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

Keren Isaev1,2, Marta Paczkowska1, 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](mailto:Juri.Reimand@utoronto.ca)

**ABSTRACT**

Long non-coding RNAs (lncRNAs) are as numerous as protein-coding genes, however their biological functions and translational 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 5,691 lncRNAs across 8,594 patient tumors to characterize their prognostic potential, genomic alterations and cellular function. We found 166 lncRNAs whose transcript abundance profiles robustly modeled cancer patient risk and validated 5/41 lncRNAs using datasets available externally. 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, developmental pathway activation and a tumor malignancy gradient. Our analysis highlights cancer-associated lncRNAs as attractive candidates for biomarker development and functional experiments.

**INTRODUCTION**

The human genome encodes numerous long non-coding RNA (lncRNA) genes that lack protein-coding potential [1–4](https://paperpile.com/c/7JGcTo/9JulU+VV3ap+hEI4v+svL6g). lncRNAs are likely as frequent as protein-coding (CDS) genes as ~20,000 high-confidence transcripts were described by FANTOM CAT [3](https://paperpile.com/c/7JGcTo/hEI4v),[5](https://paperpile.com/c/7JGcTo/DyphN). lncRNAs have tissue-specific expression at lower levels than mRNAs, however recent studies have increasingly characterized their roles in diverse biological processes [2,18,19](https://paperpile.com/c/7JGcTo/zrKWI+Ebz4Q+VV3ap), including chromatin remodeling [6,7](https://paperpile.com/c/7JGcTo/ytkek+ZI7eI), post-transcriptional gene regulation [6,8](https://paperpile.com/c/7JGcTo/ytkek+vWbw9), epigenetic silencing through interactions with DNA and proteins [7](https://paperpile.com/c/7JGcTo/ZI7eI), and proximal and distal transcriptional regulation [9–11](https://paperpile.com/c/7JGcTo/uPMFu+vQd6X+b2SUo). lncRNAs are also involved in cancer-related processes of cellular differentiation and genome integrity [12–15](https://paperpile.com/c/7JGcTo/DtEj7+x4fQz+Ap1Ys+3gKgE). Recent developments have enabled computational annotation of lncRNAs using their protein-binding motifs and interactions with transcription factors, RNA-binding proteins and microRNAs [16,17](https://paperpile.com/c/7JGcTo/Am7J8+Yx4xq). However, lncRNAs remain understudied compared to CDS genes and most of our biological knowledge is based on a few well-studied examples. Similarly, translational and clinical utility of lncRNAs is limited to a handful of prominent examples. For example, *PCA3* in prostate cancer and *HOTAIR* in breast cancer are established diagnostic biomarkers with increased transcript abundance compared to healthy tissues [20,21](https://paperpile.com/c/7JGcTo/WcsR8+pPc4y). Human Pathology Atlas and others have assessed the impact of mRNA abundance on cancer outcomes [23](https://paperpile.com/c/7JGcTo/cmu9d). However, systematic evaluation of lncRNAs as molecular biomarkers has been quite limited to date.

To address these challenges, 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)(Cancer Genome Atlas Research Network et al. 2013). We evaluated the lncRNA profiles using cross-validation experiments and in an independent patient cohort from the International Cancer Genome Consortium (ICGC) (PCAWG Transcriptome Core Group et al. 2017). 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 and 28 cancer types with matched RNA-seq data and clinical annotations from TCGA PanCanAtlas [24](https://paperpile.com/c/7JGcTo/GlfIv) (**Supplementary Table 1**). We selected 5,785 high-confidence lncRNAs that were annotated by FANTOM CAT[3](https://paperpile.com/c/7JGcTo/hEI4v) as well as Ensembl and showed above-baseline transcript abundance pan-cancer (**Supplementary Table 2**). Unsupervised clustering of the lncRNA transcriptome revealed a robust clustering of tumor samples by organ systems that was comparable to protein-coding genes, confirming earlier observations in cell lines and smaller patient cohorts [2](https://paperpile.com/c/7JGcTo/VV3ap) **(Figure 1a, Supplementary Figure 1a)**. 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 16 pairs of tumor and healthy tissue types, using 6,927 transcriptomes of healthy tissues of the Genotype-Tissue Expression (GTEx) project with RNA-seq data for 5,691/5,785 lncRNAs[25](https://paperpile.com/c/7JGcTo/WEdOy). Many lncRNAs (1,651/5,691 or 29%) showed specific differences in lncRNA abundance between tumors and healthy tissues, defined as median normalized ranks differing by one quartile (Wilcoxon rank-sum test, adjusted *P < 0.05*) (**Figure 1b)**. Comparison of cervical tumours and normal tissues was the only one to not reveal any lncRNAs with this significance. Meanwhile, the remaining tumour types were characterized by hundreds of such lncRNAs (mean: 264, range: 156-509), yet the majority of these lncRNAs were highly tissue-specific (**Supplementary Figure 1b**). Half of these lncRNAs (832/1,651 or 50%) strongly differentiated in one type of tumor relative to the adjacent normal tissue and only showed minor differences in other tumor types. At the other extreme, 131 of these lncRNAs differed significantly in transcript abundance in 6-15 cancer types, including nine lncRNAs with strong changes in transcript abundance in all cancer types (seven increased and two reduced), indicating their potential for biomarker discovery and functional characterization across human cancers (**Supplementary Table 3**). As positive controls, we found the well-known diagnostic lncRNAs *PCA3*[*20*](https://paperpile.com/c/7JGcTo/WcsR8) and *PCAT1*[*20,26*](https://paperpile.com/c/7JGcTo/WcsR8+n6eGo)with strongly increased transcript abundance in prostate tumors relative to healthy prostate tissue samples (0.95 vs 0.3 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 Cox proportional-hazards (PH) models [23](https://paperpile.com/c/7JGcTo/cmu9d). The majority of lncRNAs (3,470/5,785 or 60%) were significantly associated with patient overall survival in at least one cancer type (Wald test, adjusted *P* <0.1) (**Figure 1d**). Patient risk groups were defined for each cancer type using two strategies: median dichotomization for the majority of cases (80%) with above-zero lncRNA transcript abundance and zero-based dichotomization for 20% of cases where the lncRNA was transcriptionally silent in most tumors 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 tumors.

In agreement with their tissue-specific transcription patterns, the majority of lncRNAs were associated with survival in only one cancer type (2,494/3,985 or 63%). The most survival-associated lncRNAs were found in clear cell renal carcinoma (2,489) and low-grade glioma (1,827) and 789 of those were found in both cancer types. To interpret the large number of survival-associated lncRNA transcripts, we performed a correlation analysis of XX pairs of survival-associated lncRNAs and found 32% with significantly positive and 25% with negative correlation (Spearman rho, adjusted *P* < 0.05) (**Supplementary Figure 1c**). 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.

**Machine learning identifies 166 lncRNAs with robust prognostic performance (FIGURE2)**

To identify multivariate prognostic signatures of lncRNAs, we adapted a machine-learning strategy comprising proportional-hazards regression models with elastic-net regularization [27](https://paperpile.com/c/7JGcTo/JOBi+zCfF)-[30](https://paperpile.com/c/7JGcTo/PpmqX+sIj7s)and 100-fold cross-validations 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 [31](https://paperpile.com/c/7JGcTo/9ZDrG). To prioritize high-confidence lncRNAs, selected the ones that appeared as predictors in at least half of cross-validation models. For benchmarking, we fitted alternative survival models with clinical predictors (e.g., tumour stage, grade, patient age) and with both lncRNA and clinical predictors. Multivariate lncRNA-based survival models significantly outperformed models with just clinical predictors (**Figure 2b**), supporting their value as candidate prognostic biomarkers.

Analysis of frequently selected features from cross-validation experiments revealed 166 lncRNAs, including 159 in a single cancer type and seven in two cancer types. The majority (114/166 or 69%) were unfavorable prognostic markers where poor outcome 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* [*32*](https://paperpile.com/c/7JGcTo/rAGMz) and *CAHM* [*33*](https://paperpile.com/c/7JGcTo/zsl4i) as prognostic markers in clear cell renal carcinoma. Comparison of different iterations of cross-validation indicated the high prognostic value of each lncRNA. The majority of lncRNA-based univariate models (156/166 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 improvement in model performance: combined models outperformed clinical-only prognostic models for 148/166 of lncRNAs (89%), leading to median c-index increase of 0.03 (range c-index improvement: 0.005-0.25) (**Figure 2c**). All 166 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=0.85) compared to liver cancer models with clinical variables only (c=0.5). A minority of lncRNAs (10/166 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

We hypothesized that differences of lncRNAs transcript abundance between high-risk and low-risk patients would match corresponding differences between tumors and healthy tissues. We thus asked whether increased lncRNA transcript abundance in high-risk patient tumors was coupled with reduced lncRNA abundance in healthy tissues, and similarly, whether reduced lncRNA abundance in high-risk patient tumors was coupled with increased lncRNA abundance in healthy tissues. We found that nearly all lncRNA candidates (103/110 or 94%) showed differential ranking of transcript abundance as expected (Wilcoxon rank-sum test, adjusted *P* < 0.05) while of these, 23/110 (20%) lncRNAs also had an absolute median rank change of 0.2. **(Figure 2f)**, highlighting opportunities for developing prognostic and diagnostic biomarkers.

To exclude confounding effects, we compared 84/166 (50%) lncRNA profiles with PCGs that were located in adjacent DNA sequence within +/-10 kbp of gene bodies, including 41 divergent lncRNAs that shared promoters with CDS genes and 15 antisense lncRNAs that were co-transcribed. Only few CDS genes were also significantly prognostic in the same cancer type (14/126 or 11% at *P* <0.05) and in 11 of those cases the lncRNA-based survival model outperformed the equivalent CDS-based model (median lncRNA c-index:0.58 versus median PCG c-index: 0.53) (**Figure 2d**). For example, the lncRNA *CAHM* provided a superior prognostic model compared to *QKI*, an adjacent cancer gene recorded as an unfavorable prognostic marker in Human Pathology Atlas (**Supplementary Figure 2a**). Thus, the lncRNAs represent unique prognostic signals and potential functionality that are not confounded by the protein-coding genome.

**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/TCGA PCAWG project. We tested 41/166 (25%) lncRNAs for which samples of relevant cancer types were available and where transcript abundance exceeded baseline detection levels (**Supplementary Table 3**). The validation dataset was powered at 6.6-74.5% to detect significant associations (*P*=0.05) with a hazard ratio of 2.0 (or 0.5), based on risk/non-risk patient ratios and total events occurred, indicating limitations of current validation data (**Supplementary Figure 3a)**.

We compared the discovery cohort (TCGA) and the validation cohort (PCAWG) and found that hazard ratios of lncRNA-based models were positively correlated (Pearson rho=0.37, *P* = 0.014), suggesting that lncRNAs showed consistent prognostic signals in the two datasets **(Figure 3a)**. lncRNA-based survival models in the validation cohort showed improved c-index values compared to models with clinical variables of patient age and sex (median lncRNA concordance: 0.6 vs median clinical variables concordance = 0.55**, Supplementary Figure 3b),** howeverlimited clinical annotations of the validation cohort warrant caution in interpreting these results. Five lncRNAs (12%) showed significant prognostic capacity in the validation dataset (Wald *P*<0.05): *AC114803.3* and *CTC-297N7.5* in liver carcinoma, *LINC00524* in kidney renal carcinoma, *AC073130.1* in breast adenocarcinoma and *CTD-2316B1.2* in lung adenocarcinoma(c-index range of 0.58-0.7). Five additional lncRNAs showed sub-significant trends (0.05<*P*<0.15) (**Figure 3b-f**). Three lncRNAs showed zero-median transcript abundance in both discovery and validation cohorts, suggesting aberrant lncRNA activation in fewer high-risk patients and complete inhibition in low-risk patients. The lncRNAs were adjacent to CDS genes, however all lncRNA-based models exceeded the performance of related CDS-based models **(Figure 2e)**. External validation of these novel lncRNAs prioritizes 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 genetic alterations driving lncRNA transcript abundance in patient risk groups, we studied genomic copy number alterations and DNA methylation patterns of lncRNAs. We associated focal somatic copy number alterations (CNAs) with transcript abundance of each lncRNA using two complementary statistical tests (*FDR*<0.05) (**Figure 3b-f**).

SCNAs were associated with transcript abundance of 15/65 (25%) our prognostic lncRNAs for which CNA data were available. High transcript abundance of eight unfavorable lncRNAs was associated with copy number gains, and low transcript abundance of seven favorable lncRNAs was associated with losses (U-test, *FDR*<0.05 and Spearman, *p*<0.05; **Figure 4a**). For patients with available copy number data, SCNAs explained a fraction of patients dichotomized by lncRNA transcript abundance (median 30%, range 10-56 %). For example, copy number and transcript abundance of lncRNA *RP13-1032I1.7* was associated with favorable prognosis in cervical cancer. Out of the 95 patients with SCNA information, 18 risk patients were impacted by a deletion while 11 non-risk patients were impacted by an amplification suggesting about a third of these patients’ lncRNA expression was influenced by copy number changes (**Figure 4b)**. As another example, the antisense lncRNA *RP13-1032I1.7* was affected by the same SCNAs as CDS genes *HGS* and *SLC25A10.* However, transcript abundance of the lncRNA was a better prognostic predictor compared to that of the two PCGs (lncRNA c = 0.63, *HGS* c= 0.55, *SLC25A10 c=*0.6). Although the SCNA affected all three genes, lncRNA transcript abundance appeared the best prognostic feature. Overall, these results suggest that transcript abundance of prognostic lncRNAs is often influenced by SCNAs in tumours, suggesting potential genetic driver mechanisms affecting the lncRNAs or other adjacent genes.

We also evaluated the methylation of probes overlapping lncRNA transcripts. Overall, 73 lncRNAs (73/166, 44%) overlapped at least one CpG probe. We found 29/73 (40%) 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 (Kruskal-Wallis, adjusted *P* < 0.05)(**Figure 4c**). Further, we evaluated the correlation between each significant probe and lncRNA abundance (FPKM-UQ). In cases of multiple probes overlapping a lncRNA, we selected the one with the highest absolute correlation. We found all 29 lncRNA candidates had a significant correlation between methylation (beta values) and lncRNA transcript abundance (FPKM-UQ) (Spearman, *P* < 0.05). Specifically, we found 14/29 (48%) of CpG probes were negatively correlated with lncRNA abundance. For example, in liver cancer, the favorable lncRNA *CTC-297N7.5* was associated with a methylated CpG site in the risk group (defined as beta values >= 0.75), located within 1,332 bases from the transcription start site of the gene (**Figure 13D**). 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 transcription suppression. The remainder of significantly differentially methylated probes (n=15) were positively correlated suggesting potential methylation of lncRNA gene bodies. The median distance from probes to the transcription start site was 2,091 bases for negatively correlated probes and 8,984 bases for positively correlated probes. Thus, in addition to the well-established negative association between promoter methylation and lower gene expression, lncRNAs may also be impacted by gene body methylation leading to higher gene expression as previously described in cancer [34](https://paperpile.com/c/7JGcTo/rdZQi). Overall, we found a sizeable fraction of potentially prognostic lncRNAs that are regulated transcriptionally by either copy number aberrations or DNA methylation (42/166, 25%). Methylation differences were more widespread than SCNAs potentially because some cancer types are more susceptible to recurrent SCNAs than others [35,36](https://paperpile.com/c/7JGcTo/u0fJQ+6h9fW).

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

Next, we characterized the lncRNAs in the context of defined clinical and molecular tumor subtypes. Systematic analysis comprised 38 lncRNAs (23% of 166 lncRNAs) across 11/22 cancer types for which data were available, using data curated by the R package TCGABiolinks 49. We found 267/2,735 instances where lncRNA transcript was significantly associated with a clinical or molecular tumor subtype (Chisq-test, FDR < 0.05) **(Supplementary Table 5)**. For all remaining significant associations (n=113), we also compared univariate survival models (lncRNAs *vs.* tumor features) (**Figure 5a)**. As expected, the subtypes that associated with our prognostic lncRNAs were also prognostic features of respective tumor types. For example, transcript abundance of prognostic lncRNAs in low grade glioma was often associated with IDH mutation status, prominent SCNAs such as gains of chromosome 19 and 20, and methylation statusof*MGMT* promoter **(Supplementary Figure 5a)**.

We asked whether combining individual prognostic lncRNAs and tumor subtypes would lead to improved accuracy. Indeed, adding dichotomized lncRNA transcript abundance status to subtype-based survival models improved accuracy (c-index) and model fit 94% of the time (106/113 associations, **Figure 5b**). Median improvement of accuracy was 0.05 (range 0.01 – 0.25). For example, survival models combining the lncRNA *AC022311.1* transcript abundance with tumour stage lead to a significantly improved survival model of renal papillary cell carcinoma (Anova, *P* = 3.6e-04) and an increase in concordance of 0.25. Low grade glioma (LGG) candidate *RP5-1086K13.1’*s expression when combined with the co-gain of chromosome 19 and 20 (a prognostic feature on its own) also improved its concordance significantly by 0.16 (Anova, adjusted *P* = 1.18e-07) **(Figure 5c)**.

We focused on nine prognostic lncRNAs in low-grade glioma that were all associated with *IDH1/2* gene mutation status, a well-established molecular prognostic feature in LGG (Lu et al. 2012; SongTao et al. 2012). (**Figure 5a**). Interestingly, several of these lncRNAs including *HOXB-AS2* and *HOXA-AS4* improved the fit of survival models when added to *IDH1/2* mutation status (Anova, adjusted *P* < 0.05). Patients with no *IDH1/2* 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 *HOXA-AS4* 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 and all tumors were wildtype of *IDH1*/*2*. 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 [37](https://paperpile.com/c/7JGcTo/kvpQ9). Large transcriptional differences were apparent in lncRNA-associated high-risk patient tumours compared those with better prognosis. The majority of lncRNA-associated risk groups (146/166 or 88%) were characterized by differential expression of at least 10 PCGs (median group by 555 PCGs) (fold change > 2, FDR <0.05). Further, we identified 269 (269/8,243 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 3,792 unique pathways identified in total (*FDR* <0.01 from g:Profiler) (**Figure 6a**). The majority of pathways were enriched in few lncRNA-associated risk groups (2,208/3,792 or 58% associated with five lncRNAs at most) while a minority of processes (22/3,792 or 0.006%) including ion channel activity and extracellular matrix were associated with 40/166 (24%) lncRNAs at most. 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 lncRNAs *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 [38](https://paperpile.com/c/7JGcTo/QeMRE). Fusion of mRNA signatures across the nine lncRNAs revealed 2,163 PCGs with significant differences in mRNA abundance in one or several lncRNA-associated risk groups. These genes were enriched in 483 GO biological processes and Reactome pathways (**Figure 6b**). The majority of pathways (397/483 or 82%) 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 synaptic signaling, mitotic checkpoint, angiogenesis, brain development and T cell proliferation, indicating that prognostic lncRNAs are associated with global transcriptional de-regulation of cancer hallmarks and brain developmental pathways in LGG risk groups (**Figure 6b)**. A similar analysis for liver cancer highlighted additional PCGs and pathways (Supplementary note 1, **Supplementary Figure 6a**).

We focused on a group of four brain developmental processes that were enriched in eight of nine lncRNA-associated risk groups. The enriched pathways included 185 genes and mostly showed differential mRNA abundance in patient risk groups. These included eight HOX transcription factors involved in organism development (*HOXA, HOXB* and *HOXC* clusters) and 18 Cancer Gene Census genes (including *MYC*, OLIG*2* and *GLI1*, Fisher’s exact test P = 0.0001, 7 genes expected). Transcriptional correlation analysis of prognostic lncRNAs and known cancer genes revealed two clusters. In the first cluster, transcript abundance of the prognostic lncRNA *WAC-AS1* was positively correlated with cancer genes *MYC*, *OLIG2* and others and lower transcript abundance of the lncRNA associated with improved prognosis. In the second cluster, 12 known cancer genes and eight HOX family TFs were positively correlated with unfavorable lncRNAs *HOXA10-AS* and *HOXB-AS2*, and genes of the second cluster were negatively correlated with genes of the first cluster (**Figure 6c**). To exclude potential confounding effects, we confirmed that the prognostic performance of lncRNAs exceeded those of *HOX* genes and known cancer genes in these clusters **(Supplementary Figure 6b**).

To characterize the lncRNA candidates in the context of glioblastoma, 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 6d)**. 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 6c**).

Two prognostic HOX-associated lncRNAs *HOXB-AS2* and *HOXA10-AS* were transcriptionally co-activated with HOX TF genes and known cancer genes *EZH2*, *EGFR* and *GLI1* high-risk LGGs and GBMs. *EZH2* has been previously associated with brain cancer and was shown to contribute to angiogenesis [43](https://paperpile.com/c/7JGcTo/IDYo). The two lncRNAs remain to be characterized functionally. However, 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 [41](https://paperpile.com/c/7JGcTo/DdVaf), 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 [15,18,19,44](https://paperpile.com/c/7JGcTo/9JulU+3gKgE+zrKWI+Ebz4Q+P9dxP). 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 32,34,41, .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 [48](https://paperpile.com/c/7JGcTo/owMAp). 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 [49](https://paperpile.com/c/7JGcTo/aksSF). 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. 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. 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[50](https://paperpile.com/c/7JGcTo/IAx5) was uniformly processed as previously described [51](https://paperpile.com/c/7JGcTo/TXx3). 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 0.2.

t-SNE analysis:

We produced the t-SNE plot in FIgure 1C using the RTsne [52](https://paperpile.com/c/7JGcTo/Kapn) 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.

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.

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 41/166 (25%) of our lncRNAs candidate. As before, we fit univariate Cox-PH models with binary lncRNA predictors and plotted their Kaplan-Meier survival curves.

Comparing lncRNA and nearby protein-coding genes:

Using bedtools [53](https://paperpile.com/c/7JGcTo/FOGm) , 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). 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 153/166 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 65 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 chi-square test and correlation test, (adjusted *p* < 0.05).

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 [54](https://paperpile.com/c/7JGcTo/0CRW6) 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[37](https://paperpile.com/c/7JGcTo/kvpQ9) 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 [38,54](https://paperpile.com/c/7JGcTo/0CRW6+QeMRE). 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 [55](https://paperpile.com/c/7JGcTo/IVBf). For LGG, we extracted the genes within a set of developmental pathways to evaluate their correlation using the “corrplot” package in R [56](https://paperpile.com/c/7JGcTo/56Yb) . We generated heatmaps to summarize the expression of these genes for LGG and GBM using the “ComplexHeatmap” package [57](https://paperpile.com/c/7JGcTo/v0tg).

**CITATIONS**

1. [Iyer, M. K. *et al.* The landscape of long noncoding RNAs in the human transcriptome. *Nat. Genet.* **47,** 199–208 (2015).](http://paperpile.com/b/7JGcTo/9JulU)

2. [Brunner, A. L. *et al.* Transcriptional profiling of long non-coding RNAs and novel transcribed regions across a diverse panel of archived human cancers. *Genome Biol.* **13,** R75 (2012).](http://paperpile.com/b/7JGcTo/VV3ap)

3. [Hon, C.-C. *et al.* An atlas of human long non-coding RNAs with accurate 5’ ends. *Nature* **543,** 199–204 (2017).](http://paperpile.com/b/7JGcTo/hEI4v)

4. [Gibb, E. A. *et al.* Human Cancer Long Non-Coding RNA Transcriptomes. *PLoS One* **6,** e25915 (2011).](http://paperpile.com/b/7JGcTo/svL6g)

5. [Kung, J. T. Y., Colognori, D. & Lee, J. T. Long noncoding RNAs: past, present, and future. *Genetics* **193,** 651–669 (2013).](http://paperpile.com/b/7JGcTo/DyphN)

6. [Engreitz, J. M. *et al.* The Xist lncRNA exploits three-dimensional genome architecture to spread across the X chromosome. *Science* **341,** 1237973 (2013).](http://paperpile.com/b/7JGcTo/ytkek)

7. [Tsai, M.-C. *et al.* Long noncoding RNA as modular scaffold of histone modification complexes. *Science* **329,** 689–693 (2010).](http://paperpile.com/b/7JGcTo/ZI7eI)

8. [Gonzalez, I. *et al.* A lncRNA regulates alternative splicing via establishment of a splicing-specific chromatin signature. *Nat. Struct. Mol. Biol.* **22,** 370–376 (2015).](http://paperpile.com/b/7JGcTo/vWbw9)

9. [Yang, L. *et al.* lncRNA-dependent mechanisms of androgen-receptor-regulated gene activation programs. *Nature* **500,** 598–602 (2013).](http://paperpile.com/b/7JGcTo/uPMFu)

10. [Zhang, B. *et al.* The lncRNA Malat1 Is Dispensable for Mouse Development but Its Transcription Plays a cis-Regulatory Role in the Adult. *Cell Rep.* **2,** 111–123 (2012).](http://paperpile.com/b/7JGcTo/vQd6X)

11. [Wang, Z. *et al.* lncRNA Epigenetic Landscape Analysis Identifies EPIC1 as an Oncogenic lncRNA that Interacts with MYC and Promotes Cell-Cycle Progression in Cancer. *Cancer Cell* **33,** 706–720.e9 (2018).](http://paperpile.com/b/7JGcTo/b2SUo)

12. [Gong, X. *et al.* Novel lincRNA SLINKY is a prognostic biomarker in kidney cancer. *Oncotarget* **8,** 18657–18669 (2017).](http://paperpile.com/b/7JGcTo/DtEj7)

13. [Ali, M. M. *et al.* PAN-cancer analysis of S-phase enriched lncRNAs identifies oncogenic drivers and biomarkers. *Nat. Commun.* **9,** 883 (2018).](http://paperpile.com/b/7JGcTo/x4fQz)

14. [Berger, A. C. *et al.* A Comprehensive Pan-Cancer Molecular Study of Gynecologic and Breast Cancers. *Cancer Cell* **33,** 690–705.e9 (2018).](http://paperpile.com/b/7JGcTo/Ap1Ys)

15. [Liu, J., Lin, J., Li, Y., Zhang, Y. & Chen, X. Prognostic role of lncRNA TUG1 for cancer outcome: Evidence from 840 cancer patients. *Oncotarget* **8,** (2017).](http://paperpile.com/b/7JGcTo/3gKgE)

16. [Kirk, J. M. *et al.* Functional classification of long non-coding RNAs by k-mer content. *Nat. Genet.* (2018). doi:](http://paperpile.com/b/7JGcTo/Am7J8)[10.1038/s41588-018-0207-8](http://dx.doi.org/10.1038/s41588-018-0207-8)

17. [Chiu, H.-S. *et al.* Pan-Cancer Analysis of lncRNA Regulation Supports Their Targeting of Cancer Genes in Each Tumor Context. *Cell Rep.* **23,** 297–312.e12 (2018).](http://paperpile.com/b/7JGcTo/Yx4xq)

18. [Liu, S. J. *et al.* CRISPRi-based genome-scale identification of functional long noncoding RNA loci in human cells. *Science* **355,** (2017).](http://paperpile.com/b/7JGcTo/zrKWI)

19. [Jiang, C. *et al.* Identifying and functionally characterizing tissue-specific and ubiquitously expressed human lncRNAs. *Oncotarget* **7,** 7120–7133 (2016).](http://paperpile.com/b/7JGcTo/Ebz4Q)

20. [Tomlins, S. A. *et al.* Urine TMPRSS2:ERG fusion transcript stratifies prostate cancer risk in men with elevated serum PSA. *Sci. Transl. Med.* **3,** 94ra72 (2011).](http://paperpile.com/b/7JGcTo/WcsR8)

21. [Gupta, R. A. *et al.* Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature* **464,** 1071–1076 (2010).](http://paperpile.com/b/7JGcTo/pPc4y)

22. [McFerrin, L. G. *et al.* Analysis and visualization of linked molecular and clinical cancer data by using Oncoscape. *Nat. Genet.* **50,** 1203–1204 (2018).](http://paperpile.com/b/7JGcTo/pdKyi)

23. [Uhlen, M. *et al.* A pathology atlas of the human cancer transcriptome. *Science* **357,** (2017).](http://paperpile.com/b/7JGcTo/cmu9d)

24. [The Cancer Genome Atlas Home Page. *The Cancer Genome Atlas - National Cancer Institute* (2011). Available at:](http://paperpile.com/b/7JGcTo/GlfIv) <https://cancergenome.nih.gov/.> [(Accessed: 26th September 2018)](http://paperpile.com/b/7JGcTo/GlfIv)

25. [GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45,** 580–585 (2013).](http://paperpile.com/b/7JGcTo/WEdOy)

26. [Guo, H. *et al.* Modulation of long noncoding RNAs by risk SNPs underlying genetic predispositions to prostate cancer. *Nat. Genet.* **48,** 1142–1150 (2016).](http://paperpile.com/b/7JGcTo/n6eGo)

27. [Huang, H.-L. *et al.* Discovery of prognostic biomarkers for predicting lung cancer metastasis using microarray and survival data. *BMC Bioinformatics* **16,** 54 (2015).](http://paperpile.com/b/7JGcTo/JOBi)

28. [Xiong, J., Bing, Z., Su, Y., Deng, D. & Peng, X. An integrated mRNA and microRNA expression signature for glioblastoma multiforme prognosis. *PLoS One* **9,** e98419 (2014).](http://paperpile.com/b/7JGcTo/zCfF)

29. [Yu, K.-H. *et al.* Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. *Nat. Commun.* **7,** 12474 (2016).](http://paperpile.com/b/7JGcTo/PpmqX)

30. [Jain, R. *et al.* Outcome prediction in patients with glioblastoma by using imaging, clinical, and genomic biomarkers: focus on the nonenhancing component of the tumor. *Radiology* **272,** 484–493 (2014).](http://paperpile.com/b/7JGcTo/sIj7s)

31. [Uno, H., Cai, T., Pencina, M. J., D’Agostino, R. B. & Wei, L. J. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Stat. Med.* **30,** 1105–1117 (2011).](http://paperpile.com/b/7JGcTo/9ZDrG)

32. [Tian, F., Xu, J., Xue, F., Guan, E. & Xu, X. TINCR expression is associated with unfavorable prognosis in patients with hepatocellular carcinoma. *Biosci. Rep.* **37,** (2017).](http://paperpile.com/b/7JGcTo/rAGMz)

33. [Yang, Y., Junjie, P., Sanjun, C. & Ma, Y. Long non-coding RNAs in Colorectal Cancer: Progression and Future Directions. *J. Cancer* **8,** 3212–3225 (2017).](http://paperpile.com/b/7JGcTo/zsl4i)

34. [Arechederra, M. *et al.* Hypermethylation of gene body CpG islands predicts high dosage of functional oncogenes in liver cancer. *Nat. Commun.* **9,** 3164 (2018).](http://paperpile.com/b/7JGcTo/rdZQi)

35. [Hansen, K. D. *et al.* Increased methylation variation in epigenetic domains across cancer types. *Nat. Genet.* **43,** 768–775 (2011).](http://paperpile.com/b/7JGcTo/u0fJQ)

36. [Ciriello, G. *et al.* Emerging landscape of oncogenic signatures across human cancers. *Nat. Genet.* **45,** 1127 (2013).](http://paperpile.com/b/7JGcTo/6h9fW)

37. [Reimand, J., Kull, M., Peterson, H., Hansen, J. & Vilo, J. g:Profiler--a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Res.* **35,** W193–200 (2007).](http://paperpile.com/b/7JGcTo/kvpQ9)

38. [Paczkowska, M. *et al.* Integrative pathway enrichment analysis of multivariate omics data. (2018). doi:](http://paperpile.com/b/7JGcTo/QeMRE)[10.1101/399113](http://dx.doi.org/10.1101/399113)

39. [Rodenas-Cuadrado, P., Ho, J. & Vernes, S. C. Shining a light on CNTNAP2: complex functions to complex disorders. *Eur. J. Hum. Genet.* **22,** 171–178 (2014).](http://paperpile.com/b/7JGcTo/Zb6CM)

40. [Roghanian, A. *et al.* Antagonistic Human FcγRIIB (CD32B) Antibodies Have Anti-Tumor Activity and Overcome Resistance to Antibody Therapy In Vivo. *Cancer Cell* **27,** 473–488 (2015).](http://paperpile.com/b/7JGcTo/TIVvT)

41. [Wang, X. Q. D. & Dostie, J. Reciprocal regulation of chromatin state and architecture by HOTAIRM1 contributes to temporal collinear HOXA gene activation. *Nucleic Acids Res.* **45,** 1091–1104 (2017).](http://paperpile.com/b/7JGcTo/DdVaf)

42. [van Bokhoven, H. *et al.* MYCN haploinsufficiency is associated with reduced brain size and intestinal atresias in Feingold syndrome. *Nat. Genet.* **37,** 465–467 (2005).](http://paperpile.com/b/7JGcTo/GtOi)

43. [Yan, K.-S. *et al.* EZH2 in Cancer Progression and Potential Application in Cancer Therapy: A Friend or Foe? *Int. J. Mol. Sci.* **18,** (2017).](http://paperpile.com/b/7JGcTo/IDYo)

44. [Mondal, T. *et al.* Sense-Antisense lncRNA Pair Encoded by Locus 6p22.3 Determines Neuroblastoma Susceptibility via the USP36-CHD7-SOX9 Regulatory Axis. *Cancer Cell* **33,** 417–434.e7 (2018).](http://paperpile.com/b/7JGcTo/P9dxP)

45. [Marusyk, A., Almendro, V. & Polyak, K. Intra-tumour heterogeneity: a looking glass for cancer? *Nat. Rev. Cancer* **12,** 323–334 (2012).](http://paperpile.com/b/7JGcTo/H38U0)

46. [Gerstung, M. *et al.* Precision oncology for acute myeloid leukemia using a knowledge bank approach. *Nat. Genet.* **49,** 332–340 (2017).](http://paperpile.com/b/7JGcTo/E7eH)

47. [Du, Z. *et al.* Integrative genomic analyses reveal clinically relevant long noncoding RNAs in human cancer. *Nat. Struct. Mol. Biol.* **20,** 908–913 (2013).](http://paperpile.com/b/7JGcTo/KHoO6)

48. [Liu, J. *et al.* Abstract 3287: An integrated TCGA pan-cancer clinical data resource to drive high quality survival outcome analytics. *Cancer Res.* **78,** 3287–3287 (2018).](http://paperpile.com/b/7JGcTo/owMAp)

49. [Colaprico, A. *et al.* TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res.* **44,** e71–e71 (2015).](http://paperpile.com/b/7JGcTo/aksSF)

50. [Campbell, P. J. *et al.* Pan-cancer analysis of whole genomes. (2017). doi:](http://paperpile.com/b/7JGcTo/IAx5)[10.1101/162784](http://dx.doi.org/10.1101/162784)

51. [PCAWG Transcriptome Core Group *et al.* Genomic basis for RNA alterations revealed by whole-genome analyses of 27 cancer types. (2017). doi:](http://paperpile.com/b/7JGcTo/TXx3)[10.1101/183889](http://dx.doi.org/10.1101/183889)

52. [Krijthe, J. H. Rtsne: T-distributed stochastic neighbor embedding using Barnes-Hut implementation. *R package version 0. 13, URL https://github. com/jkrijthe/Rtsne* (2015).](http://paperpile.com/b/7JGcTo/Kapn)

53. [Quinlan, A. R. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. *Curr. Protoc. Bioinformatics* **47,** 11.12.1–34 (2014).](http://paperpile.com/b/7JGcTo/FOGm)

54. [Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43,** e47 (2015).](http://paperpile.com/b/7JGcTo/0CRW6)

55. [Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13,** 2498–2504 (2003).](http://paperpile.com/b/7JGcTo/IVBf)

56. [Wei, T. & Simko, V. R package ‘corrplot’: visualization of a correlation matrix (version 0.84).’. *Retrived from https://github. com/taiyun/corrplot* (2017).](http://paperpile.com/b/7JGcTo/56Yb)

57. [Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32,** 2847–2849 (2016).](http://paperpile.com/b/7JGcTo/v0tg)