Main changes

* 162/166 (168/173 pairs) to only keep FDR sig ones → change numbers throughout text and figures
* Update figure 4 (new cnas and methylation barplots)
* Supplementary note on establishing baseline benchmark and calibration of our method (elastic net)
* New GTEx analysis, 54 pairs because now looking at actual regions within tissue not just all samples from the same organ
* IDH accounted for in LGG differential expression analysis
  + New network
  + New text
  + New heatmap
  + HOXA10-AS manuscript and more background on HOXA10 in glioblastoma stem cells
* **Added updated ordered citations**

–––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––

**Systematic analysis of lncRNAs reveals candidate prognostic biomarkers in multiple cancer types**

Keren Isaev1,2, Jüri Reimand1,2,\*

1. Ontario Institute for Cancer Research, Toronto, Ontario, Canada

2. Department of Medical Biophysics, University of Toronto, Toronto, Ontario, Canadaˇ

\* Correspondence Juri.Reimand@utoronto.ca

**ABSTRACT**

Long non-coding RNAs (lncRNAs) are nearly as abundant as protein-coding genes however, their biological functions and translation potential remain largely unexplored. In cancer, a few well-studied lncRNAs are known as molecular biomarkers and oncogenic modulators, suggesting potential for further translational discoveries. We performed a systematic pan-cancer analysis of 5,785 lncRNAs across 8,594 patient tumours to characterize their prognostic potential, genomic alterations and cellular function. We found 162 lncRNAs whose transcript abundance profiles robustly modeled cancer patient risk and validated 5/29 lncRNAs in an additional independent dataset. lncRNA transcript abundance associated with somatic copy number alterations, differential DNA methylation and pathway deregulation, suggesting potential activation mechanisms. We found that digital transcriptional activation of HOXA10-AS and HOXB-AS2 in gliomas associated with poor patient prognosis, brain developmental pathway activation and a tumour malignancy gradient. Our analysis highlights cancer-associated lncRNAs as promising candidates for biomarker development and functional experiments.

**INTRODUCTION**

The human genome encodes numerous long non-coding RNAs (lncRNAs) that lack protein-coding potential [(Iyer et al. 2015; Brunner et al. 2012; Hon et al. 2017; Gibb et al. 2011)](https://paperpile.com/c/Q7v0TQ/68t1+PgCt+d2jV+TEPk). lncRNAs are likely to be as frequently occurring as protein-coding genes (PCGs) as more than 20,000 high-confidence transcripts were recently annotated in FANTOM CAT [(Hon et al. 2017; Kung, Colognori, and Lee 2013)](https://paperpile.com/c/Q7v0TQ/d2jV+YFLP). lncRNAs are relatively less abundant than mRNAs but are often exceptionally tissue-specific [(Kung, Colognori, and Lee 2013; Iyer et al. 2015)](https://paperpile.com/c/Q7v0TQ/YFLP+68t1). Recent studies have increasingly characterized their roles in diverse biological processes including chromatin remodeling [(Engreitz et al. 2013)](https://paperpile.com/c/Q7v0TQ/Kce7), post-transcriptional gene regulation [(Gonzalez et al. 2015)](https://paperpile.com/c/Q7v0TQ/aAnk) and epigenetic silencing through interactions with DNA and proteins [(Tsai et al. 2010)](https://paperpile.com/c/Q7v0TQ/WYIW). lncRNAs are also involved in cancer-related processes of cellular differentiation and genome integrity [(B. Zhang et al. 2012; L. Yang et al. 2013)](https://paperpile.com/c/Q7v0TQ/N8F8+NnIg). Recent developments have enabled computational annotation of lncRNAs using their protein-binding motifs and interactions with transcription factors, RNA-binding proteins and microRNAs [(Kirk et al. 2018; Chiu et al. 2018; Ali et al. 2018)](https://paperpile.com/c/Q7v0TQ/U3Rd+qnLL+75M5). However, lncRNAs remain understudied compared to protein-coding genes (PCGs) and most of our biological knowledge is based on a few well-studied examples. For example, *PCA3* in protstate cancer and *HOTAIR* in breast cancer are established diagnostic biomarkers as they exhibit increased transcript abundance in tumours [(Tomlins et al. 2011; Gupta et al. 2010)](https://paperpile.com/c/Q7v0TQ/CDOa+Cd8r) compared to healthy tissues while functional studies have shown that they may regulate tumour cell proliferation [(Y. Wang, Liu, and Yao 2014)](https://paperpile.com/c/Q7v0TQ/xbRq) and epigenetic processes through interactions with PRC2 complex[(Hajjari and Salavaty 2015)](https://paperpile.com/c/Q7v0TQ/xA2C) respectively. The Human Pathology Atlas and others have previously assessed the impact of mRNA abundance on cancer outcomes [(Uhlen et al. 2017)](https://paperpile.com/c/Q7v0TQ/byKz). However, systematic evaluation of lncRNAs as molecular prognostic biomarkers has been limited to date.

To address these limitations, we conducted a comprehensive machine-learning analysis to model lncRNA transcript abundance levels as prognostic biomarkers of human cancers. We compiled a robust catalogue of prognostic lncRNAs across nearly 9,000 patient samples of 28 cancer types from the Cancer Genome Atlas (TCGA) [(“Website” n.d.)](https://paperpile.com/c/Q7v0TQ/lRCB). We evaluated the lncRNA profiles using cross-validation experiments and in an independent patient cohort from the International Cancer Genome Consortium (ICGC) [(Campbell et al. 2017)](https://paperpile.com/c/Q7v0TQ/wzALR). We found genetic and epigenetic alterations potentially responsible for these lncRNA signatures and annotated these further with gene regulatory networks and biological processes. Our study highlights the potential translational utility of the human non-coding transcriptome and provides a catalogue of high-confidence lncRNAs for future studies of biomarkers and oncogenic mechanisms.

**RESULTS**

**lncRNAs show tissue-specific transcript abundance and associate with patient survival (FIGURE 1)**

We first characterized transcript abundance of lncRNAs across 8,594 patients from 28 cancers types with matched RNA-sequencing (RNA-seq) data and clinical annotations from the TCGA PanCanAtlas dataset [(Liu et al. 2018)](https://paperpile.com/c/Q7v0TQ/1c0E) (Supplementary Table 1). We selected 5,785 high-confidence lncRNAs that were annotated by FANTOM CAT [(Hon et al. 2017)](https://paperpile.com/c/Q7v0TQ/d2jV)as well as Ensembl [(Zerbino et al. 2018)](https://paperpile.com/c/Q7v0TQ/7s31) and showed above baseline transcript abundance (Supplementary Figure 1a**,** Supplementary Table 2). Unsupervised clustering of lncRNA transcriptome revealed a robust clustering of tumour samples by organ systems that was comparable to protein-coding gene based analysis, confirming earlier observations in smaller patient cohorts (**Figure 1a,** Supplementary Figure 1b). These data elucidate the potential for translational discoveries in the non-coding transcriptome.

To investigate tissue-specific gene regulation of lncRNAs, we compared rank-normalized lncRNA transcript abundance across 54 pairs of tumour and healthy tissue types, using 6,927 transcriptomes of healthy tissues from the Genotype-Tissue Expression (GTEx) project [(Carithers et al. 2015; GTEx Consortium 2013)](https://paperpile.com/c/Q7v0TQ/eQl4+bTlr) (Supplementary Table 3). Many lncRNAs (3,487/5,785 or 60%) showed differences in lncRNA abundance in at least one comparison of tumours and healthy tissues, defined as median normalized ranks differing by one quartile (Wilcoxon rank-sum test, adjusted P < 0.05)(Supplementary Table 4). This included 3,487 unique lncRNAs overall, some of which were upregulated in a given cancer types while being downregulated in another cancer type relative to their matched normal tissues. Overall, tumour types were characterized by hundreds of such lncRNAs (mean: 439, range: 160-1967), where nearly half of these unique lncRNAs were tissue-specific (1,577/3,487 or 50%) (**Figure 1b,** Supplementary Figure 1c). While accounting for the same directionality of expression change, 347 of lncRNAs differed in 6-22 cancer types, including nine lncRNAs with strong changes in transcript abundance in all tumour-tissue comparisons evaluated here (six upregulated and three downregulated consistently), indicating their potential for biomarker discovery and functional characterization across human cancers (Supplementary Table 5). As positive controls, we found the well-known diagnostic lncRNAs *PCA3* [(Leyten et al. 2014)](https://paperpile.com/c/Q7v0TQ/DboC) and *PCAT1* [(Guo et al. 2016)](https://paperpile.com/c/Q7v0TQ/nYz7) with strongly increased transcript abundance in prostate tumours relative to healthy prostate tissue samples (0.95 versus 0.32 median ranks, Wilcoxon rank-sum test P < 1.93e-62). Thus, a subset of lncRNAs have tissue- and disease-specific transcriptional profiles that can be exploited for molecular biomarker development.

As a pilot study of lncRNAs as prognostic markers in human cancers, we associated transcript abundance with overall patient survival using multivariate Cox proportional-hazards (PH) models. After accounting for standard clinical variables such as age, sex, tumour stage and grade, we found that a majority of lncRNAs (3,662 of 5,785, 63%) were significantly associated with overall patient survival in at least one cancer type (Wald test, adjusted P < 0.1) (**Figure 1c**). Patient risk groups were defined for each cancer type using two strategies: median dichotomization for the majority of cases (63%) with above-zero lncRNA transcript abundance and zero-based dichotomization for 37% of cases where the lncRNA was transcriptionally silent in most tumours and highly expressed in others. Interestingly, high lncRNA transcript abundance was primarily associated with poor prognosis, suggesting that aberrant activation of silenced lncRNAs is often indicative of aggressive tumours (73% unfavourable and 27% favourable lncRNA associations across all cancer types).

In agreement with their tissue-specific transcription patterns, the majority of lncRNAs were associated with survival in only one cancer type (2,480/3,662 or 68%). The most survival-associated lncRNAs were found in clear cell renal carcinoma (1,222) and low-grade glioma (459) while 603 were found in both cancer types (**Supplementary Figure 1d**). To interpret this large number of survival associated lncRNA transcripts, we performed a correlation analysis of 2,489,697 pairs of survival-associated lncRNAs (looking at cancer specific prognostic lncRNA pairs) and found 71% with positive and 9% with negative correlations (Speraman , adjusted P < 0.05) (**Figure 1d**). Thus, the majority of survival-associated lncRNAs are likely transcriptionally co-regulated and a systematic strategy is needed to distinguish representative and robust prognostic biomarkers.

**Elastic net survival models identified 162 lncRNAs with robust prognostic performance (FIGURE 2)**

To identify multivariate prognostic signatures of lncRNAs, we adapted a machine-learning strategy comprising Cox proportional-hazards regression models with elastic net regularization [(Huang et al. 2015; Xiong et al. 2014; Yu et al. 2016; Jain et al. 2014)](https://paperpile.com/c/Q7v0TQ/nyv8+Nvsw+HRc8+NIYT)[(Huang et al. 2015; Xiong et al. 2014; Yu et al. 2016; Jain et al. 2014; Yuan et al. 2014)](https://paperpile.com/c/Q7v0TQ/nyv8+Nvsw+HRc8+NIYT+WmRx) and 100-fold-cross-validation with 70/30% random split of training and testing data for each cancer type (**Figure 2a**). Initial multivariate models included as predictors all lncRNAs that were univariately survival-associated in the training set (Wald test, *P* < 0.05). Feature selection during model fitting and regularization determined a non-redundant subset of lncRNAs as predictors in the training data. Subsequent cross-validation evaluated the models using concordance index (c-index), an accuracy measure extended to survival analysis [(Uno et al. 2011)](https://paperpile.com/c/Q7v0TQ/5zMv). To prioritize high-confidence lncRNAs, we selected the ones that appeared as chosen predictors in at least half of cross-validation models. For benchmarking, we fit alternative survival models with clinical predictors (e.g., tumour stage, grade, patient age) and with both lncRNA and clinical predictors (methods) through a similar process. Multivariate lncRNA-based survival models often significantly outperformed models with clinical predictors alone (**Figure 2b**), supporting their potential value as candidate prognostic biomarkers. To benchmark our analysis, we generated 100 random datasets for each cancer type and shuffled survival time/survival status (outcome). Running the same process as described above on each random dataset, we compared the distribution of c-indices from the real datasets and those obtained from shuffled data (Supplementary Figure 2).

Analysis of frequently selected features from cross-validation experiments revealed 168 lncRNAs, including 162 in a single cancer type and six in two cancer types (Supplementary table 6). The majority (117/168 or 70%) were unfavorable prognostic markers where poor outcome was associated with increased transcript abundance (median HR = 2.54) while 51 lncRNAs showed favorable transcript abundance (median HR = 0.42). As positive controls, we found *TINCR* [*(Tian et al. 2017)*](https://paperpile.com/c/Q7v0TQ/vVm9) and *CAHM* [*(Pedersen et al. 2014)*](https://paperpile.com/c/Q7v0TQ/WOyb) as prognostic markers in kidney clear cell renal carcinoma (KIRC). Comparison of different iterations of cross-validation indicated the high prognostic value of individual lncRNAs. The majority of lncRNA-based univariate models (156/162 or 94%) showed significantly higher prognostic accuracy across cross-validations compared to models with clinical variables alone (median increase of 0.12 in c-index; Wilcoxon rank-sum test, *FDR* <0.05) (**Figure 2c**). Further, combining individual lncRNAs and clinical features as predictors lead to an even greater improvement in model performance: combined models outperformed clinical-only prognostic models for 148/162 of lncRNAs (90%), leading to median c-index increase of 0.03 (range c-index improvement: 0.005-0.25). All 162 lncRNAs significantly exceeded the prognostic accuracy of lncRNAs selected randomly from our dataset of 5,785 lncRNAs (median of 0.73 vs median of 0.52, *FDR*<0.05, **Figure 2c)**. For example, the lncRNA *RP4-630C24.3* showed high prognostic accuracy in liver cancer (median c-index=0.85) compared to liver cancer models with clinical variables only (median c-index=0.5) or lncRNA and clinical combined models (median c-index = 0.55). A minority of lncRNAs (10/162 or 6%) did not show significant improvement in univariate models compared to models with clinical variables yet showed high accuracy overall. For example, the lncRNA *RP4-604G5.1* in thyroid cancer showed a median c-index of 0.91 in a univariate model and c-index of 0.94 when combined with patient age, sex and race. In contrast, survival models with randomly selected lncRNAs showed significantly lower accuracy in thyroid cancer (median c=0.53, Wilcoxon rank-sum test, adjusted *P* < 2.2e-16). Thus, these lncRNAs show comparable or better prognostic performance to clinical variables in cross-validation experiments.

To exclude confounding effects, we compared 84/162 (52%) lncRNA profiles with PCGs that were located in adjacent DNA sequence within +/- 10 kilo base pairs of gene bodies, including 40 divergent lncRNAs that share promoters with PCGs and 15 lncRNAs antisense to PCGs. Only few PCGs were also significantly prognostic in the same cancer type (14/126 or 11% at Wald *P* < 0.05) and in 11 of those cases, the lncRNA-based survival model outperformed the equivalent PCG-based model (median lncRNA c-index: 0.58 versus median PCG c-index: 0.53)(**Figure 2d**).

**Validation in additional datasets identifies high-confidence prognostic lncRNAs (Figure 3)**

To assess the generalizability of prognostic lncRNAs to unseen data, we used an external validation cohort from the ICGC/PCAWG project.We were able to evaluate 29/162 (18%) lncRNA candidates for which independent samples of relevant cancer types were available and where transcript abundance exceeded baseline detection levels (overall or non-zero group median of 0.05 FPKM-UQ) (Supplementary Table 7).

We first compared lncRNA-based hazard-ratios obtained in the discovery cohort (TCGA) and the validation cohort (PCAWG) and found that hazard ratios of lncRNA-based models were significantly positively correlated (Pearson rho = 0.42, *P* = 0.023), suggesting that lncRNAs showed consistent prognostic signals in the two datasets (**Figure 3a**). Importantly, we found 5/29 (17%) of these lncRNA candidates to be significantly associated with survival in an independent cohort (Wald test, P-value < 0.05) where the direction of the hazard was consistent in both cohorts. These included: *AC114803.3* (HR = 3.64 ; 95% CI, 1.60-8.27, *p* = 0.004)and *CTC-297N7.5* (HR = 0.41 ; 95% CI, 0.18-0.94, *p* = 0.03)in liver hepatocellular carcinoma, *RP11-731F5.2* (HR = 0.21 ; 95% CI, 0.06 - 0.80, *p* = 0.014) in head and neck squamous cell carcinoma, *LINC00524* (HR = 2.39 ; 95% CI, 1.20 - 4.76, *p* = 0.0163) in renal clear cell carcinoma and *CTD-2316B1.2* (HR = 4.19 ; 95% CI, 1.23 - 14.31, *p* = 0.038) in lung adenocarcinoma (**Figure 3b-f**) . Three of these lncRNAs showed zero-median transcript abundance in both discovery and validation cohorts, suggesting aberrant lncRNA activation in fewer high-risk patients and complete absence of expression in low-risk patients. Although, the two liver cancer candidates were adjacent to PCGs, all lncRNA-based models exceeded the performance of of related PCG based models (**Figure 2e**). Further, for this subset of lncRNAs, lncRNA-based survival models in the validation cohort showed improved c-index values compared to models with clinical variables such as patient age and sex (median lncRNA concordance: 0.62 versus median clinical variables concordance = 0.54). However, limited clinical annotations of the validation cohort warrant caution in interpreting these results. Meanwhile, two additional lncRNAs showed sub-significant trends (0.05 < P < 0.15) including *LINC00324* in liver cancer and *RP11-875011.1* in head and neck cancer. Overall, external validation of these novel lncRNA candidates prioritizes these targets for clinical translation and lends confidence to our analysis.

**Prognostic lncRNAs are enriched in copy number alterations and DNA methylation changes (FIGURE 4)**

To investigate potential genetic or epigenetic events driving lncRNA transcript abundance in patient risk groups, we studied genomic copy number alterations and DNA methylation patterns of lncRNA candidates. We first compared lncRNA transcript abundance of 71/162 candidates with available somatic copy number data (SCNA) between patients with deletions, neutral copy number and amplifications (segment means: <-0.3, -0.3 - 0.3, >=0.3 respectively) using a U-test. Further, we focused on lncRNAs that also had a correlation greater than 0.2 between segment mean and transcript abundance (FPKM-UQ) using Spearman correlation. We found that in 14 of these lncRNA candidates (20%), SCNAs were significantly associated with transcript abundance (U-test, FDR < 0.05 and spearman, *p* < 0.05). High transcript abundance of seven unfavourable lncRNAs was associated with copy number gains and low transcript abundance of seven favourable lncRNAs was associated with deletions (**Figure 4a, Supplementary table 8**). The correlation coefficients between SCNAs and transcript abundance ranged from 0.2 to 0.59 (median = 0.35). The number of patients whose lncRNA transcript abundance may be influenced by copy number changes was variable for each lncRNA candidate. For example, increased abundance of the lncRNA *RP13-1032I1.7* was associated with favourable prognosis in cervical cancer and out of the 95 patients with SCNA information, 18 risk patients were impacted by a deletion of the lncRNA region while 11 non-risk patients were impacted by an amplification suggesting that nearly a third of these patients’ lncRNA abundance was influenced by copy number changes **(Figure 4b)**. Overall, these results suggest that transcript abundance of prognostic lncRNAs is often influenced by SCNAs in tumours, suggesting potential genetic mechanisms affecting lncRNAs transcript abundance and patient risk.

We also evaluated the methylation of probes overlapping lncRNA transcripts. Overall, 73/162 (45%) lncRNA candidates overlapped at least one CpG probe. We found 20/73 (27%) lncRNA candidates with significantly different distributions of transcript abundance (FPKM-UQ) between patients with high methylation (beta >= 0.75), low methylation (beta <= 0.25) and nonspecific methylation (0.25 < beta < 0.75) of CpG sites located in lncRNA promoters or gene bodies (U-test, adjusted P < 0.05)(**Figure 4c, Supplementary table 9**). Further, we evaluated the correlation between each significant probe and lncRNA abundance (FPKM-UQ) to maintain only those candidates with absolute correlations greater than 0.2. In cases of multiple probes overlapping a lncRNA, we selected the one with the highest absolute correlation. Specifically, we found 10/20 (50%) of CpG probes were negatively correlated with lncRNA abundance. For example, in liver cancer, the transcript abundance of the favourable lncRNA *CTC-297N7.5* was associated with a methylated CpG site in the risk group (defined as beta values >= 0.75), located 1,332 bases upstream from the transcription start site (**Figure 4d**). Thus, patients with higher risk had lower transcript abundance and higher methylation at this CpG site while those with lower risk may have been unaffected by this form of transcriptional suppression. The median distance from probes to the transcription start site was 2,872 bases for negatively correlated probes and 11,730 bases for positively correlated probes. Overall, we found a sizeable fraction of potentially prognostic lncRNAs that are regulated transcriptionally by either copy number aberrations or DNA methylation (34/162, 21%). Further, additional lncRNAs may be driven by somatic SNVs and indels however, limitations in the availability of whole genome sequences of tumours in TCGA restricts a comprehensive analysis of such aberrations.

**lncRNA transcript abundance is often associated with previously defined molecular and clinical subtypes (FIGURE5)**

We further characterized the prognostic lncRNA candidates in the context of defined clinical and molecular tumor subtypes. Systematic analysis comprised 38 lncRNAs (23% of 162 lncRNAs) across 11/22 cancer types for which additional subtype data were available, using data curated by the R package TCGABiolinks[(Colaprico et al. 2015a)](https://paperpile.com/c/Q7v0TQ/TNhd). We found 109 instances (methods) where lncRNA transcript was significantly associated with a clinical or molecular tumor subtype across 10 cancer types (Chisq-test, adjusted *P* < 0.05) **(Supplementary Table 10)**. We compared univariate Cox-PH models fit using individual lncRNA candidates versus clinical or molecular features (**Figure 5a)**. As expected, many of these features that associated with our prognostic lncRNAs were also prognostic features within their respective tumor types. For example, transcript abundance of prognostic lncRNAs in low grade glioma (LGG) was often associated with IDH mutation status, prominent SCNAs such as gains of chromosome 19 and 20, and methylation statusof*MGMT* promoter, all documented prognostic features on their own.

We asked whether combining individual prognostic lncRNAs and tumor subtypes would lead to improved survival model accuracy. Indeed, adding dichotomized lncRNA transcript abundance status to subtype-based survival models improved accuracy (c-index) and model fit 94% of the time (102/109 associations (**Figure 5b**). Median improvement of accuracy was 0.05 (range 0.0005 – 0.24). For example, survival models combining the lncRNA *RP11-279F6.3* transcript abundance with clinical-stage lead to a significantly improved survival model in renal papillary cell carcinoma (Anova, adjusted *P* = 2.42e-05) with a concordance of 0.92 compared to 0.85 in the clinical-stage only based model. LGG candidate *RP5-1086K13.1’*s transcript abundance when combined with signal from the co-gain of chromosome 19 and 20 (a prognostic feature on its own) also increased the concordance of the clinical only model by 0.12 (Anova, adjusted *P* = 2.12e-07) **(Figure 5c)**.

We further focused on nine prognostic lncRNAs in low-grade glioma that were all associated with *IDH* gene mutation status, a well-established molecular prognostic feature in LGG [(Lu et al. 2012; SongTao et al. 2012)](https://paperpile.com/c/Q7v0TQ/7QGT+lc8x) (**Figure 5a**). Interestingly, several of these lncRNAs including the unfavourable (HR >1) *HOXB-AS2* and *HOXA10-AS* improved the fit of survival models when combined with *IDH* mutation status (Anova, adjusted *P* < 0.05). Patients with no *IDH* mutations and high lncRNA abundance were clearly stratified as the highest risk group compared to all other groups of patients (**Figure 5c**). These lncRNA candidates remained significantly prognostic when accounting for tumour purity and transcriptomic subtypes (classical, neural, mesenchymal and proneural). *HOXB-AS2* and *HOXA10-AS* were characterized by bimodal transcript abundance with abrupt increase in a small subset of samples and absent transcription in the majority of samples. In glioblastoma (GBM), we observed a much smaller fraction of patients with no detected levels of these two lncRNAs where nearly all tumors were IDH wild type in our cohort (**Supplementary Figure 5**). The two lncRNAs may thus be molecular markers of advanced gliomas and indicative of dismal outcome. In summary, integrating the abundance measures of lncRNAs can significantly improve the prognostic ability of previously established tumor subtypes.

**lncRNA-based patient risk groups are associated with differential transcript abundance of known cancer genes and pathways (FIGURE 6)**

To investigate the molecular pathways and processes associated with prognostic lncRNAs, we determined PCGs with differential abundance in patient risk groups and interpreted them using pathway enrichment analysis [(Reimand et al. 2007, 2019)](https://paperpile.com/c/Q7v0TQ/e5gj9+fsK2). Large transcriptional differences were apparent in lncRNA-associated high-risk patient tumours compared those with better prognosis. The majority of lncRNA-associated risk groups (101/168 or 60%) were characterized by differential expression of at least 30 PCGs (median group by 267 PCGs) (fold change > 2, adjusted *P* <0.05) (**Supplementary Table 11**). Further, we identified 269 (256/7,897 total differentially expressed PCGs or 3%) PCGs annotated previously as cancer genes.

Pathway enrichment analysis highlighted a large extent of functional diversity of these PCG signatures with 2,587 unique pathways identified in total composed of Reactome and Gene Ontology pathways (*FDR* <0.01 from g:Profiler) (**Figure 6a**) (**Supplementary Table 12**). The majority of pathways were enriched in few lncRNA-associated risk groups (1,835/2,587 or 71% associated with five lncRNAs at most) while a minority of processes (28/2,587 or 0.01%) including ion channel activity and extracellular matrix were associated with more than 20 lncRNAs. The pathway analysis suggests that the prognostic lncRNAs are potential markers of many diverse aspects of tumor biology that may define tumour progression and patient outcomes.

**Prognostic** l**ncRNAs *HOXA10-AS* and *HOXB-AS2* associate with brain developmental pathways and a malignancy gradient across LGG and GBM**

We hypothesized that transcriptional correlations of prognostic lncRNAs and protein-coding genes would converge on a common set of molecular pathways and processes indicative of tumor biology. We focused on LGG and performed integrative pathway analysis across differential mRNA abundance signatures of nine lncRNAs using our new pathway analysis method ActivePathways [(Paczkowska et al. 2018)](https://paperpile.com/c/Q7v0TQ/gniO). Fusion of differentially expressed mRNA signatures across the nine LGG lncRNA candidates revealed 718 PCGs with significant differences in mRNA abundance in one or several lncRNA-associated risk groups. These genes were enriched in 124 GO biological processes and Reactome pathways (**Figure 6b**). The majority of pathways (75/124 or 60%) were associated with transcript abundance of multiple prognostic lncRNAs. However, different pathway genes were involved in each lncRNA-associated risk group, indicating convergence of prognostic signatures at the pathway level. The most prominent processes and pathways included ion transport, synaptic signaling and brain development indicating that prognostic lncRNAs are associated with global transcriptional de-regulation of cancer hallmarks and brain developmental pathways in LGG risk groups (**Figure 6b)**.

We focused on a group of three brain developmental processes that were enriched in two of nine lncRNA-associated risk groups, *HOXB-AS2* and *HOXA10-AS*. The enriched pathways included 50 PCGs that showed differential mRNA abundance in patient risk groups. These included nine HOX transcription factors involved in organism development (*HOXA, HOXB* and *HOXC* clusters) and one Cancer Gene Census gene, *NKX2-1*. Several of the differentially expressed mRNAs were also prognostic in low grade glioma in a similar manner to our lncRNA candidates including *DMRTA2*, *CENPF*, *SFRP2* and several of the HOX cluster genes including *HOXA4*, *HOXC4* and *HOXA3*. These particular genes have not been described as prognostic in glioma before and there is potential in integrating lncRNA transcript abundance in addition to associated pathway gene to improve patient stratification based on underlying biological subtypes that may act as functional units.

To characterize the lncRNA candidates in the context of glioblastoma (GBM), a highly malignant type of brain tumours, we performed hierarchical clustering of the transcription profiles of lncRNAs and developmental genes across all low-grade glioma and glioblastoma samples **(Figure 6c)**. The clustering showed that lncRNA transcription correlated with a malignancy gradient: high-risk LGG patients with increased lncRNA transcript abundance clustered with the aggressive GBMs whereas the latter tumors showed even higher transcript abundance of these lncRNAs. Similar correlations and anti-correlations were apparent for brain developmental genes and cancer genes we identified in the pathway analysis. Although our LGG lncRNA candidates were not individually prognostic in GBM, these showed even higher transcript abundance levels in GBMs compared to LGGs (**Supplementary Figure 6b**).

Two prognostic HOX-associated lncRNAs *HOXB-AS2* and *HOXA10-AS* were transcriptionally co-activated with HOX TF genes and known cancer genes *NKX2-1*, *OTP* and *WNT2* high-risk LGGs and GBMs. *OTP* has been associated with brain development and cellular differentiation and previously shown to be prognostic in pulmonary carcinoids [(Swarts et al. 2013; Acampora et al. 1999)](https://paperpile.com/c/Q7v0TQ/wLXi+PSpn) . *WNT2* mRNA was previously shown to be upregulated in gliomas [(H. Zhang et al. 2018)](https://paperpile.com/c/Q7v0TQ/E8Hr). While *HOXB-AS2* is yet to be functionally described, cellular knockdown of *HOXA10-AS* previouslyshowed to reduce glioma cell line proliferation [(Dong et al. 2018)](https://paperpile.com/c/Q7v0TQ/Btmv). However, the biological mechanisms of both of these lncRNAs remains to be described in detail. Meanwhile, a similar lncRNA *HOTAIRM1* located upstream of *HOXA10-AS* is known to contribute to the three-dimensional chromatin organization and transcriptional activation of adjacent HOXA family TFs [(X. Q. D. Wang and Dostie 2017)](https://paperpile.com/c/Q7v0TQ/hyme4), suggesting that *HOXA10-AS* and *HOXB-AS2* may also activate developmental genes and thus, directly or indirectly contribute to more aggressive tumour phenotypes. Overall these findings suggest potential functional roles of prognostic lncRNA candidates with regard to known cancer genes and pathways that contribute to aggressive and high-risk tumours.

**DISCUSSION**

Our knowledge of cancer driver genes and molecular classifiers is primarily derived from the protein-coding genome while the vast majority of the genome represents non-coding sequence and remains largely understudied. Our findings of lncRNAs as prognostic factors in multiple cancer types are consistent with the increasing appreciation of lncRNAs in diverse cellular processes and human diseases [(Mondal et al. 2018)](https://paperpile.com/c/Q7v0TQ/RhXS). We highlight a facet of the non-coding genome that has great potential for basic and translational discoveries. Our machine learning analysis identified a subset of lncRNAs as robust predictors of patient survival in cross-validation and experiments and external data, suggesting that these transcripts should be further evaluated as molecular prognostic biomarkers in diverse molecular datasets and ultimately clinical trials. Specific transcript abundance patterns of lncRNAs in tumors and healthy tissues also indicates potential for discovering diagnostic lncRNAs. However, our analysis remains inconclusive to whether these candidate lncRNAs are functional in cancer cells or represent passive indicators of underlying transcriptional activity. On the one hand, lncRNAs may be modulated transcriptionally or epigenetically as side effects of regulatory programs that control hallmark cancer pathways such as proliferation. For example, sharp digital activation of some hazardous lncRNAs suggests disrupted transcriptional repression. On the other hand, a subset prognostic lncRNAs may be functional in cells and drive oncogenesis or inhibit tumor suppression [(Tian et al. 2017; Y. Yang et al. 2017)](https://paperpile.com/c/Q7v0TQ/vVm9+HC0v) .In support of this, we found several lncRNAs that were affected by copy number alterations or DNA methylation patterns and showed prognostic potentials beyond those of adjacent protein-coding genes and common molecular subtypes. However, our prognostic lncRNA candidates are ultimately limited by the transcriptional and clinical information that was available for inference and validation. Tumor cohorts that we studied were under-represented in rare and early-stage malignancies and were limited in recorded clinical variables and patient follow-up. It is plausible that lncRNA transcription is associated with additional environmental, genetic and phenotypic variables that confounded our inference. We used RNA-seq datasets that had been optimized for mRNA quantification and thus additional lncRNAs likely remain uncharacterized or below the detection limit. Future multi-omics datasets with deep clinical profiles of patients will enable further discoveries and validation. Our study is a step towards systematic characterization of the non-coding transcriptome as molecular biomarkers and functional regulators of oncogenesis.

**METHODS**

Data Collection:

We downloaded RNA-Seq data for 28 tumor types from the Genome Data (https://portal.gdc.cancer.gov). Overall survival data was downloaded from the latest publication of the PanCanAtlas [(Liu et al. 2018)](https://paperpile.com/c/Q7v0TQ/1c0E). SCNA and methylation data were downloaded for the 22 cancer types in which we identified lncRNA candidates using Firehose (https://gdac.broadinstitute.org). Additional clinical data such as alcohol, smoking status and molecular subtypes was downloaded using the R package TCGABiolinks [(Colaprico et al. 2015b)](https://paperpile.com/c/Q7v0TQ/PceD4). We downloaded processed GTEx V7 (https://gtexportal.org/home/) RNA-Seq data in TPMs for 11,688 samples from 53 tissue types across 714 patient donors.

Specifically, Affymetrix Genome-Wide Human SNP Array 6.0 was the platform from which SCNA data was obtained and Illumina Human DNA Methylation 450K array was the platform from which DNA methylation data was obtained. Illumina HiSeq 2000 RNA Sequencing was the platform from which gene expression of lncRNAs and mRNAs was obtained in TCGA. We intersected clinical data and gene expression data for each cancer type to retain cohorts that had both levels of information.

Eight of the 28 cancer types analyzed using data from TCGA were also analyzed for RNA-Sequencing in PCAWG with substantial cohort sizes[(Campbell et al. 2017)](https://paperpile.com/c/Q7v0TQ/wzALR). These eight cohorts were used for external validation of prognostic candidates. At the beginning of the study, we excluded all patient samples from TCGA that also appeared in PCAWG to maintain the external validation set as large as possible.

For lncRNA annotations, we downloaded the latest, comprehensive annotation set of 5’ lncRNA CAGE peaks from FANTOM - CAT [(Hon et al. 2017)](https://paperpile.com/c/Q7v0TQ/d2jV). We studied 5,785 lncRNAs overall that were annotated by FANTOM and for which genes expression was annotated in TCGA and PCAWG RNA-Seq data.

RNA-Seq data processing:

For TCGA data, we retrieved processed RNA-Seq files in FPKM-UQ and raw counts. Tumour RNA-Seq data from the ICGC-PCAWG cohort[(Campbell et al. 2017)](https://paperpile.com/c/Q7v0TQ/wzALR) was uniformly processed as previously described[(PCAWG Transcriptome Core Group et al. 2017)](https://paperpile.com/c/Q7v0TQ/YSvHM). We retrieved a patient gene matrix in FPKM-UQ values for our analysis.

Processing of lncRNA RNA-Seq data

lncRNAs are often very lowly expressed and we first removed those lncRNAs that had were not detected in any patient across all cohorts in TCGA RNA-Seq data (n=94). Further, we evaluated median abundance for each lncRNA in every cancer type. We considered all lncRNAs with median FPKM-UQ greater than 0 or those lncRNAs that had at least 15 patients with FPKM-UQ values greater than 100 FPKM-UQ if the median was 0.

Evaluating differences in lncRNA expression between tumors and normal tissues:

We compared tumour RNA-Seq data from TCGA with RNA-Seq data from GTEx to compare relative lncRNA abundance. Since the processing of these two datasets was done differently and these are not patient matched tissues, we conducted an analysis of the ranked gene expression values for the candidate lncRNAs. Within each cancer/tissue cohort, all genes (including protein-coding genes and lncRNAs) were ranked from least expressed to most expressed for each patient sample. These ranks were then used to compare the distribution of the relative expression of each lncRNA candidate between the tumor cohort from TCGA and the normal samples from GTEx. We used the two-sided U test to obtain a *P*-value for each lncRNA candidate and further evaluated those with adjusted *P* < 0.05 and a median rank difference greater than 0.25.

t-SNE analysis:

We produced the t-SNE plot in **Figure 1a** using the RTsne [(Krijthe 2015)](https://paperpile.com/c/Q7v0TQ/SajQL) package with default parameters. FPKM-UQ values were log1p transformed for all detectable lncRNAs to generate this plot and labelled by tumour type.

Training survival models and evaluating generalizability

For each cancer type, we evaluated the survival association between the list of 6,000 lncRNAs and overall survival outcome. We also evaluated the association between available clinical variables and overall survival for comparison. We split samples into two groups: 70% as the training set and 30% as the test set (for each cancer type). Patients within each training cohort were median dichotomized by each lncRNA’s expression. If the median was 0, patients with expression greater than 0 were labelled as high and those with 0 expression were labelled as low. When evaluating lncRNA predictors, we first set a pre-filtering criterion within the training set such that only lncRNAs with a univariate Cox model, likelihood ratio test with P < 0.05 would be kept. Next, we used a Cox proportional hazards model with Elastic Net (for feature selection) to train the survival models with lncRNA expression profiles as predictors. All univariate models were built using the R package “survival” while the Elastic Net was performed using the R package “glmnet” where the penalty parameter λ was determined by fivefold cross-validation within each training set. The multivariate Cox Elastic Net models were then applied to the remaining 30% of the test set to obtain a concordance index (c-index) using the R package “survcomp”. Meanwhile, clinical variables that were available for each cancer types (age was always available while tumor stage, grade and ethnicity were rarer) were used to build a multivariate model using the training set and applied on test set in a similar manner. Lastly, the available clinical variables were integrated with the lncRNA expression profiles selected by Elastic Net into one multivariate model also applied on the test set. Thus, there were three c-indices obtained overall for each round of training. The whole outlined process was repeated 100 times, randomly splitting the data each time. For each cancer type we then compared the three distributions of c-indices using the two sided U test with clinical variables as the reference. To assess model performance on random data, we shuffled survival outcome while maintaining all predictor variables (lncRNAs and clinical variables) as is. We generated 100 random datasets and conducted 100 cross-validations on each one. We conducted the analysis as described above and compared c-indices between models fit using shuffled outcome data and real outcome data using a two sided U-test.

Extracting most clinically relevant lncRNA candidates:

In our cross validations, we fit Elastic Net Cox models using training data obtained from random splits of the patient cohort. At each such iteration, lncRNA candidates were selected to be kept in the model and the lncRNAs that were selected at each iteration may vary based on the composition of the training set. To select the more clinically relevant lncRNAs for our further analysis, we summarized the number of times each lncRNA was kept in an elastic net survival model. We chose to include lncRNAs within each cancer type that were included in at least 50% of the iterations to obtain the most consistent candidates. This list of lncRNAs was further evaluate on a one by one basis. For example, we fit multivariate Cox PH models using each lncRNA candidate and available clinical variables in respective cancer cohorts to ensure that the prognostic effect of lncRNAs remained present when these additional variables were accounted for. Further, we evaluated schoenfeld residuals to confirm that the proportionality assumption of the Cox-PH model was met, they are displayed for each lncRNA candidate in Supplementary Table 6.

Internal and external validation of most clinically relevant lncRNAs:

To obtain a sense of the performance of individual lncRNA candidates within the TCGA dataset, we conducted a second round of internal cross-validation. Looking at one lncRNA candidate at a time, we split the respective cancer cohort into 70% training and 30% for testing as before. Univariate Cox PH models were fit and evaluated on the test set to obtain a distribution of c-indices for each lncRNA candidate. Similarly, we conducted internal cross-validation of clinical variables, fitting multivariate Cox PH models and evaluating their performance on test sets using the c-index to compare this distribution to the one obtained by each lncRNA candidate in the respective cancer cohort. These distributions were compared using the two-sided U test where multiple testing was corrected for using Benjamini-Hochberg.

To externally evaluate the ability of our lncRNA candidates to stratify patients by risk, we used patient data from independent cohorts available in the ICGC-PCAWG (https://docs.icgc.org/pcawg/). Out of the 22 cancer types in which we identified lncRNA candidates, eight had external RNA-Seq and clinical data. This included lncRNA candidates from ovarian, breast and liver and pancreatic cancer, as well as clear cell renal carcinoma, papillary renal carcinoma, lung adenocarcinoma and uterine endometrial carcinoma. For those cancer types for which data was available in the two data-sets, we evaluated the performance of individual lncRNA candidates within their respective cancer using two approaches. For each lncRNA candidate, we dichotomized patients by expression as before. We further considered only those lncRNAs with at least five patients with FPKM-UQ values greater than 0.05. This allowed us to validate 29/162 (18%) of our lncRNAs candidate. As before, we fit univariate Cox-PH models with binary lncRNA predictors and plotted their Kaplan-Meier survival curves using the ‘Survival’ package in R.

Comparing lncRNA and nearby protein-coding genes:

Using bedtools [(Quinlan 2014)](https://paperpile.com/c/Q7v0TQ/ziNl2), we first intersected our lncRNA candidate coordinates with PCGs to capture all PCGs within 5,000 base pairs up/downstream of lncRNAs. From this we identified 110 lncRNA-PCG pairs that we evaluated further for differences in survival associations. For each pair, we fit univariate Cox-PH models using binary lncRNA abundance labels and compared the fit of this model to Cox-PH models fit using binary PCG abundance labels. We compared the two models using Concordance measures Wald test P-values. We also fit a multivariate model using both PCG and lncRNA labels and compared this to a univariate PCG model using anova. Multiple testing was corrected for using Benjamini-Hochberg Procedure.

Evaluating genetic and epigenetic aberrations associated with lncRNA candidates:

To study the potential underlying differences between risk groups, we first intersected all copy number segments from the Affymetrix SNP 6.0 array with candidate lncRNA coordinates (GENCODE V22)[(Derrien et al. 2012)](https://paperpile.com/c/Q7v0TQ/cinx). Candidates that overlapped segments were then evaluated for the association between the lncRNAs SCNA and gene expression. We considered only segments less than 20 megabases to avoid whole chrosomrome deletions or duplciations. We were able to evaluate 71/162 of our lncRNA candidates as the others did not overlap any segments from the Affymetrix array. Further not all patients within a cancer cohort had SCNA information and we restricted our analysis to those with at least 20 patients. This resulted in a total of 71 lncRNA candidates that we analyzed. For each lncRNA candidate, we labelled patients as high or low expression based on dichotomization as described before. We established a cutoff for segment mean where patients with segment means greater than 0.3 were labelled as amplification and those with segment means less than -0.3 as deletions. This allowed us to compare counts of patients with high expression and amplifications versus those with low expression and deletions using a chi-square test. Specifically, we expected that groups with high lncRNA expression that were at higher risk would also have statistically more amplifications of the genes while groups with low lncRNA expression that were at higher risk would have more deletions. We also evaluated the correlation between lncRNA segment mean (SCNA) and expression (FPKM-UQ) in each risk group using Spearman correlation. After correcting for multiple testing, we kept only lncRNA candidates that obtained significance in both the U-test and correlation test, (adjusted *p* < 0.05, *p* < 0.05, respectively).

Utilizing methylation data from Infinium Human Methylation 450K arrays, we evaluated differential methylation between lncRNA risk groups for each candidate and cancer type for which this data was available. We mapped CpG probe coordinates to lncRNA candidates and found 88 lncRNAs with nearby or overlapping CpGs and evaluated these further. We removed probes that had beta values of 0 or were “NA” for all patients in a given cohort. Patients with available data were labelled as methylated (beta value > =0.5) or unmethylated (beta value < 0.5). lncRNA abundance (FPKM-UQ) was compared between these two groups using the two-sided U test. Spearman correlation was also evaluated between beta and FPKM-UQ values. Multiple testing correction was applied using the Benjamini-Hochberg Procedure.

Pathway and network analysis of lncRNA-associated PCGs

For each lncRNA candidate, patients were first labelled as high or low expressing patients as before. We conducted differential expression analysis to identify protein-coding genes that may also be associated with lncRNA abundance. For this, we utilized raw counts from RNA-Seq and utilized LIMMA’s [(Ritchie et al. 2015)](https://paperpile.com/c/Q7v0TQ/UPjD1) pipeline for RNA-Seq data differential expression analysis. We considered all PCGs with absolute fold change > 2 and adjusted *P* < 0.05. We ranked differentially expressed PCGs for each lncRNA by their adjusted p-values and ran gProfiler [(Reimand et al. 2007)](https://paperpile.com/c/Q7v0TQ/e5gj9) to identify which pathways were enriched by each lncRNA PCG signature. We considered all gene sets greater than 10 and less than 250 with a minimum required overlap of 5 genes. Further, to evaluate PCGs across lncRNA candidates within a cancer type context, we implemented an integrative pathways enrichment analysis tool Active Pathways [(Ritchie et al. 2015; Paczkowska et al. 2018)](https://paperpile.com/c/Q7v0TQ/UPjD1+gniO). For a given cancer type, we integrated each lncRNA’s list of differentially expressed PCGS using LIMMA unadjusted p-values to identify which pathways were associated to all lncRNAs or a subset of them and which pathways became significantly enriched only when all lncRNAs were considered using Brown’s merged p-values. Pathway enrichment maps were built in Cytoscape[(Shannon et al. 2003)](https://paperpile.com/c/Q7v0TQ/KTZmd). For LGG, we extracted the genes within a set of developmental pathways to evaluate their correlation using the “corrplot” package in R[(Wei and Simko 2017)](https://paperpile.com/c/Q7v0TQ/dLZnU) . We generated heatmaps to summarize the expression of these genes for LGG and GBM using the “ComplexHeatmap” package [(Gu, Eils, and Schlesner 2016)](https://paperpile.com/c/Q7v0TQ/tE3X4).

**CITATIONS**

[Acampora, D., M. P. Postiglione, V. Avantaggiato, M. Di Bonito, F. M. Vaccarino, J. Michaud, and A. Simeone. 1999. “Progressive Impairment of Developing Neuroendocrine Cell Lineages in the Hypothalamus of Mice Lacking the Orthopedia Gene.” *Genes & Development* 13 (21): 2787–2800.](http://paperpile.com/b/Q7v0TQ/PSpn)

[Ali, Mohamad Moustafa, Vijay Suresh Akhade, Subazini Thankaswamy Kosalai, Santhilal Subhash, Luisa Statello, Matthieu Meryet-Figuiere, Jonas Abrahamsson, Tanmoy Mondal, and Chandrasekhar Kanduri. 2018. “PAN-Cancer Analysis of S-Phase Enriched lncRNAs Identifies Oncogenic Drivers and Biomarkers.” *Nature Communications* 9 (1): 883.](http://paperpile.com/b/Q7v0TQ/75M5)

[Brunner, Alayne L., Andrew H. Beck, Badreddin Edris, Robert T. Sweeney, Shirley X. Zhu, Rui Li, Kelli Montgomery, et al. 2012. “Transcriptional Profiling of Long Non-Coding RNAs and Novel Transcribed Regions across a Diverse Panel of Archived Human Cancers.” *Genome Biology* 13 (8): R75.](http://paperpile.com/b/Q7v0TQ/PgCt)

[Campbell, Peter J., Gad Getz, Joshua M. Stuart, Jan O. Korbel, Lincoln D. Stein, and - ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Net. 2017. “Pan-Cancer Analysis of Whole Genomes.” https://doi.org/](http://paperpile.com/b/Q7v0TQ/wzALR)[10.1101/162784](http://dx.doi.org/10.1101/162784)[.](http://paperpile.com/b/Q7v0TQ/wzALR)

[Carithers, Latarsha J., Kristin Ardlie, Mary Barcus, Philip A. Branton, Angela Britton, Stephen A. Buia, Carolyn C. Compton, et al. 2015. “A Novel Approach to High-Quality Postmortem Tissue Procurement: The GTEx Project.” *Biopreservation and Biobanking* 13 (5): 311–19.](http://paperpile.com/b/Q7v0TQ/eQl4)

[Chiu, Hua-Sheng, Sonal Somvanshi, Ektaben Patel, Ting-Wen Chen, Vivek P. Singh, Barry Zorman, Sagar L. Patil, et al. 2018. “Pan-Cancer Analysis of lncRNA Regulation Supports Their Targeting of Cancer Genes in Each Tumor Context.” *Cell Reports* 23 (1): 297–312.e12.](http://paperpile.com/b/Q7v0TQ/qnLL)

[Colaprico, Antonio, Tiago C. Silva, Catharina Olsen, Luciano Garofano, Claudia Cava, Davide Garolini, Thais S. Sabedot, et al. 2015a. “TCGAbiolinks: An R/Bioconductor Package for Integrative Analysis of TCGA Data.” *Nucleic Acids Research* 44 (8): e71–e71.](http://paperpile.com/b/Q7v0TQ/TNhd)

[———. 2015b. “TCGAbiolinks: An R/Bioconductor Package for Integrative Analysis of TCGA Data.” *Nucleic Acids Research* 44 (8): e71–e71.](http://paperpile.com/b/Q7v0TQ/PceD4)

[Derrien, Thomas, Rory Johnson, Giovanni Bussotti, Andrea Tanzer, Sarah Djebali, Hagen Tilgner, Gregory Guernec, et al. 2012. “The GENCODE v7 Catalog of Human Long Noncoding RNAs: Analysis of Their Gene Structure, Evolution, and Expression.” *Genome Research* 22 (9): 1775–89.](http://paperpile.com/b/Q7v0TQ/cinx)

[Dong, Cheng‑ya, Jiayue Cui, Dian‑he Li, Qi Li, and Xin‑yu Hong. 2018. “HOXA10‑AS: A Novel Oncogenic Long Non‑coding RNA in Glioma.” *Oncology Reports*. https://doi.org/](http://paperpile.com/b/Q7v0TQ/Btmv)[10.3892/or.2018.6662](http://dx.doi.org/10.3892/or.2018.6662)[.](http://paperpile.com/b/Q7v0TQ/Btmv)

[Engreitz, Jesse M., Amy Pandya-Jones, Patrick McDonel, Alexander Shishkin, Klara Sirokman, Christine Surka, Sabah Kadri, et al. 2013. “The Xist lncRNA Exploits Three-Dimensional Genome Architecture to Spread across the X Chromosome.” *Science* 341 (6147): 1237973.](http://paperpile.com/b/Q7v0TQ/Kce7)

[Gibb, Ewan A., Emily A. Vucic, Katey S. S. Enfield, Greg L. Stewart, Kim M. Lonergan, Jennifer Y. Kennett, Daiana D. Becker-Santos, et al. 2011. “Human Cancer Long Non-Coding RNA Transcriptomes.” *PloS One* 6 (10): e25915.](http://paperpile.com/b/Q7v0TQ/TEPk)

[Gonzalez, Inma, Roberto Munita, Eneritz Agirre, Travis A. Dittmer, Katia Gysling, Tom Misteli, and Reini F. Luco. 2015. “A lncRNA Regulates Alternative Splicing via Establishment of a Splicing-Specific Chromatin Signature.” *Nature Structural & Molecular Biology* 22 (5): 370–76.](http://paperpile.com/b/Q7v0TQ/aAnk)

[GTEx Consortium. 2013. “The Genotype-Tissue Expression (GTEx) Project.” *Nature Genetics* 45 (6): 580–85.](http://paperpile.com/b/Q7v0TQ/bTlr)

[Guo, Haiyang, Musaddeque Ahmed, Fan Zhang, Cindy Q. Yao, Side Li, Yi Liang, Junjie Hua, et al. 2016. “Modulation of Long Noncoding RNAs by Risk SNPs Underlying Genetic Predispositions to Prostate Cancer.” *Nature Genetics* 48 (10): 1142–50.](http://paperpile.com/b/Q7v0TQ/nYz7)

[Gupta, Rajnish A., Nilay Shah, Kevin C. Wang, Jeewon Kim, Hugo M. Horlings, David J. Wong, Miao-Chih Tsai, et al. 2010. “Long Non-Coding RNA HOTAIR Reprograms Chromatin State to Promote Cancer Metastasis.” *Nature* 464 (7291): 1071–76.](http://paperpile.com/b/Q7v0TQ/Cd8r)

[Gu, Zuguang, Roland Eils, and Matthias Schlesner. 2016. “Complex Heatmaps Reveal Patterns and Correlations in Multidimensional Genomic Data.” *Bioinformatics*  32 (18): 2847–49.](http://paperpile.com/b/Q7v0TQ/tE3X4)

[Hajjari, Mohammadreza, and Abbas Salavaty. 2015. “HOTAIR: An Oncogenic Long Non-Coding RNA in Different Cancers.” *Cancer Biology & Medicine* 12 (1): 1–9.](http://paperpile.com/b/Q7v0TQ/xA2C)

[Hon, Chung-Chau, Jordan A. Ramilowski, Jayson Harshbarger, Nicolas Bertin, Owen J. L. Rackham, Julian Gough, Elena Denisenko, et al. 2017. “An Atlas of Human Long Non-Coding RNAs with Accurate 5’ Ends.” *Nature* 543 (7644): 199–204.](http://paperpile.com/b/Q7v0TQ/d2jV)

[Huang, Hui-Ling, Yu-Chung Wu, Li-Jen Su, Yun-Ju Huang, Phasit Charoenkwan, Wen-Liang Chen, Hua-Chin Lee, William Cheng-Chung Chu, and Shinn-Ying Ho. 2015. “Discovery of Prognostic Biomarkers for Predicting Lung Cancer Metastasis Using Microarray and Survival Data.” *BMC Bioinformatics* 16 (February): 54.](http://paperpile.com/b/Q7v0TQ/nyv8)

[Iyer, Matthew K., Yashar S. Niknafs, Rohit Malik, Udit Singhal, Anirban Sahu, Yasuyuki Hosono, Terrence R. Barrette, et al. 2015. “The Landscape of Long Noncoding RNAs in the Human Transcriptome.” *Nature Genetics* 47 (3): 199–208.](http://paperpile.com/b/Q7v0TQ/68t1)

[Jain, Rajan, Laila M. Poisson, David Gutman, Lisa Scarpace, Scott N. Hwang, Chad A. Holder, Max Wintermark, et al. 2014. “Outcome Prediction in Patients with Glioblastoma by Using Imaging, Clinical, and Genomic Biomarkers: Focus on the Nonenhancing Component of the Tumor.” *Radiology* 272 (2): 484–93.](http://paperpile.com/b/Q7v0TQ/NIYT)

[Kirk, Jessime M., Susan O. Kim, Kaoru Inoue, Matthew J. Smola, David M. Lee, Megan D. Schertzer, Joshua S. Wooten, et al. 2018. “Functional Classification of Long Non-Coding RNAs by K-Mer Content.” *Nature Genetics* 50 (10): 1474–82.](http://paperpile.com/b/Q7v0TQ/U3Rd)

[Krijthe, J. H. 2015. “Rtsne: T-Distributed Stochastic Neighbor Embedding Using Barnes-Hut Implementation.” *R Package Version 0. 13, URL Https://github. com/jkrijthe/Rtsne*.](http://paperpile.com/b/Q7v0TQ/SajQL)

[Kung, Johnny T. Y., David Colognori, and Jeannie T. Lee. 2013. “Long Noncoding RNAs: Past, Present, and Future.” *Genetics* 193 (3): 651–69.](http://paperpile.com/b/Q7v0TQ/YFLP)

[Leyten, Gisele H. J. M., Daphne Hessels, Sander A. Jannink, Frank P. Smit, Hans de Jong, Erik B. Cornel, Theo M. de Reijke, et al. 2014. “Prospective Multicentre Evaluation of PCA3 and TMPRSS2-ERG Gene Fusions as Diagnostic and Prognostic Urinary Biomarkers for Prostate Cancer.” *European Urology* 65 (3): 534–42.](http://paperpile.com/b/Q7v0TQ/DboC)

[Liu, Jianfang, Tara Lichtenberg, Katherine A. Hoadley, Laila M. Poisson, Alexander J. Lazar, Andrew D. Cherniack, Albert J. Kovatich, et al. 2018. “An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics.” *Cell* 173 (2): 400–416.e11.](http://paperpile.com/b/Q7v0TQ/1c0E)

[Lu, Chao, Patrick S. Ward, Gurpreet S. Kapoor, Dan Rohle, Sevin Turcan, Omar Abdel-Wahab, Christopher R. Edwards, et al. 2012. “IDH Mutation Impairs Histone Demethylation and Results in a Block to Cell Differentiation.” *Nature* 483 (7390): 474–78.](http://paperpile.com/b/Q7v0TQ/7QGT)

[Mondal, Tanmoy, Prasanna Kumar Juvvuna, Agnete Kirkeby, Sanhita Mitra, Subazini Thankaswamy Kosalai, Larissa Traxler, Falk Hertwig, et al. 2018. “Sense-Antisense lncRNA Pair Encoded by Locus 6p22.3 Determines Neuroblastoma Susceptibility via the USP36-CHD7-SOX9 Regulatory Axis.” *Cancer Cell* 33 (3): 417–34.e7.](http://paperpile.com/b/Q7v0TQ/RhXS)

[Paczkowska, Marta, Jonathan Barenboim, Nardnisa Sintupisut, Natalie C. Fox, Helen Zhu, Diala Abd-Rabbo, Paul C. Boutros, Juri Reimand, and PCAWG Network and Pathway Analysis Group. 2018. “Integrative Pathway Enrichment Analysis of Multivariate Omics Data.” https://doi.org/](http://paperpile.com/b/Q7v0TQ/gniO)[10.1101/399113](http://dx.doi.org/10.1101/399113)[.](http://paperpile.com/b/Q7v0TQ/gniO)

[PCAWG Transcriptome Core Group, Claudia Calabrese, Natalie R. Davidson, Nuno A. Fonseca, Yao He, André Kahles, Kjong-Van Lehmann, et al. 2017. “Genomic Basis for RNA Alterations Revealed by Whole-Genome Analyses of 27 Cancer Types.” https://doi.org/](http://paperpile.com/b/Q7v0TQ/YSvHM)[10.1101/183889](http://dx.doi.org/10.1101/183889)[.](http://paperpile.com/b/Q7v0TQ/YSvHM)

[Pedersen, Susanne K., Susan M. Mitchell, Lloyd D. Graham, Aidan McEvoy, Melissa L. Thomas, Rohan T. Baker, Jason P. Ross, et al. 2014. “CAHM, a Long Non-Coding RNA Gene Hypermethylated in Colorectal Neoplasia.” *Epigenetics: Official Journal of the DNA Methylation Society* 9 (8): 1071–82.](http://paperpile.com/b/Q7v0TQ/WOyb)

[Quinlan, Aaron R. 2014. “BEDTools: The Swiss-Army Tool for Genome Feature Analysis.” *Current Protocols in Bioinformatics / Editoral Board, Andreas D. Baxevanis ... [et Al.]* 47 (September): 11.12.1–34.](http://paperpile.com/b/Q7v0TQ/ziNl2)

[Reimand, Jüri, Ruth Isserlin, Veronique Voisin, Mike Kucera, Christian Tannus-Lopes, Asha Rostamianfar, Lina Wadi, et al. 2019. “Pathway Enrichment Analysis and Visualization of Omics Data Using g:Profiler, GSEA, Cytoscape and EnrichmentMap.” *Nature Protocols* 14 (2): 482–517.](http://paperpile.com/b/Q7v0TQ/fsK2)

[Reimand, Jüri, Meelis Kull, Hedi Peterson, Jaanus Hansen, and Jaak Vilo. 2007. “g:Profiler--a Web-Based Toolset for Functional Profiling of Gene Lists from Large-Scale Experiments.” *Nucleic Acids Research* 35 (Web Server issue): W193–200.](http://paperpile.com/b/Q7v0TQ/e5gj9)

[Ritchie, Matthew E., Belinda Phipson, Di Wu, Yifang Hu, Charity W. Law, Wei Shi, and Gordon K. Smyth. 2015. “Limma Powers Differential Expression Analyses for RNA-Sequencing and Microarray Studies.” *Nucleic Acids Research* 43 (7): e47.](http://paperpile.com/b/Q7v0TQ/UPjD1)

[Shannon, Paul, Andrew Markiel, Owen Ozier, Nitin S. Baliga, Jonathan T. Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. 2003. “Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks.” *Genome Research* 13 (11): 2498–2504.](http://paperpile.com/b/Q7v0TQ/KTZmd)

[SongTao, Qi, Yu Lei, Gui Si, Ding YanQing, Han HuiXia, Zhang XueLin, Wu LanXiao, and Yao Fei. 2012. “IDH Mutations Predict Longer Survival and Response to Temozolomide in Secondary Glioblastoma.” *Cancer Science* 103 (2): 269–73.](http://paperpile.com/b/Q7v0TQ/lc8x)

[Swarts, Dorian R. A., Mieke E. R. Henfling, Leander Van Neste, Robert-Jan van Suylen, Anne-Marie C. Dingemans, Winand N. M. Dinjens, Annick Haesevoets, et al. 2013. “CD44 and OTP Are Strong Prognostic Markers for Pulmonary Carcinoids.” *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research* 19 (8): 2197–2207.](http://paperpile.com/b/Q7v0TQ/wLXi)

[Tian, Feng, Jian Xu, Fangxi Xue, Encui Guan, and Xiaoguang Xu. 2017. “TINCR Expression Is Associated with Unfavorable Prognosis in Patients with Hepatocellular Carcinoma.” *Bioscience Reports* 37 (4).](http://paperpile.com/b/Q7v0TQ/vVm9)

[Tomlins, Scott A., Sheila M. J. Aubin, Javed Siddiqui, Robert J. Lonigro, Laurie Sefton-Miller, Siobhan Miick, Sarah Williamsen, et al. 2011. “Urine TMPRSS2:ERG Fusion Transcript Stratifies Prostate Cancer Risk in Men with Elevated Serum PSA.” *Science Translational Medicine* 3 (94): 94ra72.](http://paperpile.com/b/Q7v0TQ/CDOa)

[Tsai, Miao-Chih, Ohad Manor, Yue Wan, Nima Mosammaparast, Jordon K. Wang, Fei Lan, Yang Shi, Eran Segal, and Howard Y. Chang. 2010. “Long Noncoding RNA as Modular Scaffold of Histone Modification Complexes.” *Science* 329 (5992): 689–93.](http://paperpile.com/b/Q7v0TQ/WYIW)

[Uhlen, Mathias, Cheng Zhang, Sunjae Lee, Evelina Sjöstedt, Linn Fagerberg, Gholamreza Bidkhori, Rui Benfeitas, et al. 2017. “A Pathology Atlas of the Human Cancer Transcriptome.” *Science* 357 (6352).](http://paperpile.com/b/Q7v0TQ/byKz)

[Uno, Hajime, Tianxi Cai, Michael J. Pencina, Ralph B. D’Agostino, and L. J. Wei. 2011. “On the C-Statistics for Evaluating Overall Adequacy of Risk Prediction Procedures with Censored Survival Data.” *Statistics in Medicine* 30 (10): 1105–17.](http://paperpile.com/b/Q7v0TQ/5zMv)

[Wang, Xue Q. D., and Josée Dostie. 2017. “Reciprocal Regulation of Chromatin State and Architecture by HOTAIRM1 Contributes to Temporal Collinear HOXA Gene Activation.” *Nucleic Acids Research* 45 (3): 1091–1104.](http://paperpile.com/b/Q7v0TQ/hyme4)

[Wang, Yue, Xiao-Jun Liu, and Xu-Dong Yao. 2014. “Function of PCA3 in Prostate Tissue and Clinical Research Progress on Developing a PCA3 Score.” *Chinese Journal of Cancer Research = Chung-Kuo Yen Cheng Yen Chiu* 26 (4): 493–500.](http://paperpile.com/b/Q7v0TQ/xbRq)

[“Website.” n.d. Accessed March 27, 2019.](http://paperpile.com/b/Q7v0TQ/lRCB) [The Cancer Genome Atlas Home Page. The Cancer Genome Atlas - National Cancer Institute (2011). Available at: https://cancergenome.nih.gov/. (Accessed: 26th September 2018)](about:blank)[.](http://paperpile.com/b/Q7v0TQ/lRCB)

[Wei, T., and V. Simko. 2017. “R Package ‘corrplot’: Visualization of a Correlation Matrix (version 0.84).’.” *Retrived from Https://github. Com/taiyun/corrplot*.](http://paperpile.com/b/Q7v0TQ/dLZnU)

[Xiong, Jie, Zhitong Bing, Yanlin Su, Defeng Deng, and Xiaoning Peng. 2014. “An Integrated mRNA and microRNA Expression Signature for Glioblastoma Multiforme Prognosis.” *PloS One* 9 (5): e98419.](http://paperpile.com/b/Q7v0TQ/Nvsw)

[Yang, Liuqing, Chunru Lin, Chunyu Jin, Joy C. Yang, Bogdan Tanasa, Wenbo Li, Daria Merkurjev, et al. 2013. “lncRNA-Dependent Mechanisms of Androgen-Receptor-Regulated Gene Activation Programs.” *Nature* 500 (7464): 598–602.](http://paperpile.com/b/Q7v0TQ/NnIg)

[Yang, Yongzhi, Peng Junjie, Cai Sanjun, and Yanlei Ma. 2017. “Long Non-Coding RNAs in Colorectal Cancer: Progression and Future Directions.” *Journal of Cancer* 8 (16): 3212–25.](http://paperpile.com/b/Q7v0TQ/HC0v)

[Yuan, Yuan, Eliezer M. Van Allen, Larsson Omberg, Nikhil Wagle, Ali Amin-Mansour, Artem Sokolov, Lauren A. Byers, et al. 2014. “Assessing the Clinical Utility of Cancer Genomic and Proteomic Data across Tumor Types.” *Nature Biotechnology* 32 (7): 644–52.](http://paperpile.com/b/Q7v0TQ/WmRx)

[Yu, Kun-Hsing, Ce Zhang, Gerald J. Berry, Russ B. Altman, Christopher Ré, Daniel L. Rubin, and Michael Snyder. 2016. “Predicting Non-Small Cell Lung Cancer Prognosis by Fully Automated Microscopic Pathology Image Features.” *Nature Communications* 7 (August): 12474.](http://paperpile.com/b/Q7v0TQ/HRc8)

[Zerbino, Daniel R., Premanand Achuthan, Wasiu Akanni, M. Ridwan Amode, Daniel Barrell, Jyothish Bhai, Konstantinos Billis, et al. 2018. “Ensembl 2018.” *Nucleic Acids Research* 46 (D1): D754–61.](http://paperpile.com/b/Q7v0TQ/7s31)

[Zhang, Bin, Gayatri Arun, Yuntao S. Mao, Zsolt Lazar, Gene Hung, Gourab Bhattacharjee, Xiaokun Xiao, et al. 2012. “The lncRNA Malat1 Is Dispensable for Mouse Development but Its Transcription Plays a Cis-Regulatory Role in the Adult.” *Cell Reports* 2 (1): 111–23.](http://paperpile.com/b/Q7v0TQ/N8F8)

[Zhang, Hao, Yanhua Qi, Decheng Geng, Yi Shi, Xu Wang, Rutong Yu, and Xiuping Zhou. 2018. “Expression Profile and Clinical Significance of Wnt Signaling in Human Gliomas.” *Oncology Letters* 15 (1): 610–17.](http://paperpile.com/b/Q7v0TQ/E8Hr)