Is evolutionary conservation a useful predictor for cancer long noncoding RNAs? Insights from the Cancer lncRNA Census 3

Adrienne Vancura (1,2,3,4), Alejandro H. Gutierrez (1,3), Thorben Hennig (4), Carlos Pulido-Quetglas (1,2,3), Frank Slack (4), Rory Johnson (1,3,5,6), Simon Häfliger (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.

[4] HMS Initiative for RNA Medicine, Department of Pathology, Beth Israel Deaconess Medical Center Cancer Center, Harvard Medical School, Boston, Massachusetts 02215, USA

[5] School of Biology and Environmental Science, University College Dublin, Dublin D04 V1W8, Ireland

[6] Conway Institute of Biomedical and Biomolecular Research, University College Dublin, Dublin D04 V1W8, Ireland

Correspondence: rory.johnson@ucd.ie, simon.haefliger@insel.ch

Abstract

Evolutionary conservation is a measure of gene functionality that is widely used to prioritise long noncoding RNAs (lncRNA) in cancer research. Intriguingly, while updating our Cancer LncRNA Census, we observed an inverse relationship between year of discovery and evolutionary conservation. This observation is specific to cancer over other diseases, implying a sampling bias in selection of lncRNA candidates and casting doubt on the value of evolutionary metrics for prioritisation of cancer-related lncRNAs.

Main

Long noncoding RNAs play central, functional roles in cancer and are being developed as targets for RNA therapeutics (Arun, Diermeier, and Spector 2018; Carlevaro-Fita et al. 2020; Vancura et al. 2021; Winkle et al. 2021). Given the high costs of drug discovery studies, and frequency of late-stage failures, it is imperative to collect and effectively prioritise lncRNAs with greatest therapeutic value. We here present the third version of the successful Cancer lncRNA Census (CLC3), covering publications from the period from 2019 to late 2020, comprising altogether 702 unique GENCODE-annotated lncRNAs with functional cancer roles based on a variety of evidence. CLC3 incorporates and extends previous versions (CLC1, 118 lncRNAs, CLC2 374 lncRNAs) (Carlevaro-Fita et al. 2020; Vancura et al. 2021). In addition to its size, CLC3 now incorporates for the first time lncRNAs involved in chemoresistance, with 10% of CLC lncRNAs exhibiting this functionality (**Figure 1A**).

We previously observed that CLC lncRNAs carry a range of features distinguishing them from other lncRNAs. Amongst these were elevated rates for several measures of evolutionary conservation, similar to that previously observed for protein-coding cancer genes (Carlevaro-Fita et al. 2020; Furney et al. 2008). First, we evaluated the confidence level of experimental support for CLC genes, finding these to be consistent between versions with roughly 40% of lncRNAs validated by the highest-confidence *in vivo* evidence (**Figure 1A**). Next, we comprehensively evaluated a range of features of CLC versions, comparing non-redundant gene sets to non-cancer lncRNAs ("nonCLC") (**Figure 1B**). For comparison, we also compared a collection of disease-associated lncRNAs, from which cancer genes were removed (EVlncRNA) (Zhou et al. 2021). All CLC versions and EVlncRNAs display elevated levels of gene expression, expression ubiquity, overall gene length, spliced RNA length and proximity to nearest protein-coding genes (**Figure 1B**). Surprisingly, however, we noticed that CLC3 lncRNAs are not more evolutionarily conserved compared to other lncRNAs (arrows). This is true not only for two different measures of conservation from the widely used PhastCons measure (average base-level score and percentage of exon coverage by conserved elements), but also for the promoter (average base-level), for which particularly elevated conservation has been observed in lncRNAs (Chodroff et al. 2010; Guttman et al. 2009). A more detailed gene-level inspection supported these findings (**Figure 1C-F**), showing a pronounced trend for the CLC3 lncRNAs to have comparable or even lower conservation than lncRNAs in general.

To strengthen these findings, we used an alternative method to evaluate evolutionary conservation: the existence of orthologous lncRNA genes in other species. Using the tool ConnectOR (Pulido-Quetglas 2021) we searched for orthologues of human lncRNAs in chimpanzee and mouse (see Methods). Overall, we identified orthologues for 4,102 and 4,493 lncRNAs in chimpanzee and mouse, respectively (lower rates in chimpanzee likely reflect less mature lncRNA annotations). Consistent with previous results, we observed that CLC3 lncRNAs have a significantly lower chance of having an identifiable orthologue than CLC1 and CLC2, at a level comparable to nonCLC lncRNAs (**Figure 2A**).

Given that CLC3 lncRNAs were collected most recently, we hypothesised that the observed trend arose from a relationship between conservation and the moment when the lncRNA was studied. Indeed, we observe a significant negative correlation between conservation and year of discovery (**Figure 2B, left**). This trend appears to be specific to cancer, because EVlncRNAs from other diseases do not display this behaviour (**Figure 2B, right**). In other words, as time goes on, researchers are turning their attention to less conserved lncRNAs that nevertheless play functional cancer roles.

In summary, we have presented the latest version of the Cancer lncRNA Census, a carefully curated resource of functional cancer-associated lncRNAs intended to serve as a useful true positive dataset for large-scale discovery and as a source for therapeutic development. We have made the surprising observation that evolutionary conservation of collected lncRNAs decreases with year of publication, and that recently published cancer lncRNAs have conservation levels similar to lncRNAs in general. Although previous studies have shown that protein-coding cancer genes are more conserved on average (Carlevaro-Fita et al. 2020), it remains possible that a similar phenomenon affects these genes. This phenomenon appears to be specific for cancer, since catalogues of lncRNAs playing roles in other diseases do not display the same trend. Evolutionary conservation is a longstanding and widely used criterion for selection of candidate lncRNAs for follow-up study (Iyer et al. 2015; Ponting 2017; Siepel et al. 2005; Ulitsky et al. 2011). However, there are numerous examples of functionally validated, non-conserved lncRNAs (Cao et al. 2022; Ruan et al. 2020). Supporting these findings, recent unbiased large-scale functional screens found no relationship between conservation and hits (Liu et al. 2017). The apparent specificity to cancer raises the possibility that tumours exploit lncRNA sequences that have no natural function. Indeed, a similar model was recently proposed by Adnane and colleagues (Adnane, Marino, and Leucci 2022). These findings suggest that the scientific community may have suffered an unconscious bias in selecting evolutionarily conserved lncRNAs for study, and thereby reinforcing the impression that conservation is a useful criterion for candidate selection (Carlevaro-Fita et al. 2020). Overall, these findings lead us to propose that evolutionary conservation is not a useful filter when selecting cancer lncRNAs for further study.

Methods

Literature search and LnCompare for Feature and Repeat analysis

This analysis was performed as described in CLC2 (Vancura et al. 2021) and the gene list can be found here:
https://docs.google.com/spreadsheets/d/1nT-YP8O4gkoRb9RwYAKtwpSt5C5liFLbxb0cc8mqht4/edit#gid=0

EVlncRNA non-cancer lncRNA dataset

EVlncRNAs were downloaded from https://academic.oup.com/nar/article/49/D1/D86/5998394 and were sorted for ENSG (GENCODE v28) and overlayed with CLC genes to exclude functional cancer lncRNAs.

Conservation scores

Exons were collapsed using exon info from GENCODE v28 and PhastCons exon conservation scores (PhastCons100way.UCSC.hg28) were generated according to Vancura et al., 2021 using Bioconductor Genomic Scores R package.

PhastConsElements 100way were downloaded from genome.ucsc.edu using the table browser. PhastConsElements were intersected with datasets using intersectBed. Statistical evaluation was performed using Wilcoxon test.

Publication year

PMID years for each lncRNA were extracted using the code from https://www.ncbi.nlm.nih.gov/books/NBK179288/ and earliest publication years was used for subsequent analysis.

Ortholog prediction

Ortholog prediction was performed using ConnectOR (https://github.com/Carlospq/ConnectOR) based on LiftOver of syntenic regions from human (hg38) to mouse (mm10) or chimpanzee (panTro3). ConnectOR results “not lifted” and “one to none” were characterized as no orthology prediction. Statistical evaluation was performed using Fisher’s one-sided t-test.

We acknowledge funding from the NIH (R35 CA232105) to FJS.

References

Adnane, Sara, Alessandro Marino, and Eleonora Leucci. 2022. "LncRNAs in Human Cancers: Signal from Noise." *Trends in Cell Biology* 32(7): 565–73. http://www.cell.com/article/S0962892422000071/fulltext (July 14, 2022).

Arun, Gayatri, Sarah D. Diermeier, and David L. Spector. 2018. "Therapeutic Targeting of Long Non-Coding RNAs in Cancer." *Trends in molecular medicine* 24(3): 257–77. https://pubmed.ncbi.nlm.nih.gov/29449148/ (December 17, 2021).

Cao, Tingyi et al. 2022. "Cis-Regulated Expression of Non-Conserved LincRNAs Associates with Cardiometabolic Related Traits." *Journal of Human Genetics 2022 67:5* 67(5): 307–10. https://www.nature.com/articles/s10038-022-01012-5 (June 20, 2022).

Carlevaro-Fita, Joana et al. 2020. "Cancer LncRNA Census Reveals Evidence for Deep Functional Conservation of Long Noncoding RNAs in Tumorigenesis." *Communications Biology*.

Chodroff, Rebecca A. et al. 2010. "Long Noncoding RNA Genes: Caaonservation of Sequence and Brain Expression among Diverse Amniotes." *Genome Biology* 11(7): 1–16. https://genomebiology.biomedcentral.com/articles/10.1186/gb-2010-11-7-r72 (June 20, 2022).

Furney, Simon J. et al. 2008. "Distinct Patterns in the Regulation and Evolution of Human Cancer Genes." *In Silico Biology*.

Guttman, Mitchell et al. 2009. "Chromatin Signature Reveals over a Thousand Highly Conserved Large Non-Coding RNAs in Mammals." *Nature*.

Iyer, Matthew K. et al. 2015. "The Landscape of Long Noncoding RNAs in the Human Transcriptome." *Nature Genetics*.

Liu, S. John et al. 2017. "CRISPRi-Based Genome-Scale Identification of Functional Long Noncoding RNA Loci in Human Cells." *Science*.

Ponting, Chris P. 2017. "Biological Function in the Twilight Zone of Sequence Conservation." *BMC Biology* 15(1): 1–9. https://bmcbiol.biomedcentral.com/articles/10.1186/s12915-017-0411-5 (December 27, 2021).

Pulido-Quetglas, Carlos. 2021. "GitHub - Carlospq/ConnectOR: Multiple Species Orthology Finder." https://github.com/Carlospq/ConnectOR (June 20, 2022).

Ruan, Xiangbo et al. 2020. "In Vivo Functional Analysis of Non-Conserved Human LncRNAs Associated with Cardiometabolic Traits." *Nature Communications 2020 11:1* 11(1): 1–13. https://www.nature.com/articles/s41467-019-13688-z (December 27, 2021).

Siepel, Adam et al. 2005. "Evolutionarily Conserved Elements in Vertebrate, Insect, Worm, and Yeast Genomes." *Genome research* 15(8): 1034–50. https://pubmed.ncbi.nlm.nih.gov/16024819/ (December 17, 2021).

Ulitsky, Igor et al. 2011. "Conserved Function of LincRNAs in Vertebrate Embryonic

Development despite Rapid Sequence Evolution." *Cell*.

Vancura, Adrienne et al. 2021. "Cancer LncRNA Census 2 (CLC2): An Enhanced Resource Reveals Clinical Features of Cancer LncRNAs." *NAR Cancer* 3(2). https://academic.oup.com/narcancer/article/3/2/zcab013/6225859 (December 27, 2021).

Winkle, Melanie, Sherien M. El-Daly, Muller Fabbri, and George A. Calin. 2021. "Noncoding RNA Therapeutics — Challenges and Potential Solutions." *Nature Reviews Drug Discovery* 20(8): 629–51.

Zhou, Bailing et al. 2021. "EVLncRNAs 2.0: An Updated Database of Manually Curated Functional Long Non-Coding RNAs Validated by Low-Throughput Experiments." *Nucleic acids research* 49(D1): D86–91. https://pubmed.ncbi.nlm.nih.gov/33221906/ (December 17, 2021).

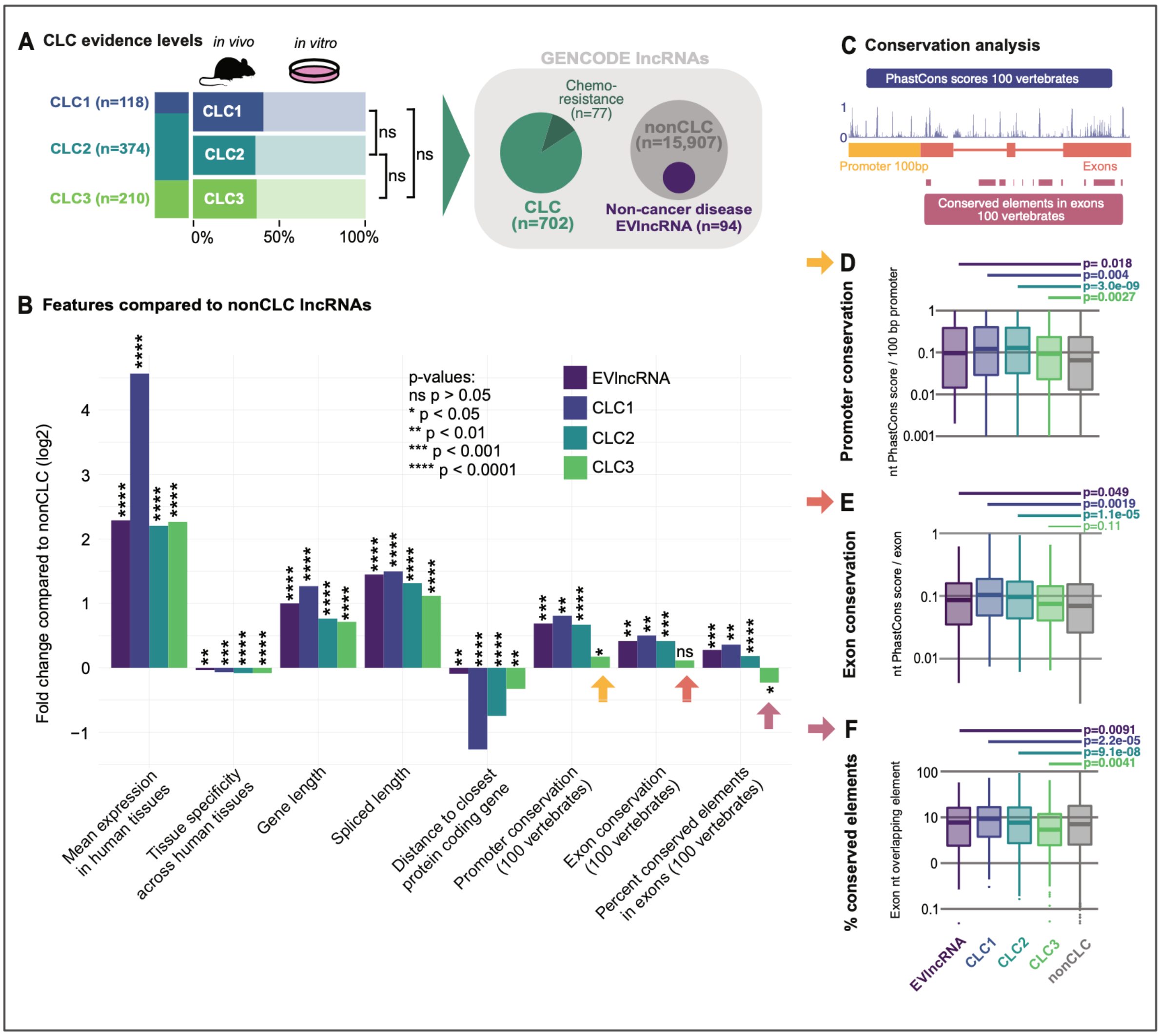

Figure 1: A) Evidence levels for functional lncRNAs in CLC database versions. Dark color indicates number of lncRNAs tested in an *in vivo* setting. No significant (ns) difference of in vivo enrichment is observed across the datasets. The full CLC consists of 702 lncRNAs with 77 lncRNAs exhibiting chemoresistance mechanisms. GENCODE lncRNAs are subdivided in CLC and nonCLC genes for further comparison. Non-cancer disease EVlncRNAs are nonCLC genes indicating a disease functionality but not represented in the CLC database. **B)** Features in datasets compared to nonCLC lncRNAs using LnCompare. **C)** Overview of conservation analysis using 100 vertebrates comparisons. **D)** Promoter conservation analysis for all datasets. **E)** Exon analysis for all datasets. **F)** Conserved elements analysis for all datasets.

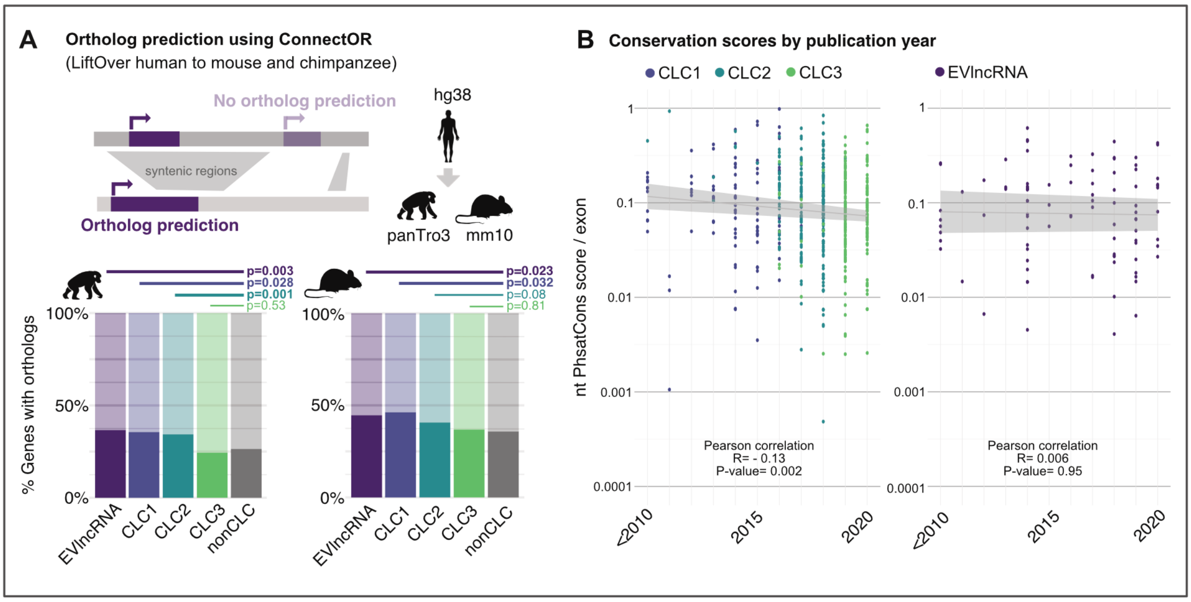

Figure 2: A) Ortholog prediction using ConnectOR for chimpanzee (left) and mouse (right). **B)** Exon conservation scores by publication year for CLC versions (left) and EVlncRNAs (right).