Pan-cancer systematic identification of lncRNAs associated with cancer prognosis

Matthew Ung[1], Evelien Schaafsma[1], Daniel Mattox[2], George L. Wang[1] and Chao Cheng[1,3,4,5]

[1] Department of Molecular and Systems Biology, Dartmouth College, Hanover, NH, USA
[2] Department of Computer Science, Dartmouth College, Hanover, NH, USA
[3] Department of Medicine, Baylor College of Medicine, Houston, TX, USA
[4] The Institute for Clinical and Translational Research, Baylor College of Medicine, Houston, TX, USA
[5] Department of Biomedical Data Science, Geisel School of Medicine at Dartmouth, Lebanon, NH, USA

ABSTRACT

Background: The "dark matter" of the genome harbors several non-coding RNA species including Long non-coding RNAs (lncRNAs), which have been implicated in neoplasia but remain understudied. RNA-seq has provided deep insights into the nature of lncRNAs in cancer but current RNA-seq data are rarely accompanied by longitudinal patient survival information. In contrast, a plethora of microarray studies have collected these clinical metadata that can be leveraged to identify novel associations between gene expression and clinical phenotypes.
Methods: In this study, we developed an analysis framework that computationally integrates RNA-seq and microarray data to systematically screen 9,463 lncRNAs for association with mortality risk across 20 cancer types.
Results: In total, we identified a comprehensive list of associations between lncRNAs and patient survival and demonstrate that these prognostic lncRNAs are under selective pressure and may be functional. Our results provide valuable insights that facilitate further exploration of lncRNAs and their potential as cancer biomarkers and drug targets.

Submitted 21 November 2019
Accepted 25 February 2020
Published 24 March 2020

Corresponding author
Chao Cheng,
chao.cheng@dartmouth.edu

Academic editor
Stephen Piccolo

Additional Information and Declarations can be found on page 17

DOI **10.7717/peerj.8797**

Subjects Bioinformatics, Computational Biology, Genomics, Oncology
Keywords LncRNA, Prognosis, Microarray, RNA-seq, TCGA

INTRODUCTION

Long non-coding RNAs (lncRNAs) constitute a relatively unexplored repertoire of gene products that exhibit diverse functions and are involved in several biological processes. As such, the ENCODE consortium reported that 80% of the genome is transcribed into a variety of functional products including non-coding RNAs (*The ENCODE Project Consortium, 2012*). Several high-level characteristics of lncRNAs provide evidence that they are indeed functional, including their association with chromatin signatures of active transcription, being transcribed by RNA polymerase II, and undergoing post-transcriptional modifications such as polyadenylation and alternative splicing (*Wang & Chang, 2011*; *Rinn & Chang, 2012*; *Kung, Colognori & Lee, 2013*).
The mechanisms by which lncRNAs regulate biological processes have not been studied

in detail but evidence suggest that they can function at the transcriptional, post-transcriptional and post-translational level by acting as biological signals, decoys, guides and scaffolds (*Wang & Chang, 2011*; *Rinn & Chang, 2012*; *Kung, Colognori & Lee, 2013*). Moreover, the organization of lncRNAs across the genome is quite diverse in that they can be transcribed from intergenic regions, sites anti-sense to protein coding genes, bi-directional promoters, or within gene introns (*Ponting, Oliver & Reik, 2009*; *Kung, Colognori & Lee, 2013*).

Having been previously referred to as transcriptomic noise or "junk" DNA, lncRNAs are now being investigated as molecular players in several disease processes including cancer (*Mattick & Makunin, 2006*; *Esteller, 2011*; *The ENCODE Project Consortium, 2012*; *Sahu, Singhal & Chinnaiyan, 2015*; *Schmitt & Chang, 2016*; *Bartonicek, Maag & Dinger, 2016*; *Evans, Feng & Chinnaiyan, 2016*). In this particular context, lncRNAs have been implicated in all hallmarks of cancer including sustaining proliferative signaling, evading growth suppressors, enabling replicative immortality, activating invasion and metastasis, inducing angiogenesis and resisting cell death (*Hanahan & Weinberg, 2000*; *Gutschner & Diederichs, 2012*; *Ali et al., 2018*; *Chiu et al., 2018*). Aberrant expression of lncRNAs might be due to their close association with certain key driver genes (*Ashouri et al., 2016*) or the establishment of cancer-specific genomic features in lncRNA loci itself, including mutational events, methylation, copy number and SNP events (*Iyer et al., 2015*; *Yan et al., 2015*). Several studies have performed pan-cancer screens for lncRNAs involved in the disease and found that several of them were differentially expressed compared to normal samples, revealing their potential as biomarker candidates (*Cabanski et al., 2015*; *Yan et al., 2015*; *Byron et al., 2016*; *Ching et al., 2016*). For instance, PCA3 is a lncRNA that is currently approved for clinical use as a prostate cancer diagnostic biomarker and can be detected in patient urine samples (*De Kok et al., 2002*). Thus, dissecting the molecular characteristics of these understudied RNAs and their associations with disease phenotypes may yield findings that can be translated into the clinic.

In light of these findings, there is a paucity of patient samples with matched RNA-seq data and clinical information which limits the ability to perform pan-cancer screening for prognostic lncRNAs. Furthermore, few of these matched datasets contains sufficiently long follow-up times which limits statistical power when performing survival analyses, especially in cancer types where patients exhibit high survival rates (*Clark et al., 2003*). In stark contrast, there is a plethora of microarray gene expression data that are available, many of which are accompanied by comprehensive clinical information with long follow-up times.

Thus, using primarily protein-coding gene expression from microarray to infer the expression of their non-coding counterparts can re-purpose these valuable data and generate novel hypotheses about lncRNAs associated with patient mortality across several cancer types. To this end, multiple studies have attempted to utilize data from microarray to make inferences about lncRNA activity and their clinical relevance. *Du et al. (2013)* re-annotated probes from microarray data to identify prognostic transcriptional activity for ~10,000 lncRNAs in prostate cancer, glioblastoma, ovarian cancer and lung squamous

cell carcinoma. From this screen, they identified novel lncRNA markers of mortality risk and validated several of them experimentally. Furthermore, *Guo, Yao & Jiang (2016)* performed a "guilt-by-association" analysis whereby lncRNAs that share an edge with prognostic protein coding genes in a biological network defined a priori were also considered prognostic. Although these studies have provided valuable insights into lncRNA biology, the reannotation of microarray probes might have missed prognostic lncRNAs not captured by microarray probes. In addition, lncRNA inference based on known protein coding target genes might bias lncRNA expression if not all target genes are known.

Therefore, we introduce a lncRNA inference approach that generates cancer-specific weighted lncRNA regulon network profiles de novo using RNA-seq data from The Cancer Genome Atlas (TCGA), and apply them to infer lncRNA expression in the PRECOG (*Gentles et al., 2015*) and METABRIC (*Curtis et al., 2012*) microarray compendia, which provide expression of protein-coding genes but not for most lncRNAs. Afterwards, we systematically interrogated each lncRNA to identify those that significantly associate with patient prognosis using clinical metadata included in the microarray studies. In total we screened 9,463 unique lncRNAs across 20 different cancer types to identify novel associations.

MATERIALS AND METHODS

Data source and pre-processing

The lncRNA gene list with Ensembl IDs was derived from the TANRIC resource (*Li et al., 2015*). Level 3 RNA-seq count data from tumor samples encompassing 23 different cancer types along with corresponding clinical information were downloaded from the National Cancer Institute's Genomic Data Commons data portal (https://portal.gdc.cancer.gov/). The count data was normalized by library size and subjected to a variance stabilizing transformation implemented using DESeq2 (*Love, Huber & Anders, 2014*). This transforms the expression values so that they are homoskedastic by fitting the dispersion to a negative binomial distribution. A total of 141 microarray gene expression datasets across 20 cancer types were downloaded from the PRECOG resource (*Gentles et al., 2015*). Normalized breast cancer gene expression and copy number alteration (CNA) datasets from METABRIC (n = 1,992) were downloaded from the European Genome-Phenome Archive (http://www.ebi.ac.uk/ega/) under the accession number EGAS00000000083. CRISPRi screening data on functional lncRNAs in MDA-MB-231 and K562 cell lines were downloaded from *Liu et al. (2017)*.

Construction of cancer-specific regulons

The ARACNe-AP algorithm was applied to each processed TCGA RNA-seq cancer dataset using the TANRIC lncRNA Ensembl gene IDs as the regulator mapping set. Briefly, ARACNe-AP calculates the mutual information between a lncRNA and potential target genes and removes edges that are unlikely to represent a biological link using the concept of data processing inequality (*Margolin et al., 2006*; *Lachmann et al., 2016*). We implemented the algorithm using 100 bootstrap iterations and a p-value threshold of

0.01. Each regulon in the cancer-specific network consisted of a lncRNA and its associated genes. Each edge in the regulon was assigned a weight using the mutual information scores outputted by ARACNe-AP (*Alvarez et al., 2016*). The mutual information scores were divided by the maximum score within each regulon so that they had a range from 0 to 1. The sign of the edge was assigned by computing a Pearson correlation coefficient between the lncRNA's expression and the associated gene's expression across the samples. Since genes in a regulon are positively or negatively correlated with the corresponding lncRNA in a specific cancer type, their expression can be used to impute the expression level of the lncRNA.

Inference of lncRNA expression in microarray datasets

For each cancer-specific regulon, we defined a pair of profiles—the genes with a positive weight were assigned to an "up-regulated" profile and the genes with a negative weight were assigned to a "down-regulated" profile. In the up-regulated profile, all genes that had a negative weight were assigned a value of 0 and all genes in the down-regulated profile that had a positive weight were assigned a value of 0. The values in the down-regulated profile were then forced to be positive. Only profiles with 20 or more associated genes were used. Thus, each lncRNA was assigned two regulon weight profiles that capture the magnitude and direction of the genes it was associated with. Genes with higher weights in the two profiles will contribute more to the imputation of lncRNA expression.

After constructing the regulon weight profiles, lncRNA expression was inferred in microarray datasets by using the regulon weight profile derived from the same cancer type (or most related cancer type) as the microarray experiment (Supplemental Results). To apply the regulon weight profiles to infer lncRNA expression in microarray samples, we utilized the BASE algorithm (*Cheng et al., 2007*) which outputs a predicted expression value for each lncRNA in every patient sample. BASE imputes the relative expression level of a lncRNA based on the expression of genes that it correlates with (i.e., regulon genes). Specifically, the algorithm sorts each patient's gene expression profile from highest to lowest expressed genes and weights them using the two regulon weight profiles. BASE then calculates a running sum statistic by moving down the profile and calculating a foreground function which captures the weighted enrichment of the lncRNA's associated genes at the top and bottom of the patient's gene expression profile. The foreground function is then compared to a background function and the maximum deviation between the foreground and background functions is computed. The maximum deviation calculated from the down-regulated profile is subtracted from the maximum deviation calculated from the up-regulated profile to yield a pre-inferred lncRNA expression value (pre-iExpr). For a lncRNA, if positively associated genes tend to be highly expressed (at the top of the expression profile) in a tumor sample while negatively associated genes tend to be lowly expressed (at the bottom of the expression profile), we will observe a high pre-iExpr value. The patient's gene expression profile is then randomly permuted and the procedure is repeated; this is performed 1,000 times to yield a null pre-iExpr distribution. The pre-iExpr score is then normalized by dividing by the mean of the null pre-iExpr values to yield

the final inferred expression of the lncRNA (iExpr). The formulas describing the details of this algorithm are provided in the Supplemental Methods.

Systematic inference of prognostic lncRNAs

A univariate Cox proportional hazards model was fit to the inferred and actual expression values for each lncRNA, separately for TCGA, PRECOG and METABRIC datasets. Actual expression was available for a small set of lncRNAs in the PRECOG and METABRIC datasets and were used for downstream validation. From the models, *z*-scores were calculated by dividing the Cox regression coefficient by its standard error. A *z*-score < 0 indicates that a lncRNA is protective and positively associated with survival. Conversely, a *z*-score > 0 indicates that a lncRNA is hazardous and negatively associated with survival.

In the PRECOG dataset, there were several microarray datasets belonging to the same cancer type. After computing the inferred expression for all lncRNAs within each dataset, we fitted a univariate Cox regression model to measure the association between a lncRNA and all-cause or disease-specific mortality (if available). *z*-scores were extracted from the fitted models and a meta *z*-score was calculated for each lncRNA across all the microarray datasets belonging to the same cancer type. The meta *z*-score was calculated using weighted Stouffer's *z*-score method using the dataset sample size as weights. A meta *z*-score < 1 indicates a positive association and a meta *z*-score > 1 indicates a negative association with survival. In addition, robust meta *z*-scores were calculated for each lncRNA by leaving out the dataset yielding the most significant association and repeating the procedure. Meta *p*-values were calculated from the meta *z*-scores by referring to the standard normal distribution. Meta *p*-values were adjusted for each cancer type using Benjamini–Hochberg and Bonferroni multiple testing correction methods. Kaplan–Meier analysis of lncRNAs was performed by dichotomizing patients into high (>0) and low (<0) inferred lncRNA expression groups and performing a log-rank test to calculate statistical significance.

In the METABRIC dataset, a multiple Cox regression model was applied and included age at diagnosis, tumor size, stage, ER and HER2 status as covariates. Disease-specific mortality was used as the outcome.

Validation of survival analysis

To compare survival results across datasets, we performed two validation analyses: (1) Cross-dataset analysis comparing Cox regression results using actual lncRNA expression from TCGA with results using inferred lncRNA expression from PRECOG and (2) Within-dataset comparison of survival results generated by models fitted to inferred or actual lncRNA expression in PRECOG and METABRIC. Pearson correlation was used to evaluate the consistency between lncRNA regression *z*-scores derived from actual and inferred expression within and between datasets. A one-sided Fisher's exact test was used to compute the enrichment of prognostic TCGA lncRNAs (actual lncRNA expression) in the set of prognostic PRECOG lncRNAs (inferred lncRNA expression). Prognostic lncRNAs were selected using FDR < 0.05 and non-prognostic lncRNAs were

selected using a FDR > 0.1. Protective (hazard ratio < 1) and hazardous (hazard ratio > 1) lncRNAs were analyzed using separate enrichment tests.

Revealing lncRNA-based subtypes in breast cancer

In the METABRIC dataset, feature selection was performed by selecting the top 500 lncRNAs with the highest variation of inferred expression across patients. The inferred expression levels were then *z*-transformed across patients and gene-wise unsupervised clustering was performed using Euclidean distance and complete linkage.

Analysis of prognostic and essential lncRNAs

Hazardous lncRNAs identified from the PRECOG meta-analysis of breast cancer and hematopoietic cancer datasets were selected using a *z*-score cutoff of >0 and *p*-value cutoff of ≤0.1. LncRNA functional screening data were downloaded from *Liu et al. (2017)* and contained averaged phenotype scores derived from systematic CRISPRi knockout of lncRNAs. Essential lncRNAs were defined as those that when knocked down, result in ablation of cell proliferation and cell death and was quantified by a phenotype score included in the dataset. Essential lncRNAs in the MDA-MB-231 (breast) or K562 (hematopoietic) cell lines were selected using an average phenotype score cutoff of <0 and a *p*-value cutoff of ≤0.1. The average phenotype score measured the growth effect on the cell line when a particular lncRNA has been knocked down; a value <0 indicated essentiality and a value >0 indicated non-essentiality. The enrichment overlap between essential MDA-MB-251 lncRNAs and hazardous breast cancer lncRNAs was computed using a one-sided Fisher's exact test. The same test was used to calculate the enrichment overlap between essential K562 lncRNAs and hazardous hematopoietic cancer lncRNAs.

Prognostic lncRNAs and CNAs

Long non-coding RNAs associated with prognosis in the METABRIC dataset were mapped to the genome for each patient. Hazardous lncRNAs were selected using *z*-score > 0 and FDR ≤ 0.01 as the cutoff. Protective lncRNAs were selected using *z*-score < 0 and FDR ≤ 0.01 as the cutoff. The CNA dataset provides the copy number signal of genomic segments throughout the genome for each patient along with binary calls indicating amplification (1) or deletion (−1). For each patient, a Fisher's exact test was performed to measure significant enrichment of hazardous lncRNAs (compared to protective lncRNAs) in genomic segments that had undergone copy number amplification or deletion. When constructing the contingency table for a Fisher's exact test, every cell had to have at least five counts in order for the test to be performed for the patient to ensure robust enrichment results. In total, 1,595 METABRIC patients were used to test enrichment of hazardous lncRNAs in amplified regions and 901 patients were used to test enrichment of protective lncRNAs in deleted regions. The Benjamini–Hochberg procedure was used to adjust for multiple hypothesis testing. When calculating the CNA signal corresponding to each lncRNA, the average copy number signal of all segments overlapping the gene region (transcription start site to the termination site) was used.

When performing the CNA enrichment analysis in the TCGA dataset, lncRNAs associated with prognosis in glioblastoma or ovarian cancer were selected using an unadjusted *p*-value cutoff of <0.05. An FDR cutoff of 0.1 was used to identify prognostic lncRNAs in pancreatic cancer and lung adenocarcinoma. These significance cutoffs were chosen to ensure a sufficient number of prognostic lncRNAs for enrichment analysis. Segments were selected using a CNA signal of >0 and <0 for amplification and deletion, respectively.

RESULTS

Overview of analysis

To systematically identify lncRNAs associated with patient prognosis, we applied the ARACNe-AP algorithm (*Lachmann et al., 2016*) to 23 TCGA RNA-seq datasets to generate lncRNA regulons for each cancer type. ARACNe-AP calculates the mutual information between a lncRNA and potential target genes and removes edges that are unlikely to represent a biological link using the concept of data processing inequality (*Margolin et al., 2006*; *Lachmann et al., 2016*). The resulting regulons represent a network where the edges encode the magnitude and direction of association between lncRNAs and other genes based on their gene expression across samples (See "Methods").
A lncRNA's expression can be inferred within a microarray dataset lacking lncRNA probes by analyzing the aggregate expression of the protein coding genes composing that lncRNA's regulon. In total, we generated cancer-specific lncRNA regulons for 23 different cancer types using TCGA RNA-seq datasets. Once these regulons were generated, we transformed them into weight profiles and validated their predictive accuracy in TCGA. We then extended our analysis by inferring lncRNA expression in microarray data compendia from PRECOG and METABRIC using the regulon weight profiles and the BASE algorithm (*Cheng et al., 2007*). The BASE algorithm outputs a predicted expression value for each lncRNA in every patient sample by imputing the relative expression level of a lncRNA based on the expression of genes that it correlates with (i.e., regulon genes). Regulon weight profiles were selected to interrogate microarray data based on matched cancer type. After inferring the expression of thousands of lncRNAs, we performed a systematic pan-cancer screen for prognostic lncRNAs using survival information included in the microarray gene expression data compendia (Fig. 1).

Inferred lncRNA expression strongly correlates with actual expression

By implementing the ARACNe-AP algorithm in TCGA RNA-seq datasets, we constructed thousands of lncRNA regulons for each TCGA cancer type (Fig. 2A). Each regulon contains a lncRNA and its associated genes, which can be used as features to infer that specific lncRNA's expression. An example of an inferred lncRNA expression pattern and the expression of its associated genes is provided in Fig. S1. The number of regulons varied across cancer types depending on whether any genes were found to have high mutual information with any given lncRNA based on expression signal. To confirm that the inferred expression of the lncRNAs was indeed accurate, we correlated each lncRNA's inferred expression with its actual expression in all TCGA RNA-seq datasets. We observed

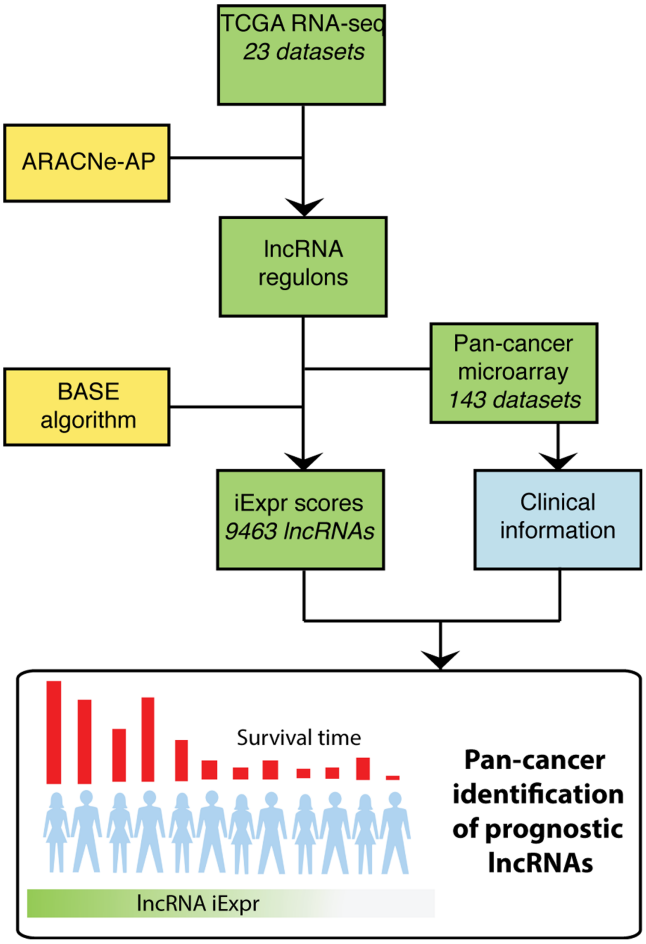

Figure 1 Overview of analysis. TCGA RNA-seq data from 23 cancer types were used as input into the ARACNe algorithm to generate cancer type specific lncRNA-target gene regulons. These regulons were used with the BASE algorithm to infer lncRNA expression in PRECOG and METABRIC microarray datasets. The BASE algorithm infers the expression of lncRNAs in microarray data using the aggregate expression of the lncRNAs' associated protein coding genes. Lastly, a systematic pan-cancer analysis of 9,463 lncRNAs was carried out to identify prognostic lncRNAs across 20 different cancer types in the microarray data compendia. Full-size DOI: 10.7717/peerj.8797/fig-1

that for the majority of lncRNAs, their inferred and actual expression were highly correlated across 23 cancer types as shown by the left-skewed distribution of correlation coefficients (Fig. 2B). These results indicate that it is possible to infer the expression levels of lncRNAs based on the aggregate expression of its associated genes.

Furthermore, we inferred lncRNA expression in the METABRIC dataset and compared the inferred and actual expression of 95 lncRNAs, which had probes present in the microarray platform. We observed that 82 of these lncRNAs had inferred expression values positively correlated with probe expression with 59 having significant associations (Fig. 3A; $p \leq 0.05$). As an example, the correlation between the inferred and actual expression of HOTAIR and PVT1 was 0.54 and 0.60, respectively (Figs. 3B and 3C). This analysis was repeated in each PRECOG dataset and we again observed that the correlation coefficient distributions were left-skewed indicating that approximately 95% of the

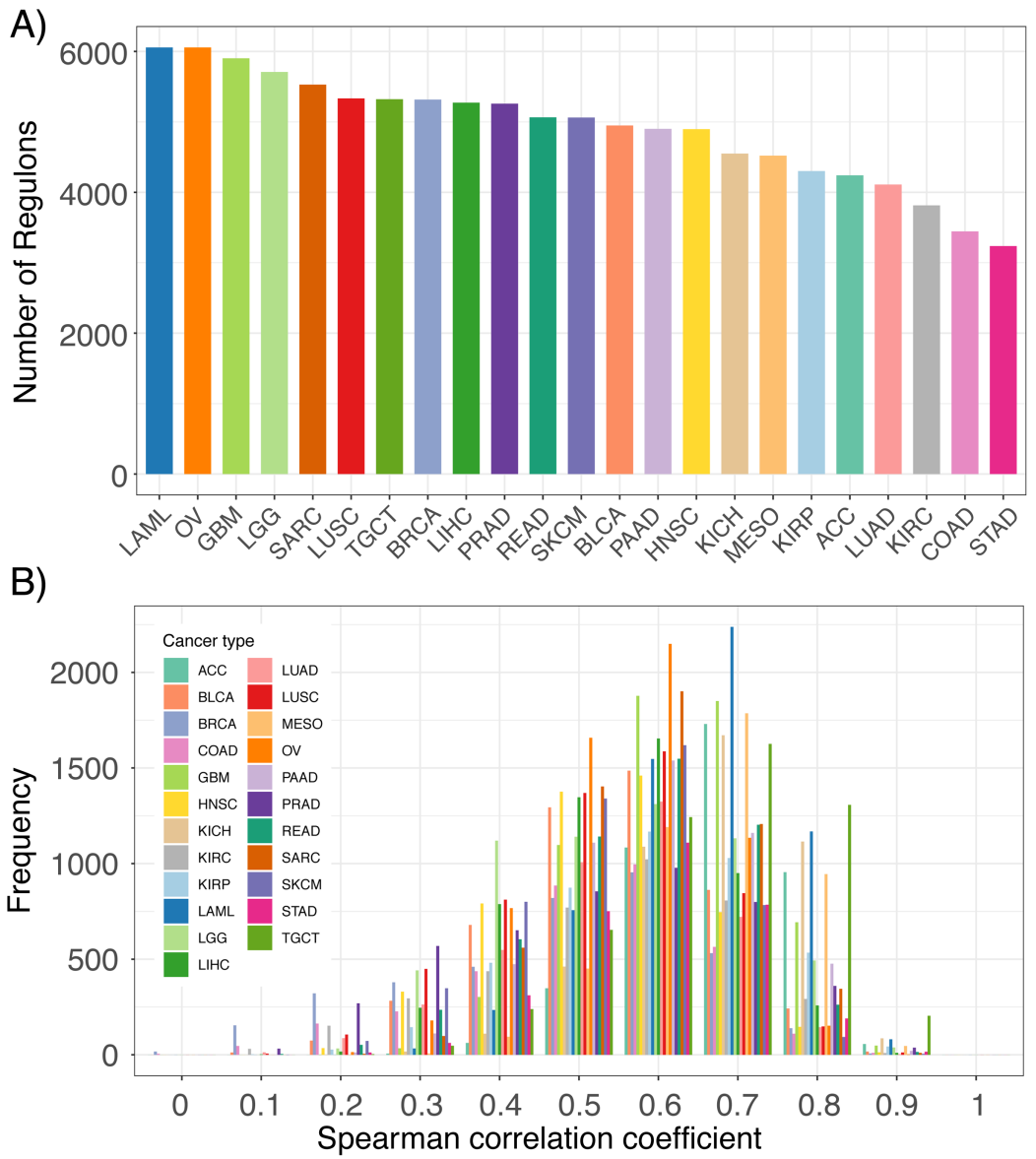

Figure 2 Comparison of inferred lncRNA expression and actual lncRNA expression. (A) Number of lncRNA regulons identified in 23 TCGA cancer types from the ARACne algorithm. Each regulon consists of a lncRNA and its associated protein coding genes. (B) Distribution of Spearman correlation coefficients from comparing inferred lncRNA expression with its actual expression using RNA-seq data from 23 TCGA cancer types. Full-size DOI: 10.7717/peerj.8797/fig-2

inferred lncRNA expression values were positively correlated with actual probe expression (Figs. 3D–3F), with a median correlation coefficient of 0.6. These results demonstrate that lncRNA expression can be inferred using the expression of protein coding genes in microarray datasets. Furthermore, we show our lncRNA inference platform is robust and can be generalized to different datasets as demonstrated by our analysis of TCGA, PRECOG and METABRIC.

Previous studies have shown that the expression patterns of lncRNAs recapitulate the four well-known molecular subtypes in breast cancer (*Su et al., 2014*), which are associated

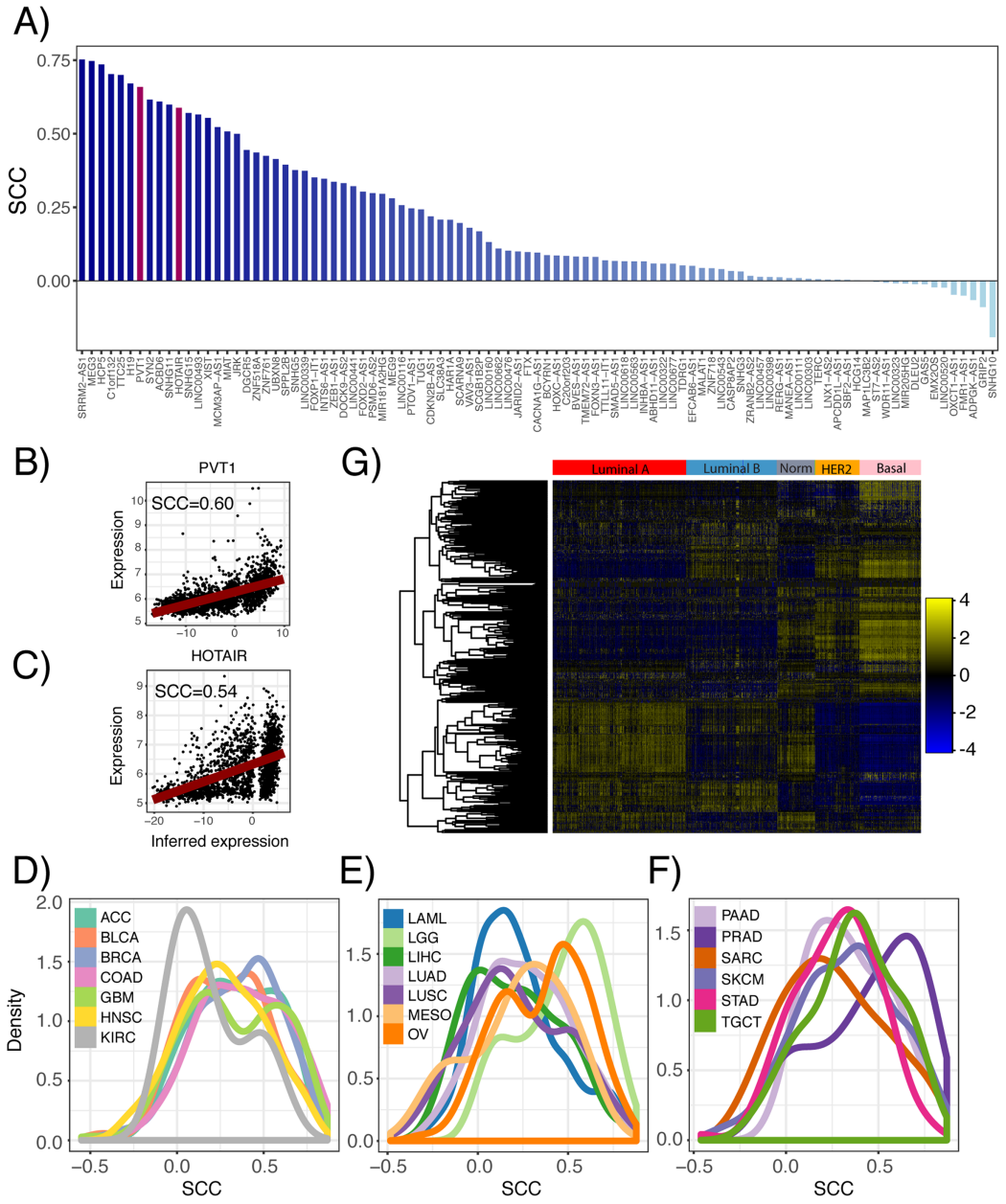

Figure 3 Comparison of inferred and actual lncRNA expression in METABRIC using available probes. (A) Waterfall plot showing correlation of inferred lncRNA expression and lncRNA probe expression in the METABRIC microarray dataset. Each lncRNA that had an available probe in the METABRIC microarray platform was selected to compare its inferred expression with its actual expression using Spearman correlation. Scatterplots show correlation of inferred and actual expression for (B) HOTAIR and (C) PVT1. (D–F) Distribution of correlation coefficients between inferred lncRNA expression and actual probe expression for 141 microarray datasets across 20 cancer types in the PRECOG compendium. Dashed vertical line indicates no correlation. Panels separated to increase legibility. (G) Heatmap showing inferred lncRNA expression differences between Luminal A, Luminal B, Normal-like, HER2-enriched and Basal breast cancer subtypes. Color bar shows *z*-score spectrum.

Full-size DOI: 10.7717/peerj.8797/fig-3

with tumor behavior and patient prognosis (*Perou et al., 2000*). We sought to confirm whether inferred lncRNA expression could similarly distinguish between the different breast cancer subtypes. Using the METABRIC dataset, we performed hierarchical clustering of the genes using the inferred expression values of 500 lncRNAs with highest variance across samples and indeed found subtype-specific differences in inferred lncRNA expression (Fig. 3G). This finding implies that lncRNA activity varies across breast cancer molecular subtypes and may play a role in tumor behavior.

Exploring the prognostic landscape of lncRNAs across 20 tumor types

In light of the current dearth of RNA-seq datasets with survival metadata and the expansive trove of microarray datasets that do have this valuable clinical information, we first used the PRECOG compendia to systematically infer lncRNA expression. Datasets in PRECOG all include patient survival information, and many patients within these datasets have been followed-up for longer periods of time compared to those recorded in TCGA, offering greater statistical power when performing survival analyses (*Clark et al., 2003*). We carried out a systematic inference of prognostic lncRNAs in PRECOG datasets with a sufficient number of probes (500) and performed a meta-analysis by combining the results within each tissue type. From this systematic screen, we identified a number of prognostic lncRNAs associated with both increased and decreased patient mortality risk in 13 PRECOG cancer types (Fig. 4A). These results suggest that lncRNAs may play a substantial role in the progression of all neoplasia. An overview of all lncRNAs and their associations with prognosis is available in the Supplemental Results.

Furthermore, we found that some lncRNAs including HOTAIR and H19 were associated with poor survival across multiple cancer types (Fig. 4A). These results are in accordance with several reports implicating these lncRNAs in neoplastic progression and metastasis across different cancer types (*Matouk et al., 2007*; *Gupta et al., 2010*). It has been suggested that HOTAIR promotes cancer invasiveness and metastasis by the induction of a more embryonic-like state (*Gupta et al., 2010*), which leads to increased resistance to known therapies and is a marker for poor prognosis in almost all cancer types (*Ge et al., 2017*; *Shibue & Weinberg, 2017*). However, some lncRNAs like TINCR were associated with both good and poor prognosis depending on the cancer type (Fig. 4B). Particularly, TINCR was associated with unfavorable prognosis in breast cancer, which is consistent with the implication of TINCR in promoting breast cancer tumorigenesis (*Xu et al., 2017*; *Liu et al., 2018*). TINCR can stabilize mRNA by preventing Staufen-mediated mRNA decay of differentiation genes in epidermal tissue (*Kretz et al., 2013*), but it is unclear whether this mechanism plays a role in tumor evolution.

To provide evidence that the identified prognostic lncRNAs are functionally relevant, we performed an integrative analysis of CRISPRi data generated from a high-throughput, systematic screen for lncRNAs that are essential for cancer cell growth (*Liu et al., 2017*). We calculated the enrichment of functional lncRNAs identified in MDA-MB-231 and K562 cell lines in the set of prognostic lncRNAs identified in breast and hematopoietic cancers, respectively. We observed that hazardous lncRNAs in hematopoietic cancers were enriched in essential lncRNAs (Odds ratio = 2.26, p = 7.1E−5)

A)

		lncRNA role	Brain	Glioma	Lung	Breast	Hemato-poietic	Ovarian	Liver	Germcell tumors	Bladder	Prostate	Melanoma	Head and neck	Sarcoma
Benj. meta-FDR	0.05	Haz.	2222	2511	2484	2171	1951	1939	893	647	227	273	229	103	1
		Pro.	2318	2549	2479	2190	1951	1978	717	668	315	398	435	102	0
	0.01	Haz.	1516	2217	2148	1917	1213	1235	273	207	10	142	57	0	0
		Pro.	1601	2245	2143	1936	1213	1276	176	169	20	234	169	0	0
Robust Benj. FDR	0.05	Haz.	1342	0	2155	1953	941	892	0	668	0	398	2	0	0
		Pro.	1261	0	2160	1934	941	854	0	647	0	273	1	0	0
	0.01	Haz.	561	0	1715	1664	473	408	0	169	0	234	0	0	0
		Pro.	523	0	1720	1645	473	366	0	207	0	142	0	0	0

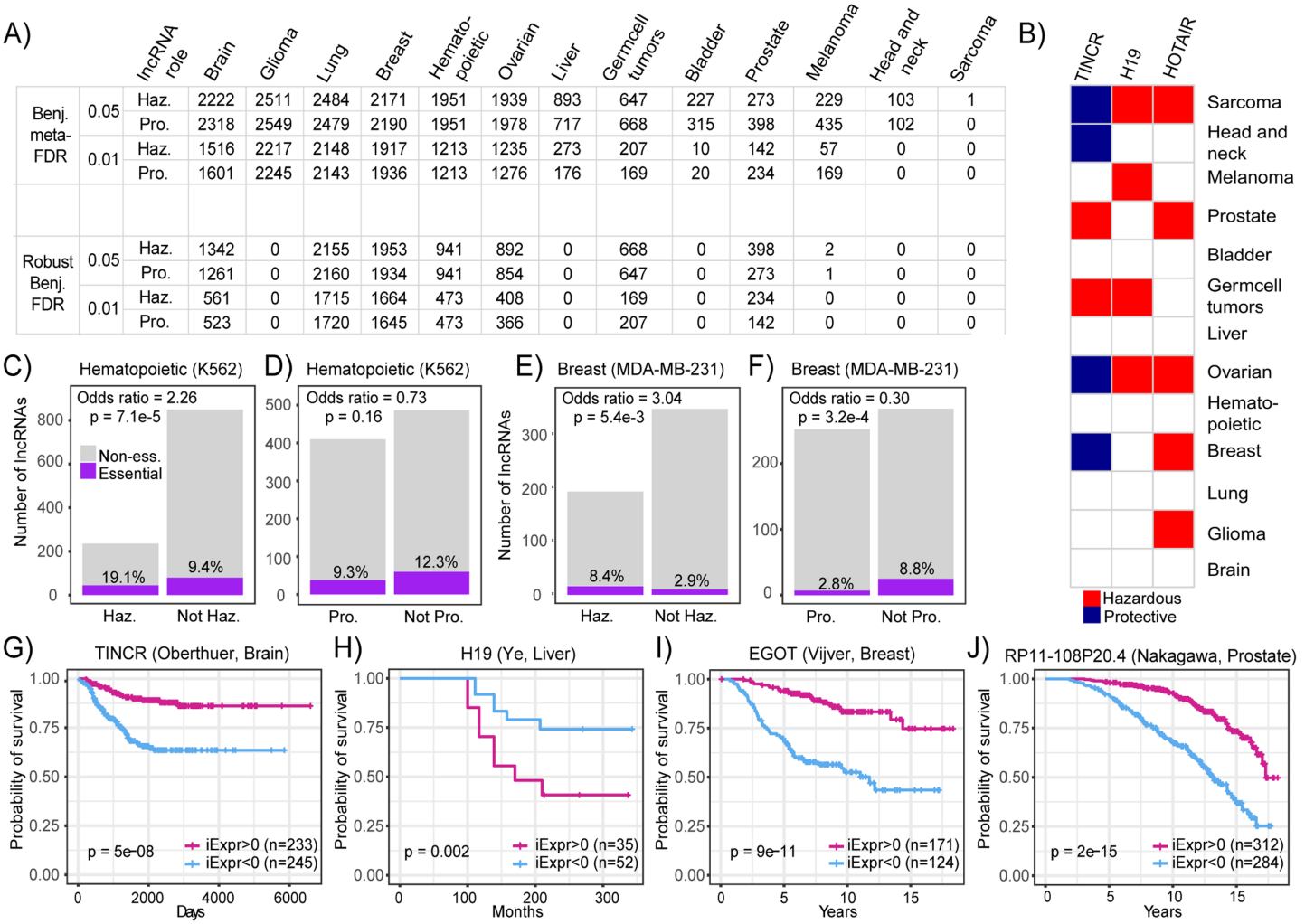

Figure 4 Systematic screening of prognostic lncRNAs in PRECOG compendium. (A) Table showing the number of lncRNAs identified to be associated with patient prognosis across PRECOG cancer types using standard and robust meta-analysis methods. Different cutoffs for the Benjamini meta-FDR and Robust Benjamini-FDR are displayed. Haz. indicates hazardous and Pro. indicates protective. (B) Selected prognostic lncRNAs (Adjusted $p < 0.001$) and their association with prognosis in each cancer type. White cells indicate associations with $p > 0.05$ or lncRNAs whose expression cannot be inferred in that cancer type. Bar plots showing odds ratio indicating enrichment overlap between essential lncRNAs in (C) K562 cells and (D) lncRNAs associated with prognosis in hematopoietic cancers in PRECOG. Bar plots showing odds ratio indicating enrichment overlap between essential lncRNAs in (E) MDA-MB-231 cells and (F) lncRNAs associated with prognosis in breast cancer in PRECOG. Haz. indicates hazardous and Pro. indicates protective. Kaplan–Meier plots showing association of (G) TINCR; (H) H19; (I) EGOT and (J) RP11-108P20.4 inferred expression with patient prognosis in selected brain, liver, breast cancer and prostate cancer datasets from PRECOG.

Full-size DOI: 10.7717/peerj.8797/fig-4

and protective lncRNAs were depleted in essential lncRNAs (Odds ratio = 0.73, $p = 0.16$) indicating that hazardous lncRNAs in hematopoietic cancers tend to be required for cancer cell growth compared to protective or non-prognostic lncRNAs (Figs. 4C and 4D). Likewise, we discovered that hazardous lncRNAs in breast cancer was also enriched in essential lncRNAs (Odds ratio = 3.04, $p = 5.4E{-}3$) and protective lncRNAs were depleted in essential lncRNAs (Odds ratio = 0.30, $p = 3.2E{-}4$), again suggesting that hazardous lncRNAs are more likely to be essential because they contribute to cell growth and are thus associated with increased mortality risk (Figs. 4E and 4F). Together, these results indicate

that hazardous lncRNAs identified in our analysis of breast and hematopoietic cancers are functionally relevant, at least in the context of in vitro cancer cell growth.

In addition to known cancer-associated lncRNAs, our screen also generated novel hypotheses about lncRNAs that have not been well-studied in certain cancer types. To highlight, high TINCR inferred expression was associated with improved survival in patients with brain tumors (Fig. 4G; p = 5E−8). Moreover, high H19 inferred expression was associated with decreased mortality risk among patients with liver tumors (Fig. 4H; p = 0.002). EGOT inferred expression was associated with decreased mortality risk in patients with breast cancer (Fig. 4I; p = 9E−11). This is consistent with a previous study, which showed that downregulation of EGOT correlates with worse clinicopathological features and poor prognosis in breast cancer (*Xu et al., 2015*). Lastly, we found that high inferred RP11-108P20.4 expression was associated with improved survival in prostate cancer (Fig. 4J; p = 2E−15), which coincides with a recent report introducing RP11-108P20.4 as part of a four lncRNA gene prognostic risk signature for prostate cancer (*Huang et al., 2002*). These results demonstrate that novel prognostic lncRNAs can be identified across several cancer types from common microarray datasets.

Furthermore, to assess the reproducibility of our screen, we performed a survival analysis of 23 TCGA cancer types to identify prognostic lncRNAs using their actual expression in each dataset. From this screen, prognostic lncRNAs (FDR < 0.05) were identified in five cancer types (LUAD, LGG, BLCA, LIHC and LAML). We stratified the lncRNAs into protective or hazardous and computed their enrichment, respectively, in protective or hazardous lncRNAs predicted from the PRECOG datasets. We identified significant overlap between the two sets of prognostic lncRNAs in all five cancer types (Fig. 5A). Moreover, we compared the Cox regression z-scores (TCGA) and meta z-scores (PRECOG) for all lncRNAs within lung adenocarcinoma, low-grade glioma and bladder cancer datasets and observed significant correlations (Figs. 5B–5D). These z-scores were calculated by dividing the Cox regression coefficient by its standard error and the meta z-scores were calculated using weighted Stouffer's z-score method using the dataset sample size as weights. Several well-studied lncRNAs including H19, BCAR4, GAS5, XIST, HOTAIR and EGOT had concordant z-scores (Figs. 5B–5D).

lncRNAs associated with prognosis localize to genomic regions under selective pressure

Operating under the hypothesis that genomic amplifications and deletions indicate regions of positive and negative selective pressure by the tumor, respectively (*Zack et al., 2013*), we aimed to provide further evidence that lncRNAs associated with prognosis are also linked to genomic structural abnormalities that confer a selective advantage to neoplastic cells. Thus, we analyzed CNA together with inferred lncRNA expression of each patient sample in the METABRIC data set. Strikingly, we observed a significant enrichment of lncRNAs associated with poor prognosis (hazardous) in amplified regions of the genome in 448 METABRIC patient tumors (Fig. 6A). In comparison, we only observed 54 patient tumors where amplified regions were significantly depleted of hazardous lncRNAs (Fig. 6B). Likewise, we observed 47 patients with deleted genomic regions enriched in

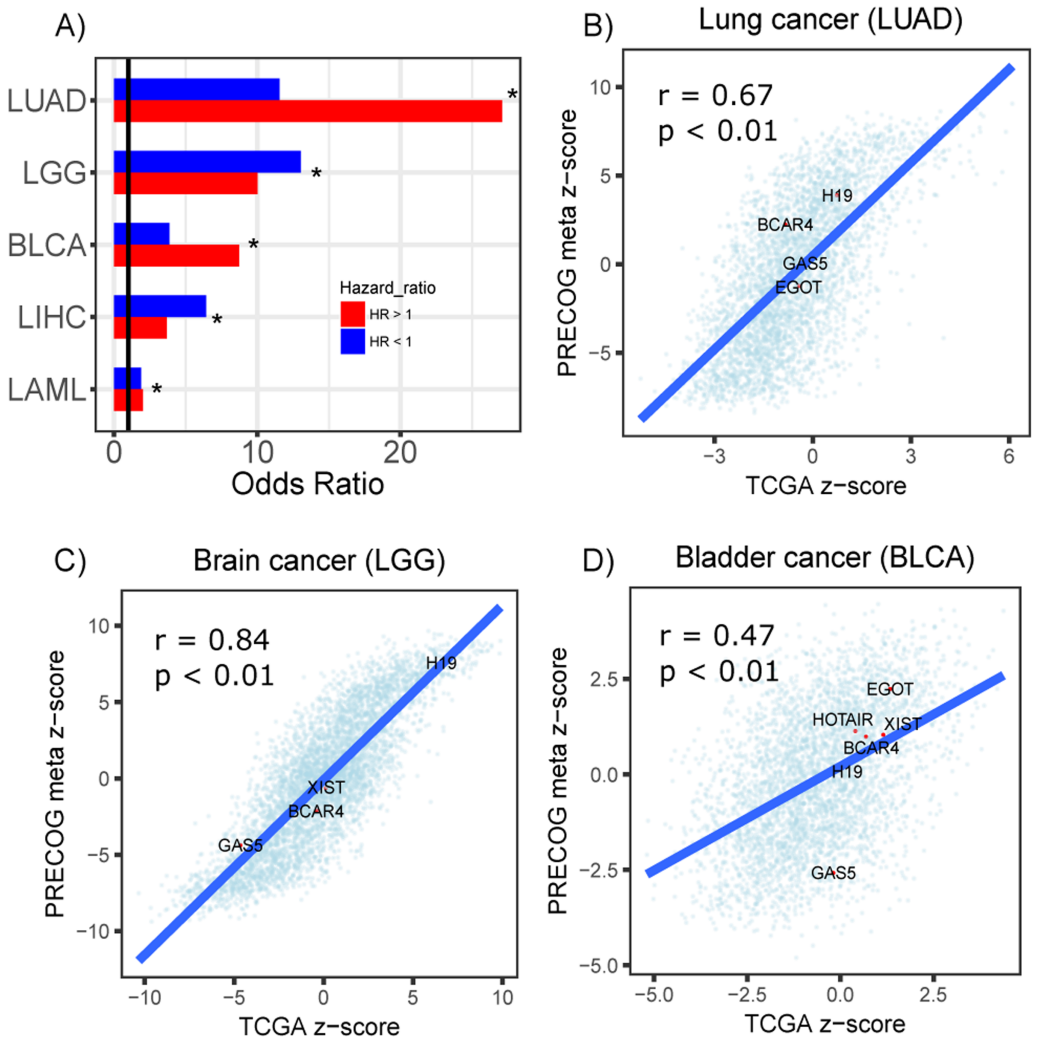

Figure 5 Prognostic lncRNAs identified in TCGA and PRECOG. (A) Barplots showing odds ratios from enrichment analysis of prognostic lncRNAs identified in TCGA and PRECOG for LUAD, LGG, BLCA, LIHC and LAML. Enrichment analysis was performed separately for lncRNAs with hazard ratios >1 (Red) and <1 (Blue). Vertical black line denotes an odds ratio of 1. Scatterplots showing correlation of *z*-scores and meta *z*-scores for all lncRNAs screened in TCGA and PRECOG, respectively, in (B) lung cancer, (C) brain cancer and (D) bladder cancer. Labeled points denote lncRNAs that have been characterized in previous literature. Full-size DOI: 10.7717/peerj.8797/fig-5

lncRNAs associated with decreased mortality (protective), compared to 17 patients who had protective lncRNAs depleted in deleted genomic regions. We also explored whether prognostic lncRNAs were enriched in amplified or deleted regions of the genome in pancreatic cancer, lung adenocarcinoma and glioblastoma TCGA datasets and observed consistent results (Fig. S2). In summary, these results indicate that prognostic lncRNAs localize to genomic regions that undergo CNA suggesting that they are under both positive and negative selective pressure by the tumor.

To demonstrate that lncRNAs associated with mortality risk are under selection, we highlight JRK and CADM3-AS1. In our analysis of JRK, we discovered that patients with

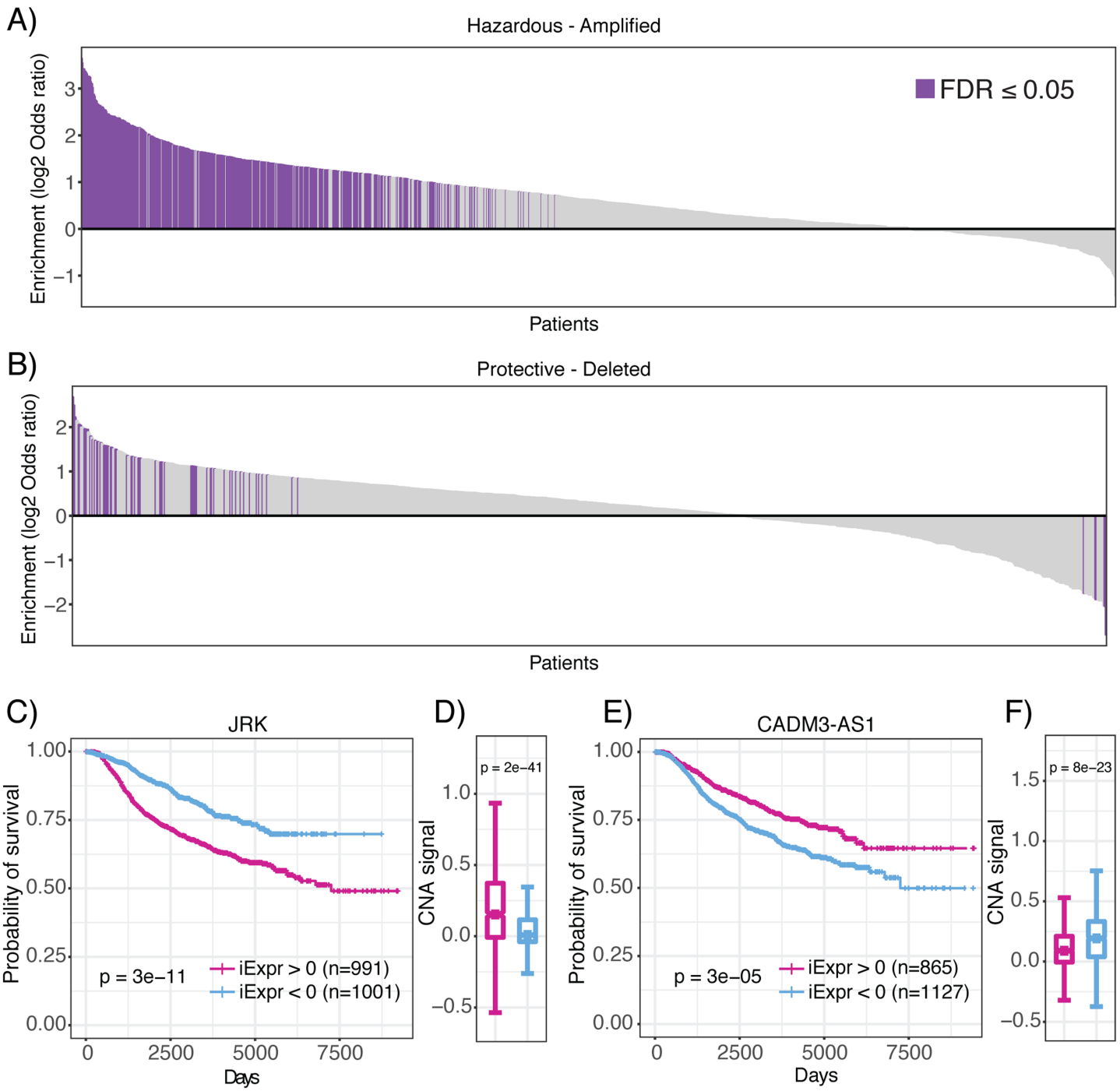

Figure 6 Enrichment of prognostic lncRNAs in genomic regions with copy number alterations. Waterfall plots showing enrichment of (A) hazardous and (B) protective lncRNAs in amplified and deleted regions of the genome for each patient, respectively. Log_2 odds ratio >0 indicates enrichment and log_2 odds ratio <0 indicates depletion. Purple bars indicate statistically significant enrichment. (C) High JRK inferred expression is associated with poor prognosis. (D) High JRK inferred expression is concentrated in regions with higher copy number signal across all METABRIC patients. (E) High CADM3-AS1 inferred expression is associated with favorable prognosis. (F) High CADM3-AS1 inferred expression is concentrated in genomic regions with lower copy number signal across all METABRIC patients. Full-size DOI: 10.7717/peerj.8797/fig-6

high inferred JRK expression exhibited a higher mortality rate compared to patients with low inferred JRK expression (Fig. 6C). In conjunction with this result, we also observed higher amplification signal of the region harboring JRK in patients with high inferred JRK expression compared to patients with low inferred JRK expression (Fig. 6D). These results suggest that JRK exhibits a pro-oncogenic effect because it is under positive selection by breast tumors, which consistently coincides with their association with increased mortality risk. In contrast, we found that high CADM3-AS1 inferred expression was associated with a more favorable prognosis compared to patients with low CADM3-AS1 inferred expression (Fig. 6E). In agreement with our prediction, we found that the genomic region harboring CADM3-AS1 was significantly more amplified in patients with low CADM3-AS1 expression compared to patients with high CADM3-AS1 expression (Fig. 6F). Together, these results suggest that due to CADM3-AS1's association with decreased mortality risk, it exhibits anti-tumor effects that are not selected for by breast neoplasms. Our analysis of prognostic lncRNAs in the context of CNA indicates that they are under selective pressure and provides evidence that they are functionally involved in cancer development. This hypothesis compliments a recent publication showing that somatic copy number variations in lncRNA loci were predictive of target gene expression and might be responsible for the dysregulation of dozens of cancer-associated genes (*Chiu et al., 2018*).

DISCUSSION

Investigation into lncRNAs using integrative and systematic approaches can help provide insight into the genome's "dark matter" and how it may influence disease and ultimately patient prognosis in cancer. Studies are now underway to characterize and dissect the intricacies of lncRNA regulatory mechanisms in several biological contexts which may revise our current understanding of genome regulation (*Cech & Steitz, 2014*). Hypothesis generating projects are essential for guiding the biomedical community towards investigating more promising leads as to accelerate the discovery of novel drug targets and biomarkers for cancer and other diseases. We have proposed a novel analysis framework to infer lncRNA expression in microarray gene expression data compendia and subsequently carry out systematic survival analysis to identify prognostic lncRNAs across 20 cancer types. Our approach is novel in that we utilize expression information from TCGA RNA-seq data to generate cancer-specific lncRNA regulon profiles that capture the lncRNA-gene relationships within a specific tissue context. We then apply these profiles to microarray gene expression data using a sensitive enrichment algorithm, BASE, to infer lncRNA expression based primarily on protein coding gene expression. We perform this analysis at a pan-cancer scale to identify new prognostic lncRNAs that have global and tissue-specific associations with survival. In contrast to other prognostic lncRNA pan-cancer analyses, we evaluated lncRNA expression in a large number of microarray datasets, providing us with a more comprehensive view of prognostic lncRNAs in cancer.

In particular, we were able to validate that the inferred lncRNA expression values are accurate and reproducible within and across several datasets. We identified novel associations between lncRNAs and patient mortality risk across 13 (out of 20 total) cancer

types that can be further evaluated in more detail. We confirmed that associations between lncRNA expression and mortality risk were consistent regardless of whether inferred or actual expression were used and showed that prognostic lncRNAs identified in breast and hematopoietic cancers were significantly enriched in functional lncRNAs required for cell growth. Furthermore, we demonstrated that hazardous lncRNAs were enriched within regions under positive selective pressure.

In spite of the evidence we provide, this study does have limitations that are imposed by the data. First, in each microarray dataset, we used the regulon profile that was generated from the TCGA cancer type that was the best match based on tissue. However, it was not always possible to find an exact cancer type match for each microarray dataset. Thus, using an inappropriately matched regulon profile may yield false associations. Second, univariate Cox regression models were used to screen for prognostic lncRNAs, which do not account for other clinical or demographic factors that may modify the associations. These may include age, gender, race, histological marker status, stage and grade. However, as an initial screen our framework can be further improved to include multivariate analyses in follow up studies if more specific hypotheses are to be tested (*McNamee, 2005*). Third, not all lncRNAs are poly-adenylated and are thus captured in poly-(A)-enriched RNA-seq or microarray studies. Due to this, we likely did not include all known lncRNAs in our study. Lastly, our analysis does not account for all cancer subtypes to address the issue of molecular heterogeneity within the same cancer type (*Gerdes et al., 2014*). As a result, certain associations between lncRNA expression and prognosis may only be valid in certain subtypes of the same cancer type. As stated previously, future analyses can address this issue by performing subgroup analyses within specific subtypes.

CONCLUSIONS

Our approach can be extended to other microarray gene expression datasets by utilizing our regulon profiles to infer lncRNA expression. As a result, it is possible to identify novel associations between lncRNAs and other disease phenotypes other than survival. Moreover, is possible to generate regulon profiles for other non-coding RNA species and infer their expression in microarray datasets. In summary, our systematic analysis introduces new avenues to investigate clinically relevant lncRNAs and demonstrate that these long, diverse transcripts constitute a new source of gene products that can serve as novel drug targets or biomarkers.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This work was supported by American Cancer Society Research grant #IRG-82-003-30, the National Center for Advancing Translational Sciences of the National Institutes of Health under Award Number KL2TR001088, the Rosaline Borison Memorial Pre-doctoral fellowship provided to Matthew H. Ung, the Cancer Prevention Research Institute of Texas (CPRIT) (RR180061 to Chao Cheng) and the National Cancer Institute of the National Institutes of Health (1R21CA227996 to Chao Cheng). Chao Cheng is a CPRIT

Scholar in Cancer Research. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
American Cancer Society Research: #IRG-82-003-30.
National Center for Advancing Translational Sciences of the National Institutes of Health: KL2TR001088.
Rosaline Borison Memorial Pre-Doctoral Fellowship.
Cancer Prevention Research Institute of Texas (CPRIT): RR180061.
National Cancer Institute of the National Institutes of Health: 1R21CA227996.
CPRIT Scholar in Cancer Research.

Competing Interests

The authors declare that they have no competing interests.

Author Contributions

- Matthew Ung conceived and designed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.
- Evelien Schaafsma analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.
- Daniel Mattox analyzed the data, prepared figures and/or tables, and approved the final draft.
- George L. Wang analyzed the data, prepared figures and/or tables, and approved the final draft.
- Chao Cheng conceived and designed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

All significant associations between inferred lncRNA expression and patient prognosis are available as a Supplemental File.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.8797#supplemental-information.

REFERENCES

Ali MM, Akhade VS, Kosalai ST, Subhash S, Statello L, Meryet-Figuiere M, Abrahamsson J, Mondal T, Kanduri C. 2018. PAN-cancer analysis of S-phase enriched lncRNAs identifies oncogenic drivers and biomarkers. *Nature Communications* **9**:883 DOI 10.1038/s41467-018-03265-1.

Alvarez MJ, Shen Y, Giorgi FM, Lachmann A, Ding BB, Ye BH, Califano A. 2016. Functional characterization of somatic mutations in cancer using network-based inference of protein activity. *Nature Genetics* **48(8)**:838–847 DOI 10.1038/ng.3593.

Ashouri A, Sayin VI, Van den Eynden J, Singh SX, Papagiannakopoulos T, Larsson E. 2016. Pan-cancer transcriptomic analysis associates long non-coding RNAs with key mutational driver events. *Nature Communications* **7**:13197 DOI 10.1038/ncomms13197.

Bartonicek N, Maag JLV, Dinger ME. 2016. Long noncoding RNAs in cancer: mechanisms of action and technological advancements. *Molecular Cancer* **15(1)**:43 DOI 10.1186/s12943-016-0530-6.

Byron SA, Keuren-Jensen KRV, Engelthaler DM, Carpten JD, Craig DW. 2016. Translating RNA sequencing into clinical diagnostics: opportunities and challenges. *Nature Reviews Genetics* **17(5)**:257–271 DOI 10.1038/nrg.2016.10.

Cabanski CR, White NM, Dang HX, Silva-Fisher JM, Rauck CE, Cicka D, Maher CA. 2015. Pan-cancer transcriptome analysis reveals long noncoding RNAs with conserved function. *RNA Biology* **12(6)**:628–642 DOI 10.1080/15476286.2015.1038012.

Cech TR, Steitz JA. 2014. The noncoding RNA revolution—trashing old rules to forge new ones. *Cell* **157(1)**:77–94 DOI 10.1016/j.cell.2014.03.008.

Cheng C, Yan X, Sun F, Li LM. 2007. Inferring activity changes of transcription factors by binding association with sorted expression profiles. *BMC Bioinformatics* **8(1)**:452 DOI 10.1186/1471-2105-8-452.

Ching T, Peplowska K, Huang S, Zhu X, Shen Y, Molnar J, Yu H, Tiirikainen M, Fogelgren B, Fan R, Garmire LX. 2016. Pan-cancer analyses reveal long intergenic non-coding RNAs relevant to tumor diagnosis, subtyping and prognosis. *EBioMedicine* **7**:62–72 DOI 10.1016/j.ebiom.2016.03.023.

Chiu H-S, Somvanshi S, Patel E, Chen T-W, Singh VP, Zorman B, Patil SL, Pan Y, Chatterjee SS, Sood AK, Gunaratne PH, Sumazin P, Cancer Genome Atlas Research Network. 2018. Pan-cancer analysis of lncRNA regulation supports their targeting of cancer genes in each tumor context. *Cell Reports* **23**:297–312 DOI 10.1016/j.celrep.2018.03.064.

Clark TG, Bradburn MJ, Love SB, Altman DG. 2003. Survival analysis part I: basic concepts and first analyses. *British Journal of Cancer* **89(2)**:232–238 DOI 10.1038/sj.bjc.6601118.

The ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489(7414)**:57–74 DOI 10.1038/nature11247.

Curtis C, Shah SP, Chin S-F, Turashvili G, Rueda OM, Dunning MJ, Speed D, Lynch AG, Samarajiwa S, Yuan Y, Gräf S, Ha G, Haffari G, Bashashati A, Russell R, McKinney S, Langerød A, Green A, Provenzano E, Wishart G, Pinder S, Watson P, Markowetz F, Murphy L, Ellis I, Purushotham A, Børresen-Dale A-L, Brenton JD, Tavaré S, Caldas C, Aparicio S, METABRIC Group. 2012. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**:346–352 DOI 10.1038/nature10983.

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 Research* **62(9)**:2695–2698.

Du Z, Fei T, Verhaak RGW, Su Z, Zhang Y, Brown M, Chen Y, Liu XS. 2013. Integrative genomic analyses reveal clinically relevant long noncoding RNAs in human cancer. *Nature Structural & Molecular Biology* **20(7)**:908–913 DOI 10.1038/nsmb.2591.

Esteller M. 2011. Non-coding RNAs in human disease. *Nature Reviews Genetics* **12(12)**:861–874 DOI 10.1038/nrg3074.

Evans JR, Feng FY, Chinnaiyan AM. 2016. The bright side of dark matter: lncRNAs in cancer. *Journal of Clinical Investigation* **126(8)**:2775–2782 DOI 10.1172/JCI84421.

Ge Y, Gomez NC, Adam RC, Nikolova M, Yang H, Verma A, Lu CP-J, Polak L, Yuan S, Elemento O, Fuchs E. 2017. Stem cell lineage infidelity drives wound repair and cancer. *Cell* **169(4)**:636–650 DOI 10.1016/j.cell.2017.03.042.

Gentles AJ, Newman AM, Liu CL, Bratman SV, Feng W, Kim D, Nair VS, Xu Y, Khuong A, Hoang CD, Diehn M, West RB, Plevritis SK, Alizadeh AA. 2015. The prognostic landscape of genes and infiltrating immune cells across human cancers. *Nature Medicine* **21(8)**:938–945 DOI 10.1038/nm.3909.

Gerdes MJ, Sood A, Sevinsky C, Pris AD, Zavodszky MI, Ginty F. 2014. Emerging understanding of multiscale tumor heterogeneity. *Frontiers in Oncology* **4(Suppl. 6)**:366 DOI 10.3389/fonc.2014.00366.

Guo L, Yao L, Jiang Y. 2016. A novel integrative approach to identify lncRNAs associated with the survival of melanoma patients. *Gene* **585(2)**:216–220 DOI 10.1016/j.gene.2016.03.036.

Gupta RA, Shah N, Wang KC, Kim J, Horlings HM, Wong DJ, Tsai M-C, Hung T, Argani P, Rinn JL, Wang Y, Brzoska P, Kong B, Li R, West RB, Van de Vijver MJ, Sukumar S, Chang HY. 2010. Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature* **464(7291)**:1071–1076 DOI 10.1038/nature08975.

Gutschner T, Diederichs S. 2012. The hallmarks of cancer: a long non-coding RNA point of view. *RNA Biology* **9(6)**:703–719 DOI 10.4161/rna.20481.

Hanahan D, Weinberg RA. 2000. The hallmarks of cancer. *Cell* **100(1)**:57–70 DOI 10.1016/S0092-8674(00)81683-9.

Huang K-C, Rao PH, Lau CC, Heard E, Ng S-K, Brown C, Mok SC, Berkowitz RS, Ng S-W. 2002. Relationship of XIST expression and responses of ovarian cancer to chemotherapy. *Molecular Cancer Therapeutics* **1**:769–776.

Iyer MK, Niknafs YS, Malik R, Singhal U, Sahu A, Hosono Y, Barrette TR, Prensner JR, Evans JR, Zhao S, Poliakov A, Cao X, Dhanasekaran SM, Wu Y-M, Robinson DR, Beer DG, Feng FY, Iyer HK, Chinnaiyan AM. 2015. The landscape of long noncoding RNAs in the human transcriptome. *Nature Genetics* **47(3)**:199–208 DOI 10.1038/ng.3192.

Kretz M, Siprashvili Z, Chu C, Webster DE, Zehnder A, Qu K, Lee CS, Flockhart RJ, Groff AF, Chow J, Johnston D, Kim GE, Spitale RC, Flynn RA, Zheng GXY, Aiyer S, Raj A, Rinn JL, Chang HY, Khavari PA. 2013. Control of somatic tissue differentiation by the long non-coding RNA TINCR. *Nature* **493(7431)**:231–235 DOI 10.1038/nature11661.

Kung JTY, Colognori D, Lee JT. 2013. Long noncoding RNAs: past, present, and future. *Genetics* **193(3)**:651–669 DOI 10.1534/genetics.112.146704.

Lachmann A, Giorgi FM, Lopez G, Califano A. 2016. ARACNe-AP: gene network reverse engineering through adaptive partitioning inference of mutual information. *Bioinformatics* **32(14)**:2233–2235 DOI 10.1093/bioinformatics/btw216.

Li J, Han L, Roebuck P, Diao L, Liu L, Yuan Y, Weinstein JN, Liang H. 2015. TANRIC: an interactive open platform to explore the function of lncRNAs in cancer. *Cancer Research* **75(18)**:3728–3737 DOI 10.1158/0008-5472.CAN-15-0273.

Liu Y, Du Y, Hu X, Zhao L, Xia W. 2018. Up-regulation of ceRNA TINCR by SP1 contributes to tumorigenesis in breast cancer. *BMC Cancer* **18(1)**:367 DOI 10.1186/s12885-018-4255-3.

Liu SJ, Horlbeck MA, Cho SW, Birk HS, Malatesta M, He D, Attenello FJ, Villalta JE, Cho MY, Chen Y, Mandegar MA, Olvera MP, Gilbert LA, Conklin BR, Chang HY, Weissman JS, Lim DA. 2017. CRISPRi-based genome-scale identification of functional long noncoding RNA loci in human cells. *Science* **355(6320)**:eaah7111 DOI 10.1126/science.aah7111.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15(12)**:550 DOI 10.1186/s13059-014-0550-8.

Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Favera RD, Califano A. 2006. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics* **7(Suppl. 1)**:S7 DOI 10.1186/1471-2105-7-S1-S7.

Matouk IJ, DeGroot N, Mezan S, Ayesh S, Abu-lail R, Hochberg A, Galun E. 2007. The H19 non-coding RNA is essential for human tumor growth. *PLOS ONE* **2(9)**:e845 DOI 10.1371/journal.pone.0000845.

Mattick JS, Makunin IV. 2006. Non-coding RNA. *Human Molecular Genetics* **15(1)**:R17–R29 DOI 10.1093/hmg/ddl046.

McNamee R. 2005. Regression modelling and other methods to control confounding. *Occupational and Environmental Medicine* **62(7)**:500–506 DOI 10.1136/oem.2002.001115.

Perou CM, Sørlie T, Eisen MB, Van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, Fluge O, Pergamenschikov A, Williams C, Zhu SX, Lønning PE, Børresen-Dale A-L, Brown PO, Botstein D. 2000. Molecular portraits of human breast tumours. *Nature* **406(6797)**:747–752 DOI 10.1038/35021093.

Ponting CP, Oliver PL, Reik W. 2009. Evolution and functions of long noncoding RNAs. *Cell* **136(4)**:629–641 DOI 10.1016/j.cell.2009.02.006.

Rinn JL, Chang HY. 2012. Genome regulation by long noncoding RNAs. *Annual Review of Biochemistry* **81(1)**:145–166 DOI 10.1146/annurev-biochem-051410-092902.

Sahu A, Singhal U, Chinnaiyan AM. 2015. Long noncoding RNAs in cancer: from function to translation. *Trends in Cancer* **1(2)**:93–109 DOI 10.1016/j.trecan.2015.08.010.

Schmitt AM, Chang HY. 2016. Long noncoding RNAs in cancer pathways. *Cancer Cell* **29(4)**:452–463 DOI 10.1016/j.ccell.2016.03.010.

Shibue T, Weinberg RA. 2017. EMT, CSCs, and drug resistance: the mechanistic link and clinical implications. *Nature Reviews Clinical Oncology* **14(10)**:611–629 DOI 10.1038/nrclinonc.2017.44.

Su X, Malouf GG, Chen Y, Zhang J, Yao H, Valero V, Weinstein JN, Spano J-P, Meric-Bernstam F, Khayat D, Esteva FJ. 2014. Comprehensive analysis of long non-coding RNAs in human breast cancer clinical subtypes. *Oncotarget* **5**:9864–9876.

Wang KC, Chang HY. 2011. Molecular mechanisms of long noncoding RNAs. *Molecular Cell* **43(6)**:904–914 DOI 10.1016/j.molcel.2011.08.018.

Xu S, Kong D, Chen Q, Ping Y, Pang D. 2017. Oncogenic long noncoding RNA landscape in breast cancer. *Molecular Cancer* **16(1)**:129 DOI 10.1186/s12943-017-0696-6.

Xu S-P, Zhang J-F, Sui S-Y, Bai N-X, Gao S, Zhang G-W, Shi Q-Y, You Z-L, Zhan C, Pang D. 2015. Downregulation of the long noncoding RNA EGOT correlates with malignant status and poor prognosis in breast cancer. *Tumour Biology* **36(12)**:9807–9812 DOI 10.1007/s13277-015-3746-y.

Yan X, Hu Z, Feng Y, Hu X, Yuan J, Zhao SD, Zhang Y, Yang L, Shan W, He Q, Fan L, Kandalaft LE, Tanyi JL, Li C, Yuan C-X, Zhang D, Yuan H, Hua K, Lu Y, Katsaros D, Huang Q, Montone K, Fan Y, Coukos G, Boyd J, Sood AK, Rebbeck T, Mills GB, Dang CV, Zhang L. 2015. Comprehensive genomic characterization of long non-coding RNAs across human cancers. *Cancer Cell* **28(4)**:529–540 DOI 10.1016/j.ccell.2015.09.006.

Zack TI, Schumacher SE, Carter SL, Cherniack AD, Saksena G, Tabak B, Lawrence MS, Zhsng C-Z, Wala J, Mermel CH, Sougnez C, Gabriel SB, Hernandez B, Shen H, Laird PW, Getz G, Meyerson M, Beroukhim R. 2013. Pan-cancer patterns of somatic copy number alteration. *Nature Genetics* **45(10)**:1134–1140 DOI 10.1038/ng.2760.