Unique genomic features and deeply-conserved functions of long non-coding RNAs in the Cancer LncRNA Census (CLC)

Joana Carlevaro-Fita* (1,2)
Andrés Lanzós* (3,4,5)
Lars Feuerbach (6)
Chen Hong (6)
David Mas-Ponte (3,4,5)
PCAWG Working Group 2-5-9-14
Roderic Guigó (3,4,5)
Jakob Skou Pedersen (7)
Rory Johnson (1,2)

1. Department of Clinical Research, University of Bern, 3008 Bern, Switzerland
2. Department of Medical Oncology, Inselspital, University Hospital and University of Bern, 3010 Bern, Switzerland
3. Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Dr. Aiguader 88, Barcelona 08003, Spain
4. Universitat Pompeu Fabra (UPF), Barcelona, Spain.
5. Institut Hospital del Mar d'Investigacions Mèdiques (IMIM), Dr. Aiguader 88, 08003 Barcelona, Catalonia, Spain.
6. Applied Bioinformatics, Deutsches Krebsforschungszentrum, 69120 Heidelberg, Germany
7. Department for Molecular Medicine, Aarhus University Hospital, Palle Juul-Jensens Boulevard 99, 8200 Aarhus N, Denmark.

*** Equal contribution**

Correspondence to rory.johnson@dkf.unibe.ch

Keywords: long non-coding RNA; lncRNA; GENCODE; cancer driver; driver prediction; manual curation; pancancer analysis of whole genomes; PCAWG.

Abstract

Long non-coding RNAs (lncRNAs) that drive tumorigenesis are a growing focus of cancer genomics studies. To facilitate further discovery, we have created the "Cancer LncRNA Census" (CLC), a manually-curated and strictly-defined compilation of lncRNAs with causative roles in cancer. CLC has two principle applications: first, as a resource for training and benchmarking *de novo* identification methods; and second, as a dataset for studying the fundamental properties of these genes.

CLC Version 1 comprises 122 lncRNAs implicated in 31 distinct cancers. LncRNAs are included based on functional or genetic evidence for different causative roles in cancer progression. All belong to the GENCODE reference annotation, to facilitate integration across projects and datasets. For each entry, the evidence type, biological activity (oncogene or tumour suppressor), source reference and cancer type are recorded. CLC genes are significantly enriched amongst *de novo* predicted driver genes from PCAWG. CLC genes are distinguished from other lncRNAs by a series of features consistent with biological function, including gene length, expression and sequence conservation of both exons and promoters. We identify a trend for CLC genes to be co-localised with known protein-coding cancer genes along the human genome. Finally, by integrating data from transposon-mutagenesis functional screens, we show that mouse orthologues of CLC genes tend also to be cancer driver genes.

Thus CLC represents a valuable resource for research into long non-coding RNAs in cancer. Their evolutionary and genomic properties have implications for understanding disease mechanisms and point to conserved functions across ~80 million years of evolution.

Introduction

Tumorigenesis is driven by a series of genetic mutations that promote cancer phenotypes and consequently experience positive selection (Yates & Campbell 2012). The systematic discovery of such driver mutations, and the genes whose functions they alter, has been made possible by tumour genome sequencing. By collecting the entirety of such genes for every cancer type, we aim to develop a comprehensive view of underlying processes and pathways, and thereby formulate effective, targeted therapeutic strategies.

The cast of genetic elements implicated in tumorigenesis has recently grown as diverse new classes of non-coding RNAs and regulatory features have been discovered. These include the long non-coding RNAs (lncRNAs), of which tens of thousands have been catalogued (Derrien et al. 2012; Cabili et al. 2011; Guttman et al. 2009; Jia et al. 2010). LncRNAs are >200 nt long transcripts with no protein-coding capacity. Their evolutionary conservation and regulated expression, combined with a number of well-characterised examples, have together led to the view that lncRNAs are *bona fide* functional genes (Grote et al. 2013; Sauvageau et al. 2013; Ulitsky & Bartel 2013; Liu et al. 2017). Current thinking holds that lncRNAs function by forming complexes with proteins and RNA both inside and outside the nucleus (Guttman & Rinn 2012; Johnson & Guigó 2014).

LncRNAs have been shown to play important roles in various cancers. For example, *MALAT1*, a potent oncogene across numerous cancers, is restricted to the nucleus and plays a housekeeping role in splicing (Gutschner & Diederichs 2012; Engreitz et al. 2014). *MALAT1* is overexpressed in a variety of cancer types, and its knockdown potently reduces not only proliferation but also metastasis *in vivo* (Gutschner et al. 2013).The *MALAT1* gene is subjected to elevated mutational rates in human tumours, suggesting that genetic alteration in its function is an important step in tumorigenesis (Lanzós et al. 2017). LncRNAs may also function as tumour suppressors. *LincRNA-p21* acts as a downstream effector of p53 regulation through recruitment of the repressor hnRNP-K (Huarte et al. 2010). These and other examples of lncRNAs linked to cancer, raise the question of how many more remain to be found amongst the ~99% of lncRNAs that are presently uncharacterised (Derrien et al. 2012; Quek et al. 2015; Iyer et al. 2015).

Recent tumour genome sequencing studies, in step with advanced bioinformatic driver-gene prediction methods, have yielded hundreds of new candidate protein-coding driver genes (Tamborero et al. 2013). For economic reasons, these studies initially restricted their attention to “exomes” or the ~2% of the genome covering protein-coding exons (Chang et al. 2013). Unfortunately such a strategy ignores mutations in the remaining ~98% of genomic sequence, home to the majority of lncRNAs (Gutschner & Diederichs 2012; Derrien et al. 2012). Driver gene identification methods rely on statistical models that make a series of assumptions about and simplifications of complex tumour mutation patterns (Lawrence et al. 2014). It is critical to test the performance of such methods using true-positive lists of known driver genes. For protein-coding genes, this role has been fulfilled by the Cancer Gene Census (CGC) (Futreal et al. 2004), collected and regularly updated by manual annotators based on experimentally-validated cases. Comparison of driver predictions to CGC genes facilitates further method refinement and comparison between methods (Sjoblom et al. 2006; Redon et al. 2006; Mularoni et al. 2016).

In addition to its benchmarking role, the CGC resource has also been useful in identifying unique biological features of cancer genes (Furney et al. 2006; Furney et al. 2008). For example, CGC genes tend to be more conserved and longer. Furthermore, they are enriched for genes with transcription regulator activity and nucleic acid binding functions (Furney et al. 2006; Furney et al. 2008).

Until very recently, efforts to discover cancer lncRNAs have depended on classical functional genomics approaches of differential expression using microarrays or RNA sequencing (Huarte et al. 2010; Iyer et al. 2015). While valuable, differential expression *per se* is not direct evidence for causative roles in tumour evolution. To more directly identify lncRNAs that drive cancer progression, methods have recently been developed to search for signals of positive selection using mutation maps created using tumour genome sequences. OncodriveFML utilises nucleotide-level functional impact scores inferred from predicted changes in RNA secondary structure (Sabarinathan et al. 2013) together with an empirical significance estimate, to identify lncRNAs with an excess of high-impact mutations (Mularoni et al. 2016). Another method, ExInAtor, identifies candidates with elevated mutational load, using trinucleotide-adjusted local background (Lanzós et al. 2017). A clear impediment in both cases has been the lack of true-positive set of known lncRNA driver genes, analogous to CGC. Although there do exist databases of cancer lncRNAs, notably LncRNADisease (Chen et al. 2013) and Lnc2Cancer (Ning et al. 2016), they mix unfiltered data from numerous sources, resulting in inconsistent criteria for inclusion (including expression changes), and inconsistent gene identifiers.

To facilitate the future discovery of cancer lncRNAs, and gain insights into their biology, we have compiled a highly-curated set of cases with roles in cancer processes. Here we present the *Cancer LncRNA Census* (CLC), the first compendium of lncRNAs with direct functional or genetic evidence for cancer roles. We demonstrate the utility of CLC in assessing the performance of driver lncRNA predictions. Through analysis of this geneset, we demonstrate that cancer lncRNAs have a unique series of features that may in future be used to assist *de novo* predictions. Finally, we show that CLC genes have conserved cancer roles across the approximately 80 million years of evolution separating humans and rodents.

Results

Definition of cancer related lncRNAs

As part of recent efforts to identify driver lncRNAs within the Pancancer Analysis of Whole Genomes (PCAWG) project, we discovered the need for a high-confidence set of cancer-related lncRNA genes (CRL) (Lanzós et al. 2017), which we henceforth refer to as "cancer lncRNAs", and which we here present as Version 1 of the *Cancer LncRNA Census* (CLC).

Cancer lncRNAs were identified from the literature using defined and consistent criteria, being direct experimental or genetic evidence for roles in cancer progression or phenotypes (see Materials and Methods). Alterations in lncRNA expression alone were not considered sufficient evidence. Importantly, only lncRNAs with GENCODE identifiers were included. For every cancer lncRNA, one or more associated cancer types were collected.

Attesting to the value of this approach, we identified several cases in semi-automatically annotated cancer lncRNA databases of lncRNAs that were misassigned GENCODE identifiers, usually that of their overlapping protein-coding gene (Chen et al. 2013). We also excluded a number of published lncRNAs for which we could not find evidence to meet our criteria, for example CONCR, SRA1 and KCNQ1OT1 (Marchese et al. 2016; Lanz et al. 1999; Higashimoto et al. 2006).

Version 1 of CLC contains a total of 122 genes, or 0.76% of a total of 15,941 lncRNA gene loci annotated in GENCODE v24 (Derrien et al. 2012; Harrow et al. 2012) (Figure 1) (eight of the CLC genes are annotated as pseudogenes rather than lncRNAs by GENCODE). The entire remaining set of 15,827 lncRNA loci is henceforth referred to as "nonCLC" (Figure 1). The full CLC dataset is found in Supplementary Table 1. The CLC set is smaller than the *Lnc2Cancer* database (n= 667) (Ning et al. 2016), the *LncRNADisease* Database (n=121) (Chen et al. 2013) and *lncRNAdb* (n=191) (Quek et al. 2015). CLC covers between 17% and 31% of these databases (*Lnc2Cancer* and *LncRNADisease* respectively) but none of these resources contain the complete list of genes presented here (Figure 2). CLC is substantially larger than a previous "Cancer Related LncRNAs" set we produced (Lanzós et al. 2017)(Figure 2B). For comparison, the Cancer Gene Census (CGC) (COSMIC v78, downloaded Oct, 3, 2016) lists 561 or 2.8% of protein-coding genes (Futreal et al. 2004). Using the International Classification of Diseases for Oncology (World Health Organization 2013), we reassigned the cancer types described in the original research articles to a reduced set of 31(Figure 1, Supplementary Figure 1).

Altogether, CLC contains 337 unique lncRNA-cancer type relationships. Out of 122 genes, 77 (63.1%) were shown to function as oncogenes, 36 (29.5%) as tumour suppressors, and 9 (7.4%) with evidence for both activities (Figure 1 and Supplementary Figure 1).

The most prolific lncRNAs, with >15 recorded cancer types, are *HOTAIR*, *MALAT1*, *MEG3* and *H19* (Figure 1 and Supplementary Figure 1). It is not clear whether this reflects their unique pan-cancer functionality, or is simply a result of their being amongst the most early-discovered and widely-studied lncRNAs.

In vitro experiments were the most frequent evidence source, usually consisting of RNAi-mediated knockdown in cultured cell lines, coupled to phenotypic assays such as proliferation or migration (Supplementary Figure 1). Far fewer have been studied *in vivo*, or have cancer-associated somatic or germline mutations. 19 lncRNAs had 3 or more independent evidence sources (Supplementary Figure 1).

CLC for benchmarking lncRNA driver prediction methods

One of the primary motivations for CLC is as a true positive set for benchmarking and comparing methods for identifying driver lncRNAs. In the domain of protein-coding driver gene predictions, the Cancer Gene Census (CGC) has become the "gold standard" training set, against which new methods are tested (Futreal et al. 2004). Typically, the predicted driver genes belonging to CGC are judged to be true positives, and the fraction of these amongst predictions is used to estimate the Positive Predictive Value (PPV), or precision. This measure can be calculated for increasing cutoff levels, to assess the optimal cutoff.

First, we used CLC to examine the performance of the lncRNA driver predictor ExInAtor (Lanzós et al. 2017) in recalling CLC genes using PCAWG tumour mutation data (PCAWG Consortium, Manuscript In Preparation). A total of 2,687 GENCODE lncRNAs were included here, of which 82 (3.1%) belong to CLC. Driver predictions on several cancers at the standard False Discovery Rate ("*q*-value") cutoff of 0.1 are shown for selected cancers in Figure 3A. That panel shows the CLC-defined precision (*y*-axis) as a function of predicted driver genes ranked by *q*-value (*x*-axis). We observe rather heterogeneous performance across cancer cohorts. This may reflect a combination of intrinsic biological differences and differences in cohort sizes, which differs widely between the datasets shown. For the merged Pancancer dataset ("PANCAN"), ExInAtor predicted three CLC genes amongst its top ten candidates (*q*-value < 0.1), a rate far in excess of the background expectation. Similar enrichments are observed for other cancer types. These results support both the predictive value of ExInAtor, and the usefulness of CLC in assessing lncRNA driver predictors.

Comprehensive CLC-based assessments of lncRNA driver discovery, across all methods and tumour cohorts in PCAWG, may be found in the main PCAWG driver prediction publication (PCAWG Consortium, Manuscript In Preparation).

Finally, we assessed the precision (i.e. positive predictive value) of PCAWG lncRNA and protein-coding driver predictions across all cancers and all prediction methods (PCAWG Consortium, Manuscript In Preparation). Using the same *q*-value cutoff of 0.1, we found that across all cancer types and methods, a total of 8 (8.5%) of lncRNA predictions are in CLC (Figure 3B), while a total of 139 (23.1%) of protein-coding predictions are CGC genes (Figure 3C). In terms of sensitivity, 9.8% and 25.1% of CLC and CGC genes are predicted as candidates, respectively. Despite the lower detection of CLC genes in comparison to CGC genes, both sensitivity rates significantly exceed the prediction rate of nonCLC and nonCGC genes, again highlighting the usefulness of the CLC geneset.

CLC genes are distinguished by function- and disease-related features

We recently found evidence, using a smaller geneset than presented here, that cancer lncRNAs are distinguished by various genomic and expression features indicative of biological function (Lanzós et al. 2017). We here extended these findings using a large series of potential gene features, to search for those features distinguishing CLC from nonCLC lncRNAs (Figure 4A).

First, associations with expected cancer-related features were tested (Figure 4B). CLC genes are significantly more likely to have their transcription start site (TSS) within 100 kb of cancer-associated germline SNPs (“Cancer SNPs 100kb TSS”), and more likely to be either differentially-expressed or epigenetically-silenced in tumours (Yan et al. 2015) (Figure 4B). Intriguingly, we observed a tendency for CLC lncRNAs to be more likely to lie within 1 kb of known cancer protein-coding genes (“CGC 1kb TSS”) – this is explored in more detail below.

We next investigated the properties of the genes themselves. As seen in Figure 4C, and consistent with our previous findings (Lanzós et al. 2017), CLC genes (“Gene length”) and their spliced products (“Exonic length”) are significantly longer than average. No difference was observed in the ratio of exonic to total length (“Exonic content”), nor repetitive sequence coverage (“Repeats coverage”), nor GC content.

CLC genes also tend to have greater evidence of function, as inferred from evolutionary conservation. Base-level conservation at various evolutionary depths was calculated for lncRNA exons and promoters (Figure 4D). Across all measures tested, using either average base-level scores or percent coverage by conserved elements, we found that CLC genes’ exons are significantly more conserved than other lncRNAs (Figure 4D). The same was observed for conservation of promoter regions.

High levels of gene expression are known to correlate with lncRNA conservation, and are hypothesized to be a reflection of functionality (Managadze et al. 2011). We found that CLC genes have consistently higher steady-state expression levels, across both healthy organs and cultured cell lines (Figure 4E). This difference tended to be more marked for cell lines compared to organ samples. It is unclear whether this is a batch effect, since cell lines data comes from ENCODE (Djebali et al. 2012) and organ samples from Illumina Bodymap, or reflects a generalised upregulation of cancer lncRNA expression in cultured cells. It should be noted that these mainly comprise immortalised cancer cell lines but also includes lines such as IMR-90 that are neither immortal nor transformed.

Evidence for genomic clustering of non-coding and protein-coding cancer genes

In light of recent evidence for colocalisation and coexpression of disease-related lncRNAs and protein-coding genes (Tan et al. 2017), we were curious whether such an effect holds for cancer-related lncRNAs and protein-coding genes. We also asked, more specifically, whether CLC genes tend to be closer to CGC genes than expected by chance, and whether this is manifested in a more co-regulated expression.

Using TSS-TSS distances, we found that CLC genes on average tend to lie moderately closer to protein-coding genes of all types, compared to nonCLC lncRNAs (median 29 kb vs 38 kb respectively) (P=0.03, Wilcoxon test) (Supplementary Figure 2A). Interestingly, this effect is also observed when assessing specifically for distance to CGC genes (median 1,122 kb vs 1,599 kb, respectively) (P= 0.0004, Wilcoxon test) (Figure 5A).

It has been widely proposed that proximal lncRNA / protein-coding gene pairs are involved in *cis*-regulatory relationships, which is reflected in expression correlation (Ponjavic et al. 2009). We next asked whether proximal CLC-CGC pairs exhibit this behaviour. An important potential confounding factor is the known positive correlation between nearby gene pairs (Marques et al. 2013), and this must be controlled for. Using gene expression data across 11 human cell lines, we observed a positive correlation between CLC-CGC gene pairs for each cell type (Figure 5B). To control for the effect of proximity on correlation, we next randomly sampled a similar number of non-CLC lncRNAs with matched distances (TSS-TSS) from the same CGC genes, and found that this correlation was lost (Figure 5B, “nonCLC-CGC”). To further control for a possible correlation arising from the simple fact that both CGC and CLC genes are involved in cancer, we randomly shuffled the CLC-CGC pairs 1000 times, again observing no correlation (Figure 5B, “Shuffled CLC-CGC”). Together these results show that genomically-proximal protein-coding/non-coding gene pairs exhibit an expression correlation that exceeds that expected by chance, even when controlling for genomic distance.

These results prompted us to further explore the genomic localization of CLC genes relative to their proximal protein-coding gene and the nature of their neighbouring genes. First, considering gene pairs as proximal when both TSS lie within 10 kb, we find, in agreement with our previous results, that CLC genes are significantly more likely to lie near to a CGC gene, compared to other lncRNAs (6.6% vs 2.1%, respectively) (P=0.004, Fisher's exact test) (Supplementary Figure 2B). Next, we observed an unexpected difference in the genomic organisation of CLC genes: when classified by orientation with respect to nearest protein-coding gene (Derrien et al. 2012), we found a significant enrichment of CLC genes immediately downstream and on the same strand as protein-coding genes (“Samestrand, pc up”, Figure 5C). Moreover, CLC genes are approximately twice as likely to lie in an upstream, divergent orientation to a protein-coding gene (“Divergent”, Figure 5C). Of these CLC genes, 20% are divergent to a CGC gene, compared to 5% for nonCLC genes (P=0.018, Fisher’s exact test) (Figure 5D), and several are divergent to protein-coding genes that have also been linked or defined to be involved in cancer, despite not being classified as CGCs (Supplementary Table 2).

Given this noteworthy enrichment of CGC genes in a divergent configuration to protein-coding genes, we next inspected the latters’ function annotation. Examining their Gene Ontology (GO) terms, molecular pathways and other gene function related terms, we found this group of genes to be enriched in GO terms for “sequence-specific DNA binding”, “DNA binding”, “tube development” and “transcriptional misregulation in cancer” (Figure 5E). These results were confirmed by another, independent GO-analysis suite (see Materials and Methods). Interestingly, three out of the top four functional groups were observed previously in a study of protein-coding genes divergent to long upstream antisense transcripts (LUAT) in primary mouse tissues (Lepoivre et al. 2013).

Thus, CLC genes appear to be non-randomly distributed with respect to protein-coding genes, and particularly their CGC subset.

Evidence for anciently conserved cancer roles of lncRNAs

Numerous studies have employed unbiased forward genetic screens to identify genes that either inhibit or promote tumorigenesis in mice (Copeland & Jenkins 2010). These studies use engineered, randomly-integrating transposons, carrying bidirectional polyadenylation sites as well as strong promoters. Insertions, or clusters of insertions, called “common insertion sites” (CIS) that are identified in sequenced tumour DNA, implicate the overlapping or neighbouring gene locus as either an oncogene or tumour-suppressor gene. Although these studies have traditionally been focused on identifying protein-coding genes, they can in principle also identify non-coding driver loci.

We reasoned that comparison of mouse CISs to orthologous human regions could yield independent evidence for the functionality of human cancer lncRNAs (Figure 6A). To test this, we collected a comprehensive set of CISs in mouse (Abbott et al. 2015), consisting of 2,906 loci from 7 distinct cancer types (Supplementary Table 4). These sites were then mapped to in orthologous regions in the human genome, resulting in 1,309 human CISs, or hCISs. 7.3% of these CISs lie outside of protein-coding gene boundaries, which were used for the following analyses.

Mapping hCISs to lncRNA annotations, we discovered altogether eight CLC genes (6.6%) carrying an insertion within their gene span: *DLEU2*, *GAS5*, *MONC*, *NEAT1*, *PINT*, *PVT1*, *SLNCR1*, *XIST* (Table 1). In contrast, just 61 (0.4%) nonCLC genes contained hCISs (Figure 6B). A good example is *SLINCR1*, shown in Figure 6C, which drives invasiveness of human melanoma cells (Schmidt et al. 2016), and whose mouse orthologue contains a CIS discovered in pancreatic cancer.

This analysis would suggest that CLC genes are enriched for hCISs; however, there remains the possibility that this is confounded by their greater length. To account for this, we performed two separate validations. First, sets of nonCLC genes with CLC-matched length were randomly sampled, and the number of intersecting hCISs per unit gene length (Mb) was counted (Figure 6D). Second, CLC genes were randomly relocated in the genome, and the number of genes intersecting at least one hCIS was counted (Figure 6E). Both analyses showed that the number of intersecting hCISs per Mb of CLC gene span is far greater than expected. In contrast, nonCLC genes show a depletion for hCIS sites (Figure 6F).

Together these analyses demonstrate that CLC genes are orthologous to mouse cancer-causing genomic loci at a rate greater than expected by random chance. These identified cases, and possibly other CLC genes, display cancer functions that have been conserved over tens of millions of years since human-rodent divergence.

Discussion

We have presented the Cancer LncRNA Census, the first controlled set of GENCODE-annotated lncRNAs with demonstrated roles in tumorigenesis or cancer phenotypes.

The present state of knowledge of lncRNAs in cancer, and indeed lncRNAs generally, remains highly incomplete. Consequently, our aim was to create a geneset with the greatest possible confidence, by eliminating the relatively large number of published “cancer lncRNAs” with as-yet unproven causative roles in disease processes. Thus, we used a rather strict definition of cancer lncRNA, being any having direct experimental or genetic evidence for a causative role in cancer phenotypes. By this measure, gene expression changes alone do not suffice. By introducing these well-defined inclusion criteria, we hope to ensure that CLC contains the highest possible proportion of *bona fide* cancer genes, giving it maximum utility for *de novo* predictor benchmarking. In addition, its basis in GENCODE ensures portability across datasets and projects. Inevitably some well-known lncRNAs did not meet these criteria (including *SRA1*, *CONCR*, *KCNQ1OT1*) (Marchese et al. 2016; Lanz et al. 1999; Higashimoto et al. 2006); these may be included in future when more validation data becomes available. We believe that CLC will complement the established lncRNA databases such as *lncRNAdb*, *LncRNADisease* and *Lnc2Cancer*, which are more comprehensive, but are likely to have a higher false-positive rate due to their more relaxed inclusion criteria (Chen et al. 2013; Quek et al. 2015; Ning et al. 2016).

De novo lncRNA driver gene discovery is likely to become increasingly important as the number of sequenced tumours grow. The creation and refinement of statistical methods for driver gene discovery will depend on the available of high-quality true positive genesets such as CLC. It will be important to continue to maintain and improve the CLC in step with anticipated growth in publications on validated cancer lncRNAs. Very recently, CRISPR-based screens (Zhu et al. 2016; Liu et al. 2017) have catalogued large numbers of lncRNAs contributing to proliferation in cancer cell lines, which will be incorporated in future versions.

We used CLC to estimate the performance of *de novo* driver lncRNA predictions from the PCAWG project, made using the ExInAtor pipeline (Lanzós et al. 2017). Supporting the usefulness of this approach, we found a strong enrichment for CLC genes amongst the top-ranked driver predictions. Extending this to the full set of PCAWG driver predictors, approximately ten percent of CLC genes (9.8%) are called as drivers by at least one method (PCAWG Consortium, Manuscript In Preparation), which is lower to the rate of CGC genes identified (25.1%). This difference may be due to technical or biological factors. For example, CLC genes that are not identified as drivers by PCAWG may be false negatives, whose identification simply awaits greater statistical power afforded by more genomes; or else many CLC genes may contribute to cancer phenotypes, but are not targeted by somatic mutations during tumorigenesis. Addressing these questions will be a key objective in future studies.

Analysis of CLC geneset has broadened our understanding of the unique features of cancer lncRNAs, and generally supports the notion that lncRNAs have intrinsic biological functionality. Cancer lncRNAs are distinguished by a series of features that are consistent with both (a) roles in cancer (eg tumour expression changes), and (b) general biological functionality (eg high expression, evolutionary conservation). Elevated evolutionary conservation in the exons of CLC genes would appear to support their functionality as a mature RNA transcript, in contrast to the act of their transcription alone (Latos et al. 2012). Another intriguing observation has been the colocalisation of cancer lncRNAs with known protein-coding cancer genes: these are genomically proximal and exhibit elevated expression correlation, even after normalising for proximity. This points to a regulatory link between cancer lncRNAs and protein-coding genes, perhaps through chromatin looping, as described in previous reports for *CCAT1* and *MYC*, for example (Xiang et al. 2014).

One important caveat for all features discussed here is ascertainment bias: almost all lncRNAs discussed have been curated from published, single-gene studies. It is entirely possible that selection of genes for initial studies was highly non-random, and influenced by a number of factors – including high expression, evolutionary conservation and proximity to known cancer genes – that could bias our inference of lncRNA features. This may be the explanation for the observed excess of cancer lncRNAs in divergent configuration to protein-coding genes. However, the general validity of some of the CLC-specific features described here – including high expression and evolutionary conservation – were also observed recent unbiased genome-wide screens (Lanzós et al. 2017; Liu et al. 2017), suggesting that they are genuine.

In principle, lncRNAs may play two distinct roles in cancer: first, as “driver genes” whose mutations are early and positively-selected events in tumorigenesis; or second, as “downstream genes”, which do make a genuine contribution to cancer phenotypes, but through non-genetic alterations in cellular networks resulting from changes in expression, localisation or molecular interactions. These downstream genes may not be expected to display positively-selected mutational patterns, but would be expected to display cancer-specific alterations in expression. A key question for the future is how lncRNAs break down between these two categories, and whether this ratio is distinct from protein-coding genes. Despite the care taken to annotate CLC, it might be expected to contain both types of cancer lncRNA.

Despite the relatively low concordance of CLC genes with PCAWG driver predictions, the results of this study strongly support the value and key cancer role of identified lncRNAs in cancer. Most notably, the existence of a core set of eight lncRNAs with independently-identified mouse orthologues with similar cancer functions, is powerful evidence that these genes are *bona fide* cancer genes, whose overexpression or silencing can drive tumour formation. To our knowledge this is the most direct demonstration to date of anciently-conserved functions and disease roles for lncRNAs. It will be intriguing to investigate in future whether more human-mouse orthologous lncRNAs have been identified in such screens.

Materials and Methods

Manual Curation

All lncRNAs in lncRNAdb and those listed in Schmitt and Chang's recent review article were collected (Quek et al. 2015; Schmitt et al. 2016). To these were added all cases from *LncRNADisease* and *Lnc2Cancer* databases (Chen et al. 2013; Ning et al. 2016). This primary list formed the basis for a manual literature search: all available publications for each gene were identified by keyword search in Pubmed. If publications were found conforming to at least one of the inclusion criteria (below) and the gene has a GENCODE ID, then it was added to CLC, with appropriate information on the associated cancer, biological activity. For the numerous cases where no GENCODE ID was supplied in the original publication, any available ID, or primer or siRNA sequence was used to identify the gene using the UCSC Genome Browser Blat tool (Kent et al. 2002).

Inclusion criteria sufficient to define a cancer lncRNA and link it to a cancer type were:

1. Class t: *In vitro* demonstration that their knockdown and/or overexpression in cultured cancer cells results in changes to cancer-associated phenotypes. These typically include proliferation rates, migration, sensitivity to apoptosis, or anchorage-independent growth.

2. Class v: *In vivo* demonstration that their knockdown and/or overexpression in cancer cells alters their tumorigenicity when injected into animal models.

3. Class g: Germline mutations or variants that predispose humans to cancer.

4. Class s: Somatic mutations that show evidence for positive selection during tumour formation.

An additional criterion was allowed to link an lncRNA to a cancer type, only if one of the above criteria was already met for another cancer:

5. Class p: Prognosis: The lncRNAs expression is statistically linked to disease progression or response to treatment.

If an lncRNA was found to promote tumorigenesis or cancer phenotype, it was defined as "oncogene" (og). Conversely those found to inhibit such phenotypes were defined as "tumour suppressor" (tsg). Several lncRNAs were found to have both activities recorded in different cancer types, and were given both labels (og/tsg). For every lncRNA-cancer association, a single representative publication is recorded. Finally, it is important to note that no lncRNAs were included based on evidence from previous driver gene discovery studies of the types represented by OncodriveFML, ExInAtor, ncdDetect or others described in PCAWG (Mularoni et al. 2016; Lanzós et al. 2017; Juul et al. 2017) (PCAWG Consortium, Manuscript In Preparation).

CLC set at this stage relies on GENCODE v24 annotation, and therefore all CLC genes have a GENCODE v24 ID assigned. However, data relative to GENCODE v24 was not available for all types of data used in this study (ie. all data relative to PCAWG is based on GENCODE v19). Thus, for some analysis only genes also present in GENCODE v19 could be used (specified in the corresponding methods section) and the total number of genes analysed in these cases is slightly lower (107 instead of 122 CLC genes and 13503 instead of 15827 nonCLC).

LncRNA and protein-coding driver prediction analysis

LncRNA and protein-coding predictions for ExInAtor and the rest of PCAWG methods, as well as the combined list of drivers, were extracted from the consortium database (PCAWG Consortium, Manuscript In Preparation). Parameters and details about each individual methods and the combined list of drivers can be found on the main PCAWG driver publication (PCAWG Consortium, Manuscript In Preparation) and false discovery rate correction was applied on each individual cancer type for each individual method in order to define candidates (*q*-value cutoff of 0.1). To calculate sensitivity (percentage of true positives that are predicted as candidates) and precision (percentage of predicted candidates that are true positives) for lncRNA and protein-coding predictions we used the CLC and CGC (COSMIC v78, downloaded Oct, 3, 2016) sets, respectively. To assess the statistical significance of sensitivity rates, we used Fisher's exact test.

Feature Identification

We compiled several quantitative and qualitative traits of GENCODE lncRNAs and used them to compare CLC genes to the rest of lncRNAs (referred to as "nonCLC"). Analysis of quantitative traits were performed using Wilcoxon test while qualitative traits were tested using Fisher' exact test. These methods principally refer to Figure 4.

Sequence / gene properties: Exonic positions of each gene were defined as the projection of the union of exons from all its transcripts. Introns were defined as all remaining non-exonic nucleotides within the gene span. Repeats coverage refers to the percent of exonic nucleotides of a given gene overlapping repeats and low complexity DNA sequence regions obtained from RepeatMasker data housed in the UCSC Genome Browser (Tyner et al. 2017). Exonic content refers to the fraction of total gene span covered by exons. For this section we used GENCODE v19.

Evolutionary conservation: Two types of PhastCons conservation data were used: base-level scores and conserved elements. These data for different multispecies alignments (GRCh38/hg38) were downloaded from UCSC genome browser (Tyner et al. 2017). Mean scores and percent overlap by elements were calculated for exons and promoter regions (GENCODE v24). Promoters were defined as the 200nt region centred on the annotated gene start.

Expression: We used polyA+ RNA-seq data from 10 human cell lines produced by ENCODE (Djebali et al. 2012; ENCODE Project Consortium et al. 2012) and from various human tissues by the Illumina Human Body Map Project (HBM) (www.illumina.com; ArrayExpress ID: E-MTAB-513) (GENCODE v19).

Cancer SNPs: On October, 4, 2016, we collected all 2,192 SNPs related to “cancer”, “tumour” and “tumor” terms in the NHGRI-EBI Catalog of published genome-wide association studies (Hindorff et al. 2009; Welter et al. 2014) (https://www.ebi.ac.uk/gwas/home). Then we calculated the closest SNP to each lncRNA TSS using *closest* function from Bedtools v2.19 (GENCODE v24).

Epigenetically-silenced lncRNAs: We obtained a published list of 203 cancer-associated epigenetically-silenced lncRNA genes present in GENCODE v24 (Yan et al. 2015). These candidates were identified due to DNA methylation alterations in their promoter regions affecting their expression in several cancer types.

Differentially expressed in cancer: We collected a list of 3,533 differentially-expressed lncRNAs in cancer (Yan et al. 2015) (GENCODE v24).

Distance to protein-coding genes and CGC genes: For each lncRNA we calculated the TSS to TSS distance to the closest protein-coding gene (GENCODE v24) or CGC gene (downloaded on October, 3, 2016 from Cosmic database)(Futreal et al. 2004) using *closest* function from Bedtools v2.19.

Coexpression with closest CGC gene: We took CLC-CGC gene pairs whose TSS-TSS distance was <200kb. RNAseq data from 11 human cell lines from ENCODE was used to assess expression levels (Djebali et al. 2012; ENCODE Project Consortium et al. 2012). ENCODE RNAseq data were obtained from ENCODE Data Coordination Centre (DCC) in September 2016, https://www.encodeproject.org/matrix/?type=Experiment. All data is relative to GENCODE v24. We calculated the expression correlation of gene pairs within each of the 11 cell lines, using the Pearson measure. To control for the effect of proximity, we randomly sampled a subset of nonCLC-CGC pairs matching the same TSS-TSS distance distribution as above, and performed the same expression correlation analysis (“nonCLC-CGC”). Finally, to further control for the fact that CLC and CGC are both cancer genes, which may influence their expression correlation, we shuffled CLC-CGC pairs 1000 times, and tested expression correlation for each set (“Shuffled CLC-CGC”). Finally, we tested if correlation coefficients of the three groups were significantly different (Kolmogorov Smirnov test).

Genomic classification: We used an in-house script to classify lncRNA transcripts into different genomic categories based on their orientation and proximity to the closest protein-coding gene

(GENCODE v24): a 10 kb distance was used to distinguish "genic" from "intergenic" lncRNAs. When transcripts belonging to the same gene had different classifications, we used the category represented by the largest number of transcripts.

Functional enrichment analysis: The list of protein-coding genes (GENCODE v24) that are divergent and closer than 10 kb to CLC genes (or nonCLC) was used for a functional enrichment analysis (20 unique genes in the case of CLC analysis and 1202 in the case of nonCLC analysis). We show data obtained using g:Profiler web server (Reimand et al. 2016), g:GOSt, with default parameters for functional enrichment analysis of protein-coding genes divergent to CLC and using Bonferroni correction for protein-coding gene divergent to nonCLC. For CLC analysis we performed the same test with independent methods: Metascape (http://metascape.org) (Tripathi et al. 2015) and GeneOntoloy (Panther classification system)(Mi et al. 2013; Mi et al. 2017). In both cases similar results were found.

Mouse mutagenesis screen analysis

We extracted the genomic coordinates of transposon common insertion sites (CISs) in Mouse (GRCm38/mm10) http://ccgd-starrlab.oit.umn.edu/about.php (Abbott et al. 2015). This database contains target sites identified by transposon-based forward genetic screens in mice. LiftOver (Kent et al. 2002) was used at default settings to obtain aligned human genome coordinates (hCISs) (GRCh38/hg38). We discarded hCIS regions longer than 1000 nucleotides and intersecting protein coding genes, and intersected the remainders with the genomic coordinates CLC and nonCLC genes.

To correctly assess the statistical enrichment of CLC in hCIS regions, we performed two control analyses:

Randomly repositioning of CLC and nonCLC genes: We randomly relocated CLC/nonCLC genes 10,000 times within the whole genome using the tool *shuffle* from BedTools v19. In each iteration, we calculated the number of genes that intersected at least one hCIS, and created the distribution of these simulated values. Finally, we calculated an empirical *p*-value by counting how many of the simulated values were higher or equal than the real values. This analysis was performed separately for CLC and nonCLC genes.

Length-matched sampling: To calculate if the enrichment of hCIS intersecting genes in CLC set is higher and statistically different from nonCLC set, while controlling by gene length, we created 1000 samples of nonCLC genes with the same gene length distribution as CLC genes. Each sample was intersected with hCIS, and the number of intersecting hCISs per Mb of gene length was calculated. To create the nonCLC samples we used the *matchDistribution* script: https://github.com/julienlag/matchDistribution.

Acknowledgements

We wish to thanks Julien Lagarde (CRG) for help and advice in bioinformatic analysis. We acknowledge Romina Garrido (CRG), Deborah Re (DKF), Silvia Roesselet (DKF) and Marianne Zahn (Inselspital) for administrative support. We thank Ivo Buchhalter (DKFZ) and Sandra Koser (DKFZ) for preprocessing the SNV and expression data for the integrated analysis. Iñigo Martincorena (Sanger Institute) kindly provided the script for analysing driver prediction sensitivity. A.L. is supported by pre-doctoral fellowship FPU14/03371. This research was supported by the NCCR RNA & Disease funded by the Swiss National Science Foundation.

Contributions

RJ conceived the project, performed manual annotation of CLC, and supervised with advice and suggestions of JS-P, RG, LF and CH. JCF and AL performed the feature analysis and evolutionary analysis. AL performed mutation analysis, with help from IM. RJ, AL and JCF drafted the manuscript and prepared the figures and supplementary material. All authors read and approved the final draft.

Figure Legends

Figure 1: Overview of the Cancer LncRNA Census. Rows represent the 122 CLC genes, columns represent 31 cancer types. Blue cells indicate evidence for the involvement of a given lncRNA in that cancer type. Left column indicates functional classification: tumour suppressor (TSG), oncogene (OG) or both (OG/TSG). Above and to the right, barplots indicate the count totals of each column / row. The piechart shows the fraction that CLC within GENCODE v24 lncRNAs. Note that 8 CLC genes are classified as "pseudogenes" by GENCODE. "nonCLC" refers to all other GENCODE-annotated lncRNAs, which are used as background in comparative analyses.

Figure 2: Intersection of CLC with public databases. (A) Shown are the total number of unique human lncRNAs contained in each of the three indicated databases (note that for LncRNADisease, numbers refer only to cancer-related genes) and in the CRL set (cancer-related lncRNA genes defined by Lanzós et al. 2017). Colours represent the number of genes that: are present in CLC (green); belong to GENCODE v24 annotation but not CLC (pink); do not belong to GENCODE annotation, and therefore wouldn't meet the criteria for CLC (red). (B) Barplot shows the inverse overlap, being the proportion of CLC genes present in the other databases.

Figure 3: CLC as benchmark for cancer driver predictions. (A) CLC benchmarking of ExInAtor driver lncRNA predictions using PCAWG whole genome tumours at *q*-value (false discovery rate) cutoff of 0.1. Genes sorted increasingly by *q*-value are ranked on *x* axis. Percentage of CLC genes amongst cumulative set of predicted genes at each step of the ranking (precision), are shown on the *y* axis. Black line shows the baseline, being the percentage of CLC genes in the whole list of genes tested. Coloured dots represent the number of candidates predicted under the *q*-value cutoff of 0.1. "n" in the legend shows the number of CLC and total candidates for each cancer type. (B) Rate of driver gene predictions amongst CLC and nonCLC genesets (*q*-value cutoff of 0.1) by all the individual methods and the combined list of drivers developed in PCAWG. *p*-value is calculated using Fisher's exact test for the difference between CLC and nonCLC genesets. (C) Rate of driver gene predictions amongst CGC and nonCGC genesets (*q*-value cutoff of 0.1) by all the individual methods and the combined list of drivers developed in PCAWG. *p*-value is calculated using Fisher's exact test for the difference between CGC and nonCGC genesets.

Figure 4: Distinguishing features of CLC genes. (A) Panel showing a hypothetic feature analysis example to illustrate the content of the following figures. All panels in figure 4 display features (dots), plotted by their log fold difference (Odds Ratio in case of panel "B") between CLC / nonCLC genesets (*y*-axis) and statistical significance (*x*-axis). In all plots dark and light green dashed lines indicate 0.05 and 0.01 significance thresholds, respectively. *(*B) Cancer-related data from indicated sources: *y*-axis shows the log2 of the Odds Ratio obtained by comparing CLC to nonCLC

by Fisher's exact test; *x*-axis displays the estimated *p*-value from the same test. "CGC 1 kb TSS" refers to the fraction of genes that have a nearby known CGC cancer protein-coding gene. This is explored in more detail in the next Figure. (C) Sequence and gene properties: *y*-axis shows the log2 fold-difference of CLC / nonCLC means; *x*-axis represents the *p*-value obtained. (D) Evolutionary conservation: "Phastc mean" indicates average base-level PhastCons score; "Elements" indicates percent coverage by PhastCons conserved elements (see Materials and Methods). Colours distinguish exons (blue) and promoters (purple). (E) Gene expression: cell lines (blue) and tissues (purple). For B-D, statistical significance was calculated using Wilcoxon test.

Figure 5: Evidence for genomic clustering of non-coding and protein-coding cancer genes. (A) Cumulative distribution of the genomic distance of lncRNA transcription start site (TSS) to the closest Cancer Gene Census (CGC) (protein-coding) gene TSS. *p*-value was calculated using Wilcoxon test. (B) Boxplot shows the distribution of the gene expression correlation between CLC and their closest CGC genes in 11 human cell lines, including two control analyses (distance-matched nonCLC-CGC pairs, and shuffled CLC-CGC pairs). Correlation was calculated for gene pairs within each cell type, using Pearson method. *p*-value for Kolmogorov-Smirnov test is shown. (C) Genomic classification of lncRNAs. Genes are classified according to distance and orientation to the closest protein-coding gene, and these are grouped into three categories: genes closer than 10kb to closest protein-coding gene, genes overlapping a protein-coding gene and intergenic genes (>10kb from closest protein-coding gene). *p*-values for Fisher's exact tests are shown. (D) The percentage of divergent CLC (left bar) and nonCLC (right bar) genes divergent to a cancer protein-coding gene (CGC). Numbers represent numbers of genes with which the percentage is calculated. *p*-value for Fisher's exact test is shown. (E) Functional annotations of the 20 protein-coding genes divergent to CLC genes from Panel C. Bars indicate the –log10 (corrected) *p*-value (see Materials and Methods) and are coloured based on the "enrichment": the number of genes that contain the functional term divided by the total number of queried genes. Numbers at the end of the bars correspond to the number of genes that fall into the category.

Figure 6: Evidence for ancient conserved cancer roles of lncRNAs. (A) Functional conservation of human CLC genes was inferred by the presence of Common Insertion Sites (CIS), identified in transposon mutagenesis screens, at orthologous regions in the mouse genome. Orthology was inferred from Chain alignments and identified using LiftOver utility. (B) Number of CLC and nonCLC genes that contain human orthologous common insertion sites (hCIS) (see Table 1). Significance was calculated using Fisher's exact test. (C) UCSC browser screenshot of a CLC gene (*SLNCR1*, ENSG00000227036) intersecting a CIS (yellow arrow). (D) Distribution of the number of intersecting hCIS per Megabase (Mb) of total gene length, for 1000 subsets of nonCLC genes with

same length distribution as CLC genes (grey). Vertical blue line represents the overall value for CLC geneset: 1.42 hCIS sites per Mb of gene span. (E) Distribution of the number of genes overlapping a hCIS after 10,000 genomic randomizations of CLC genes (blue). Vertical black line represents the observed number of CLC genes (8) that intersect a hCIS. (F) Distribution of the number of intersecting genes with a hCIS after 10,000 genomic randomizations of nonCLC genes (grey). Vertical black line represents the observed number of nonCLC genes that intersect a hCIS (64).

Supplementary Figure Legends

Supplementary Figure 1: CLC summary statistics. (A) Barplot showing the non-redundant number of genes in CLC broken down by supporting evidence types. p: prognostic; t: in vitro; v: in vivo; g: germline mutations; s: somatic mutations. (B) Similar as previous, but with (redundant) number of genes per individual evidence type. (C) Histogram of genes broken down by their number of associated cancer types. (D) Histogram of cancer types, by their (redundant) number of associated lncRNAs.

Supplementary Figure 2: CLC lncRNAs tend to be closer to protein-coding genes. (A) Cumulative distribution of the genomic distance from CLC and nonCLC genes, to the closest protein-coding gene (NB this may be a CGC gene or not). Distances are defined as the distance of the annotated transcription start site (TSS) of each gene in the pair. *p*-value for Wilcoxon test is shown. (B) Number of CLC and nonCLC genes that are proximal to the nearest CGC protein-coding gene. Proximity is defined as a TSS-TSS distance of <10 kb. Significance was calculated using Fisher's exact test.

Supplementary Table Legends:

Supplementary Table 1: full CLC set.

Supplementary Table 2: CLC – protein-coding pairs.

Supplementary Table 3: GO analysis for protein coding genes divergent to nonCLC genes.

Supplementary Table 4: Counts of mouse CIS per cancer type.

References

Abbott, K.L. et al., 2015. The candidate cancer gene database: A database of cancer driver genes from forward genetic screens in mice. *Nucleic Acids Research*, 43(D1), pp.D844–D848.

Cabili, M.N. et al., 2011. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes & development*, 25(18), pp.1915–27. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21890647 [Accessed March 7, 2017].

Chang, K. et al., 2013. The Cancer Genome Atlas Pan-Cancer analysis project. *Nature genetics*, 45(10), pp.1113–1120. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24071849.

Chen, G. et al., 2013. LncRNADisease: a database for long-non-coding RNA-associated diseases. *Nucleic acids research*, 41(Database issue), pp.D983-6. Available at: http://www.ncbi.nlm.nih.gov/pubmed/23175614 [Accessed March 2, 2017].

Copeland, N.G. & Jenkins, N.A., 2010. Harnessing transposons for cancer gene discovery. *Nature reviews. Cancer*, 10(10), pp.696–706. Available at: http://www.nature.com/doifinder/10.1038/nrc2916 [Accessed March 7, 2017].

Derrien, T. et al., 2012. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome research*, 22(9), pp.1775–89. Available at: http://genome.cshlp.org/content/22/9/1775.long [Accessed May 23, 2014].

Djebali, S. et al., 2012. Landscape of transcription in human cells. *Nature*, 489(7414), pp.101–108.

ENCODE Project Consortium, T. et al., 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 488.

Engreitz, J.M. et al., 2014. RNA-RNA interactions enable specific targeting of noncoding RNAs to nascent Pre-mRNAs and chromatin sites. *Cell*, 159(1), pp.188–99. Available at: http://www.ncbi.nlm.nih.gov/pubmed/25259926 [Accessed March 2, 2017].

Furney, S. et al., 2006. Structural and functional properties of genes involved in human cancer. *BMC Genomics*, 7(1), p.3. Available at: http://www.ncbi.nlm.nih.gov/pubmed/16405732 [Accessed February 24, 2017].

Furney, S.J. et al., 2008. Distinct patterns in the regulation and evolution of human cancer genes. *In Silico Biol*, 8(December 2007), pp.33–46.

Futreal, P. et al., 2004. A census of human cancer genes. *Nat Rev Cancer*, 4(3), pp.177–183.

Grote, P. et al., 2013. *The Tissue-Specific lncRNA Fendrr Is an Essential Regulator of Heart and Body Wall Development in the Mouse*, Available at: http://www.sciencedirect.com/science/article/pii/S1534580712005862 [Accessed April 25, 2017].

Gutschner, T. et al., 2013. The noncoding RNA MALAT1 is a critical regulator of the metastasis phenotype of lung cancer cells. *Cancer research*, 73(3), pp.1180–9. Available at: http://www.ncbi.nlm.nih.gov/pubmed/23243023 [Accessed March 2, 2017].

Gutschner, T. & Diederichs, S., 2012. The hallmarks of cancer: a long non-coding RNA point of view. *RNA biology*, 9(6), pp.703–19. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22664915 [Accessed June 28, 2016].

Guttman, M. et al., 2009. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature*, 458(7235), pp.223–7. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19182780 [Accessed March 7, 2017].

Guttman, M. & Rinn, J.L., 2012. Modular regulatory principles of large non-coding RNAs. *Nature*, 482(7385), pp.339–46. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22337053 [Accessed March 2, 2017].

Harrow, J. et al., 2012. GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Research*, 22(9), pp.1760–1774. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22955987 [Accessed April 25, 2017].

Higashimoto, K. et al., 2006. Imprinting disruption of the CDKN1C/KCNQ1OT1 domain: the molecular mechanisms causing Beckwith-Wiedemann syndrome and cancer. *Cytogenetic and genome research*, 113(1–4), pp.306–12. Available at: http://www.ncbi.nlm.nih.gov/pubmed/16575194 [Accessed April 27, 2017].

Hindorff, L. a et al., 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences of the United States of America*, 106(23), pp.9362–7. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19474294.

Huarte, M. et al., 2010. A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response. *Cell*, 142(3), pp.409–19. Available at: http://www.ncbi.nlm.nih.gov/pubmed/20673990 [Accessed February 25, 2017].

Iyer, M.K. et al., 2015. The landscape of long noncoding RNAs in the human transcriptome. *Nature GeNetics*, 47(3). Available at: https://www.nature.com/ng/journal/v47/n3/pdf/ng.3192.pdf [Accessed April 19, 2017].

Jia, H. et al., 2010. Genome-wide computational identification and manual annotation of human long noncoding RNA genes. *RNA (New York, N.Y.)*, 16(8), pp.1478–87. Available at: http://rnajournal.cshlp.org/cgi/doi/10.1261/rna.1951310 [Accessed March 7, 2017].

Johnson, R. & Guigó, R., 2014. The RIDL hypothesis: transposable elements as functional domains of long noncoding RNAs. *RNA (New York, N.Y.)*, 20(7), pp.959–76. Available at:

http://www.ncbi.nlm.nih.gov/pubmed/24850885 [Accessed March 2, 2017].

Juul, M. et al., 2017. Non-coding cancer driver candidates identified with a sample- and position-specific model of the somatic mutation rate. *eLife*, 6. Available at: http://www.ncbi.nlm.nih.gov/pubmed/28362259 [Accessed April 27, 2017].

Kent, W.J. et al., 2002. The human genome browser at UCSC. *Genome research*, 12(6), pp.996–1006. Available at: http://www.ncbi.nlm.nih.gov/pubmed/12045153 [Accessed April 20, 2017].

Lanz, R.B. et al., 1999. A steroid receptor coactivator, SRA, functions as an RNA and is present in an SRC-1 complex. *Cell*, 97(1), pp.17–27. Available at: http://www.ncbi.nlm.nih.gov/pubmed/10199399 [Accessed April 25, 2017].

Lanzós, A. et al., 2017. Discovery of Cancer Driver Long Noncoding RNAs across 1112 Tumour Genomes: New Candidates and Distinguishing Features. *Scientific Reports*, 7, p.41544. Available at: http://www.ncbi.nlm.nih.gov/pubmed/28128360 [Accessed February 24, 2017].

Latos, P.A. et al., 2012. Airn Transcriptional Overlap, But Not Its lncRNA Products, Induces Imprinted Igf2r Silencing. *Science*, 338(6113), pp.1469–1472. Available at: http://www.sciencemag.org/cgi/doi/10.1126/science.1228110 [Accessed April 25, 2017].

Lawrence, M.S. et al., 2014. Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature*, 505(7484), pp.495–501. Available at: http://www.nature.com/doifinder/10.1038/nature12912 [Accessed January 13, 2017].

Lepoivre, C. et al., 2013. Divergent transcription is associated with promoters of transcriptional regulators. *BMC Genomics*, 14(1), p.914. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24365181 [Accessed April 25, 2017].

Liu, S.J. et al., 2017. CRISPRi-based genome-scale identification of functional long noncoding RNA loci in human cells. *Science*, 355(6320), p.eaah7111. Available at: http://www.sciencemag.org/lookup/doi/10.1126/science.aah7111 [Accessed April 25, 2017].

Managadze, D. et al., 2011. Negative Correlation between Expression Level and Evolutionary Rate of Long Intergenic Noncoding RNAs. *Genome Biology and Evolution*, 3(0), pp.1390–1404. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22071789 [Accessed March 26, 2017].

Marchese, F.P. et al., 2016. A Long Noncoding RNA Regulates Sister Chromatid Cohesion. *Molecular Cell*, 63(3), pp.397–407. Available at: http://www.ncbi.nlm.nih.gov/pubmed/27477908 [Accessed April 25, 2017].

Marques, A.C. et al., 2013. Chromatin signatures at transcriptional start sites separate two equally populated yet distinct classes of intergenic long noncoding RNAs. *Genome biology*, 14(11), p.R131. Available at: http://genomebiology.biomedcentral.com/articles/10.1186/gb-2013-14-

11-r131 [Accessed March 20, 2017].

Mi, H. et al., 2013. Large-scale gene function analysis with the PANTHER classification system. *Nature Protocols*, 8(8), pp.1551–1566. Available at: http://www.ncbi.nlm.nih.gov/pubmed/23868073 [Accessed April 23, 2017].

Mi, H. et al., 2017. PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. *Nucleic Acids Research*, 45(D1), pp.D183–D189. Available at: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkw1138 [Accessed April 23, 2017].

Mularoni, L. et al., 2016. OncodriveFML: a general framework to identify coding and non-coding regions with cancer driver mutations. *Genome biology*, 17(1), p.128. Available at: http://www.ncbi.nlm.nih.gov/pubmed/27311963 [Accessed June 28, 2016].

Ning, S. et al., 2016. Lnc2Cancer: a manually curated database of experimentally supported lncRNAs associated with various human cancers. *Nucleic acids research*, 44(D1), pp.D980-5. Available at: http://www.ncbi.nlm.nih.gov/pubmed/26481356 [Accessed March 2, 2017].

Ponjavic, J. et al., 2009. Genomic and transcriptional co-localization of protein-coding and long non-coding RNA pairs in the developing brain. Y. Hayashizaki, ed. *PLoS genetics*, 5(8), p.e1000617. Available at: http://dx.plos.org/10.1371/journal.pgen.1000617 [Accessed March 20, 2017].

Quek, X.C. et al., 2015. lncRNAdb v2.0: expanding the reference database for functional long noncoding RNAs. *Nucleic acids research*, 43(Database issue), pp.D168-73. Available at: http://www.ncbi.nlm.nih.gov/pubmed/25332394 [Accessed March 2, 2017].

Redon, R. et al., 2006. Global variation in copy number in the human genome. *Nature*, 444(7118), pp.444–54. Available at: http://dx.doi.org/10.1038/nature05329.

Reimand, U. et al., 2016. g:Profiler—a web server for functional interpretation of gene lists (2016 update). *Nucleic Acids Research*. Available at: http://biit.cs.ut.ee/gprofiler/doc/papers/gprofiler_nar_2016.pdf [Accessed April 20, 2017].

Sabarinathan, R. et al., 2013. RNAsnp: Efficient Detection of Local RNA Secondary Structure Changes Induced by SNPs. *Human Mutation*, 34(4), p.n/a-n/a. Available at: http://doi.wiley.com/10.1002/humu.22273 [Accessed April 19, 2017].

Sauvageau, M. et al., 2013. Multiple knockout mouse models reveal lincRNAs are required for life and brain development. *eLife*, 2, p.e01749. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24381249 [Accessed April 25, 2017].

Schmidt, K. et al., 2016. The lncRNA SLNCR1 Mediates Melanoma Invasion through a Conserved SRA1-like Region. *Cell reports*, 15(9), pp.2025–37. Available at:

http://linkinghub.elsevier.com/retrieve/pii/S2211124716304314 [Accessed March 20, 2017].

Schmitt, A.M. et al., 2016. Long Noncoding RNAs in Cancer Pathways. *Cancer Cell*, 29(4), pp.452–463. Available at: http://www.ncbi.nlm.nih.gov/pubmed/27070700 [Accessed April 20, 2017].

Sjoblom, T. et al., 2006. The Consensus Coding Sequences of Human Breast and Colorectal Cancers. *Science*, 314(5797), pp.268–274. Available at: http://www.ncbi.nlm.nih.gov/pubmed/16959974 [Accessed February 24, 2017].

Tamborero, D. et al., 2013. Comprehensive identification of mutational cancer driver genes across 12 tumor types. *Scientific reports*, 3, p.2650. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24084849 [Accessed June 28, 2016].

Tan, J.Y. et al., 2017. cis -Acting Complex-Trait-Associated lincRNA Expression Correlates with Modulation of Chromosomal Architecture. *Cell Reports*, 18(9), pp.2280–2288. Available at: http://www.ncbi.nlm.nih.gov/pubmed/28249171 [Accessed April 25, 2017].

Tripathi, S. et al., 2015. Meta- and Orthogonal Integration of Influenza &quot;OMICs&quot; Data Defines a Role for UBR4 in Virus Budding. *Cell host & microbe*, 18(6), pp.723–35. Available at: http://linkinghub.elsevier.com/retrieve/pii/S1931312815004564 [Accessed April 23, 2017].

Tyner, C. et al., 2017. The UCSC Genome Browser database: 2017 update. *Nucleic Acids Research*, 45. Available at: https://oup.silverchair-cdn.com/oup/backfile/Content_public/Journal/nar/45/D1/10.1093_nar_gkw1134/3/gkw1134.pdf?Expires=1493043886&Signature=VkpgKibgiT9kty13UqDbQabocXaBUaciIshI~KU4D010dRoa-9n7qLF0kT3eT2HZRiqrI7W74jyxPg1eyKhuPrlOzHFwCJWxa3tO3wh95deEodxSyEwvjA64rFAZFbmv0EtdUoWRqL5nhsJqLSPCZmPXukDpSxBH7SrrmYX33UqRcVo6jq-ICvde4XmPNDgacH4BwRTU2K0~D-OeV~kzq6s-zshWmPjUIJM-XCB7Mpx2kd5JVIN7lPVCt0vs8gK~BDHHdryJEkWLf9L4ZbFxzvLvtulvQUZnrNSNZFLRPPDHFnJ5C7YZM8-U5g27ma2eGbwIQyjZ1qqbcxOg6QLkLw__&Key-Pair-Id=APKAIUCZBIA4LVPAVW3Q [Accessed April 20, 2017].

Ulitsky, I. & Bartel, D.P., 2013. lincRNAs: Genomics, Evolution, and Mechanisms. *Cell*, 154(1), pp.26–46. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0092867413007599 [Accessed November 20, 2016].

Welter, D. et al., 2014. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Research*, 42(D1), pp.1001–1006.

World Health Organization, 2013. *International Classification of Diseases for Oncology (ICD-O). Third Edition. First Revision.*, Xiang, J.-F. et al., 2014. Human colorectal cancer-specific CCAT1-L lncRNA regulates long-range chromatin interactions at the MYC locus. *Cell Research*, 24(5), pp.513–531. Available at: http://www.nature.com/doifinder/10.1038/cr.2014.35 [Accessed April 25, 2017].

Yan, X. et al., 2015. Comprehensive Genomic Characterization of Long Non-coding RNAs across Human Cancers. *Cancer Cell*.

Yates, L.R. & Campbell, P.J., 2012. Evolution of the cancer genome. *Nature Reviews Genetics*, 13(11), pp.795–806. Available at: http://www.nature.com/doifinder/10.1038/nrg3317 [Accessed April 25, 2017].

Zhu, S. et al., 2016. Genome-scale deletion screening of human long non-coding RNAs using a paired-guide RNA CRISPR–Cas9 library. *Nature Biotechnology*, 34(12), pp.1279–1286. Available at: http://www.nature.com/doifinder/10.1038/nbt.3715 [Accessed December 21, 2016].

Tables

Table 1: List of intergenic CIS human (GRCh38) / mouse (GRCm38) gene pairs.

Human CLC Name	Human CLC ID	Chr Human	Start Human	End Human	Chr Mouse	Start Mouse	End Mouse	PubMed ID	Cancer Type Mouse
DLEU2	ENSG00000231607	chr13	50,048,971	50,049,063	chr14	61,631,880	61,631,972	24316982	Liver
DLEU2	ENSG00000231607	chr13	50,049,117	50,049,206	chr14	61,632,026	61,632,110	24316982	Liver
GAS5	ENSG00000234741	chr1	173,864,370	173,864,435	chr1	161,038,091	161,038,156	25961939	Sarcoma
MONC	ENSG00000215386	chr21	16,539,096	16,539,161	chr16	77,598,935	77,599,000	23685747	Nervous System
MONC	ENSG00000215386	chr21	16,561,654	16,561,655	chr16	77,616,439	77,616,440	24316982	Liver
NEAT1	ENSG00000245532	chr11	65,444,511	65,444,512	chr19	5,825,497	5,825,498	24316982	Liver
PINT	ENSG00000231721	chr7	131,049,455	131,049,456	chr6	31,179,149	31,179,150	22699621	Pancreatic
PVT1	ENSG00000249859	chr8	128,007,970	128,007,971	chr15	62,186,646	62,186,647	22699621	Pancreatic
SLNCR1	ENSG00000227036	chr17	72,507,275	72,507,276	chr11	113,137,613	113,137,614	22699621	Pancreatic
XIST	ENSG00000229807	chrX	73,841,539	73,841,540	chrX	103,473,862	103,473,863	24316982	Liver
XIST	ENSG00000229807	chrX	73,841,539	73,841,540	chrX	103,473,862	103,473,863	24316982	Liver

Figure 1

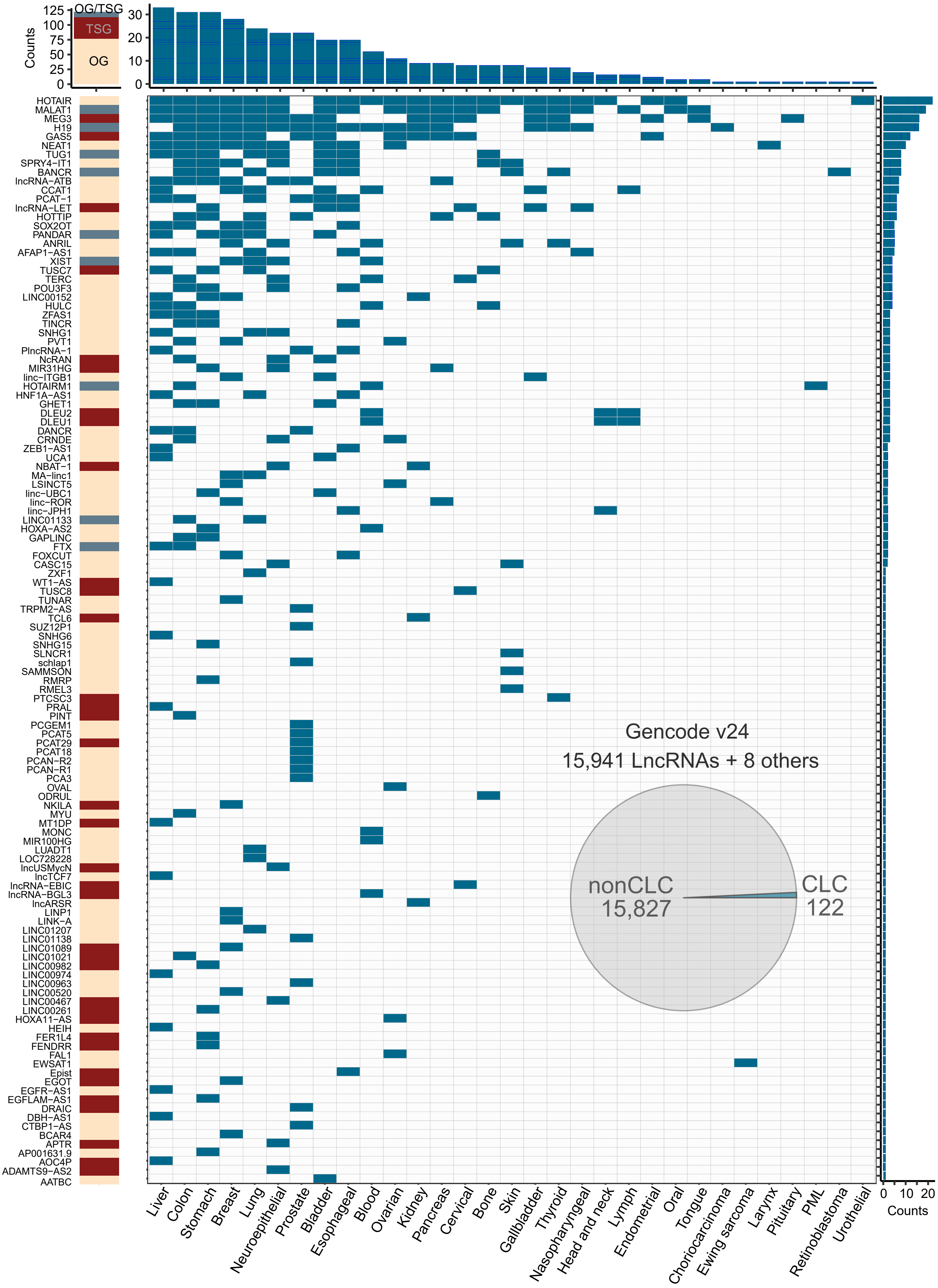

Counts
125
100
75
50
25
0
OG/TSG
TSG
OG
30
20
10
0
HOTAIR
MALAT1
MEG3
H19
GAS5
NEAT1
TUG1
SPRY4−IT1
BANCR
lncRNA−ATB
CCAT1
PCAT−1
lncRNA−LET
HOTTIP
SOX2OT
PANDAR
ANRIL
AFAP1−AS1
XIST
TUSC7
TERC
POU3F3
LINC00152
HULC
ZFAS1
TINCR
SNHG1
PVT1
PlncRNA−1
NcRAN
MIR31HG
linc−ITGB1
HOTAIRM1
HNF1A−AS1
GHET1
DLEU2
DLEU1
DANCR
CRNDE
ZEB1−AS1
UCA1
NBAT−1
MA−linc1
LSINCT5
linc−UBC1
linc−ROR
linc−JPH1
LINC01133
HOXA−AS2
GAPLINC
FTX
FOXCUT
CASC15
ZXF1
WT1−AS
TUSC8
TUNAR
TRPM2−AS
TCL6
SUZ12P1
SNHG6
SNHG15
SLNCR1
schlap1
SAMMSON
RMRP
RMEL3
PTCSC3
PRAL
PINT
PCGEM1
PCAT5
PCAT29
PCAT18
PCAN−R2
PCAN−R1
PCA3
OVAL
ODRUL
NKILA
MYU
MT1DP
MONC
MIR100HG
LUADT1
LOC728228
lncUSMycN
lncTCF7
lncRNA−EBIC
lncRNA−BGL3
lncARSR
LINP1
LINK−A
LINC01207
LINC01138
LINC01089
LINC01021
LINC00982
LINC00974
LINC00963
LINC00520
LINC00467
LINC00261
HOXA11−AS
HEIH
FER1L4
FENDRR
FAL1
EWSAT1
Epist
EGOT
EGFR−AS1
EGFLAM−AS1
DRAIC
DBH−AS1
CTBP1−AS
BCAR4
APTR
AP001631.9
AOC4P
ADAMTS9−AS2
AATBC
Liver
Colon
Stomach
Breast
Lung
Neuroepithelial
Prostate
Bladder
Esophageal
Blood
Ovarian
Kidney
Pancreas
Cervical
Bone
Skin
Gallbladder
Thyroid
Nasopharyngeal
Head and neck
Lymph
Endometrial
Oral
Tongue
Choriocarcinoma
Ewing sarcoma
Larynx
Pituitary
PML
Retinoblastoma
Urothelial
0
10
20
Counts
Gencode v24
15,941 LncRNAs + 8 others
nonCLC
15,827
CLC
122

Figure 2

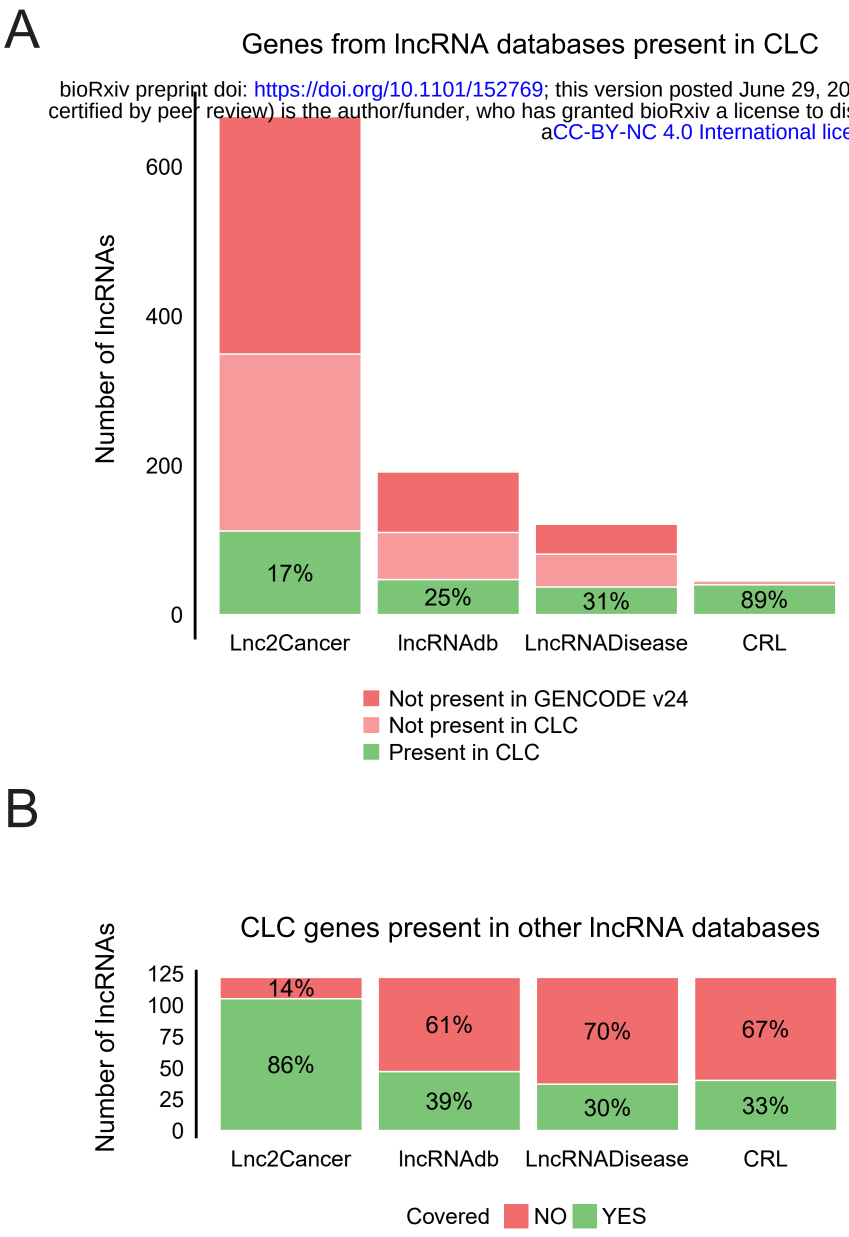

A
Genes from lncRNA databases present in CLC
Number of lncRNAs
600
400
200
0
17%
25%
31%
89%
Lnc2Cancer
lncRNAdb
LncRNADisease
CRL
🟥 Not present in GENCODE v24
🟥 Not present in CLC
🟩 Present in CLC
B
CLC genes present in other lncRNA databases
Number of lncRNAs
125
100
75
50
25
0
14%
86%
61%
39%
70%
30%
67%
33%
Lnc2Cancer
lncRNAdb
LncRNADisease
CRL
Covered 🟥 NO 🟩 YES

Figure 3

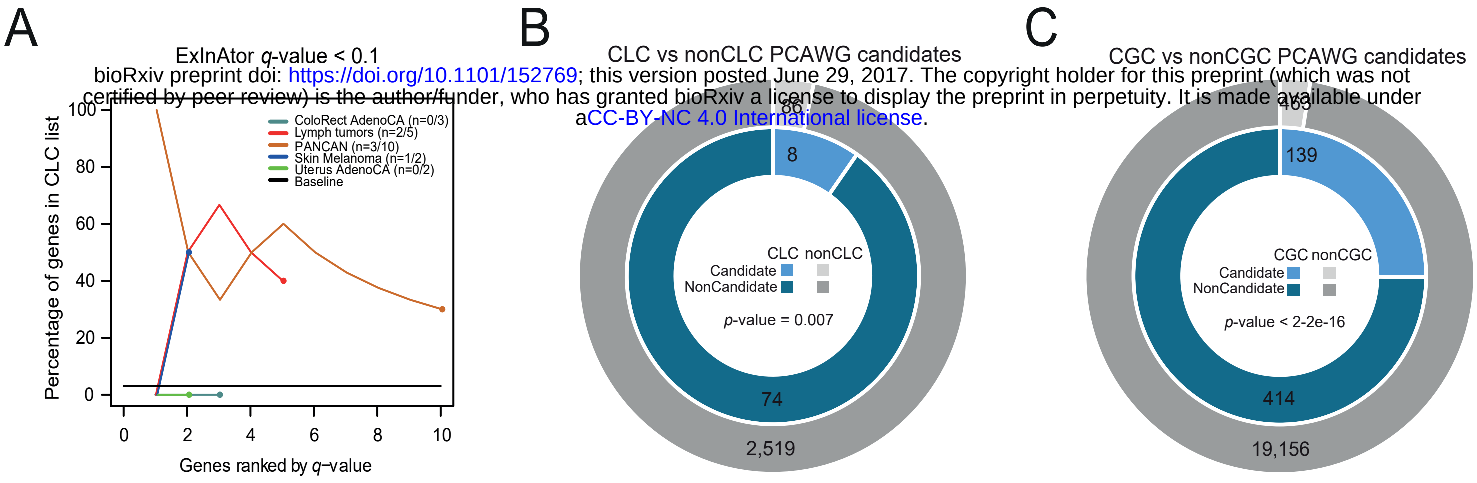

A
ExInAtor q-value < 0.1
Percentage of genes in CLC list
100
80
60
40
20
0
0
2
4
6
8
10
Genes ranked by q−value
ColoRect AdenoCA (n=0/3)
Lymph tumors (n=2/5)
PANCAN (n=3/10)
Skin Melanoma (n=1/2)
Uterus AdenoCA (n=0/2)
Baseline
B
CLC vs nonCLC PCAWG candidates
86
8
CLC nonCLC
Candidate
NonCandidate
p-value = 0.007
74
2,519
C
CGC vs nonCGC PCAWG candidates
463
139
CGC nonCGC
Candidate
NonCandidate
p-value < 2-2e-16
414
19,156

Figure 4

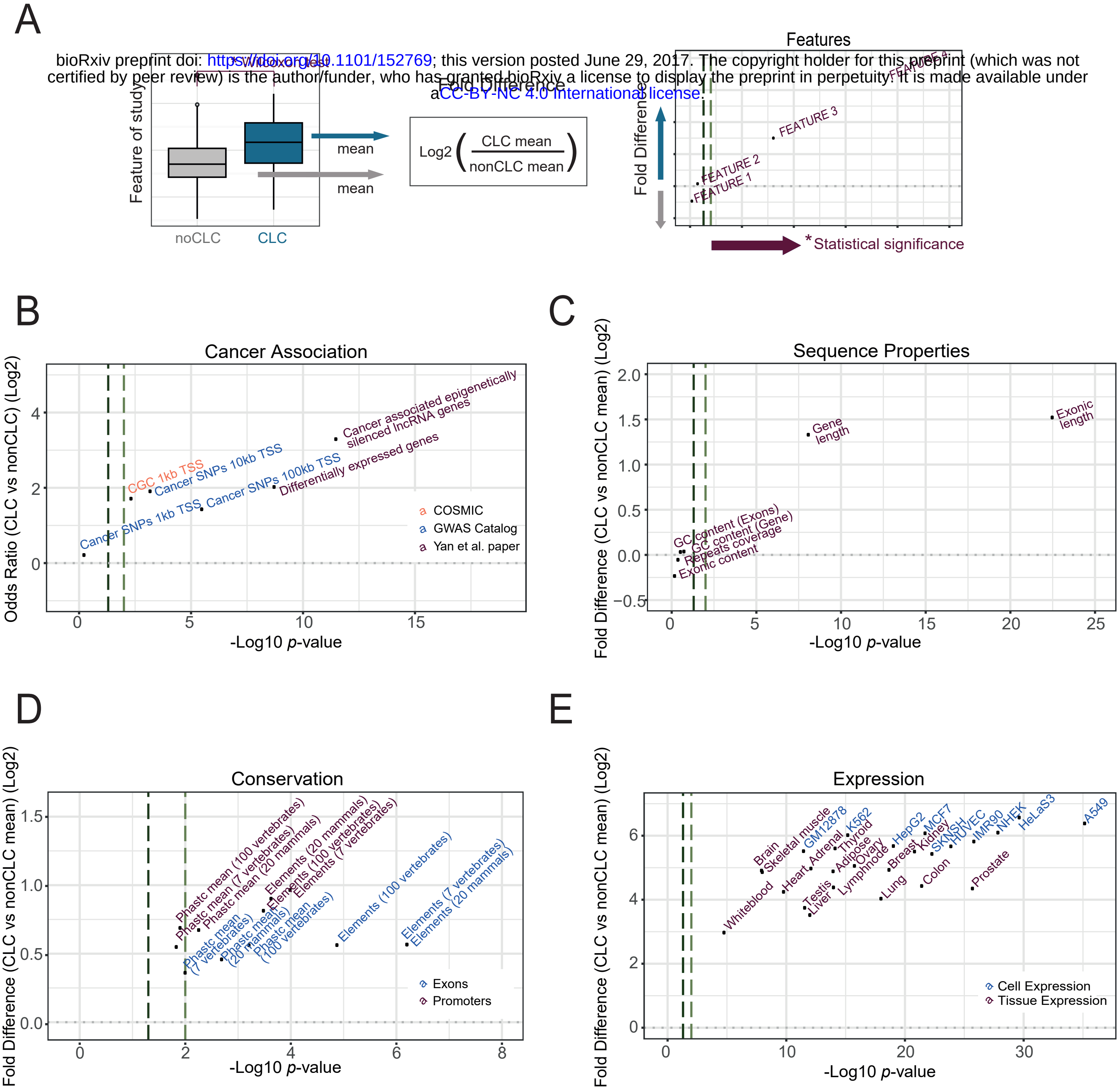

A
Features
Wilcoxon test
Feature of study
noCLC
CLC
mean
mean
Log2
CLC mean
nonCLC mean
Fold Difference
Fold Difference
FEATURE 1
FEATURE 2
FEATURE 3
FEATURE 4
*Statistical significance
B
Cancer Association
Odds Ratio (CLC vs nonCLC) (Log2)
-Log10 p-value
0
5
10
15
0
2
4
CGC 1kb TSS
Cancer SNPs 10kb TSS
Cancer SNPs 1kb TSS
Cancer SNPs 100kb TSS
Differentially expressed genes
Cancer associated epigenetically silenced lncRNA genes
a COSMIC
a GWAS Catalog
a Yan et al. paper
C
Sequence Properties
Fold Difference (CLC vs nonCLC mean) (Log2)
-Log10 p-value
−0.5
0.0
0.5
1.0
1.5
2.0
0
5
10
15
20
25
Gene length
Exonic length
GC content (Exons)
GC content (Gene)
Repeats coverage
Exonic content
D
Conservation
Fold Difference (CLC vs nonCLC mean) (Log2)
-Log10 p-value
0.0
0.5
1.0
1.5
0
2
4
6
8
Phastc mean (100 vertebrates)
Phastc mean (7 vertebrates)
Phastc mean (20 mammals)
Elements (20 mammals)
Elements (100 vertebrates)
Elements (7 vertebrates)
Phastc mean (7 vertebrates)
Phastc mean (20 mammals)
Phastc mean (100 vertebrates)
Elements (100 vertebrates)
Elements (7 vertebrates)
Elements (20 mammals)
a Exons
a Promoters
E
Expression
Fold Difference (CLC vs nonCLC mean) (Log2)
-Log10 p-value
0
2
4
6
0
10
20
30
Whiteblood
Brain
Skeletal muscle
GM12878
Heart
Adrenal
K562
Thyroid
Testis
Liver
Adipose
Lymphnode
Ovary
HepG2
Breast
Lung
MCF7
Kidney
Colon
SKNSH
HUVEC
IMR90
Prostate
NHEK
HeLaS3
A549
a Cell Expression
a Tissue Expression

Figure 5

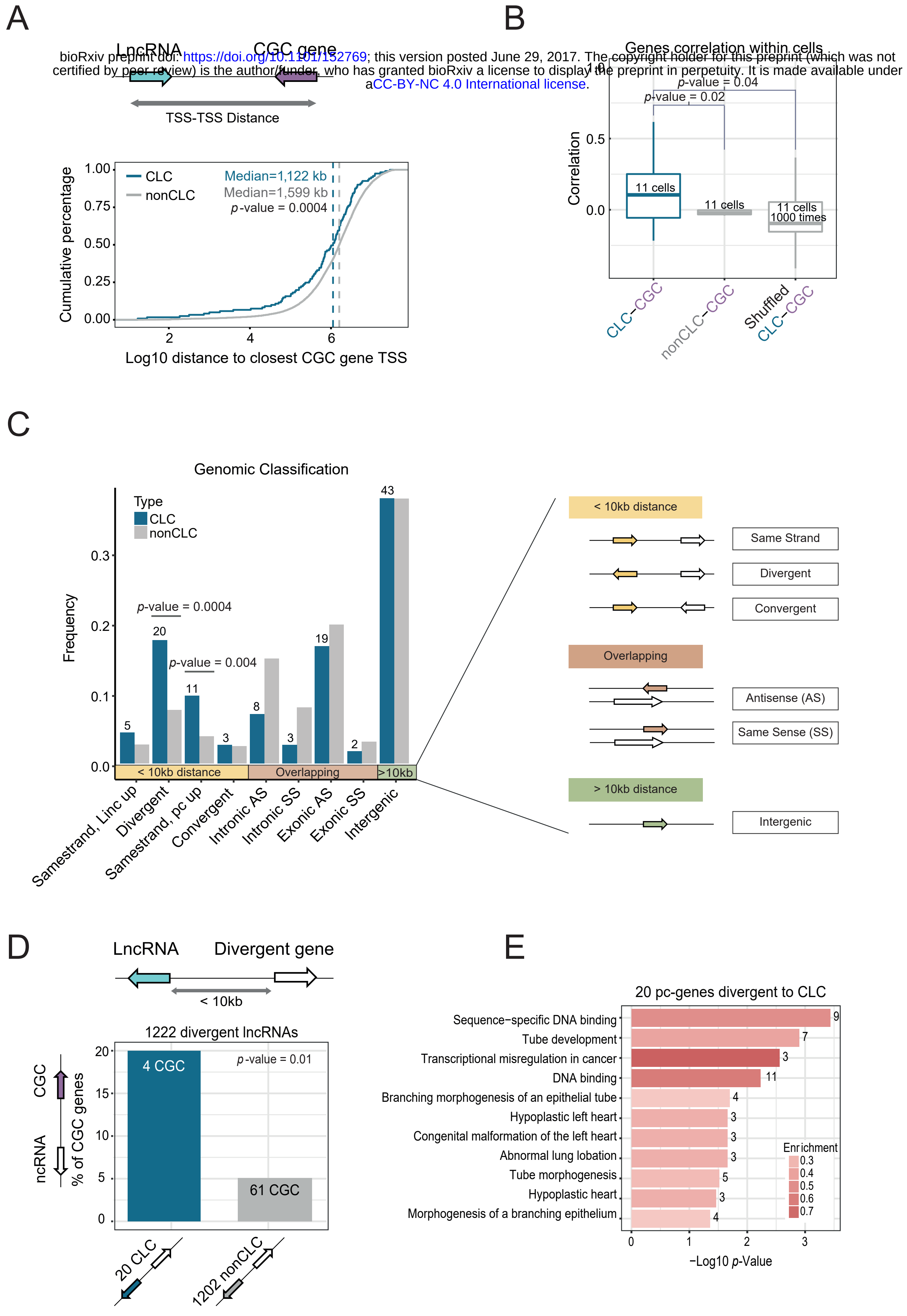
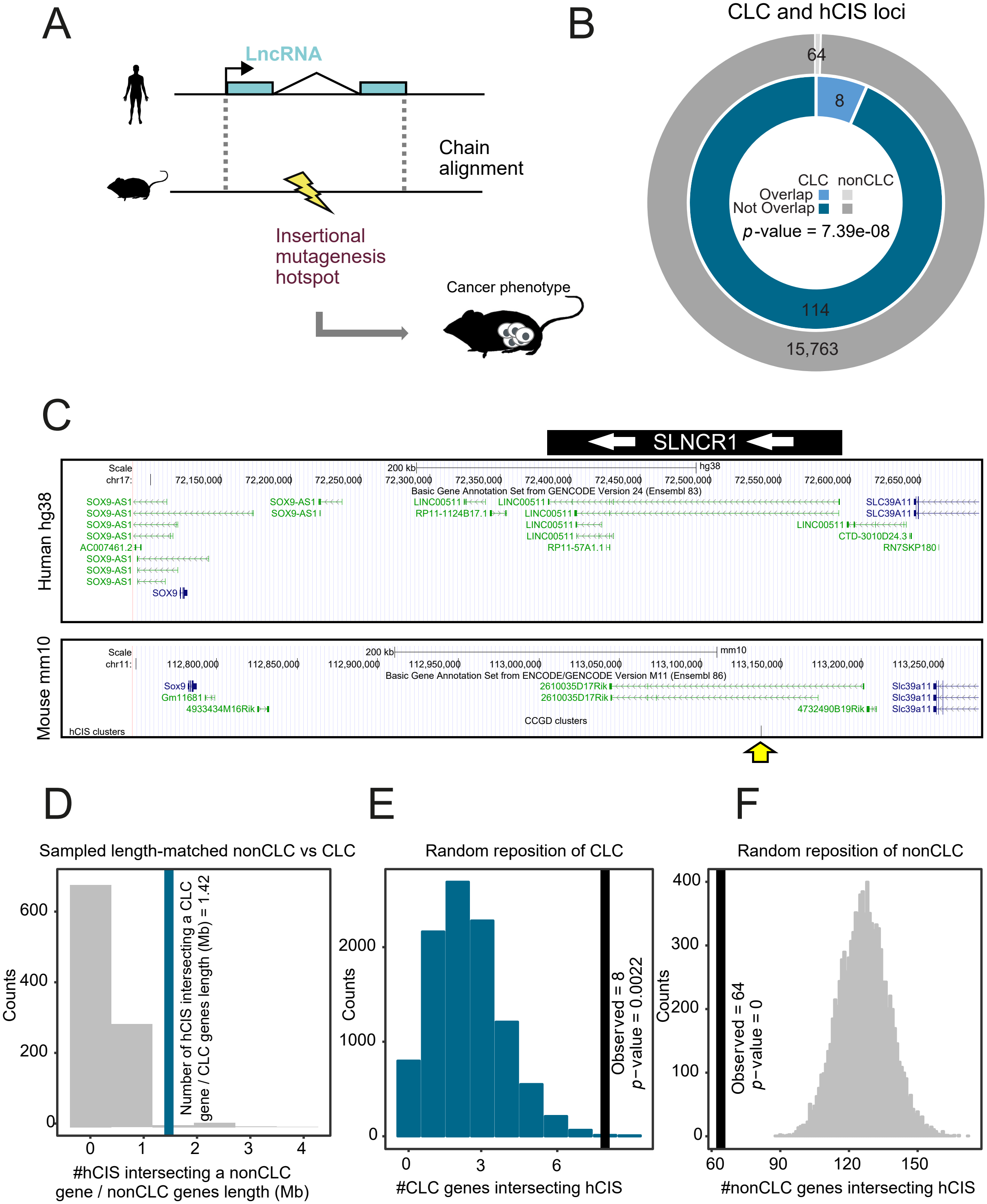

A
LncRNA
CGC gene
TSS-TSS Distance
CLC
nonCLC
Median=1,122 kb
Median=1,599 kb
p-value = 0.0004
Cumulative percentage
1.00
0.75
0.50
0.25
0.00
2
4
6
Log10 distance to closest CGC gene TSS
B
Genes correlation within cells
p-value = 0.04
p-value = 0.02
Correlation
1.0
0.5
0.0
11 cells
11 cells
11 cells
1000 times
CLC−CGC
nonCLC−CGC
Shuffled
CLC−CGC
C
Genomic Classification
Type
CLC
nonCLC
Frequency
0.3
0.2
0.1
0.0
p-value = 0.0004
p-value = 0.004
5
20
11
3
8
3
19
2
43
< 10kb distance
Overlapping
>10kb
Samestrand, Linc up
Divergent
Samestrand, pc up
Convergent
Intronic AS
Intronic SS
Exonic AS
Exonic SS
Intergenic
< 10kb distance
Same Strand
Divergent
Convergent
Overlapping
Antisense (AS)
Same Sense (SS)
> 10kb distance
Intergenic
D
LncRNA
Divergent gene
< 10kb
1222 divergent lncRNAs
p-value = 0.01
4 CGC
61 CGC
CGC
ncRNA
% of CGC genes
20
15
10
5
0
20 CLC
1202 nonCLC
E
20 pc-genes divergent to CLC
Sequence−specific DNA binding
Tube development
Transcriptional misregulation in cancer
DNA binding
Branching morphogenesis of an epithelial tube
Hypoplastic left heart
Congenital malformation of the left heart
Abnormal lung lobation
Tube morphogenesis
Hypoplastic heart
Morphogenesis of a branching epithelium
9
7
3
11
4
3
3
3
5
3
4
Enrichment
0.3
0.4
0.5
0.6
0.7
0
1
2
3
−Log10 p-Value

Figure 6


A
LncRNA
Chain
alignment
Insertional
mutagenesis
hotspot
Cancer phenotype
B
CLC and hCIS loci
64
8
CLC nonCLC
Overlap
Not Overlap
p-value = 7.39e-08
114
15,763
C
SLNCR1
Human hg38
Scale
chr17:
200 kb
hg38
72,150,000
72,200,000
72,250,000
72,300,000
72,350,000
72,400,000
72,450,000
72,500,000
72,550,000
72,600,000
72,650,000
Basic Gene Annotation Set from GENCODE Version 24 (Ensembl 83)
SOX9-AS1
AC007461.2
SOX9
LINC00511
RP11-1124B17.1
RP11-57A1.1
CTD-3010D24.3
RN7SKP180
SLC39A11
Mouse mm10
Scale
chr11:
200 kb
mm10
112,800,000
112,850,000
112,900,000
112,950,000
113,000,000
113,050,000
113,100,000
113,150,000
113,200,000
113,250,000
Basic Gene Annotation Set from ENCODE/GENCODE Version M11 (Ensembl 86)
Sox9
Gm11681
4933434M16Rik
2610035D17Rik
4732490B19Rik
Slc39a11
hCIS clusters
CCGD clusters
D
Sampled length-matched nonCLC vs CLC
Counts
600
400
200
0
0
1
2
3
4
Number of hCIS intersecting a CLC gene / CLC genes length (Mb) = 1.42
#hCIS intersecting a nonCLC gene / nonCLC genes length (Mb)
E
Random reposition of CLC
Counts
2000
1000
0
0
3
6
Observed = 8
p−value = 0.0022
#CLC genes intersecting hCIS
F
Random reposition of nonCLC
Counts
400
300
200
100
0
60
90
120
150
Observed = 64
p−value = 0
#nonCLC genes intersecting hCIS