prostate cancer originate close to segmental duplications. Genes, Chromosom Cancer. 2012;51(2):149–60.
395 32. Mehra R, Salami SS, Lonigro R, Bhalla R, Siddiqui J, Cao X, et al. Association of ERG/PTEN status with biochemical recurrence
396 after radical prostatectomy for clinically localized prostate cancer. Med Oncol [Internet]. 2018 Oct 5;35(12):152. Available
397 from: http://www.ncbi.nlm.nih.gov/pubmed/30291535
398 33. Krohn A, Freudenthaler F, Harasimowicz S, Kluth M, Fuchs S, Burkhardt L, et al. Heterogeneity and chronology of PTEN
399 deletion and ERG fusion in prostate cancer. Mod Pathol. 2014;27(12):1612.
400 34. Massoner P, Kugler KG, Unterberger K, Kuner R, Mueller LAJ, Fälth M, et al. Characterization of transcriptional changes in

ERG rearrangement-positive prostate cancer identifies the regulation of metabolic sensors such as neuropeptide Y. PLoS One [Internet]. 2013;8(2):e55207. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23390522

35. Network CGAR, Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet [Internet]. 2013 Oct;45(10):1113–20. Available from: http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=24071849&retmode=ref&cmd=prlinks

36. Hon C-C, Ramilowski JA, Harshbarger J, Bertin N, Rackham OJLL, Gough J, et al. An atlas of human long non-coding RNAs with accurate 5′ ends. Nature [Internet]. 2017 Mar 1;543(7644):199–204. Available from: http://dx.doi.org/10.1038/nature21374

37. Salameh A, Lee AK, Cardó-Vila M, Nunes DN, Efstathiou E, Staquicini FI, et al. PRUNE2 is a human prostate cancer suppressor regulated by the intronic long noncoding RNA PCA3. Proc Natl Acad Sci [Internet]. 2015;112(27):8403–8. Available from: http://www.pnas.org/lookup/doi/10.1073/pnas.1507882112

38. He JH, Li BX, Han ZP, Zou MX, Wang L, Lv YB, et al. Snail-activated long non-coding RNA PCA3 up-regulates PRKD3 expression by miR-1261 sponging, thereby promotes invasion and migration of prostate cancer cells. Tumor Biol [Internet]. 2016;37(12):16163–76. Available from: http://dx.doi.org/10.1007/s13277-016-5450-y

39. Lemos AEGEGEG, Ferreira LB, Batoreu NM, de Freitas PP, Bonamino MH, Gimba ERP. PCA3 long noncoding RNA modulates the expression of key cancer-related genes in LNCaP prostate cancer cells. Tumor Biol [Internet]. 2016;37(8):11339–48. Available from: http://dx.doi.org/10.1007/s13277-016-5012-3

40. Agarwal R, D'Souza T, Morin PJ. Claudin-3 and claudin-4 expression in ovarian epithelial cells enhances invasion and is associated with increased matrix metalloproteinase-2 activity. Cancer Res. 2005;65(16):7378–85.

41. Srikantan V, Zou Z, Petrovics G, Xu L, Augustus M, Davis L, et al. PCGEM1, a prostate-specific gene, is overexpressed in prostate cancer. Proc Natl Acad Sci [Internet]. 2000;97(22):12216–21. Available from: http://www.pnas.org/cgi/doi/10.1073/pnas.97.22.12216

42. Petrovics G, Zhang W, Makarem M, Street JP, Connelly R, Sun L, et al. Elevated expression of PCGEM1, a prostate-specific gene with cell growth-promoting function, is associated with high-risk prostate cancer patients. Oncogene. 2004;23(2):605.

43. Lin X, Shang X, Manorek G, Howell SB. Regulation of the epithelial-mesenchymal transition by claudin-3 and claudin-4. PLoS One. 2013;8(6):e67496.

44. Song YX, Sun JX, Zhao JH, Yang YC, Shi JX, Wu ZH, et al. Non-coding RNAs participate in the regulatory network of CLDN4 via ceRNA mediated miRNA evasion. Nat Commun [Internet]. 2017;8(1):1–16. Available from: http://dx.doi.org/10.1038/s41467-017-00304-1

45. Prensner JR, Iyer MK, Sahu A, Asangani IA, Cao Q, Patel L, et al. The long noncoding RNA SChLAP1 promotes aggressive prostate cancer and antagonizes the SWI/SNF complex. Nat Genet. 2013;45(11):1392–403.

46. Mehra R, Udager AM, Ahearn TU, Cao X, Feng FY, Loda M, et al. Overexpression of the long non-coding RNA SChLAP1 independently predicts lethal prostate cancer. Eur Urol. 2016;70(4):549–52.

47. Kumar-Sinha C, Tomlins SA, Chinnaiyan AM. Recurrent gene fusions in prostate cancer. Nat Rev Cancer. 2008;8(7):497.

48. Nam RK, Zhang WW, Klotz LH, Trachtenberg J, Jewett MAS, Sweet J, et al. Variants of the hK2 protein gene (KLK2) are associated with serum hK2 levels and predict the presence of prostate cancer at biopsy. Clin cancer Res. 2006;12(21):6452–8.

49. Cicek MS, Liu X, Casey G, Witte JS. Role of androgen metabolism genes CYP1B1, PSA/KLK3, and CYP11$α$ in prostate cancer risk and aggressiveness. Cancer Epidemiol Prev Biomarkers. 2005;14(9):2173–7.

50. Whiteland H, Spencer-Harty S, Morgan C, Kynaston H, Thomas DH, Bose P, et al. A role for STEAP2 in prostate cancer progression. Clin \& Exp metastasis. 2014;31(8):909–20.

51. Perner S, Rupp NJ, Braun M, Rubin MA, Moch H, Dietel M, et al. Loss of SLC45A3 protein (prostein) expression in prostate cancer is associated with SLC45A3-ERG gene rearrangement and an unfavorable clinical course. Int J cancer. 2013;132(4):807–12.

52. Patel N, Itakura T, Jeong S, Liao C-P, Roy-Burman P, Zandi E, et al. Expression and Functional Role of Orphan Receptor GPR158 in Prostate Cancer Growth and Progression. Robson CN, editor. PLoS One [Internet]. 2015 Feb 18;10(2):e0117758. Available from: https://dx.plos.org/10.1371/journal.pone.0117758

53. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdottir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. Bioinformatics [Internet]. 2011 Jun 15;27(12):1739–40. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btr260

54. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A [Internet]. 2005 Oct 25;102(43):15545–50. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16199517

55. Kaur HB, Guedes LB, Lu J, Maldonado L, Reitz L, Barber JR, et al. Association of tumor-infiltrating T-cell density with molecular subtype, racial ancestry and clinical outcomes in prostate cancer. Mod Pathol [Internet]. 2018;31(10):1539–52. Available from: http://www.ncbi.nlm.nih.gov/pubmed/29849114

56. Trigunaite A, Dimo J, Jørgensen TN. Suppressive effects of androgens on the immune system. Cell Immunol [Internet]. 2015 Apr;294(2):87–94. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25708485

159 57. Ylitalo EB, Thysell E, Jernberg E, Lundholm M, Crnalic S, Egevad L, et al. Subgroups of Castration-resistant Prostate Cancer
160 Bone Metastases Defined Through an Inverse Relationship Between Androgen Receptor Activity and Immune Response.
161 Eur Urol [Internet]. 2017;71(5):776–87. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27497761
162 58. Wise HM, Hermida MA, Leslie NR. Prostate cancer, PI3K, PTEN and prognosis. Clin Sci. 2017;131(3):197–210.
163 59. Lotan TL, Gurel B, Sutcliffe S, Esopi D, Liu W, Xu J, et al. PTEN protein loss by immunostaining: analytic validation and
164 prognostic indicator for a high risk surgical cohort of prostate cancer patients. Clin Cancer Res [Internet]. 2011 Oct
165 15;17(20):6563–73. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21878536
166 60. Guo Y, Li H, Guan H, Ke W, Liang W, Xiao H, et al. Dermatopontin inhibits papillary thyroid cancer cell proliferation through
167 MYC repression. Mol Cell Endocrinol. 2019;480:122–32.
168 61. Yamatoji M, Kasamatsu A, Kouzu Y, Koike H, Sakamoto Y, Ogawara K, et al. Dermatopontin: a potential predictor for
169 metastasis of human oral cancer. Int J cancer. 2012;130(12):2903–11.
170 62. Ishii K, Usui S, Sugimura Y, Yoshida S, Hioki T, Tatematsu M, et al. Aminopeptidase N regulated by zinc in human prostate
171 participates in tumor cell invasion. Int J cancer. 2001;92(1):49–54.
172 63. Hashida H, Takabayashi A, Kanai M, Adachi M, Kondo K, Kohno N, et al. Aminopeptidase N is involved in cell motility and
173 angiogenesis: its clinical significance in human colon cancer. Gastroenterology. 2002;122(2):376–86.
174 64. Sørensen KD, Abildgaard MO, Haldrup C, Ulhøi BP, Kristensen H, Strand S, et al. Prognostic significance of aberrantly
175 silenced ANPEP expression in prostate cancer. Br J Cancer. 2013;108(2):420–8.
176 65. Hossain S, Takatori A, Nakamura Y, Suenaga Y, Kamijo T, Nakagawara A. LRRN1 enhances EGF-mediated MYCN induction
177 in neuroblastoma and accelerates tumor growth in vivo. Cancer Res. 2012;72(17):4587–96.
178 66. Hossain MS, Ozaki T, Wang H, Nakagawa A, Takenobu H, Ohira M, et al. N-MYC promotes cell proliferation through a direct
179 transactivation of neuronal leucine-rich repeat protein-1 (NLRR1) gene in neuroblastoma. Oncogene. 2008;27(46):6075–
180 82.
181 67. Liao C-H, Wang Y-H, Chang W-W, Yang B-C, Wu T-J, Liu W-L, et al. Leucine-Rich Repeat Neuronal Protein 1 Regulates
182 Differentiation of Embryonic Stem Cells by Post-Translational Modifications of Pluripotency Factors. Stem Cells.
183 2018;36(10):1514–24.
184 68. Wilhelm BT, Landry J-R. RNA-Seq-quantitative measurement of expression through massively parallel RNA-sequencing.
185 Methods [Internet]. 2009 Jul;48(3):249–57. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19336255
186 69. Hardwick JP. Cytochrome P450 omega hydroxylase (CYP4) function in fatty acid metabolism and metabolic diseases.
187 Biochem Pharmacol. 2008;75(12):2263–75.

488 70. Coon TA, Glasser JR, Mallampalli RK, Chen BB. Novel E3 ligase component FBXL7 ubiquitinates and degrades Aurora A,
489 causing mitotic arrest. Cell cycle. 2012;11(4):721–9.
490 71. Katsha A, Soutto M, Sehdev V, Peng D, Washington MK, Piazuelo MB, et al. Aurora kinase A promotes inflammation and
491 tumorigenesis in mice and human gastric neoplasia. Gastroenterology [Internet]. 2013 Dec;145(6):1312-22.e1-8. Available
492 from: http://www.ncbi.nlm.nih.gov/pubmed/23993973
493 72. Fenner A. Prostate cancer: Orphan receptor GPR158 finds a home in prostate cancer growth and progression. Nat Rev Urol
494 [Internet]. 2015 Apr;12(4):182. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25753095
495 73. Li S, Zhu M, Pan R, Fang T, Cao Y-YY, Chen S, et al. The tumor suppressor PTEN has a critical role in antiviral innate immunity.
496 Nat Immunol. 2016;17(3):241.
497 74. Vidotto T, Melo CM, Castelli E, Koti M, dos Reis RB, Squire JA. Emerging role of PTEN loss in evasion of the immune response
498 to tumours. Br J Cancer [Internet]. 2020;122(12):1732–43. Available from: http://dx.doi.org/10.1038/s41416-020-0834-6
499 75. Lastwika KJ, Wilson W, Li QK, Norris J, Xu H, Ghazarian SR, et al. Control of PD-L1 Expression by Oncogenic Activation of the
500 AKT-mTOR Pathway in Non-Small Cell Lung Cancer. Cancer Res [Internet]. 2016 Jan 15;76(2):227–38. Available from:
501 http://www.ncbi.nlm.nih.gov/pubmed/26637667
502 76. Berghoff AS, Kiesel B, Widhalm G, Rajky O, Ricken G, Wöhrer A, et al. Programmed death ligand 1 expression and tumor-
503 infiltrating lymphocytes in glioblastoma. Neuro Oncol [Internet]. 2015 Aug;17(8):1064–75. Available from:
504 http://www.ncbi.nlm.nih.gov/pubmed/25355681
505 77. Martin AM, Nirschl TR, Nirschl CJ, Francica BJ, Kochel CM, van Bokhoven A, et al. Paucity of PD-L1 expression in prostate
506 cancer: innate and adaptive immune resistance. Prostate Cancer Prostatic Dis [Internet]. 2015 Dec;18(4):325–32. Available
507 from: http://www.ncbi.nlm.nih.gov/pubmed/26260996
508 78. Yuan H, Wei X, Zhang G, Li C, Zhang X, Hou J. B7-H3 over expression in prostate cancer promotes tumor cell progression. J
509 Urol [Internet]. 2011 Sep;186(3):1093–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21784485
510 79. Papanicolau-Sengos A, Yang Y, Pabla S, Lenzo FL, Kato S, Kurzrock R, et al. Identification of targets for prostate cancer
511 immunotherapy. Prostate [Internet]. 2019;79(5):498–505. Available from:
512 http://www.ncbi.nlm.nih.gov/pubmed/30614027
513 80. Yang S, Wei W, Zhao Q. B7-H3, a checkpoint molecule, as a target for cancer immunotherapy. Int J Biol Sci [Internet].
514 2020;16(11):1767–73. Available from: http://www.ncbi.nlm.nih.gov/pubmed/32398947
515 81. Benzon B, Zhao SG, Haffner MC, Takhar M, Erho N, Yousefi K, et al. Correlation of B7-H3 with androgen receptor, immune
516 pathways and poor outcome in prostate cancer: an expression-based analysis. Prostate Cancer Prostatic Dis [Internet].

2017;20(1):28–35. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27801901

82. Carver BS, Chapinski C, Wongvipat J, Hieronymus H, Chen Y, Chandarlapaty S, et al. Reciprocal feedback regulation of PI3K and androgen receptor signaling in PTEN-deficient prostate cancer. Cancer Cell. 2011;19(5):575–86.

83. Gezer U, Tiryakioglu D, Bilgin E, Dalay N, Holdenrieder S. Androgen stimulation of PCA3 and miR-141 and their release from prostate cancer cells. Cell J. 2015;16(4):488.

84. Parolia A, Crea F, Xue H, Wang Y, Mo F, Ramnarine VR, et al. The long non-coding RNA PCGEM1 is regulated by androgen receptor activity in vivo. Mol Cancer. 2015;14(1):46.

Figures and Tables

Cohort	PTEN-null	PTEN-intact	N
TCGA	95	321	416
HPFS	91	299	390
Natural History	56	151	207
Total	242	771	1,013

Table 1. Cohorts summary Table shows cohorts summary for the 3 cohorts used in this study: TCGA (only primary tumor samples with high Gistic scores were used); Health Professional Follow-up Study (all); and Natural History cohort (samples with IHC call available). PTEN-null represents samples with PTEN deletion and PTEN-intact regular primary tumors.

	PTEN-null vs PTEN-intact overall	PTEN-null vs PTEN-intact in ERG+	PTEN-null vs PTEN-intact in ERG-
Coding genes	257 (13)	226 (7)	185 (10)
Non-coding genes	264 (134)	209 (117)	179 (82)
Total	521 (137)	435 (124)	364 (92)

Table 2. Summary of differentially expressed genes between PTEN-null and PTEN-intact with logFC ≥ 1 and FDR ≤ 0.01 across different ERG backgrounds. Number in parenthesis shows the number of genes exclusive to the FANTOM-CAT annotations.

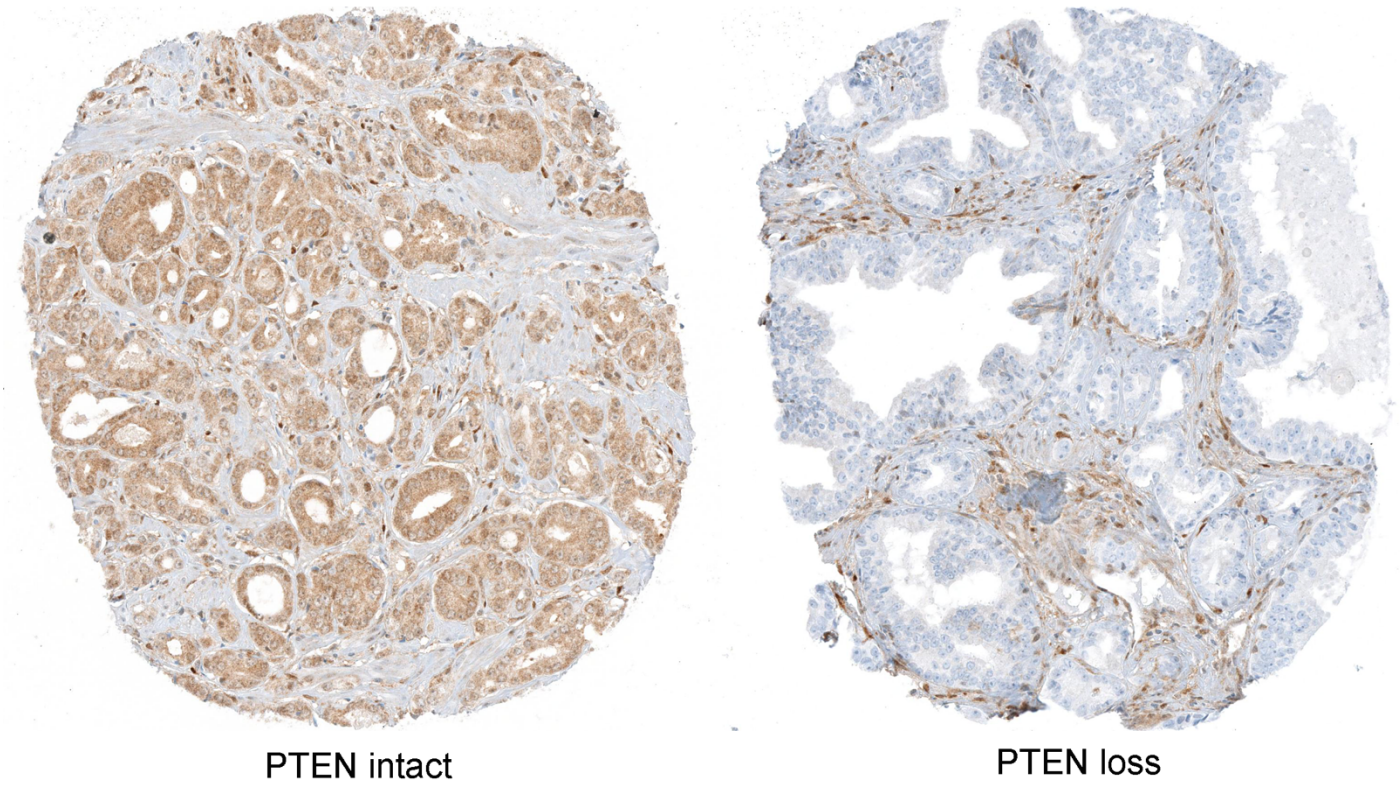


Figure 1. PTEN immunostaining in tissue microarray (TMA) spots from the Natural History Cohort. Left panel: intact PTEN protein is present in all sampled tumor glands (brown chromogen). Right panel: PTEN loss in all sampled tumor glands. Images reduced from 40X.

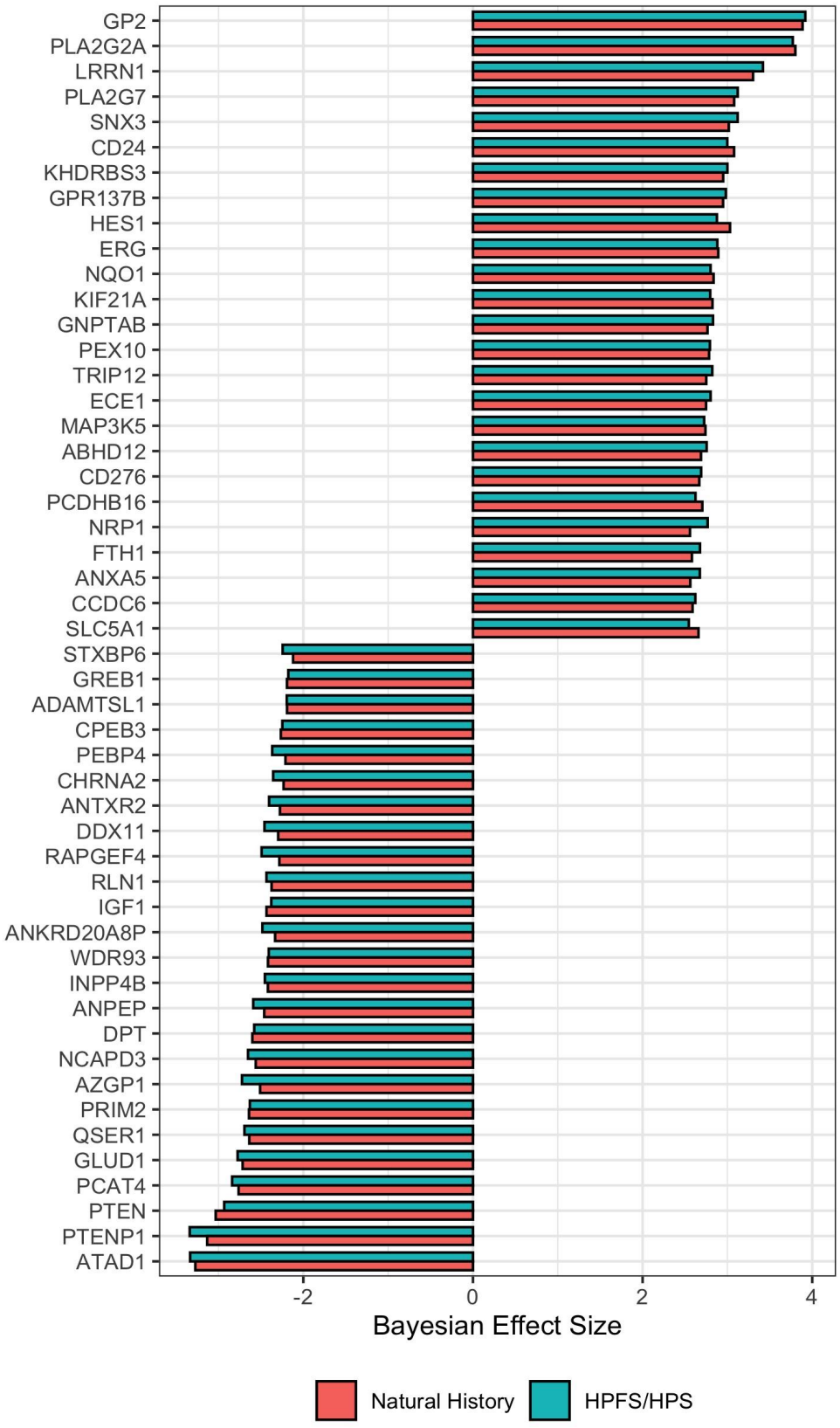


Figure 2. Cross-study meta-analysis of differential gene expression. Genes in the same loci as PTEN such as RLN1 and ATAD1 were found down-regulated. PTEN-null vs PTEN-intact meta-analysis of HPFS/PHS and NH cohorts with Bayesian Hierarchical

542 Model for DGE using XDE showing the top 25 most concordant differentially up- and down-regulated genes. PTEN status were

543 based on IHC assays.

544

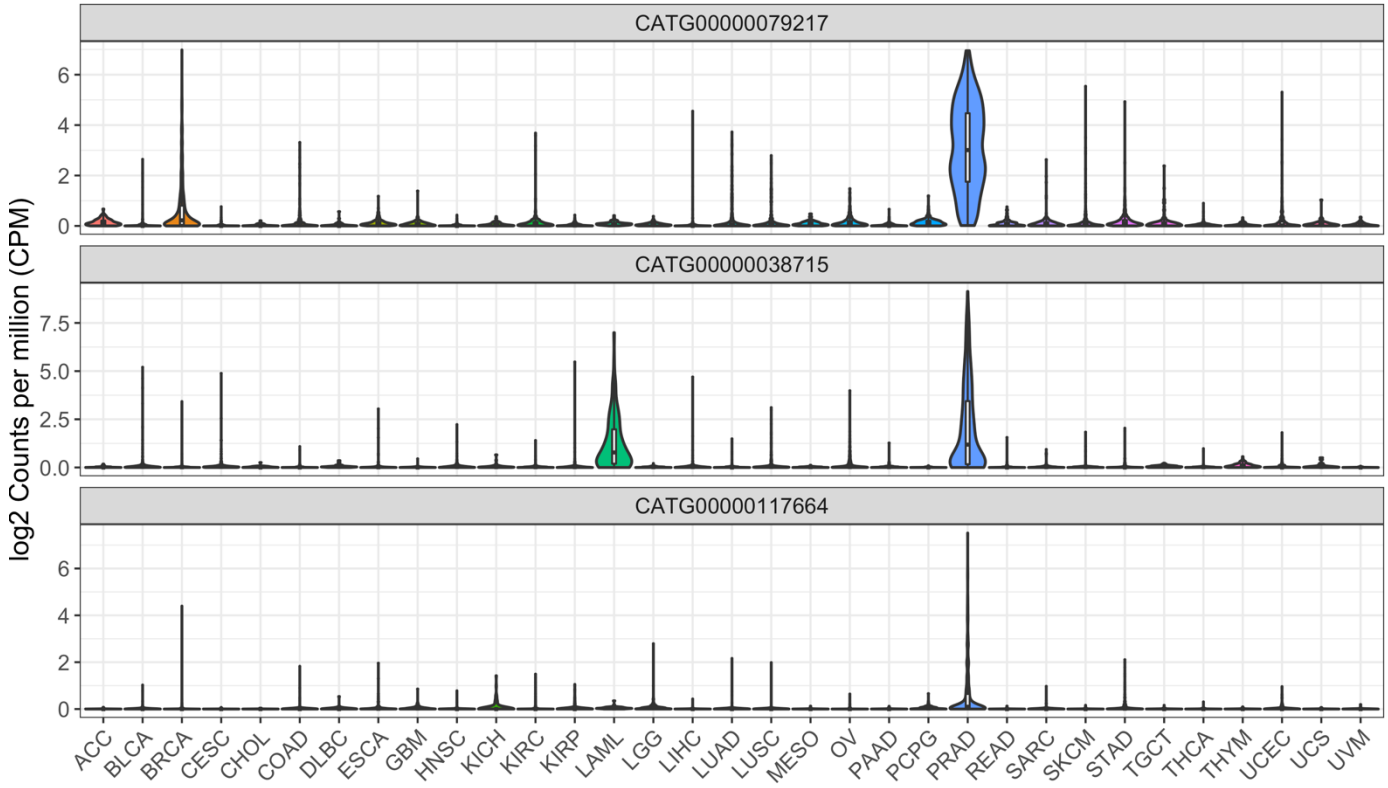


Figure 3. Expression profiles of novel FANTOM-CAT genes CATG00000038715, CATG00000079217 and CATG00000117664 across 33 cancer types. Violin-plots shows expression (log_2 CPM+1) distribution.

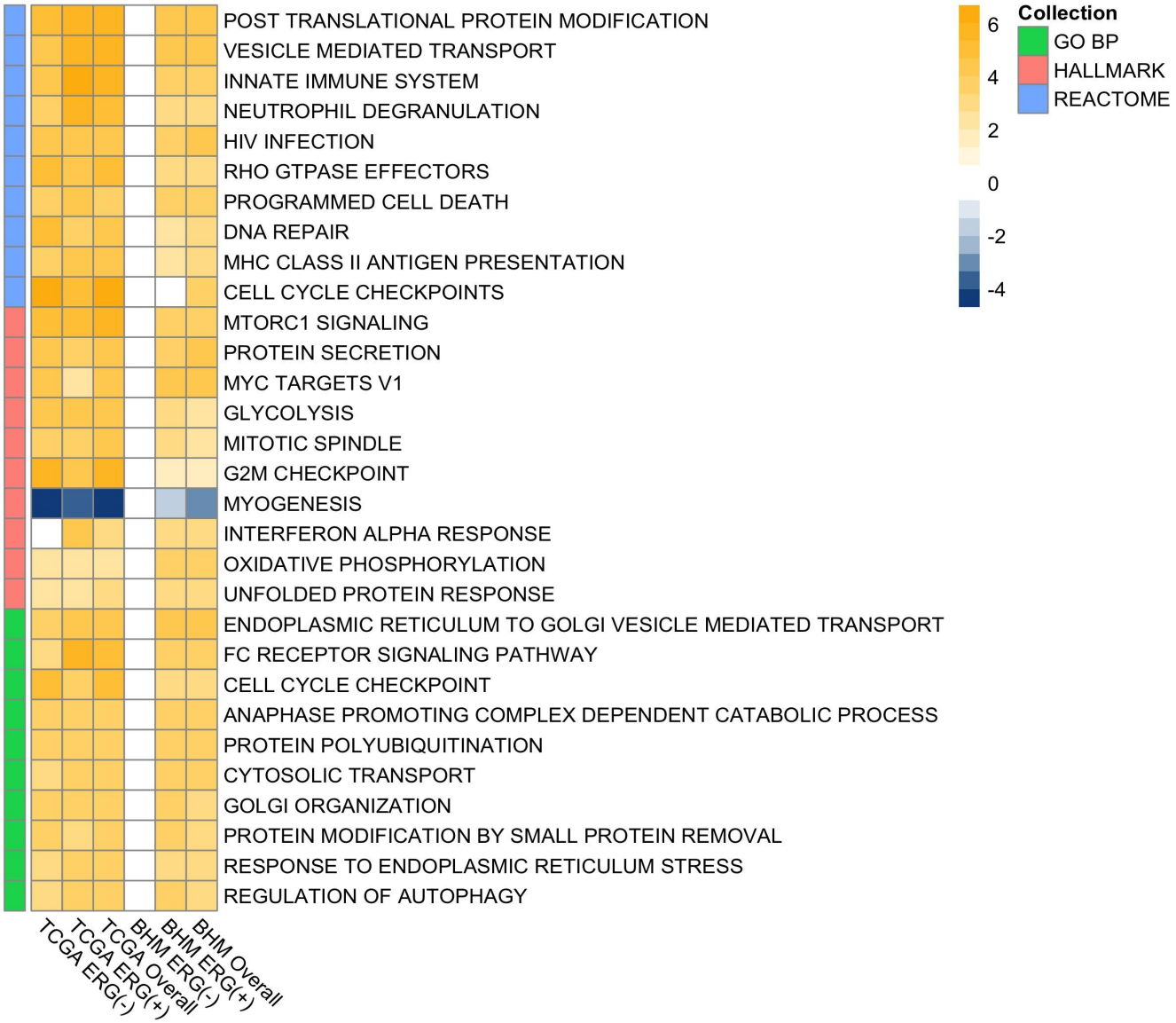


Figure 4. Top enriched gene sets enriched across PTEN-null and PTEN-intact in the TCGA and BHM cohorts stratified by ERG status and overall. Heatmap of mean-centered $\log_2$ signed p-values (normalized enriched score multiplied by $\log_{10}$ of p-value) showing the top 10 enriched gene sets of each collection (ranked by signed p-value).

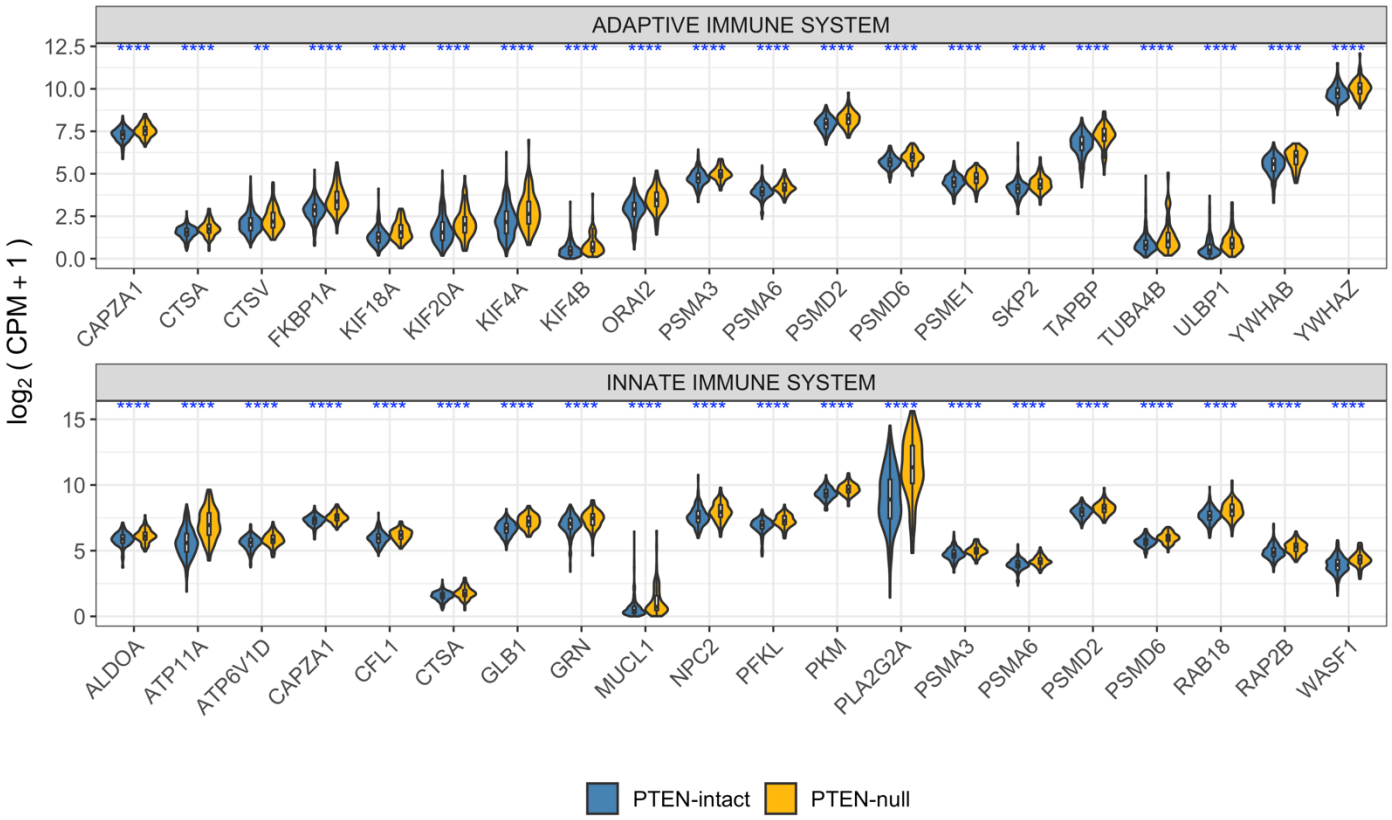
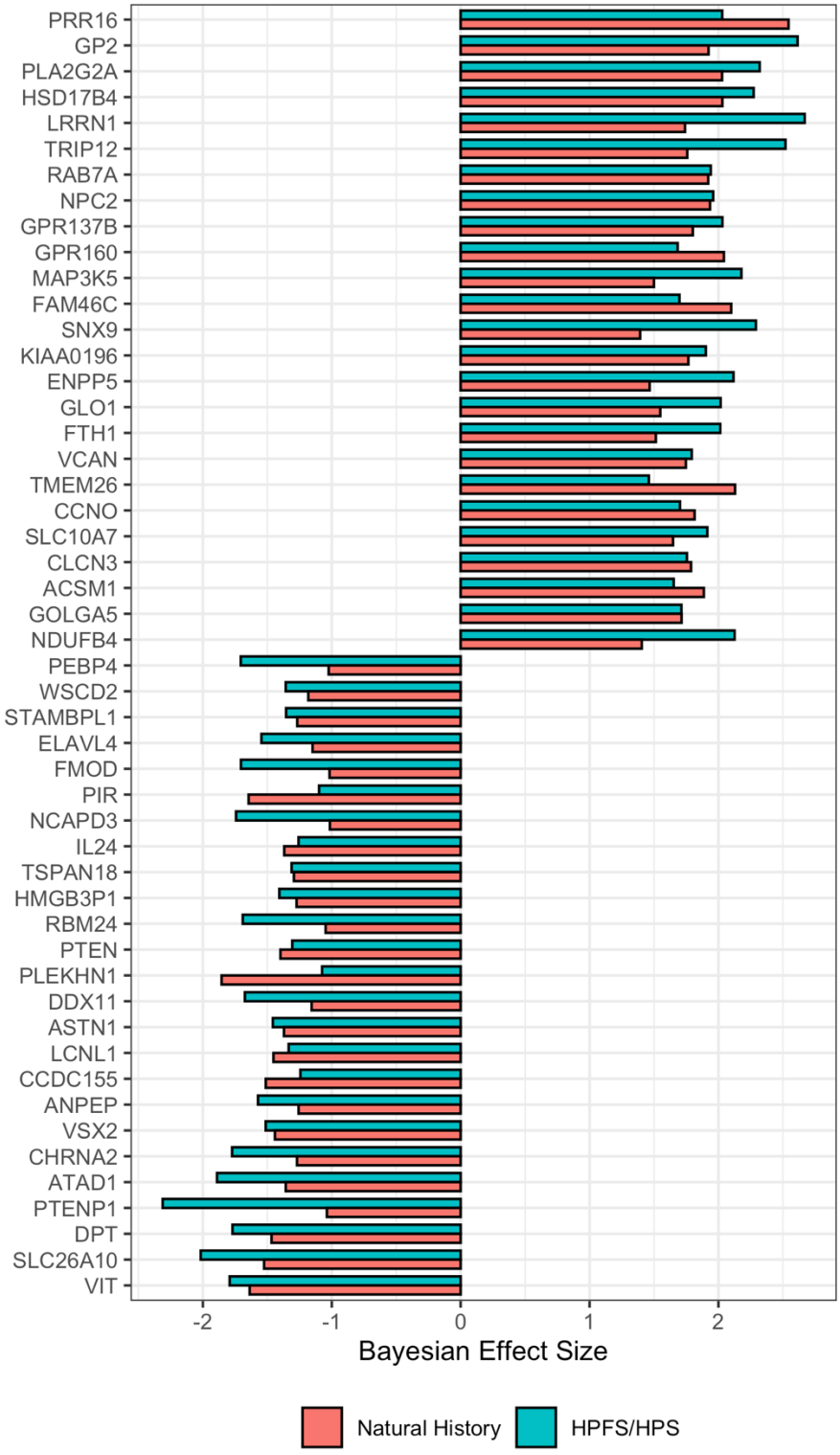


Figure 5. Expression of immune-related genes stratified by PTEN status. Top 20 were selected based on the leading edge of the GSEA of the adaptive and innate immune system gene sets from REACTOME. Significances based on t-test between PTEN-null and PTEN-intact using $\log_2$ CPM+1 values. Significance cutoffs: *=≤ 0.05; **≤0.01; *** ≤ 0.001; ****≤0.0001.

561 **Supplementary Figures and Tables**

562



563

564 **Figure S1.** Cross-study of differential gene expression in PTEN-null vs PTEN-intact in ERG^{+} samples. Meta-analysis of HPFS/PHS

565 and NH cohorts with Bayesian Hierarchical Model for DGE using XDE showing the top 25 most concordant differentially up- and

566 down-regulated genes. PTEN status were based on IHC assays.

567

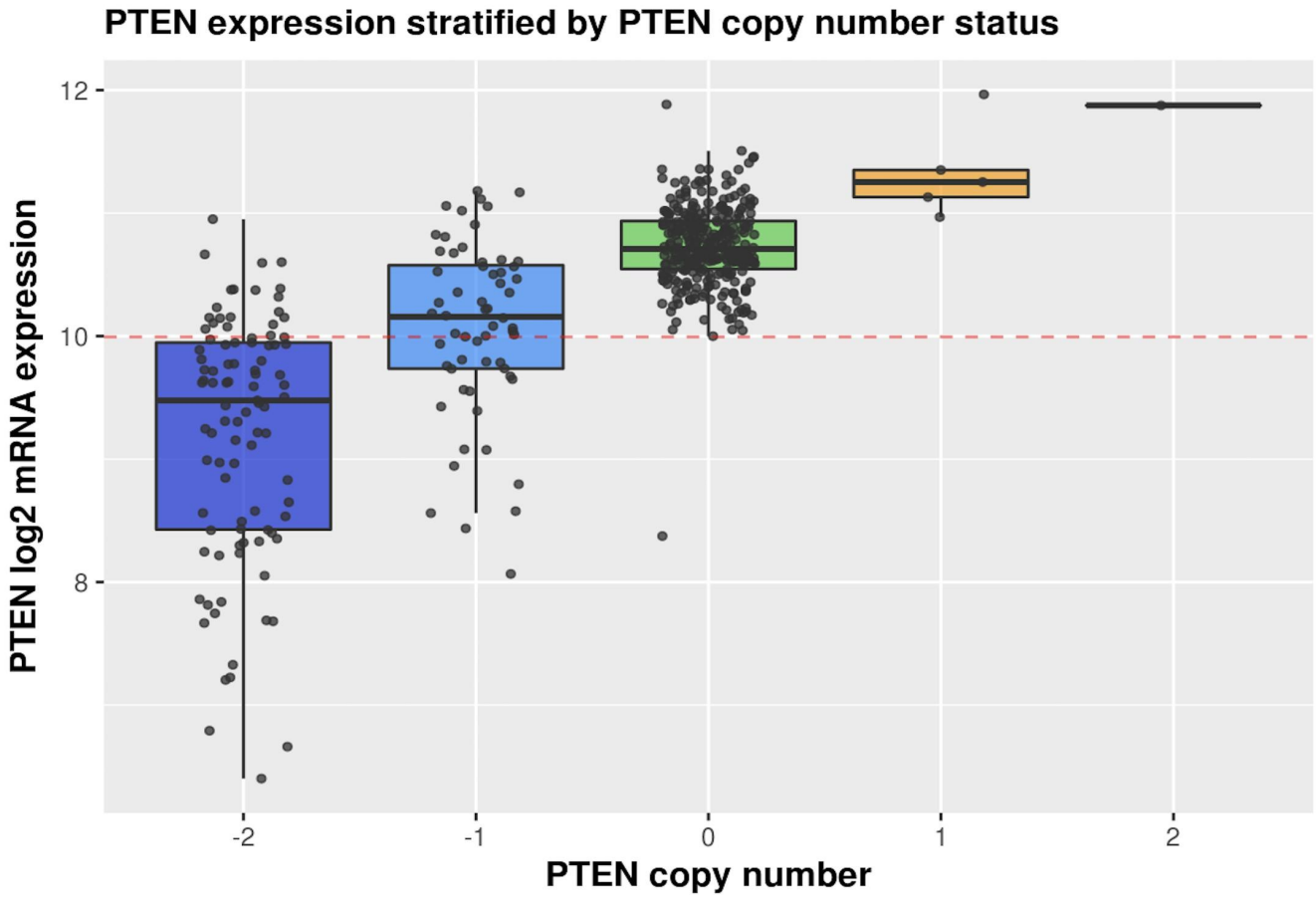
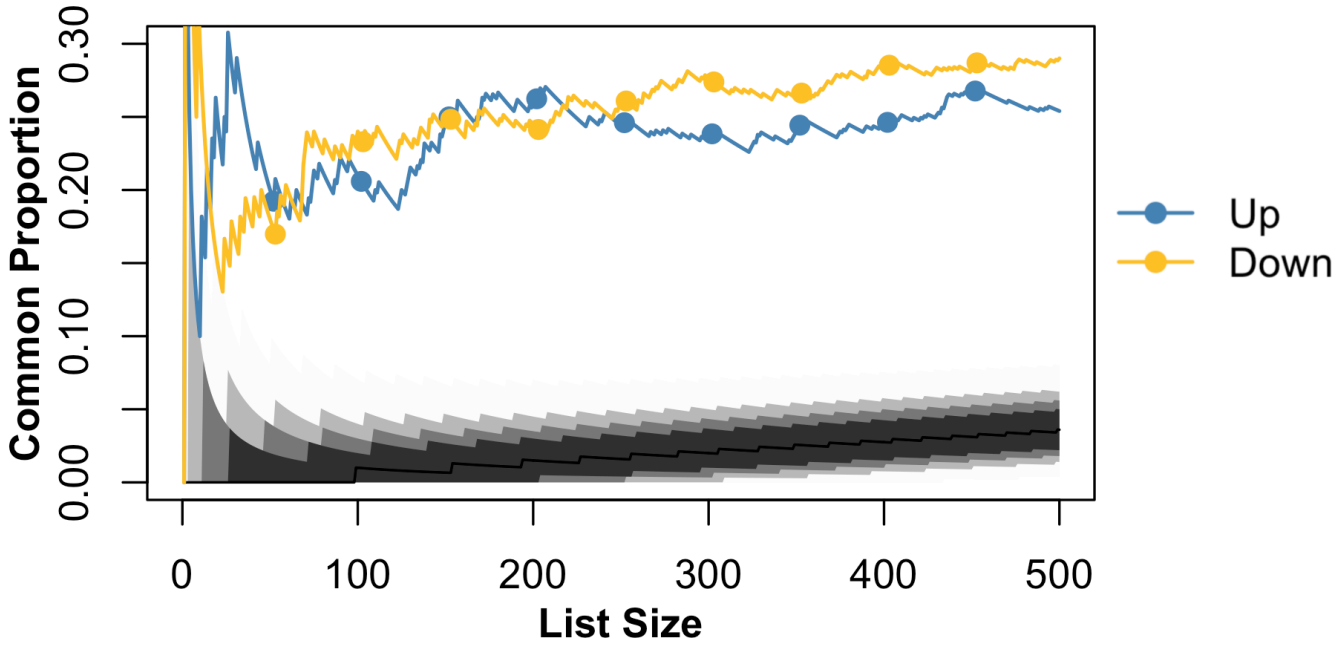


Figure S3. PTEN expression levels stratified by CNV. Figure shows PTEN expression levels distribution by copy number variation (CNV), called by GISTIC algorithm.



572

573 **Figure S4.** Correspondence-at-the-top (CAT) plot between TCGA CNV-based calls and the Bayesian Hierarchical Model approach
574 (BHM). Agreement of genes ranked by t-statistics (TCGA) and average Bayesian Effect Size (BHM). Lines represent agreement
575 between tested cohorts for PTEN-intact vs PTEN-null. Black-to-light grey shades represent the decreasing probability of agreeing
576 by chance based on the hypergeometric distribution, with intervals ranging from 0.999999 (light grey) to 0.95 (dark grey). Lines
577 outside this range represent agreement in different cohorts with a higher agreement than expected by chance.

578

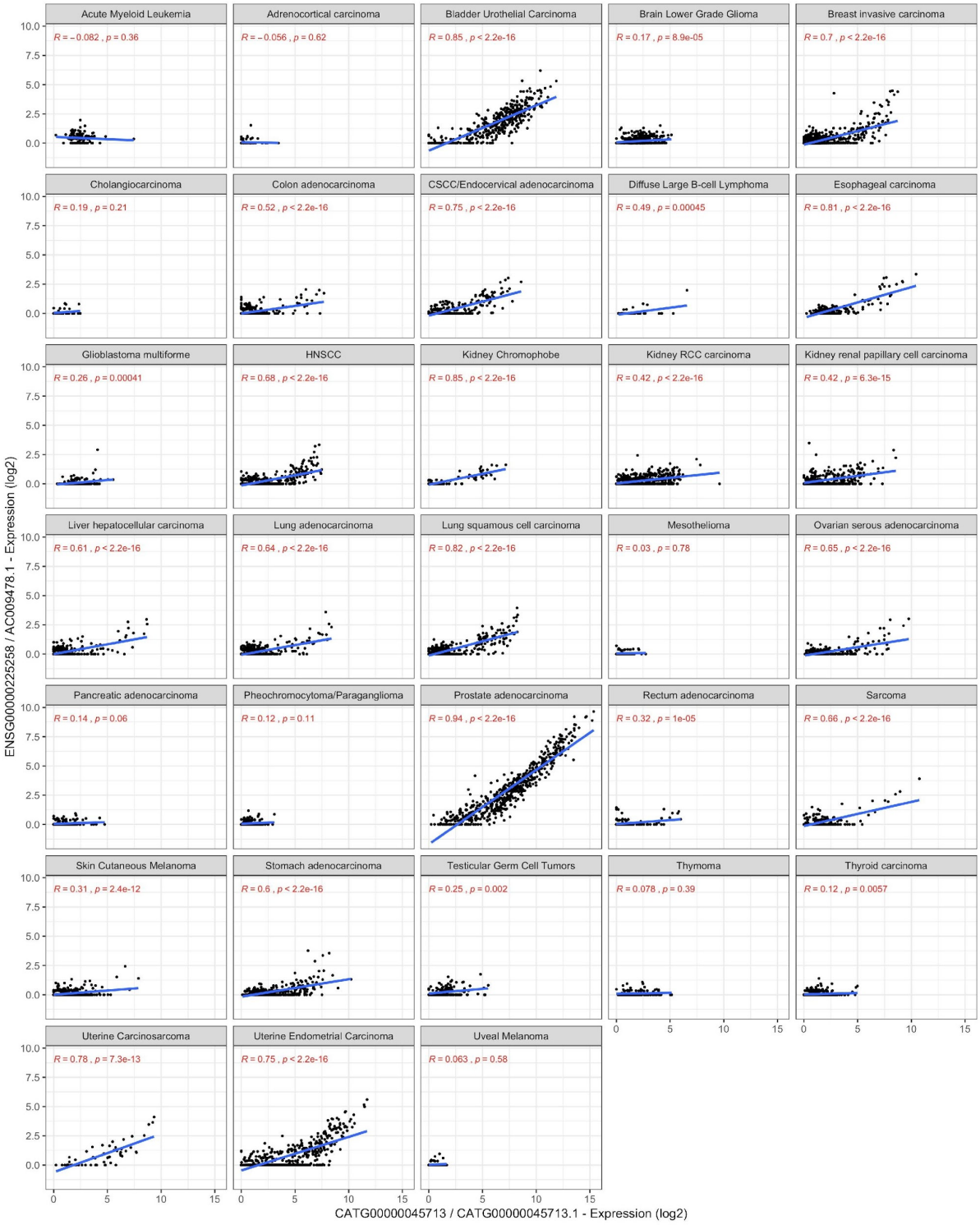


Figure S5. Expression of AC009478.1 is shown to be highly specific to PRAD, BLCA, to a lesser extent in UECA and BRCA. Figure shows raw expression values of SchLAP1 and AC009478.1 across cancer types. Pearson correlations and p-values are shown in red.

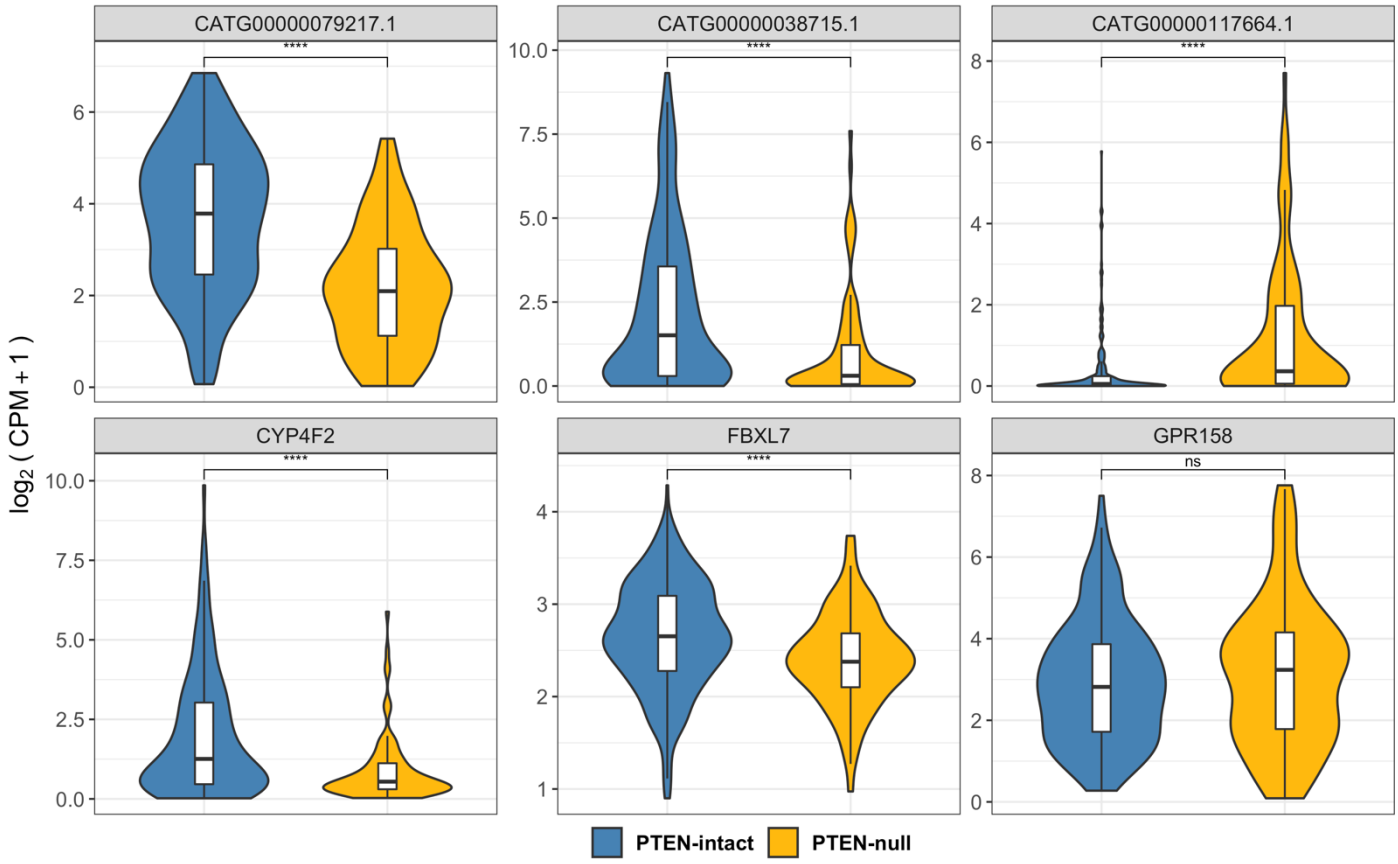


Figure S6. Expression of FANTOM-CAT lncRNAs genes (top) and close coding genes (bottom) stratified by PTEN status. Significances based on t-test between PTEN-null and PTEN-intact using log_2 CPM+1 value. Significance cutoffs: *=≤ 0.05; **≤0.01; *** ≤ 0.001; ****≤0.0001.

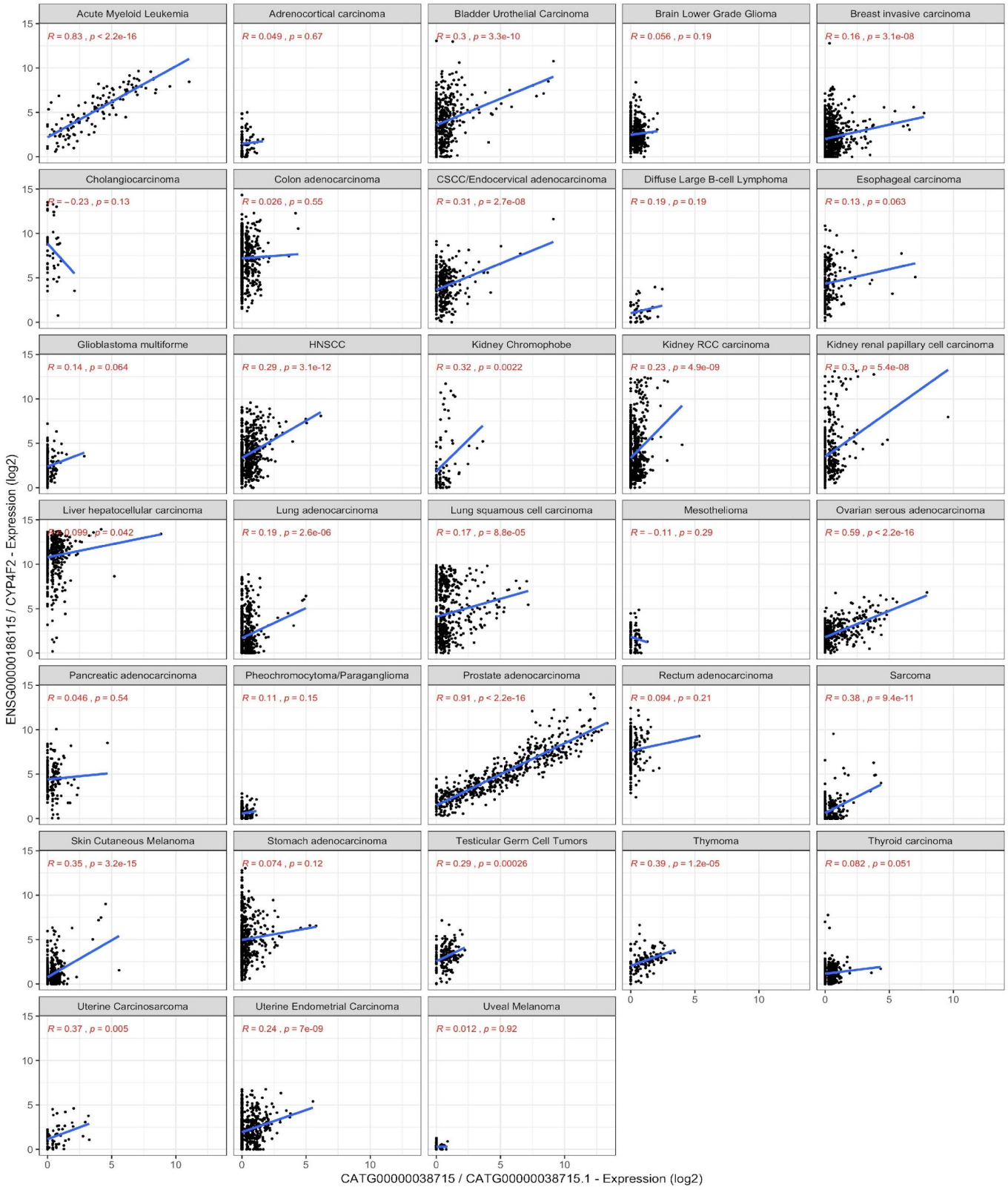


Figure S7. Person correlation gene CATG00000038715 and CYP4F2 across cancer types. CATG00000038715 and CYP4F2 expression are shown to be highly correlated in PCa. Moreover, CATG00000038715 expression is shown to be highly specific to PCa. With exception of leukemia cells, none of the other tumors expressed high levels of CATG00000038715.

594

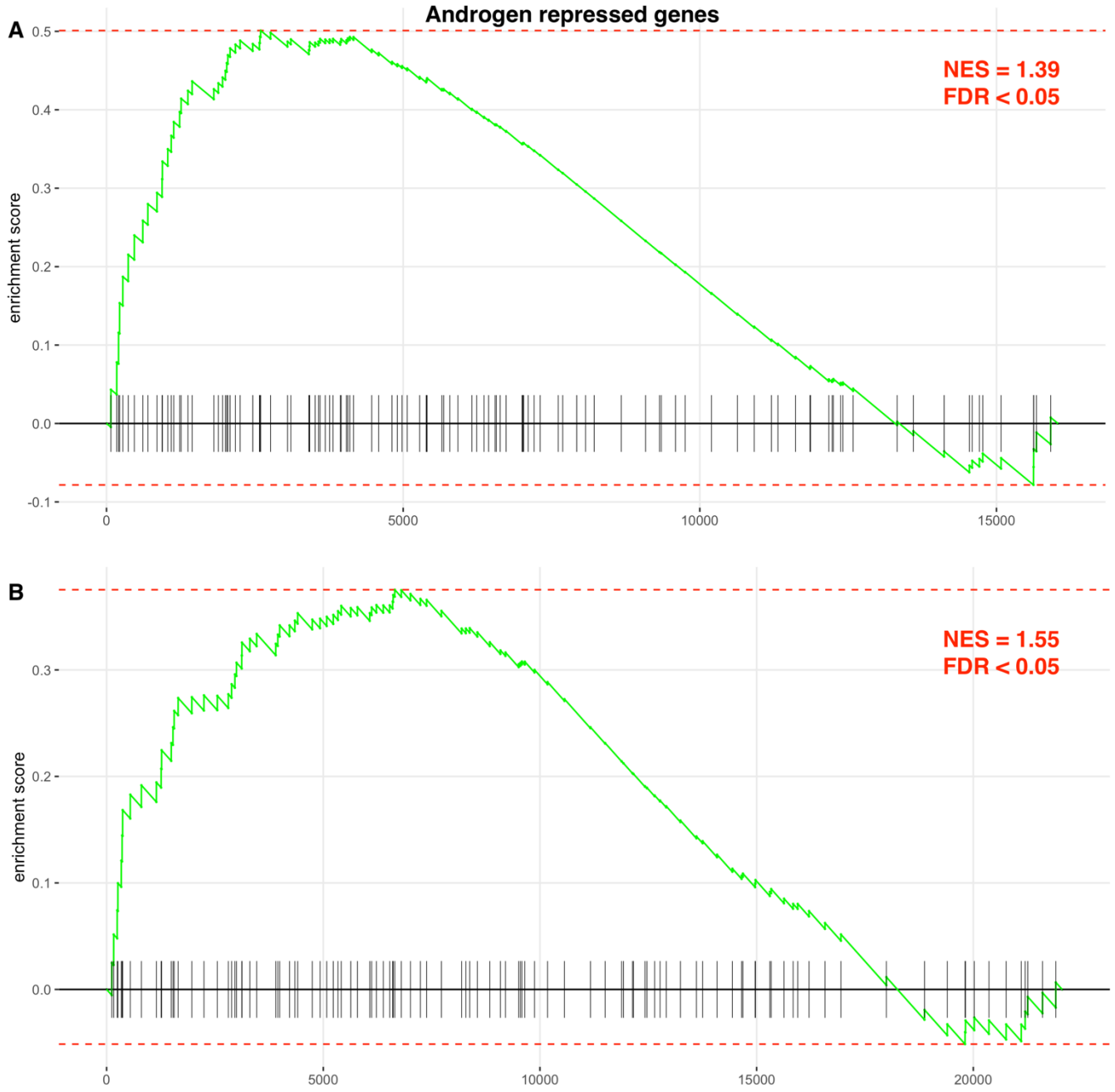


Figure S8. Gene set enrichment for Androgen repressed genes. Gene set enrichment analysis of gene signature showing positive enrichment of genes repressed by dihydrotestosterone after 6 hours of exposure obtained from Schaeffer et al.[48]. Enrichment for BHM-signature is shown in panel A and TCGA-signature in panel B.

501