Semi-automated annotation of known and novel cancer long noncoding RNAs with the Cancer LncRNA Census 2 (CLC2)

Authors: Adrienne Vancura (1,2), Andrés Lanzós (1), Núria Bosch (1,2), Monica Torres (1), Simon Häfliger (1), Rory Johnson (1,3)

Affiliations:

[1]Department of Medical Oncology, Inselspital, Bern University Hospital, University of Bern, Bern, Switzerland;

[2]Graduate School of Cellular and Biomedical Sciences, University of Bern, Bern, Switzerland.

[3]Department for BioMedical Research, University of Bern, Bern, Switzerland.

Correspondance: rory.johnson@dbmr.unibe.ch

Keywords: long noncoding RNA; long-noncoding RNA; lncRNA, cancer, curation, *DGCR5*, *CARMEN*, *CARMN*, *LINC00570*

Abstract

Long noncoding RNAs (lncRNAs) can promote or repress the cellular hallmarks of cancer. Understanding their molecular roles and realising their therapeutic potential depend on high-quality catalogues of cancer lncRNA genes. Presently, such catalogues depend on labour-intensive curation of heterogeneous data with permissive criteria, resulting in unknown numbers of genes without direct functional evidence. Here, we present an approach for semi-automated curation focused exclusively on pathogenic functionality. The result is Cancer LncRNA Census 2 (CLC2), comprising 492 gene loci in 33 cancer types. To complement manual literature curation, we develop an automated pipeline, CLIO-TIM, to identify novel cancer lncRNAs based on functional evolutionary conservation with mouse. This yields 95 novel lncRNAs, which display characteristics of known cancer genes and include *LINC00570* (*ncRNA-a5*), which we demonstrate experimentally to promote cell proliferation. The clinical importance and curation accuracy of CLC2 lncRNAs is highlighted by a range of features, including evolutionary selection, expression in tumours, and both somatic and germline polymorphisms. The entire dataset is available in a highly-curated format facilitating the widest range of downstream applications. In summary, we show how manual and automated methods can be integrated to catalogue known and novel functional cancer lncRNAs with unique genomic and clinical properties.

Introduction

Cancer arises through a series of somatic genetic mutations, leading via defined cellular phenotypic hallmarks to the formation of a tumour (Hanahan & Weinberg, 2011)(Yates & Campbell, 2012). Such mutations are principally thought to alter the function of polypeptides encoded by protein-coding genes (pc-genes) (Sondka et al., 2018) and the increased probability of suffering dysfunctional mutations defines known cancer genes (Furney et al., 2006). Datasets such as the Cancer Gene Census (CGC) collect and organise comprehensive sets of cancer pc-genes according to defined criteria, and represent invaluable and widely-used resources for scientific research and drug discovery (Sondka et al., 2018).

The past decade has witnessed the discovery of numerous non-protein-coding RNA genes in mammalian cells (Guttman et al., 2009; Uszczynska-Ratajczak et al., 2018). The most numerous but poorly understood produce long noncoding RNAs (lncRNAs), defined as transcripts >200 nt in length with no detectable protein-coding potential (Derrien et al., 2012). Although their molecular mechanisms are highly diverse, many lncRNAs have been shown to interact with other RNA molecules, proteins and DNA by structural and sequence specific (Guttman & Rinn, 2012) (Johnson & Guigó, 2014). Most lncRNAs are clade- and species-specific, but a subset display deeper evolutionary conservation in their gene structure (Ulitsky et al., 2011) and a handful have been demonstrated to have functions that were conserved across millions of years of evolution (Marín-Béjar et al., 2017; Ulitsky et al., 2011). The numbers of known lncRNA genes in human have grown rapidly, and present catalogues range from 18,000 to ~100,000 (Frankish et al., n.d.), however just a tiny fraction have been functionally characterized (Kopp & Mendell, 2018)(Ulitsky & Bartel, 2013)(Ma et al., 2019)(Quek et al., 2015). Understanding the clinical and therapeutic significance of these numerous novel genes is a key contemporary challenge.

LncRNAs have been implicated in molecular processes governing tumorigenesis (Slack & Chinnaiyan, 2019). LncRNAs may promote or oppose cancer hallmarks (Du et al., 2013b). This fact, coupled to the emergence of potent *in vivo* inhibitors in the form of antisense oligonucleotides (ASOs) (Dias & Stein, 2002), has given rise to serious interest in lncRNAs as drug targets in cancer by both academia and pharma (Tony Gutschner et al., 2013; Wahlestedt, 2013)(Kaczmarek et al., 2017)(Slack & Chinnaiyan, 2019).

Initially, cancer lncRNAs were discovered by classical functional genomics workflows employing microarray or RNA-seq expression profiling (Huarte et al., 2010; Iyer et al., 2015). More recently, CRISPR-based functional screening (Esposito et al., 2019) and bioinformatic predictions (Lanzós et al., 2017; Mularoni et al., 2016; Rheinbay et al., 2017) have also emerged as powerful tools for novel cancer gene discovery. To assess their accuracy, these approaches require accurate benchmarks in the form of curated databases of known cancer lncRNAs.

Any discussion of lncRNAs and cancer requires careful terminology. Tumours display large numbers of differentially expressed genes (Iyer et al., 2015). However, just a fraction of these are likely to functionally contribute to a relevant cellular phenotype or cancer hallmark (T Gutschner et al., n.d.; Hosono et al., 2017; Lee et al., 2016; Leucci et al., 2016; Munschauer et al., 2018). Such genes, termed here "functional cancer lncRNAs", are the focus of this study. Remaining changing genes are non-functional "bystanders", which are largely irrelevant in understanding or inhibiting the molecular processes causing cancer.

There are a number of excellent databases of cancer-associated lncRNAs: lncRNADisease (Bao et al., 2019), CRlncRNA (Wang et al., 2018), EVLncRNAs (Zhou et al., 2018) and Lnc2Cancer (Gao et al., 2019). These principally employ labour-intensive manual curation, and rely extensively on differential expression to identify candidates. On the other hand, these databases have not begun to use more recent high-confidence sources of functional cancer lncRNAs, such as high-throughput functional screens (Esposito et al., 2019)(Abbott et al., 2015). For these reasons, existing annotations likely contain unknown numbers of bystander lncRNAs, while omitting large numbers of *bona fide* functional cancer lncRNAs. Thus, studies requiring high-confidence gene sets, including benchmarking or drug discovery, call for a database focussed exclusively on functional cancer lncRNAs.

Here we address this need through the creation of the Cancer LncRNA Census 2 (CLC2). This extends our previous CLC dataset by several fold (Carlevaro-Fita et al., 2020). More importantly, CLC2 takes a major step forward methodologically, by implementing an automated curation component that utilises functional evolutionary conservation for the first time. Using this data, we present a comprehensive analysis of the genomic and clinical features of cancer lncRNAs. Most important, we present a practical and versatile dataset intended for use by basic researchers and drug discovery projects.

Results

Integrative, semi-automated cataloguing of cancer lncRNAs

We sought to develop an improved map of lncRNAs with functional roles in either promoting or opposing cancer hallmarks or tumorigenesis. Such a map should prioritise lncRNAs with genuine causative roles, and exclude false-positive "bystander" lncRNAs whose expression changes but play no functional role.

We began with conventional manual curation of lncRNAs from the scientific literature, covering the period from January 2017 (directly after the end of the first CLC (Carlevaro-Fita et al., 2020)) to the end of December 2018. The criteria for defining cancer lncRNAs were identical: genes must be annotated in GENCODE (here version 28), and cancer function must be demonstrated by *in vitro* or *in vivo* experiments or germline or somatic mutational evidence (Figure 1A). Altogether we collected 253 novel lncRNAs in this way, which added to the original CLC (n=122) amounts to 375 lncRNAs, hereafter denoted as "literature lncRNAs" (Figure 1A).

We previously showed that transposon insertion mutagenesis (TIM) hits are enriched in cancer-associated lncRNA genes, implying that cancer lncRNAs' functions can be deeply evolutionarily conserved (Carlevaro-Fita et al., 2020). We developed a pipeline to automatically identify likely human functional cancer lncRNAs by orthology to a collection of TIM hits (Abbott et al., 2015). In this way 123 lncRNAs were detected, of which 102 were not already in the literature set. These were added to the CLC2, henceforth denoted as "mutagenesis lncRNAs" (Figure 1B). This analysis is discussed in more detail in the next section.

Pooled functional screens based on CRISPR-Cas9 loss-of-function have recently emerged as a powerful means of identifying function cancer lncRNAs (Esposito et al., 2019). The most comprehensive dataset presently available comes from a CRISPR-inhibition (CRISPRi) screen of ~16,000 lncRNAs in seven human cell lines, with proliferation as a readout (Liu et al., 2017). Of the 499 hits identified, 322 are annotated by GENCODE. These hits are significantly enriched for known cancer lncRNAs from the literature search (Figure 1C). That study independently validated 21 GENCODE-annotated hits. Four (19%) of these were already mentioned in the scientific literature, and 2 (10%) were detected in the TIM screen above. Given their high-confidence, we added the remaining 15 novel lncRNAs to CLC2 ("CRISPRi lncRNAs") (Figure 1C).

Altogether, CLC2 comprises 492 unique lncRNA genes detected in 33 cancer types. *PVT1*, *SNHG1* and *SNHG12* genes are detected in all three sources. The entire CLC2 dataset is available in Supplementary Table 1 and 2. Importantly, the dataset is fully annotated with evidence information, enabling users to filter particular sets of lncRNAs in which they have greater levels of confidence.

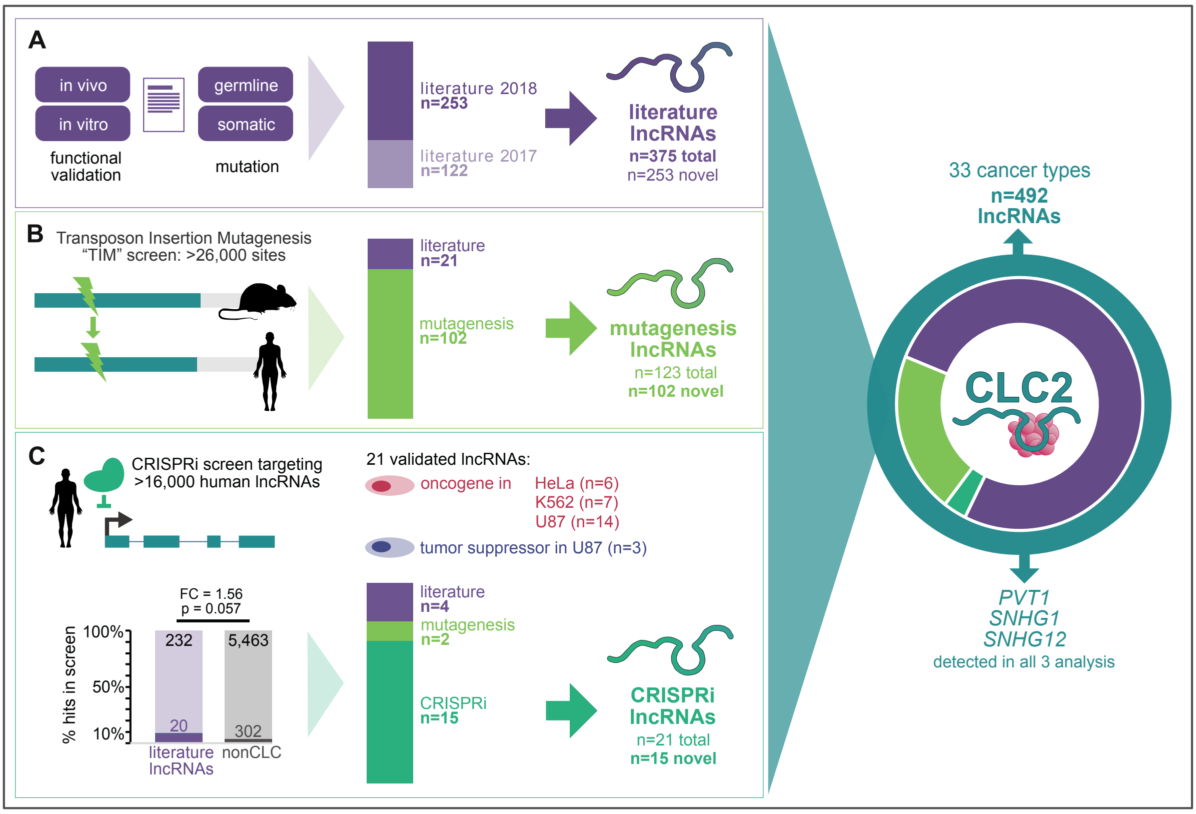

Figure 1: Functional cancer lncRNAs from three sources are integrated in the CLC2.
A) Literature curation with four criteria are used to define "literature lncRNAs". **B)** Transposon insertion mutagenesis screens identify "mutagenesis lncRNAs". **C)** Validated hits from CRISPRi proliferation screens are denoted "CRISPRi lncRNAs". Statistical significance calculated by one-sided Fisher's test.

Automated annotation of human cancer lncRNAs via functional conservation

We recently showed that transposon insertional mutagenesis (TIM) screens can be used to identify cancer lncRNAs in mouse (Copeland & Jenkins, 2010) (Carlevaro-Fita et al., 2020) , and that these often have human orthologues which are also cancer genes (Figure 2A). TIM screens identify "common insertion sites" (CIS), where multiple transposon insertions at a particular genomic location have given rise to a tumour, thereby implicating the underlying gene as an oncogene or tumour suppressor.

We reasoned that this could be extended to identify new functional cancer lncRNAs in human, and developed a pipeline for this: CLIO-TIM (cancer lncRNA identification by orthology to TIM). Briefly, CLIO-TIM uses chain alignments to map mouse CIS to orthologous regions of the human genome, and then identifies their likely target gene by proximity (see Methods) (Figure 2B) (SUPP FIG 1B).

Using 26,345 mouse CIS from public databases (Abbott et al., 2015), CLIO-TIM identifies 16,430 orthologous regions in human (hCIS) (Figure 2B) (SUPP FIG 1A). Altogether, 123 lncRNAs and 9,295 pc-genes are identified as potential cancer genes. An example is the human-mouse orthologous lncRNA locus shown in Figure 2B, comprising *Gm36495* in mouse and *LINC00570* in human. A CIS lies upstream of the mouse gene's TSS, mapping to the first intron of the human orthologue. *LINC00570* is an alternative identifier for ncRNA-a5 *cis*-acting lncRNA identified by Orom et al. (Ørom et al., 2010), that has not previously been associated with cancer or cell growth.

We expect that hCIS regions are enriched in known cancer genes. Consistent with this, the 698 pc-genes from the COSMIC Cancer Gene Census (CGC) (Sondka et al., 2018) (red in SUPP FIG 1D) are 155-fold enriched with hCIS over intergenic regions (light grey). Turning to lncRNAs, the 375 literature lncRNA are 19.5-fold enriched, supporting their disease importance (Figure 2C). Thus, CLIO-TIM predictions are enriched in genuine protein-coding and lncRNA functional cancer genes. As expected, the overall numbers of genes implicated by CLIO-TIM agree with independent analysis in the CCGD database (SUPP FIG 1C).

An additional 209 hCIS fall in intergenic regions that are neither part of pc-genes or lncRNAs, leading us to ask whether some may affect lncRNAs that are not annotated by GENCODE (Figure 2C). To test this, we utilised the large set of cancer-associated lncRNAs from miTranscriptome (Iyer et al., 2015). 186 hCIS intersect 2167 miTranscriptome genes, making these potentially novel non-annotated transcripts involved in cancer. Nevertheless, simulations indicated that this rate of overlap was no greater than expected by random chance (see Methods), making it unlikely that substantial numbers of undiscovered cancer lncRNAs remain to be discovered in intergenic regions, at least with the datasets used here (SUPP FIG 1E).

In addition to known cancer lncRNAs, CLIO-TIM identifies 102 lncRNAs not previously linked to cancer (FIG 2C, dark grey) with a 3.8-fold enrichment of insertions over intergenic space. These lncRNAs represent novel functional cancer genes, and were added to the CLC2. This makes the assumption that human orthologues of mouse cancer genes will have a conserved function. We tested this using *LINC00570*, predicted to be a cancer gene by CLIO-TIM but never previously been linked to cancer or cell proliferation. *LINC00570* expression is dysregulated in several tumour types (SUPP FIG 2A) and its high expression is a poor prognostic indicator for Glioblastoma Multiforme (GBM) patients (SUPP FIG 2A) (Tang et al., 2019). We asked whether *LINC00570* promotes cell growth in cancer cells. We found that it is robustly detected in cervical carcinoma HeLa cells (SUPP FIG 2B) and to a lesser extent in HCT116 colon carcinoma cells (SUPP FIG 2B). We designed three distinct antisense oligonucleotides (ASOs) targeting the *LINC00570* intron 2 and 3 and exon 3 of the short isoform. Transfection of these ASOs led to strong and reproducible decreases in steady state RNA levels in HeLa cells (Figure 2D). As a consequence, we observed significant decreases in cell proliferation rates (Figure 2D, SUPP FIG 2C). We observed a similar effect through CRISPRi-mediated inhibition of gene transcription by two independent guide RNAs in HeLa (Figure 2D), and using the same ASOs in HCT116 cells (SUPP FIG 2D and E). Therefore, the

CLIO-TIM pipeline correctly identified *LINC00570* as a likely new functional cancer lncRNA in human.

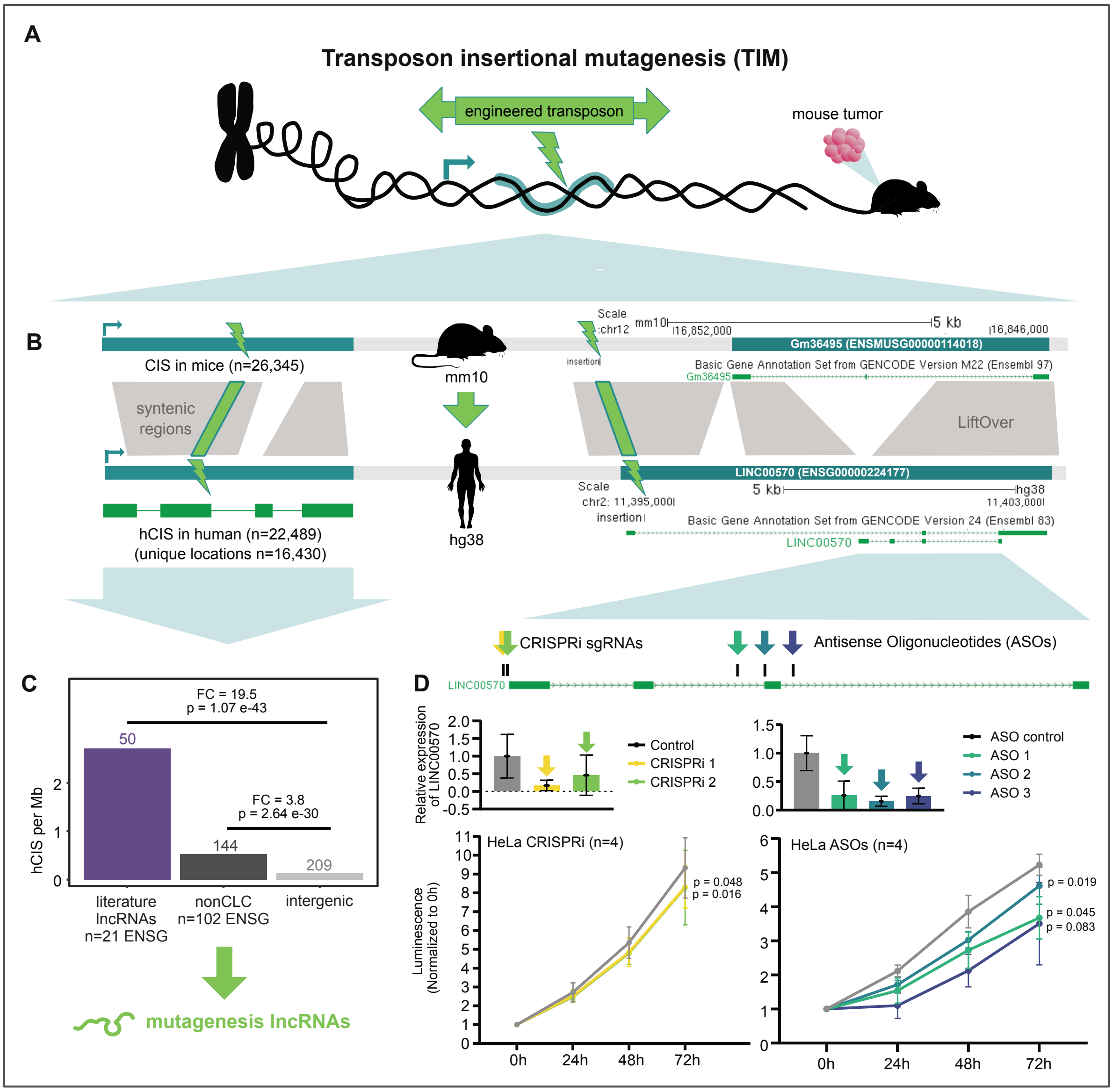

Figure 2: The CLIO-TIM pipeline identifies human cancer lncRNAs via functional evolutionary conservation.

A) Overview of transposon insertional mutagenesis (TIM) method for identifying functional cancer genes. Engineered transposons carry bidirectional cassettes capable of either blocking or upregulating gene transcription, depending on orientation. Transposons are introduced into a population of cells, where they integrate at random genomic sites. The cells are injected into a mouse. In some cells, transposons will land in and perturb expression of a cancer gene (either tumour suppressor or oncogene), giving rise to a tumour. DNA of tumour cells is sequenced to identify the exact location of the transposon insertion. Clusters of such insertions are termed Common Insertion Sites (CIS). **B)** (Left) Schematic of the CLIO-TIM pipeline used here to identify human cancer genes using mouse CIS. (Right) An example of a CLIO-TIM predicted cancer lncRNA. **C)** The density of hCIS sites, normalised by gene length, in indicated classes of lncRNAs. Statistical significance calculated by one-sided Fisher's test. **D)** Upper panels: Expression of *LINC00570* RNA in response to inhibition by CRISPRi (left) or ASOs (right). Lower panels: Measured populations of the same cells over time. Statistical significance calculated by Student's *t*-test.

Enhanced cancer lncRNA catalogue integrating manual annotation, CRISPR screens and functional conservation

We here describe some general features of the CLC2 dataset. Figure 3A shows a breakdown of the composition of CLC2 in terms of source, gene function and evidence strength. Where possible, the genes are given a functional annotation, oncogene (og) or tumour suppressor (ts), according to evidence of promoting or opposing cancer hallmarks. Oncogenes (n=275) quite considerably outnumber tumour suppressors (n=95), although it is not clear whether this reflects genuine biology or an ascertainment bias relating to scientific interest or technical issues. Smaller sets of lncRNAs are associated with both functions, or have no functional information (those from TIM screens where the functions of hits are ambiguous).

In terms of the quality of evidence sources, CLC2 represents a strong improvement over the original CLC. The fraction of lncRNAs with high quality *in vivo* evidence (defined as functional validation in mouse models or identified by mutagenesis analysis) now represent 66% compared to 24% previously (FIG 3A, SUPP FIG 3B). In total the updated CLC2 comprises 33 cancer types (vs 29) and more lncRNAs are reported for every cancer subtype (SUPP FIG 3A).

Several longstanding cancer lncRNA collections have provided an invaluable resource for the community (Figure 3B). Comparing only GENCODE v28 genes, CLC2 with 492 genes is second only to Lnc2Cancer (n= 512) (Gao et al., 2019). However Lnc2Cancer uses looser inclusion criteria, including lncRNAs without GENCODE IDs and those that are differentially expressed in tumours with no functional evidence. Remaining databases are smaller. Furthermore, CLC2 has the greatest number of unique GENCODE gene loci, compared to other databases (n=225), including numerous literature-annotated cases and also 95 novel mutagenesis lncRNAs. Just 40 lncRNAs are common to all five databases (Bao et al., 2019; Gao et al., 2019; Wang et al., 2018; Zhou et al., 2018). In summary, CLC2 is a large, practical and high-quality cancer lncRNA annotation that complements existing resources.

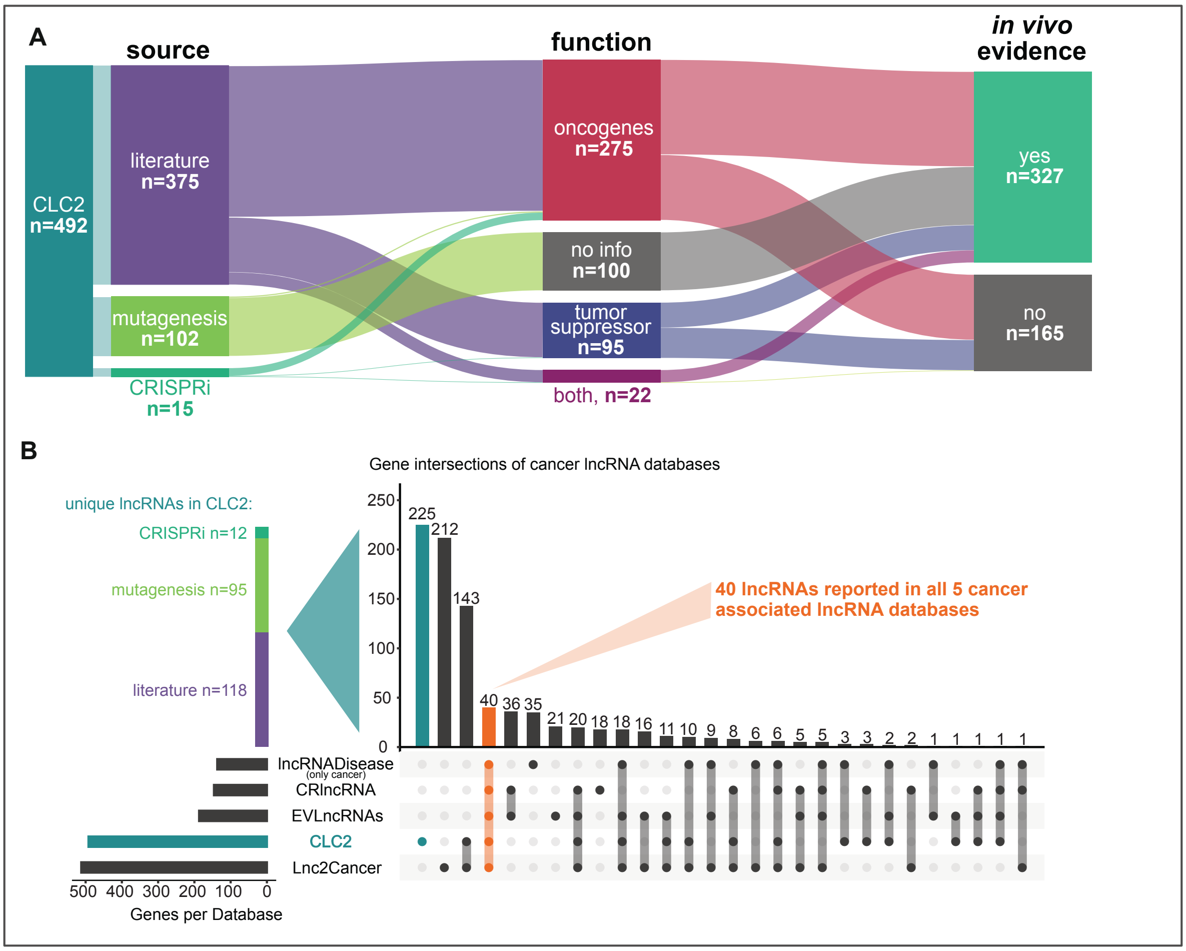

Figure 3: An overview of the CLC2 database and comparison with other lncRNA databases.

A) The CLC2 database broken down by source, function and evidence type. **B)** Comparison of CLC2 to other leading cancer lncRNA databases.

Unique genomic properties of CLC2 lncRNAs

Cancer genes, both protein-coding and not, have elevated characteristics of essentiality and clinical importance, compared to other genes (Furney et al., 2008)(Furney et al., 2006)(Du et al., 2013a)(Sondka et al., 2018). We next asked whether CLC2 lncRNAs display genomic features expected for functionally important genes.

In the following analyses, we compared gene features of CLC2 lncRNAs to all other lncRNAs. Comparison of gene sets can often be confounded by covariates such as gene length or gene expression, therefore where appropriate we used control gene sets that were matched to CLC2 by expression (denoted “nonCLCmatched”) (SUPP FIG 4A) and reported findings correcting for gene length (SUPP FIG 4B).

Using the LnCompare tool (Carlevaro-Fita et al., 2019), we find that the promoters and exons of CLC2 genes display elevated signatures of functionality, in terms of evolutionary conservation in mammalian and vertebrate phylogeny (Figure 4A) and expression in cancer cell lines (Figure 4B). Strikingly we observe a similar effect when considering the novel mutagenesis lncRNAs alone: their promoters are significantly more conserved than expected by chance, and their expression is an order of magnitude higher than other lncRNAs (Figure 4C and D).

Further, we found that CLC2 lncRNAs are enriched in repetitive elements (SUPP FIG 5C) and are more likely to house a small RNA gene, possibly indicating that some act as precursor transcripts (SUPP FIG 5B). CLC2 lncRNAs also have non-random distributions of gene biotypes, being depleted for intergenic class and enriched in divergent orientation to other genes (SUPP FIG 5A).

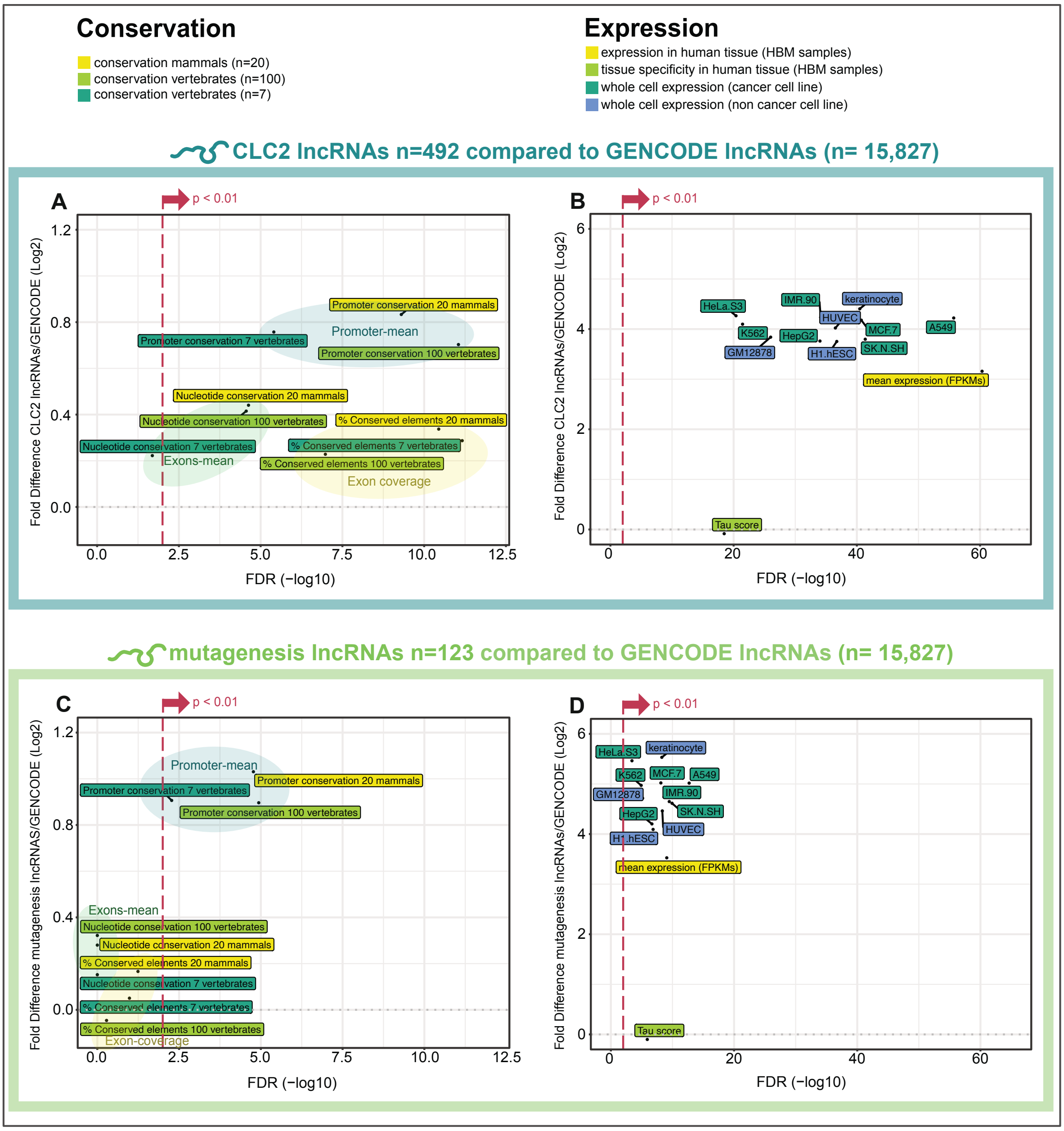

Figure 4: Features of functionality in CLC2 and mutagenesis lncRNAs.

In each panel, two gene sets are compared: the test set (either all CLC2 genes, or mutagenesis genes alone), and the set of all other lncRNAs (GENCODE v24). Y-axis: Log2 fold difference between the means of gene sets. X-axis: false-discovery rate adjusted significance, calculated by Wilcoxon test. **A)** Evolutionary conservation for all CLC2, calculated by PhastCons. **B)** Expression of all CLC2 in cell lines. **C)** Evolutionary conservation for mutagenesis lncRNAs, calculated by PhastCons. **D)** Expression of mutagenesis lncRNAs in cell lines. For (A) and (C), “Promoter mean” and “Exon mean” indicate mean PhastCons scores (7-vertebrate alignment) for those features, while “Exon-coverage” indicates percent coverage by PhastCons elements. Promoters are defined as a window of 200 nt centered on the transcription start site.

CLC2 lncRNAs display consistent tumour expression changes and prognostic properties

Although gene expression was not a criterion for inclusion, we do expect that CLC2 lncRNAs levels will be altered in tumours. Furthermore, we expect that oncogenes should be overexpressed, and tumour suppressors downregulated.

To test this, we analysed TCGA RNA-sequencing (RNA-seq) data from 686 individual tumours with matched healthy tissue (total n=1,372 analyzed samples) in 20 different cancer types (SUPP FIG 6A and B), and classed every gene as either differentially expressed (in at least one cancer subtype with a log2 Fold Change >1 and a FDR <0.05) or not. CLC2 lncRNAs are 3.4-fold more likely to be differentially expressed compared to expression-matched lncRNAs (Figure 5A). LncRNAs from each individual evidence source (literature, mutagenesis, CRISPRi) display the same trend. Similar effects were found for pc-genes (SUPP FIG 7A).

Next we asked whether the direction of expression change corresponds to gene function. Indeed, oncogenes are enriched for overexpressed genes, whereas tumor suppressors are enriched for down-regulated genes, supporting the functional labelling scheme (Figure 5B).

Cancer genes' expression is often prognostic for patient survival. By correlating expression to patient survival, we found that the expression of 392 CLC2 lncRNAs correlated to patient survival in at least one cancer type (SUPP FIG 7C). When analyzing the most significant correlation of each CLC2 lncRNA compared to expression-matched nonCLC lncRNAs, we find a weak but significant enrichment (SUPP FIG 7C), showing that CLC2 lncRNAs tend to be prognostic for patient survival.

In summary, gene expression characteristics of CLC2 genes, and subsets from different evidence sources, support their functional labels as oncogenes and tumour suppressors and is more broadly consistent with their important roles in tumorigenesis.

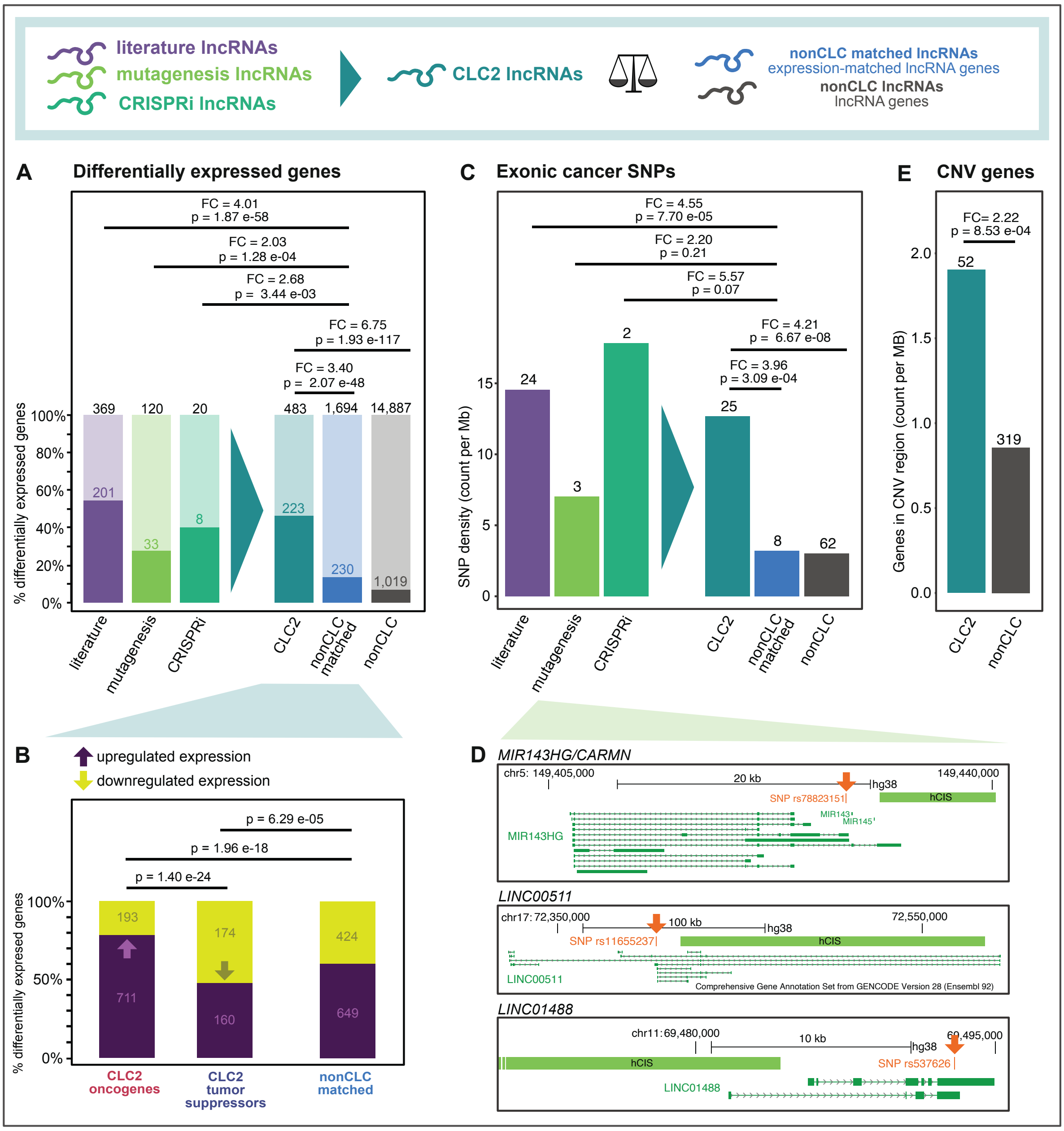

Figure 5: Clinical features of CLC2 lncRNAs.

A) The percent of indicated genes that are significantly differentially expressed in at least one tumour type from the TCGA. Statistical significance calculated by one-sided Fisher's test. **B)** Here, only differentially expressed genes from (A) are considered. LncRNAs with both tumour suppressor and oncogene labels are excluded. Remaining lncRNAs are divided by those that are up- or down-regulated (positive or negative fold change). Statistical significance calculated by one-sided Fisher's test. **C)** The density of germline cancer-associated SNPs is displayed. Only SNPs falling in gene exons are counted, and are normalised to the total length of those exons. Statistical significance calculated by one-sided Fisher's test. **D)** Examples of mutagenesis lncRNAs with an exonic cancer SNP. **E)** Length-normalised overlap rate of copy number variants (CNVs) in lncRNA gene span. Statistical significance calculated by one-sided Fisher's test.

CLC2 lncRNAs are enriched with cancer genetic mutations

Cancer genes are characterized by a range of germline and somatic mutations that lead to gain or loss of function. We hypothesised that cancer lncRNAs should be enriched with germline single nucleotide polymorphisms that have been linked to cancer predisposition (Deng et al., 2017). We obtained 5,331 cancer-associated single nucleotide polymorphisms (SNPs) from genome-wide association studies (GWAS) (Buniello et al., 2019) and mapped them to lncRNA and pc-gene exons, calculating a density score that normalises for exon length (SUPP FIG 4B). As expected, exons of known cancer pc-genes are >2-fold enriched in germline SNPs (SUPP FIG 7B). When performing the same analysis with CLC2 lncRNAs, one observes an even more pronounced enrichment of 4.0-fold when comparing to expression-matched nonCLC lncRNAs (Figure 5C). Once again, the lncRNAs from each evidence source individually show enrichment for cancer SNPs >2-fold (Figure 5C). Three mutagenesis lncRNAs, namely *miR143HG/CARMN*, *LINC00511* and *LINC01488*, exhibit an exonic cancer SNP (Figure 5D).

Cancer genes are also frequently the subject of large-scale somatic mutations, or copy number variants (CNVs). Using a collection of CNV data from LncVar (Chen et al., 2017), we calculated the gene-span length-normalized coverage of lncRNAs by CNVs. CLC2 lncRNAs are enriched for CNVs compared to all non cancer lncRNAs (Figure 5E).

In summary, CLC2 lncRNAs display germline and somatic mutational patterns consistent with known oncogenes and tumour suppressors.

Discussion

We have presented the Cancer LncRNA Census 2, a resource of lncRNAs with functional roles in cancer. We hope this dataset will be of utility to a wide range of studies, from bioinformatic identification of new disease genes, to developing a new generation of cancer therapeutics with anti-lncRNA ASOs (Amodio et al., 2018).

A key novelty of CLC2 is its use of automated gene curation using functional evolutionary conservation. This responds to the challenge arising from the rapid growth of scientific literature, which makes manual curation increasingly impractical. Other automated methods like text mining and machine learning will also be important, although it will be necessary to ensure their predictions are sufficiently accurate in identifying *bona fide* functional lncRNAs and removing bystanders. Available evidence suggests that the CLIO-TIM pipeline accurately identifies functional cancer genes: in addition to experimental validation of the *LINC00570* in two cell lines using two perturbations, the entire set of 102 predicted “mutagenesis lncRNAs” carry a range of features consistent with literature-curated cancer lncRNAs: promoter conservation, high expression, differential tumour expression and germline SNP enrichment. Removing cases that overlap other databases, this set amounts to 95 completely novel functional cancer lncRNAs, an invaluable resource for discovery of new molecular pathways and therapeutic targets.

We also integrated hits from latest CRISPRi screens into CLC2 (Liu et al., 2017) Similar to the “mutagenesis” lncRNAs, the CRISPRi hits have features consistent with cancer functionality.

We recognise that some colleagues may ascribe lower confidence to the novel genes in CLC2 originating from mutagenesis and CRISPR sources. For this reason, the CLC2 data table is organised to facilitate filtering by confidence, to extract only the 375 literature-supported cases, or indeed any other subset based on source, evidence or function as desired by the researcher.

LINC00570 is a new functional cancer lncRNA predicted by CLIO-TIM. The gene was previously studied by Orom and colleagues, as a *cis*-activating enhancer-like RNA named *ncRNA-a5* (Ørom et al., 2010). That and a subsequent study showed that perturbation by siRNA transfection affects the expression of the nearby pc-gene *ROCK2* in HeLa. However, these studies did not investigate the effect on cell proliferation. We here show by means of two independent perturbations, that *LINC00570* promotes proliferation of HeLa and HCT116 cells. These findings make *LINC00570* a potential therapeutic target for follow up.

Intriguingly, amongst the novel mutagenesis lncRNAs identified by CLIO-TIM are genes previously linked to other diseases. *miR143HG/CARMEN1* (*CARMN*) was shown to regulate cardiac specification and differentiation in mouse and human hearts (Ounzain et al., 2015). In addition to being a TIM target, CARMEN1 also contains a germline cancer SNP correlating to the risk of developing lung cancer (Park et al., 2015), adding further weight to the notion that it also plays a role in oncogenesis. Similarly, *DGCR5*, is located in the DiGeorge critical locus and has been linked to neurodevelopment and neurodegeneration (Johnson et al., 2009), and was recently implicated as a tumour suppressor in prostate cancer (Li et al., 2019). These results raise the possibility that developmental lncRNAs can also play roles in cancer.

In summary, we anticipate that CLC2 will lay the foundation for understanding how lncRNAs are integrated into the molecular events underlying tumorigenesis, and provide targets for a new generation of anti-cancer therapies.

Material and Methods

Screening lncRNAs for inclusion in the CLC2

If not stated otherwise, GENCODE v28 gene IDs (gencode.v28.annotation.gtf) were used.

Literature search. PubMed was searched for publications linking lncRNA and cancer using keywords: long noncoding RNA cancer, lncRNA cancer. The manual curation and assigning evidence levels to each lncRNA was performed in the same way as previously (Carlevaro-Fita et al., 2020) and included reports until December 2018.

CLIO-TIM Insertional mutagenesis screen analysis. From the CCGD website (http://ccgd-starrlab.oit.umn.edu/about.php, May 2018 (Abbott et al., 2015)) a table with all CIS elements was downloaded. These mouse genomic regions (mm10) were converted to homologous regions in the human genome assembly hg38 using the LiftOver tool (https://genome.ucsc.edu/cgi-bin/hgLiftOver). Settings: original Genome was Mouse GRCm38/mm10 to New Genome Human GRCh38/hg38, minMatch was 0.1 and minBlocks 0.1. For insertion sites intersecting several lncRNA genes, all the genes were reported. IntersectBed from bedtools was used to align human insertion sites to GENCODE IDs by intersecting at least 1nt and assigned to protein-coding or lncRNA gene families. Insertion sites aligning to protein-coding and lncRNA genes were always assigned to protein-coding genes. If insertion sites overlap multiple ENSGs, all genes are reported. Insertion sites not aligning to protein-coding or lncRNAs genes were added to the intergenic region.

CCGD human Entrez gene results were converted to GENCODE IDs using the “Entrez gene ids” Metadata file from https://www.gencodegenes.org/human/ to compare CLIO-TIM results with CCGD results for each gene set.

Cancer-associated MiTranscriptome: The cancer associated MiTranscriptome IDs (Iyer et al., 2015) previously used in Bergada et al. (Bergadà-Pijuan et al., 2019) were intersected with intergenic insertion sites using IntersectBed. With ShuffleBed the intergenic insertions were randomly shuffled 1000x and assigned to MiTranscriptome IDs.

CRISPRi screen analysis. We used the Supp Table 1 (Liu et al., 2017) to extract ENST IDs and gene names which are then converted to GENCODE IDs to match each guide (LH identifier in the screen). From supplementary table S4 (Liu_et_al_aah7111-TableS4) (Liu et al., 2017) we extracted genes with “hit” (validated as a hit in the screen), “LH” (unique identifiers correlating to a gene in the screen) and “lncRNA” (referring to a lncRNA gene and to exclude lncRNA hits close to a protein-coding gene (“Neighbor hit”)) resulting in 499 hits. Of these, 322 hits contain a GENCODE IDs and were used for enrichment analysis, tested by one-sided Fisher’s test.

We included n=21 CRISPRi genes to the CLC2 from Liu et al. Supp Fig 8A, cancer cell line and the effect of the CRISPRi on the growth phenotype (either promoting (tumor suppressor) or inhibiting (oncogene)) of each lncRNA was reported.

Gene family grouping. For downstream analysis protein-coding (pc) genes (GENCODE IDs) are grouped in cancer-associated pc-genes (CGC genes) and non cancer-associated pc-genes (nonCGC n=19,174). The TSV file containing the CGC data was downloaded from https://cancer.sanger.ac.uk/census with 700 ENSGs with 698 ENSG IDs detected in GENCODE v28 of which 696 are unique (CGC n=696). The same is done for lncRNAs, into CLC2 (n=492) and nonCLC genes (n= 15,314).

Matched expression analysis. Based on an in house script used for Survival analysis (section below), TCGA survival expression data for each GENCODE ID is reported and the average FPKM across all tumor samples is calculated. The count distribution of nonCGC and nonCLC gene expression to CGC and CLC2 expression, respectively, is matched using the matchDistribution.pl script (https://github.com/julienlag/matchDistribution).

Cancer lncRNA databases. The tested databases were first filtered for lncRNAs in the GENCODE v28 long noncoding annotation (n=15,767).

Lnc2cancer GENCODE IDs from datatable (http://www.bio-bigdata.com/lnc2cancer/download.html) were evaluated (n=512) (Gao et al., 2019).

CRlncRNA gene names from (http://crlnc.xtbg.ac.cn/download/) were converted to GENCODE IDs (n=146) (Wang et al., 2018).

EVlncRNAs gene names (http://biophy.dzu.edu.cn/EVLncRNAs/) were converted to GENCODE IDs (n=187) (Zhou et al., 2018).

lncRNADisease gene names from (http://www.rnanut.net/lncrnadisease/index.php/home/info/download) and only cancer-associated transcripts (carcinoma, lymphoma, cancer, leukemia, tumor, glioma, sarcoma, blastoma, astrocytoma, melanoma, meningioma) were extracted. Names were converted to GENCODE IDs (n=137) (Bao et al., 2019).

Features of CLC2 genes

Genomic classification. The genomic classification was performed as previously (Carlevaro-Fita et al., 2020) using an in house script (https://github.com/gold-lab/shared_scripts/tree/master/lncRNA.annotator).

Small RNA analysis. For this analysis “snoRNA”, “snRNA”, “miRNA” and “miscRNA” coordinates were extracted from GENCODE v28 annotation file and intersected with the genomic region of the genes (intronic and exonic regions).

Repeat elements. In total 452 CLC2 lncRNAs compared to 1693 expression-matched nonCLC lncRNAs using the LnCompare Categorical analysis (http://www.rnanut.net/lncompare/) (Carlevaro-Fita et al., 2019).

Feature analysis. In total 452 CLC2 lncRNAs and 120 mutagenesis lncRNAs are compared to the GENCODE v24 reference using LnCompare (http://www.rnanut.net/lncompare/) (Carlevaro-Fita et al., 2019).

Cancer characteristic analysis

Differential gene expression analysis (DEA). was performed using TCGA data and TCGAbiolinks. Analysis was performed as reported in manual for matching tumor and normal tissue samples using the HTseq analysis pipeline as described previously. (https://www.bioconductor.org/packages/devel/bioc/vignettes/TCGAbiolinks/inst/doc/analysis.html) (Colaprico et al., 2016). For this analysis only matched samples were used and the TCGA data was presorted for tumor tissue samples (TP with 01 in sample name) and solid tissue normal (NT with 11 in sample name). Settings used for DEA analysis: fdr.cut = 0.05 , logFC.cut = 1 for DGE output between matched TP and NT samples for 20 cancer types. CLC2 cancer types had to be converted to TCGA cancer types (Supp Fig 6A) Cancer types and number of samples used in the analysis can be found in Supp Fig 6B. DEA enrichment analysis tested with one-sided Fisher's test. For each CLC2 gene reported as true oncogene (n=275) or tumor suppressor (n=95), hence where no double function is reported (n=22), the positive and negative fold change (FC) values were counted and compared to expression-matched lncRNA genes found in the DEA.

Survival analysis. An inhouse script for extracting TCGA survival data was used to generate p values correlating to survival for each gene. Expression and clinical data from 33 cohorts from TCGA with the "TCGAbiolinks" R package (https://bioconductor.org/packages/release/bioc/html/TCGAbiolinks.html) were downloaded (Colaprico et al., 2016). P value and Hazard ratio were calculated with the Cox proportional hazards regression model from "Survival" R package (https://cran.r-project.org/web/packages/survival/survival.pdf). All scripts were adapted from here (https://www.biostars.org/p/153013/) and are available upon request. For downstream analysis, only groups with at least 20 patient samples in high or low expression group were used. The plot comprises only the most significant cancer survival p value per gene and was assessed by the Komnogorow-Smirnow-Test (ks-test).

Cancer-associated SNP analysis. SNP data linked to tumor/cancer/tumour were extracted from the GWAS page (https://www.ebi.ac.uk/gwas/docs/file-downloads) (n=5,331) and intersected with the whole exon body of the genes. SNPs were intersected to the transcript bed file and plotted per nt in each subset (SNP/nt y axis) and tested using one-sided Fisher's test.

CNV analysis. Human CNV in lncRNAs downloaded from http://bioinfo.ibp.ac.cn/LncVar/download.php (Chen et al., 2017). NONCODE IDs were converted to GENCODE IDs using NONCODEv5_hg38.lncAndGene.bed.gz. CLC2 and nonCLC ENSGs were matched to NONHSAT IDs with a significant pvalue (0.05, n=733) in the LncVAR table and tested using one-sided Fisher's test.

Code availability. Custom code are available from the corresponding author upon request.

In vitro validation

Cell culture. HeLa and HCT116 were cultured on Dulbecco's Modified Eagles Medium (DMEM) (Sigma-Aldrich, D5671) supplemented with 10% Fetal Bovine Serum (FBS) (ThermoFisher Scientific, 10500064), 1% L-Glutamine (ThermoFisher Scientific, 25030024), 1% Penicillin-Streptomycin (ThermoFisher Scientific, 15140122). Cells were grown at 37°C and 5% CO2 and passaged every two days at 1:5 dilution.

Generation of Cas9 stable cell lines. HeLa cells were infected with lentivirus carrying the Cas9-BFP (blue fluorescent protein) vector (Addgene 52962). HCT116 were transfected with the same vector using Lipofectamine 2000 (ThermoFisher Scientific, 11668019). Both cell types were selected with blasticidin (4ug/ml) for at least five days and selected for BFP-positive cells twice by fluorescence activated cell sorting.

CRISPR inhibition sgRNA pair design and cloning. sgRNA pairs targeting *LINC00570* were designed using GPP sgRNA designer (https://portals.broadinstitute.org/gpp/). The sgRNA pairs were manually selected from the output list and cloned into the pGECKO backbone (CRISPRi.1: *5' GTTACTTCCAACGTACCATG 3'*, CRISPRi.2: *5' CCTGTACCCCCATGGTACGT 3'*) (Addgene 78534; (Aparicio-Prat et al., 2015))

Antisense LNA GapmeR design. Antisense LNA GapmeR Control (*5' AACACGTCTATACGC 3')* and three Antisense LNA GapmeR Standard targeting *LINC00570* (LNA1: *5' GGAAATTGCTCTGATG 3'*, LNA2: *5' GATTGGCATTGGGATA 3'*, LNA3: *5' GAAGTGGCCTGAGAAA 3'*) were designed and purchased at Qiagen.

RT-qPCR. For each time point total RNA was extracted and reverse transcribed (Promega). Transcript levels of *LINC00570* (FP: *5' TAGGAGTGCTGGAGACTGAG 3'*, RP: *5' GTCGCCATCTTGGTTGTCTG 3'*) and housekeeping gene *HPRT1* (FP: *5' ATGACCAGTCAACAGGGGACAT 3'*, RP: *5' CAACACTTCGTGGGGTCCTTTTCA 3'*) were measured using GoTaq qPCR Master Mix (Promega, A6001) on a TaqMan Viia 7 Real-Time PCR System. Data were normalized using the ΔΔCt method (Schmittgen & Livak, 2008)).

Gene-specific RT-PCR and cDNA amplification. From the extracted total RNA, we performed a gene specific reverse transcription using the reverse primers for *LINC00570* and *HPRT1* to enrich for their cDNA. Presence or absence of transcript was detected by a regular

PCR using GoTaq® G2 DNA Polymerase (Promega, M7841) from 100ng cDNA and visualized on an agarose gel.

Viability assay. HeLa and HCT116 cells were transfected with Antisense LNA GapmeRs at a concentration of 50nM using Lipofectamine 2000 (Thermofisher) according to manufacturer's protocol. One day after, transfected cells were plated in a white, flat 96-well plate (3000 cells/well). Viability was measured in technical replicates using CellTiter-Glo 2D Kit (Promega) following manufacturer's recommendations at 0, 24, 48, 72 hours after seeding. Luminescence was detected with Tecan Reader Infinite 200. Statistical significance calculated by t-test.

For CRISPR inhibition experiments, HeLa-Cas9 and HCT116-Cas9 cells were transfected with control sgRNA plasmid and two *LINC00570* targeting plasmids. Cells were selected with puromycin (2ug/ml) for 48h. Viability assay was performed as previously described.

Acknowledgements

We gratefully acknowledge administrative support from Ana Radovanovic and Silvia Roesselet (DBMR, University of Bern). We also acknowledge Juna Carlevaro-Fita (EPFL, Lausanne) and Judith Bergada (University of Zurich) for the helpful advice and discussions as well as Roberta Esposito, Panagiotis Chouvardas, Hugo Guillen Ramirez and the other members of the Laboratory for Genomics of LncRNA and Disease for their valuable input. This work was funded by the Swiss National Science Foundation through the National Center of Competence in Research (NCCR) "RNA & Disease", by the Medical Faculty of the University and University Hospital of Bern, by the Helmut Horten Stiftung and Krebsliga Schweiz (4534-08-2018).

Contributions

RJ conceived the project. RJ, AV performed manual annotation of CLC2. AV performed the feature analysis, evolutionary analysis, mutation analysis, differential expression, GWAS SNP, CNV analysis and data integration. AL performed survival analysis. NB performed the ASO and CRISPRi KD experiments. AV, NB and MT performed the qPCR experiments. RJ, AV, AL, NB, MT and SH drafted the manuscript and prepared the figures and supplementary material. All authors read and approved the final draft.

References

Abbott, K. L., Nyre, E. T., Abrahante, J., Ho, Y.-Y., Isaksson Vogel, R., & Starr, T. K. (2015). The Candidate Cancer Gene Database: a database of cancer driver genes from forward genetic screens in mice. *Nucleic Acids Research*, *43*(D1), D844–D848. https://doi.org/10.1093/nar/gku770

Amodio, N., Stamato, M. A., Juli, G., Morelli, E., Fulciniti, M., Manzoni, M., Taiana, E., Agnelli, L., Cantafio, M. E. G., Romeo, E., Raimondi, L., Caracciolo, D., Zuccalà, V., Rossi, M., Neri, A., Munshi, N. C., Tagliaferri, P., & Tassone, P. (2018). Drugging the lncRNA MALAT1 via LNA gapmeR ASO inhibits gene expression of proteasome subunits and triggers anti-multiple myeloma activity. *Leukemia*. https://doi.org/10.1038/s41375-018-0067-3

Aparicio-Prat, E., Arnan, C., Sala, I., Bosch, N., Guigó, R., & Johnson, R. (2015). DECKO: Single-oligo, dual-CRISPR deletion of genomic elements including long non-coding RNAs. *BMC Genomics*, *16*(1), 846. https://doi.org/10.1186/s12864-015-2086-z

Bao, Z., Yang, Z., Huang, Z., Zhou, Y., Cui, Q., & Dong, D. (2019). LncRNADisease 2.0: An updated database of long non-coding RNA-associated diseases. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gky905

Bergadà-Pijuan, J., Pulido-Quetglas, C., Vancura, A., & Johnson, R. (2019). CASPR, an analysis pipeline for single and paired guide RNA CRISPR screens, reveals optimal target selection for long noncoding RNAs. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btz811

Buniello, A., Macarthur, J. A. L., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., Suveges, D., Vrousgou, O., Whetzel, P. L., Amode, R., Guillen, J. A., Riat, H. S., Trevanion, S. J., Hall, P., Junkins, H., … Parkinson, H. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gky1120

Carlevaro-Fita, J., Lanzós, A., Feuerbach, L., Hong, C., Mas-Ponte, D., Pedersen, J. S., Abascal, F., Amin, S. B., Bader, G. D., Barenboim, J., Beroukhim, R., Bertl, J., Boroevich, K. A., Brunak, S., Campbell, P. J., Carlevaro-Fita, J., Chakravarty, D., Chan, C. W. Y., Chen, K., … Johnson, R. (2020). Cancer LncRNA Census reveals evidence for deep functional conservation of long noncoding RNAs in tumorigenesis. *Communications Biology*. https://doi.org/10.1038/s42003-019-0741-7

Carlevaro-Fita, J., Liu, L., Zhou, Y., Zhang, S., Chouvardas, P., Johnson, R., & Li, J. (2019). LnCompare: gene set feature analysis for human long non-coding RNAs. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkz410

Chen, X., Hao, Y., Cui, Y., Fan, Z., He, S., Luo, J., & Chen, R. (2017). LncVar: A database of genetic variation associated with long non-coding genes. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btw581

Colaprico, A., Silva, T. C., Olsen, C., Garofano, L., Cava, C., Garolini, D., Sabedot, T. S., Malta, T. M., Pagnotta, S. M., Castiglioni, I., Ceccarelli, M., Bontempi, G., & Noushmehr, H. (2016). TCGAbiolinks: An R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkv1507

Copeland, N. G., & Jenkins, N. A. (2010). Harnessing transposons for cancer gene discovery. In *Nature Reviews Cancer*. https://doi.org/10.1038/nrc2916

Deng, N., Zhou, H., Fan, H., & Yuan, Y. (2017). Single nucleotide polymorphisms and cancer susceptibility. In *Oncotarget*. https://doi.org/10.18632/oncotarget.22372

Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., Guernec, G., Martin, D., Merkel, A., Knowles, D. G., Lagarde, J., Veeravalli, L., Ruan, X., Ruan, Y., Lassmann, T., Carninci, P., Brown, J. B., Lipovich, L., Gonzalez, J. M., … Guigó, R. (2012). The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. *Genome Research*. https://doi.org/10.1101/gr.132159.111

Dias, N., & Stein, C. A. (2002). Antisense oligonucleotides: basic concepts and mechanisms. *Molecular Cancer Therapeutics*, *1*(5), 347–355. https://doi.org/10.1016/s1357-4310(99)01638-x

Du, Z., Fei, T., Verhaak, R. G. W., Su, Z., Zhang, Y., Brown, M., Chen, Y., & Liu, X. S. (2013a). Integrative genomic analyses reveal clinically relevant long noncoding RNAs in human cancer. *Nature Structural and Molecular Biology*. https://doi.org/10.1038/nsmb.2591

Du, Z., Fei, T., Verhaak, R. G. W., Su, Z., Zhang, Y., Brown, M., Chen, Y., & Liu, X. S. (2013b). Integrative genomic analyses reveal clinically relevant long noncoding RNAs in human cancer. *Nature Structural & Molecular Biology*, *20*(7), 908–913. https://doi.org/10.1038/nsmb.2591

Esposito, R., Bosch, N., Lanzós, A., Polidori, T., Pulido-Quetglas, C., & Johnson, R. (2019). Hacking the Cancer Genome: Profiling Therapeutically Actionable Long Non-coding RNAs Using CRISPR-Cas9 Screening. *Cancer Cell*, *35*(4), 545–557. https://doi.org/10.1016/j.ccell.2019.01.019

Frankish, A., Diekhans, M., Ferreira, A.-M., Johnson, R., Jungreis, I., Loveland, J., Mudge, J. M., Sisu, C., Wright, J., Armstrong, J., Barnes, I., Berry, A., Bignell, A., Carbonell Sala, S., Chrast, J., Cunningham, F., Di Domenico, T., Donaldson, S., Fiddes, I. T., … Flicek, P. (n.d.). GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Research*, *47*(D1), D766–D773. https://doi.org/10.1093/nar/gky955

Furney, S. J., Higgins, D. G., Ouzounis, C. A., & López-Bigas, N. (2006). Structural and functional properties of genes involved in human cancer. *BMC Genomics*, *7*(1), 3. https://doi.org/10.1186/1471-2164-7-3

Furney, S. J., Madden, S. F., Kisiel, T. A., Higgins, D. G., & Lopez-Bigas, N. (2008). Distinct patterns in the regulation and evolution of human cancer genes. *In Silico Biology*.

Gao, Y., Wang, P., Wang, Y., Ma, X., Zhi, H., Zhou, D., Li, X., Fang, Y., Shen, W., Xu, Y., Shang, S., Wang, L., Wang, L., Ning, S., & Li, X. (2019). Lnc2Cancer v2.0: Updated database of experimentally supported long non-coding RNAs in human cancers. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gky1096

Gutschner, T, Hammerle, M., Eissmann, M., Hsu, J., Kim, Y., Hung, G., Revenko, A., Arun, G., Stentrup, M., Gross, M., & others. (n.d.). (2013) {The} noncoding {RNA} {MALAT}1 is a critical regulator of the metastasis phenotype of lung cancer cells. *Cancer Research*, *73*, 1180–1189.

Gutschner, Tony, Hämmerle, M., Eißmann, M., Hsu, J., Kim, Y., Hung, G., Revenko, A., Arun, G., Stentrup, M., Groß, M., Zörnig, M., MacLeod, A. R., Spector, D. L., & Diederichs, S. (2013). The noncoding RNA MALAT1 is a critical regulator of the metastasis phenotype of lung cancer cells. *Cancer Research*. https://doi.org/10.1158/0008-5472.CAN-12-2850

Guttman, M., Amit, I., Garber, M., French, C., Lin, M. F., Feldser, D., Huarte, M., Zuk, O., Carey, B. W., Cassady, J. P., Cabili, M. N., Jaenisch, R., Mikkelsen, T. S., Jacks, T., Hacohen, N., Bernstein, B. E., Kellis, M., Regev, A., Rinn, J. L., & Lander, E. S. (2009). Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature*. https://doi.org/10.1038/nature07672

Guttman, M., & Rinn, J. L. (2012). Modular regulatory principles of large non-coding RNAs. In *Nature*. https://doi.org/10.1038/nature10887

Hanahan, D., & Weinberg, R. A. (2011). Hallmarks of {Cancer}: {The} {Next} {Generation}. *Cell*, *144*(5), 646–674. https://doi.org/10.1016/j.cell.2011.02.013

Hosono, Y., Niknafs, Y. S., Prensner, J. R., Iyer, M. K., Dhanasekaran, S. M., Mehra, R., Pitchiaya, S., Tien, J., Escara-Wilke, J., Poliakov, A., Chu, S.-C., Saleh, S., Sankar, K., Su, F., Guo, S., Qiao, Y., Freier, S. M., Bui, H.-H., Cao, X., … Chinnaiyan, A. M. (2017). No Title. *Cell*, *171*(7), 1559-1572.e20. https://doi.org/10.1016/j.cell.2017.11.040

Huarte, M., Guttman, M., Feldser, D., Garber, M., Koziol, M. J., Kenzelmann-Broz, D., Khalil, A. M., Zuk, O., Amit, I., Rabani, M., Attardi, L. D., Regev, A., Lander, E. S., Jacks, T., & Rinn, J. L. (2010). A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response. *Cell*, *142*(3), 409–419. https://doi.org/10.1016/j.cell.2010.06.040

Iyer, M. K., Niknafs, Y. S., Malik, R., Singhal, U., Sahu, A., Hosono, Y., Barrette, T. R., Prensner, J. R., Evans, J. R., Zhao, S., Poliakov, A., Cao, X., Dhanasekaran, S. M., Wu, Y.-M. M., Robinson, D. R., Beer, D. G., Feng, F. Y., Iyer, H. K., & Chinnaiyan, A. M. (2015). The landscape of long noncoding RNAs in the human transcriptome. *Nature Genetics*, *47*(3), 199–208. https://doi.org/10.1038/ng.3192

Johnson, R., & Guigó, R. (2014). The RIDL hypothesis: Transposable elements as functional domains of long noncoding RNAs. *Rna*, *20*(7), 959–976. https://doi.org/10.1261/rna.044560.114

Johnson, R., Teh, C. H. L., Jia, H., Vanisri, R. R., Pandey, T., Lu, Z. H., Buckley, N. J., Stanton, L. W., & Lipovich, L. (2009). Regulation of neural macroRNAs by the transcriptional repressor REST. *RNA*. https://doi.org/10.1261/rna.1127009

Kaczmarek, J. C., Kowalski, P. S., & Anderson, D. G. (2017). Advances in the delivery of RNA therapeutics: From concept to clinical reality. In *Genome Medicine*. https://doi.org/10.1186/s13073-017-0450-0

Kopp, F., & Mendell, J. T. (2018). Functional Classification and Experimental Dissection of Long Noncoding RNAs. In *Cell*. https://doi.org/10.1016/j.cell.2018.01.011

Lanzós, A., Carlevaro-Fita, J., Palumbo, E., Reverter, F., Mularoni, L., Guigó, R., Johnson, R., Reverter, F., Palumbo, E., Guigó, R., Johnson, R., Lanzos, A., Carlevaro-Fita, J., Mularoni, L., Reverter, F., Palumbo, E., Guigo, R., Johnson, R., Lanzós, A., … Johnson, R. (2017). Discovery of Cancer Driver Long Noncoding RNAs across 1112 Tumour Genomes: New Candidates and Distinguishing Features. *Scientific Reports*, *7*(1), 41544. https://doi.org/10.1038/srep41544

Lee, S., Kopp, F., Chang, T.-C. C., Sataluri, A., Chen, B., Sivakumar, S., Yu, H., Xie, Y., & Mendell, J. T. T. (2016). Noncoding RNA NORAD Regulates Genomic Stability by Sequestering PUMILIO Proteins. *Cell*, *164*(1–2), 69–80. https://doi.org/10.1016/j.cell.2015.12.017

Leucci, E., Vendramin, R., Spinazzi, M., Laurette, P., Fiers, M., Wouters, J., Radaelli, E., Eyckerman, S., Leonelli, C., Vanderheyden, K., Rogiers, A., Hermans, E., Baatsen, P., Aerts, S., Amant, F., Van Aelst, S., van den Oord, J., De Strooper, B., Davidson, I., … Marine, J.-C. (2016). Melanoma addiction to the long non-coding RNA SAMMSON. *Nature*, *531*(7595), 518–522. https://doi.org/10.1038/nature17161

Li, B., Guo, Z., Liang, Q., Zhou, H., Luo, Y., He, S., & Lin, Z. (2019). LncRNA DGCR5 Up-regulates TGF-β1, increases cancer cell stemness and predicts survival of prostate cancer patients. *Cancer Management and Research*. https://doi.org/10.2147/CMAR.S231112

Liu, S. J., Horlbeck, M. A., Cho, S. W., Birk, H. S., Malatesta, M., He, D., Attenello, F. J., Villalta, J. E., Cho, M. Y., Chen, Y., Mandegar, M. A., Olvera, M. P., Gilbert, L. A., Conklin, B. R., Chang, H. Y., Weissman, J. S., & Lim, D. A. (2017). CRISPRi-based genome-scale identification of functional long noncoding RNA loci in human cells.

Science, *355*(6320), eaah7111. https://doi.org/10.1126/science.aah7111

Ma, L., Cao, J., Liu, L., Du, Q., Li, Z., Zou, D., Bajic, V. B., & Zhang, Z. (2019). Lncbook: A curated knowledgebase of human long non-coding rnas. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gky960

Marín-Béjar, O., Mas, A. M., González, J., Martinez, D., Athie, A., Morales, X., Galduroz, M., Raimondi, I., Grossi, E., Guo, S., Rouzaut, A., Ulitsky, I., Huarte, M., Marin-Bejar, O., Mas, A. M., Gonzalez, J., Martinez, D., Athie, A., Morales, X., … Huarte, M. (2017). The human lncRNA LINC-PINT inhibits tumor cell invasion through a highly conserved sequence element. *Genome Biology*, *18*(1), 202. https://doi.org/10.1186/s13059-017-1331-y

Mularoni, L., Sabarinathan, R., Deu-Pons, J., Gonzalez-Perez, A., & López-Bigas, N. (2016). OncodriveFML: A general framework to identify coding and non-coding regions with cancer driver mutations. *Genome Biology*, *17*(1), 128. https://doi.org/10.1186/s13059-016-0994-0

Munschauer, M., Nguyen, C. T., Sirokman, K., Hartigan, C. R., Hogstrom, L., Engreitz, J. M., Ulirsch, J. C., Fulco, C. P., Subramanian, V., Chen, J., Schenone, M., Guttman, M., Carr, S. A., & Lander, E. S. (2018). The NORAD lncRNA assembles a topoisomerase complex critical for genome stability. *Nature*, *561*(7721), 132–136. https://doi.org/10.1038/s41586-018-0453-z

Ørom, U. A., Derrien, T., Beringer, M., Gumireddy, K., Gardini, A., Bussotti, G., Lai, F., Zytnicki, M., Notredame, C., Huang, Q., Guigo, R., & Shiekhattar, R. (2010). Long noncoding RNAs with enhancer-like function in human cells. *Cell*. https://doi.org/10.1016/j.cell.2010.09.001

Ounzain, S., Micheletti, R., Arnan, C., Plaisance, I., Cecchi, D., Schroen, B., Reverter, F., Alexanian, M., Gonzales, C., Ng, S. Y., Bussotti, G., Pezzuto, I., Notredame, C., Heymans, S., Guigó, R., Johnson, R., & Pedrazzini, T. (2015). CARMEN, a human super enhancer-associated long noncoding RNA controlling cardiac specification, differentiation and homeostasis. *Journal of Molecular and Cellular Cardiology*. https://doi.org/10.1016/j.yjmcc.2015.09.016

Park, S. L., Carmella, S. G., Chen, M., Patel, Y., Stram, D. O., Haiman, C. A., Le Marchand, L., & Hecht, S. S. (2015). Mercapturic acids derived from the toxicants acrolein and crotonaldehyde in the urine of cigarette smokers from five ethnic groups with differing risks for lung cancer. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0124841

Quek, X. C., Thomson, D. W., Maag, J. L. V., Bartonicek, N., Signal, B., Clark, M. B., Gloss, B. S., & Dinger, M. E. (2015). lncRNAdb v2.0: Expanding the reference database for functional long noncoding RNAs. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gku988

Rheinbay, E., Nielsen, M. M., Abascal, F., Tiao, G., Hornshøj, H., Hess, J. M., Pedersen, R. I. I., Feuerbach, L., Sabarinathan, R., Madsen, H. T., KIM, J., Mularoni, L., Shuai, S., Camaioni, A. A. L., Herrmann, C., Maruvka, Y. E., Shen, C., Amin, S. B., Bertl, J., … Net, I. P.-C. A. of W. G. (2017). Discovery and characterization of coding and non-coding driver mutations in more than 2,500 whole cancer genomes. *BioRxiv*, 237313. https://doi.org/10.1101/237313

Schmittgen, T. D., & Livak, K. J. (2008). Analyzing real-time PCR data by the comparative CT method. *Nature Protocols*, *3*(6), 1101–1108. https://doi.org/10.1038/nprot.2008.73

Slack, F. J., & Chinnaiyan, A. M. (2019). The Role of Non-coding RNAs in Oncology. In *Cell*. https://doi.org/10.1016/j.cell.2019.10.017

Sondka, Z., Bamford, S., Cole, C. G., Ward, S. A., Dunham, I., & Forbes, S. A. (2018). The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. In *Nature Reviews Cancer*. https://doi.org/10.1038/s41568-018-0060-1

Tang, Z., Kang, B., Li, C., Chen, T., & Zhang, Z. (2019). GEPIA2: an enhanced web server for large-scale expression profiling and interactive analysis. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkz430

Ulitsky, I., & Bartel, D. P. P. P. (2013). lincRNAs: genomics, evolution, and mechanisms. *Cell*, *154*(1), 26–46. https://doi.org/10.1016/j.cell.2013.06.020

Ulitsky, I., Shkumatava, A., Jan, C. H., Sive, H., & Bartel, D. P. (2011). Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. *Cell*, *147*(7), 1537–1550. https://doi.org/10.1016/j.cell.2011.11.055

Uszczynska-Ratajczak, B., Lagarde, J., Frankish, A., Guigó, R., & Johnson, R. (2018). Towards a complete map of the human long non-coding RNA transcriptome. In *Nature Reviews Genetics*. https://doi.org/10.1038/s41576-018-0017-y

Wahlestedt, C. (2013). Targeting long non-coding RNA to therapeutically upregulate gene expression. In *Nature Reviews Drug Discovery*. https://doi.org/10.1038/nrd4018

Wang, J., Zhang, X., Chen, W., Li, J., & Liu, C. (2018). CRlncRNA: A manually curated database of cancer-related long non-coding RNAs with experimental proof of functions on clinicopathological and molecular features. *BMC Medical Genomics*. https://doi.org/10.1186/s12920-018-0430-2

Yates, L. R., & Campbell, P. J. (2012). Evolution of the cancer genome. In *Nature Reviews Genetics*. https://doi.org/10.1038/nrg3317

Zhou, B., Zhao, H., Yu, J., Guo, C., Dou, X., Song, F., Hu, G., Cao, Z., Qu, Y., Yang, Y., Zhou, Y., & Wang, J. (2018). EVLncRNAs: A manually curated database for long non-coding RNAs validated by low-throughput experiments. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkx677