AUTHORS’ RESPONSES TO REVIEW

Nature Genetics submission NG-LE49200

Widespread transposable element-driven oncogene expression in cancers

Hyo Sik Jang, Nakul M. Shah, Alan Y. Du, Zea Z. Dailey, Erica C. Pehrsson, Paula M. Godoy, David Zhang, Daofeng Li, Xiaoyun Xing, Sungsu Kim, David O’Donnell, Jeffrey I. Gordon, Ting Wang

GENERAL RESPONSES

We are extremely grateful for the reviewers’ enthusiasm for the quality of our work. Most importantly, we are in debt for the very constructive and thoughtful comments – these detailed suggestions provided us the opportunity to greatly enhance our manuscript. Through significant additional analyses and a more thorough presentation of our methods and results, we believe to have addressed all the issues raised.

We have marked edited text: new text was marked with underline, and removed text was marked with strikethrough.

We provide a brief summary of the new data and analyses included in our revised manuscript:

General Response 1. Bioinformatic Pipeline Upgrade

We thank all the reviewers for their constructive comments on our bioinformatics analysis. Our original intent was to screen for candidates and then highlight the most interesting ones by which we were able to identify AluJb-LIN28B; however, we realize more advanced analysis could help answer many interesting questions about this mechanism across many different oncogenes. Reading these comments and suggestions inspired us to make many enhancements to our pipeline to give a more comprehensive analysis on novel as well as previously published examples of this mechanism. Some of the significant changes we incorporated include: (1) A more stringent threshold of how many uniquely mapped reads must support each candidate, (2) transcript-level expression quantification to compare candidates to all other transcript isoforms of the gene, (3) protein coding and reading frame prediction of all candidates, (4) incorporation of overall survival clinical data from TCGA, and (5) analysis of FANTOM5 promoter atlas peak data. Through this expanded analysis, we were able to focus on candidates driving a significant percentage of oncogene expression and find candidates associated with cancer prognosis. In addition, we now provide a comprehensive resource of TE-derived alternative promoters with data on relative expression and coding information. Many previously published candidates were filtered out in our original analysis due to lack of high enough tumor specificity, but now we include useful metrics on every example found before our stringent filtering. Even though some of these promoters can be active in normal tissues, our FANTOM5 analysis was able to show that they have much higher expression in cancer tissues (Response 3.8). We also included tumor suppressor genes in the updated analysis and found that TE-derived oncogene promoters were more likely to drive a significant portion of overall gene expression which is further discussed in Response 1.2 and 3.4. Overall, we believe this enhanced analysis is able to answer many important questions about onco-exaptation and provides a rich resource of many interesting examples of this phenomenon.

General Response 2. Improving rigor of AluJb-LIN28B biology
We are always appreciative for reviewers’ recommendations to improve rigor and robustness of our scientific discoveries. As our astute reviewers requested, we have performed a rescue of both canonical human LIN28B and AluJb-LIN28B protein in AluJb-P KO clones of H1299 and H838. Further addressed in Responses 1.3, 1.8, 2.5 and 4.4, this effort revealed that AluJb-LIN28B retained the canonical oncogenic function of repressing let-7 levels and moderately rescued proliferation in AluJb-P KO clones.

Additionally, we apologetically acknowledge that treating WT as experimental controls for the xenograft experiment was not a sound scientific comparison for the critical reasons that Reviewer 2 pointed out in the main concerns. Therefore, we repeated the xenograft experiments by comparing H1299 AluJb-P KO and H1299 LIN28BP KO (addressed in Responses 1.3, 2.6 and 4.5). These lines underwent the same CRISPR procedure and clonal selection. LIN28BP KO behaves identically to WT in all other functional assays, including LIN28B expression level, therefore is probably better than a rescue experiment where control of expression level might not be trivial to calibrate. We feel that this is a more rigorous comparison and more to the point that we’d like to make (ie. different tumorigenic effect as a function of deleting a promoter vs “junk DNA”). We are thankful to the reviewers for the critical insights and the opportunity to enhance the scientific rigor of this manuscript.

General Response 3. Identification of additional onco-exaptation candidates

We appreciate the reviewer’s recommendations for validating additional onco-exaptation events to strengthen the paper. In brief, we incorporated FANTOM5 promoter database to inquire how many onco-exapted TEs show promoter activity. This effort identified numerous TEs that show promoter activity in various cancer cell lines, including 5 out of our top ten candidates. One caveat of FANTOM5 promoter atlas is that the promoters are often unannotated or annotated to the closest gene. To verify that the active TEs act as alternative promoters for the predicted oncogenes, we obtained H727 lung carcinoid cell line and validated two additional top 10 candidates using paired-end CAGE-seq (Tigger3a-ARID3A, L1PA2-SYT1). More thorough details can be found in Responses 1.11, 2.3 and 3.7.

In addition to validating more candidates, we wanted to interrogate the impact of DNA methylation on another TE onco-exaptation candidate. We verified the presence of onco-exapted ARID3A transcript in H727. When we measured the DNA methylation level with BSPCR-seq, the onco-exapted promoter was completely demethylated, consistent with our hypothesis. In H1299, the Tigger3a and MLT1D TEs are highly methylated thus we postulated that demethylating the TEs could resurrect promoter activity. Using SunTag-TET1CD system, we successfully demethylated the region (Response Figure 1a), but to our disappointment, we could not detect any onco-exapted ARID3A transcript via qPCR or 5’RACE. This negative result exemplifies the importance of having both the proper transcriptional factor network and epigenetic signature to activate the regulatory activity of transposable elements. Although unsubstantiated in this paper, this could reflect why certain cancers are more likely to have a certain onco-exaptation event than other cancers. We are excited to have this epigenetic “primed” system since with careful analysis, we can validate what transcription factors are essential for the activation this particular onco-exapted TEs and delve into the genetic and transcriptional evolution of this onco-exaptation candidate in future studies.

Furthermore, we attempted to methylate Tigger3a/MLT1D in H727 to examine if DNA methylation can silence promoter activity. Using the SunTag-DNMT3A system, we were only able to increase the methylation in the flanking regions of Tigger3a/MLT1D (Response Figure 1b), but not the TEs themselves. This result could suggest that there is strong selective pressure or dynamic demethylation on these TEs. Unfortunately, this example reflects the limitations of targeted methylation technology and therefore, we are not able to properly interrogate the effects of methylation on the TE promoter activity.
General Response 4. Addressing Guo et al. 2018

Our overarching mission for this manuscript is to be the first to characterize and quantify onco-exaptation events across various cancer types. We are extremely gratified that the reviewers share the same enthusiasm about the potential impact of our work for the genome regulation, transposable element and cancer biology field. During the preparation of this manuscript, Guo et al. 2018 presented a thorough body of work on its discovery and characterization of LIN28B-tumor-specific transcript (TST) with a focus on liver cancer model. We want to provide full disclosure that the LIN28B-TST identified in Guo et al. is the same transcript as AluJb-LIN28B. However, as Reviewer 3 and 4 pointed out, Guo et al. did not recognize this particular event to be part of a wide spread onco-exaptation phenomenon. Our motivation and strategic direction to interrogate the genetic and epigenetic context that define this onco-exaptation event stemmed from our passion for transposable elements. This inspiration led to complementary but distinct experiments in profiling AluJb-P’s promoter capacity that ultimately led to similar conclusions as Guo et al.. For example, Guo et al. presented that the LIN28B-TST promoter is conserved in primate but could not explain why this was a primate-specific promoter. Understanding this promoter arose from onco-exaptation event can provide insight in the evolution of this promoter, such as AluJb is a primate-specific SINE element that invaded the genome millions of years ago. Furthermore, we can perform more specific promoter deletion assays by asking whether the TE is necessary and sufficient for promoter activity instead of performing random serial 5’ deletions. The information provided by the mutations accrued in this AluJb copy led to the discovery of three new putative transcription factor binding sites that control promoter activity. For functional experiments, the question we asked was fundamentally different from Guo et al. We asked what is the consequence of
deleting a TE on cancer oncogenesis. By deleting AluJb-P, we effectively abolished LIN28B expression in lung cancer as illustrated by the Western blot. In Guo et al., the functional experiments were done in a LIN28B knock-down model in Huh-7 cells compared to our experiments with LIN28B knockout in lung cancer cell lines. We show synergistic results for impact of AluJb-LIN28B on let-7 miRNA: Guo et al. reports that ectopic expression of LIN28B-TST leads to suppression of let-7 levels while our work reveals that loss of AluJb-P (in consequence, AluJb-LIN28B) elevates let-7 levels. We perform additional functional experiments, as per reviewers’ request, and present that re-expression of AluJb-LIN28B in AluJb-P KO cells represses let-7 levels and modestly rescues proliferation. Again, the overall conclusion is matching; LIN28B is important for oncogenesis, but our work provides synergistic and robust evidence in lung cancer context that further supports Guo et al. claims. Lastly, Guo et al. reported that DNA methylation is associated with LIN28B-TST promoter activity. The authors treat cells with 5-AZA, a DNMTi drug, to show that global reduction of DNA methylation can activate LIN28B-TST promoter. In contrast, we performed targeted CRISPR epigenetic engineering experiments to pinpoint that AluJb-P demethylation alone is sufficient for promoter activity and circumvent potential artifacts from global de-methylation events. Furthermore, we present targeted methylation experiments to show that methylating AluJb-P suppresses promoter activity, providing more rigorous support for the role of DNA methylation on epigenetic control of TEs. In conclusion, we want to reemphasize that LIN28B is a high value example supporting our hypothesis, but the novelty of our manuscript draws from the large scale comprehensive screen for TE onco-exaptation events.

Below we provide point-by-point responses to Reviewers’ comments.

POINT-BY-POINT RESPONSES

Reviewer #1:

Remarks to the Author:

In this manuscript, Ting Wang and colleagues perform a pan-cancer analysis of transposable element (TE)-driven cryptic promoter activation events across 15 TCGA tumor types. They identify TE-derived activation of promoters (termed “onco-exaptation”) as a pervasive mechanism of oncogene activation. They specifically characterize the case of LIN28B, in which oncogene expression is activated by the AluJb TE. They show in cell line models and xenografts that deletion of the AluJb TE eliminates LIN28B expression and suppresses cancer cell/xenograft growth.

Overall, the data are clearly presented and the manuscript is well written. The experiments are convincing overall. This group has previously shown that hypomethylation of specific TEs within normal cell types can drive tissue-specific gene expression patterns (Xie et al., Nature Genetics, 2013). They have also shown that DNMT or HDAC inhibitors can activate transcription from cryptic transcriptional start sites encoded in the LTRs of TEs in cancer cell lines (Brocks et al, Nature Genetics, 2017). The data in this manuscript are novel for two reasons: 1) they point to a potentially wide-spread role for onco-exaptation in the activation of oncogenes in cancer; 2) they suggest onco-exaptation as a common mechanism of LIN28B activation in cancer. While there is now a large body of data implicating LIN28B in cancer, primarily through its repressive effects on let-7 miRNAs, this gene is only infrequently activated through conventional mechanisms (amplification, translocation, mutation).
We want to thank the reviewer for the encouraging comments on our work!

For these reasons, I feel the findings are of general interest and would be enthusiastic about publication in Nature Genetics, providing the issues below are addressed:

Major Issues
1. The authors restrict their analysis of onco-exaptation events to 702 oncogenes. However, it would be important to do this analysis across all genes and determine to what extent other genes are also being transcribed from TE-promoters in cancer. If this is truly an important mechanism of oncogene activation, one should expect to see an enrichment/selection for in-frame TE-driven oncogene expression, as compared with TE-driven expression of other genes (and certainly as compared with TE-driven-expression of tumor-suppressors, for example). If oncogenes and tumor-suppressors are both equally subject to cryptic transcription in cancer, the argument for this being an important mode of oncogene activation is much less compelling.

We thank the reviewer for this important consideration when evaluating whether this is mechanism specific to genes promoting oncogenesis. Due to the computational burden of attempting to expand the analysis to all genes and the ambiguity of whether certain genes are important in oncogenesis and tumor suppression, we decided to focus on tumor suppressor genes and oncogenes. We have included a list of 201 tumor suppressor genes identified from COSMIC in our second round of analysis to see if this mechanism does indeed have a preference for oncogenes. Based on the filters used in defining our initial candidate list that are highly tumor enriched, we found 129 (106/702 genes 15.1%) total events for oncogenes and 33 total events (25/201 genes 12.4%) for tumor suppressor genes. These raw percentages of genes do not show statistically significant difference; however, there are differences that arise when looking at the overall impact of the events on gene expression. First, we have graphed the average expression percentage of each candidate for both the oncogene transcripts versus the tumor suppressor gene transcripts (P<.001) (Response Figure 2a). When looking at those that on average drive >50% of expression, 56 (51/702 genes 7.3%) total candidates exist for oncogenes, and only 8 (7/201 genes 3.5%) total events exist for tumor suppressor genes. This is greater than a 2x enrichment and shows that tumor suppressor gene candidates are less likely to drive a significant percentage of expression. Furthermore, the maximum number of tumors that have a tumor suppressor gene candidate with >50% of total expression was 94 (mean = 34.4) whereas the maximum tumors for an oncogene candidate was 901 (mean = 68.3); however, the difference in means was not statistically significant (Response Figure 2b). These results indicate that exaptation events are more likely to contribute significantly to the expression of oncogenes rather than tumor suppressor genes.
Though the evidence is quite compelling, we would like to note that there are limitations in making this conclusion. First, we are not sure what the mechanism of the tumor-suppressor transcripts and potential protein-products are. These could serve as dominant negative isoforms that are actually leading to suppression of activity of the gene; previous examples of splicing variants with shorter isoforms have been found to have dominant negative activity on tumor suppressor genes\textsuperscript{3,4}. Furthermore, there is a grey area between oncogene and tumor suppressor gene definition where some genes can have dual roles; a recent paper characterized 83 such examples\textsuperscript{5}; 2 of the 8 tumor suppressor exaptation events with >50% expression are on this list (BRCA2 and PPARG). Thus, we elected to not include this enrichment analysis in the final manuscript since we believe we do not have enough evidence with our current study to confidently publish this claim.

2. There seems to be a discrepancy between the magnitude of phenotype observed with Alu-JB KO in vitro (Fig. 3d) and in vivo (Fig 3f). Can the authors explain this? Specifically, there is a total loss of growth by KO clones in the xenograft experiments and my concern is that the knockout clones were already dying or close to dying when implanted, due to decreased LIN28B expression. Are the Alu-JB knockout clones able to be maintained in culture long-term or do they die after some time due to loss of LIN28B expression? A preferable experiment would be to transduce the knockout clones with a Doxycycline-inducible LIN28B ORF, grow xenografts to a certain volume in both arms, and then withdraw Doxycycline from the mice and evaluate for tumor regression.
We are thankful to the reviewer for raising these important points about the biological impact of removing AluJb-LIN28B in cancer cell lines. To address whether AluJb KO clones are able to propagate long term in culture, we passaged all AluJb KO clones and LIN28BP KO clones every three days for 30 days. All KO clones illustrated no visible morphological defects and propagated normally during this period.

We are in complete agreement that a doxycycline-inducible LIN28B xenograft experiment would be a robust experimental design. However, due to the time-sensitive nature of this study, we cautiously hope that rescue experiments in an in-vitro system would suffice in addressing the biological impact of AluJb-LIN28B. In brief, we obtained pBABE-hLIN28B retroviral vector (as characterized in Viswanathan et al. 2009) and cloned in the CDS of AluJb-LIN28B. We decided not to include the 5'UTR and 3'UTR of the AluJb-LIN28B transcript to specifically test the impact of the AluJb-LIN28B protein on let-7 levels and proliferation rate. We re-expressed the hLIN28B and AluJb-LIN28B protein in four AluJb KO clones (2 clones from H1299 and 2 clones from H838) to address two main points (Figure 3). First, the re-expression of both LIN28B isoforms resulted in a reduction of let-7 miRNA levels and partial to full rescue of proliferation in the four AluJb clones. Interestingly, we note that the canonical isoform of LIN28B is also present in the rescue clones, potentially inferring that the canonical promoter can still express the canonical LIN28B transcript but undergoes let-7 mediated degradation. This represents the negative feedback relationship between let-7 miRNA and LIN28B. These results support the notion that AluJb-LIN28B protein has comparable function and oncogenic consequence to the canonical LIN28B protein. Second, the rescue experiments complement the finding that the loss of proliferation and latent oncogenic ability of the lung cancer cell lines is due to the loss of AluJb-LIN28B. We do note in H838 that the canonical LIN28B re-expression did not remarkable rescue the proliferation rate. This potentially could be due to interesting biological or technical issues, but we believe that our hypothesis that the loss of proliferation is due to loss of AluJb-LIN28B protein still holds true.

We agree with the reviewer that the difference in oncogenicity between in vitro and in vivo experiments is pretty drastic. As Reviewer 2 and Reviewer 4 also justifiably pointed out, evaluating the xenograft ability of AluJb KO clones to WT is not a fair comparison (General Response 2). We agree completely and apologize for this oversight. Reviewer 2 raised concerns about single clone expansions impacting the grafting and tumor growth of heterogeneous cancer cell lines. To address this concern, we repeated the H1299 xenograft experiment and included what we believe to be the proper control, LIN28BP KO clone. The LIN28BP KO clone also went through CRISPR activity and single clone expansion thus we can better answer whether the loss of tumorigenicity is due to clonal expansion or loss of LIN28B in AluJb KO clones. In addition, the nature of LIN28BP KO in essence represents a LIN28 rescue experiment in AluJb KO background. All six injections of LIN28BP #1 KO clone generated a tumor, albeit at a slower growth rate than the WT, which could reflect the adverse impact of single clone expansion. We replicated the results from previous xenograft experiment where the WT grew sizable tumors while AluJb KO clones did not after 38 days. We recognize that while this could be due to lack of proper tumor engraftment, a more likely explanation is that the AluJb KO clones grow extremely slowly. To test this hypothesis, we examined mice injected with the AluJb KO clones for addition 30 days. Indeed, 4 out of 6 mice injected with AluJb2 #1 KO clone had visible tumor growth while we still did not notice any palpable tumor in mice injected with AluJb1#1 KO clone. To reflect this knowledge, we revised the text to “Also, parental H1299 and LIN28BP KO clone established rapidly growing tumors in vivo, whereas AluJb-P KO cells exhibited a marked defect in tumor growth during the time of inspection”. 
Minor Issues

1. Line 61 – authors state 8/10 of the most prevalent onco-exaptation (OE) events form in-frame transcripts. Which two do not? For these two transcripts, is RNA expression in the OE candidate samples low? Please annotate whether OE candidates are in-frame or out-of-frame in the supplemental table.

(1.4) We agree that more systematic labeling of the predicted protein-product would enhance the completeness of our data and make it more useful for identifying potentially robust candidates. To this end, we have added a label in Figure 1d that indicates which ones are in-frame versus out-of-frame. Also, this information is included in Supplementary Table 2 and Supplementary Table 3. Due to our enhanced pipeline that takes expression information into consideration, some of the members in the top 10 have changed, and thus now there are 8 number of in-frame and 2 non-coding candidates in the top 10. The two non-coding candidates are L1PA2-XCL1 and THE1A-HMGA2. In part due to our filters requiring them to represent at least 25% of the total gene expression and >1 FPKM total gene transcription, the candidates are present in samples with relatively high expression of the gene (Figure 1d). However, when we look at percent of total gene expression, of the top 10 candidates they are in the bottom half (Response Figure 3). Though in silico prediction indicates these likely do not code for functional protein products, future functional studies may elucidate an interesting alternative functional consequence of these highly tumor-specific transcripts.

![Response Figure 3](image)

Response Figure 3. A bar plot displaying the average percentage of total oncogene expression for the top 10 candidates

2. Fig 1c. The data could be presented in a more useful manner – in the current visualization, names of the TE-oncogenes are not listed and this information needs to be cross referenced with the supplemental tables. Can the authors plot total number of samples with a given TE-oncogene event versus total number of tumor types in which that TE-oncogene event is observed? Outlier points can then be annotated with the oncogene names.
We thank the reviewer for this very helpful suggestion and agree that a different visualization of the data would be able to provide more useful information. We thus have elected to produce the figure suggested by the reviewer and have highlighted individual candidates (Response Figure 4). The top panel has also been included in the manuscript as Figure 1c.

Response Figure 4. Distribution of onco-exaptation candidates across number of cancer types. In the top panel we have a boxplot that highlights the distribution of total number of tumor samples per candidate that is present in a certain number of cancer types. We have zoomed in on 1-11 so that the distribution can be more clearly seen. We have also labeled all the outliers candidates. The bottom panel displays the number of candidates that is present at various numbers of cancer types. There were 0 candidates present in all 15 cancer types.

3. Fig 1d. One issue that is not clear to me is whether expression of a TE-oncogene precludes transcription from the native promoter. At least for the cell lines studied here for LIN28B, this appears to be the case. However, can the authors plot this systematically in the TCGA data – i.e. please show expression level of TE-oncogene and native oncogene in OE samples and non-OE samples.

We agree that relative expression is a very important consideration in evaluating candidates, and thus in our revised pipeline we have made sure to consider the overall
percentage of expression of the gene in candidate selection. We plot the distribution for our top 10 candidates in Figure 1d, and we have included this information in Supplementary Table 2 and Supplementary Table 3. All our 129 candidates on average account for at least 25% of expression across all tumor types, and the ones we have highlighted in the top 10 account for >50% of expression on average in all tumor samples they are present in.

4. Fig 3c, right. Let-7g is listed twice on the x-axis. Perhaps the middle should read “let-7b”

(1.7) We apologize for the oversight. This has been rectified.

5. Fig 3d. Please perform a rescue experiment in which LIN28B is reconstituted in the knockout clones and growth is restored to WT levels, to ensure that the proliferation effects seen are due to LIN28B loss.

(1.8) We thank the reviewer for this recommendation to improve robustness of our findings. As mentioned in Response (1.3), we re-expressed canonical human LIN28B protein and AluJb-LIN28B protein in each AluJb KO clone from H1299 and H838 and report consistent conclusions that re-expression of AluJb-LIN28B reduces levels of let-7 miRNA and rescues proliferation (Fig 3g). The extent of rescue is similar to canonical LIN28B suggesting that AluJb-LIN28B does indeed have the conserved function of LIN28B oncogene. However, it is still unclear if the 22 additional amino acids from AluJb-LIN28B provide a neo-function and would be interesting to pursue as future endeavors.

6. Fig. 4d. Is this showing relative fold change in gene expression by qPCR? Can the authors show a western blot as for the CRISPR-SunTag-TET1CD clones in Figure 4f?

(1.9) We apologize to the reviewer for not being clear in the manuscript and figure legends. We clarified in figure legends to reflect what was being measured. Furthermore, we added a Western blot of LIN28B to further support the notion that LIN28B levels are reduced.

7. Fig. S1. Can the authors specify in this figure or elsewhere how many “onco-exaptation” events were seen in normal samples? This is important to get a sense for the background level of events that are being filtered – presumably, there should be few such events in normal samples as they would be predicted to be functionally deleterious.

(1.10) To help answer this and other questions related to alternative transcripts that were filtered out, we have included Supplementary Table 2 that includes all TE-derived transcripts present in at least 1 sample (tumor or normal) before filtering was applied. Before our filtering for highly tumor enriched transcripts, we start with 625 potential TE-derived alternative transcripts, 304 of which were exclusively present in tumor samples. There are only 5 events that are only present in normal samples, and they detected in only 1 or 2 samples each. Once we apply our 10x tumor-enrichment filter and assure candidates are present in at least 4 samples, only 5 of the remaining examples are present in any normal samples, and those samples are reproduced in Response Table 1 below.

<table>
<thead>
<tr>
<th>TE</th>
<th>Gene</th>
<th>Number of Tumor Samples</th>
<th>Number of Normal Samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1PA2</td>
<td>SYT1</td>
<td>901</td>
<td>2</td>
</tr>
<tr>
<td>HERVH</td>
<td>SLCO1B3</td>
<td>779</td>
<td>1</td>
</tr>
<tr>
<td>L1PA2</td>
<td>XCL1</td>
<td>559</td>
<td>2</td>
</tr>
<tr>
<td>Tigger3a</td>
<td>ARID3A</td>
<td>362</td>
<td>2</td>
</tr>
<tr>
<td>---------</td>
<td>--------</td>
<td>-----</td>
<td>---</td>
</tr>
<tr>
<td>MER5B</td>
<td>GPM68</td>
<td>126</td>
<td>1</td>
</tr>
</tbody>
</table>

Response Table 1: Candidates with expression in normal samples.

8. *Fig S4. Can the authors show CAGE-Seq or ATAC Seq tracks for additional onco-exaptation events and confirm that in those cases also, peaks are seen in the TE-promoter but not the native promoter?*

(1.11) We thank the reviewer for this suggestion on validating additional candidates. To address this suggestion, we analyzed FANTOM5 promoter atlas, which mapped transcription start site of numerous human tissue and cancer cell lines (more detail in Response 3.7). In FANTOM5, we report that five out of the top 10 candidates TEs were annotated to be a promoter. To verify the transcript derived from TE FANTOM5 peaks splice into appropriate oncogenes, we performed paired-end CAGE-seq on H727 lung carcinoid cell line that is predicted to have L1PA2-SYT1 and Tigger3a-ARID3A onco-exaptation candidates. In H727, we detect CAGE signal in L1PA2 and MLT1D (TE flanking Tigger3a). We mapped the paired reads that initiate from these TEs to illustrate that the mate reads splice into downstream exon of the predicted oncogenes, substantiating the discovery that these TE act as alternative promoter for these oncogenes (Response Figure 5c,d). Furthermore, the H727 CAGE-seq revealed that for ARID3A oncogene, majority of the transcript is derived from the TE, while half of the transcripts are derived from L1PA2 for the SYT1 oncogene. In both cases, the TEs contribute to significant transcriptional activity of the oncogenes.
Response Figure 5. Validation of onco-exaptation candidates in lung cancer cell line. a, WashU Epigenome browser view of CAGE-seq and mate-paired reads where the forward read initiates from AluJb and the reverse read ends in the gene body of LIN28B in H1299 and b, H838. c, WashU Epigenome browser view of H727 CAGE-seq and mate-paired reads that where the forward read from L1PA2 and the reverse read ends in the gene body of SYT1. d, WashU Epigenome browser view of H727 ARID3A 5’RACE, H727 CAGE-seq and mate-paired reads where the forward read initiates from Tigger3a/MLT1D and the reverse read ends in the gene body of ARID3A. e, WashU Epigenome browser view of H727 CAGE-seq over SYT1 promoters. f, WashU Epigenome browser view of H727 CAGE-seq over ARID3A promoter.
9. Fig S6f. The authors show that KO of the LIN28BP de-represses let-7 (Fig S6e) but does not cause decreased proliferation of K562 (Fig S6f). This is somewhat surprising and distinct from what is typically seen with LIN28B knockdown in this context. It is also different from what the authors show with AluJB2-LIN28B KO in H838 and H1299 cells. Are the authors able to explain this result if the assumption is that AluJB2-LIN28B and native LIN28B share the same function?

(1.12) The intended purpose of K562 CRISPR KO was to examine whether potential off-target effects of CRISPR activity from targeting TEs could undermine cellular fitness. Our results showed that the genetic deletion of AluJb element in K562 did not impact let-7 levels and proliferation thus suggesting that off-target impact is minimal to fitness. We were also surprised to find that although deletion of canonical LIN28B promoter led to increase in let-7 levels, yet the proliferation was affected minimally. One explanation could be that we selected K562 clones that evolved during clonal expansion to escape let-7 mediated control. Another interesting observation is that K562 expresses a shorter isoform of LIN28B (Supplementary Fig. 8) that is not present in the two lung cancer lines. Currently, it's still unclear what the biological impact and function of this particular isoform of LIN28B. The shorter isoform was first identified by Guo et al. 2006 in HCC samples. Although missing the first 70 amino acids from the canonical LIN28B, the short LIN28B contains the zinc finger domain responsible for binding to let-7. Guo et al. 2006 overexpressed short LIN28B in MCF-7 and reported no effect on cell growth. Viswanathan et al. 2009 knocked down LIN28B expression via shRNA and reported impaired proliferation and ability to form colonies in K562. However, the shRNA targets exon 3, which is present in both canonical and short LIN28B transcript, thus the phenotype is associated with significant decreased expression of both isoforms of LIN28B. It's difficult to parse the isoform-specific function from this monumental work. Recently, Mizuno et al. 2018 uncovered that short LIN28B antagonizes the canonical LIN28B function by acting as a miRNA sponge for let-7, consequently preventing let-7 degradation in colon cancer cell lines. In K562, one hypothesis could be that the short LIN28B isoform, in the absence of canonical LIN28B, can act as a let-7 sponge to prevent let-7 mediated suppression of key oncogenes, such as MYC. Further work is necessary to substantiate these claims, but we believe that these experiments are out of scope for the current manuscript. We have relabeled the Western blot in the Supplementary Fig. 8d to draw attention to the smaller LIN28B isoform.

10. Refs 31-32. Authors cite these manuscripts for evidence that LIN28B represses let-7 processing leads to upregulation of let-7 target oncogenes. However, these are not the key manuscripts in this field. There is also not much discussion in the text of why the authors selected this gene for further functional validation, and what the connection is between LIN28 and let-7. Since this is a significant focus of the manuscript would suggest expanding the text on this and citing one or more of the relevant original papers (Newman et al, RNA, 2008; Viswanathan et al, Science, 2008; Rybak Nat Cell Biol 2008; Heo Mol Cell 2008; Viswanathan Nat Genet 2009; Johnson Cell 2005, etc)

(1.13) We are conscious about the large breadth of influential literature on describing LIN28B’s role in cancer and thank the reviewer for guiding us towards the key literature that shaped the LIN28B field. We incorporated the relevant original papers into the manuscript and expanded in the manuscript that we pursued this candidate for the intriguing nature of onco-exaptation mechanism and potent functional consequences.

11. Would suggest more detailed descriptions in the figure legends.
We again apologize for providing sparse figure legends. We have added more detail on the biological traits and the assays used to quantify each trait.

12. Statistics/replicates should be specified per Nature Genetics guidelines.

For each experimental assay, we have included the number of replicates and the statistical test utilized in each method section and also mentioned the statistical test in figure legends for clarity.

Reviewer #2:

Remarks to the Author:

An emerging body of evidence suggests that transposable elements function to regulate gene expression in developmental, physiological and pathological conditions. Yet the importance of retrotransposons in the pathogenesis of cancer is largely unexplored. In this manuscript, Jang et al. performed the first genome-wide analyses on retrotransposon-driven oncogene expression in 15 tumor types across thousands of human tumors, demonstrating the prevalence of TE-derived cryptic promoters that drive oncogene overexpression. The authors went on to demonstrate that one of the onco-exaptation AluJb-LIN28B produces a protein 22aa larger than the canonical LIN28B. Knockout of AluJb promoter leads to silencing of AluJb-LIN28B expression and inhibition of tumor cell growth. This manuscript gives us a first glimpse of the important roles TE may play in cancers, and characterized an interesting case where oncogene expression is induced through derepression of TEs during tumor development.

This paper reported several exciting and important findings, and would be of general interests to both the retrotransposon field and the cancer field. The genomic analyses of retrotransposon-dependent oncogene expression is particularly novel, and would be a valuable resource for the field. The characterization of AluJb-LIN28B is interesting and well documented, yet overall, the studies could use more depths and rigor.

Overall, I think this is an exciting story and should be considered for publication by Nature Genetics after addressing the following comments.

(2.1) We want to thank the reviewer for the encouraging comments on our work! By responding carefully to the reviewer’s comments we hope to have provided more depths and rigor in our analysis.

Major comments
1. In Figure 1d, please label which eight candidates are predicted to form in-frame transcripts. In addition, overall what is the percentage of onco-exaptation transcripts form in-frame transcripts and likely produce proteins? Are there cases where TE-derived oncogene transcripts yield defective protein products? What is the relative ratio between onco-exaptation transcripts and the corresponding canonical oncogene transcripts? Are they mutually exclusive when comparing control and tumor samples?

(2.2) We appreciate the constructive suggestions, and many of these points have helped shape our new round of analysis that focuses on quantifying expression distribution and utilizing that value to help distinguish candidates. First, we have labeled the in-frame candidates in Figure
and this label is also present in Supplementary Tables 2 and 3 for every candidate. Overall, for all candidates the percentage that are predicted to code in-frame proteins is 70.5% (91/129), and for candidates accounting for >50% of expression the percentage is 82.1% (46/56). There are 3 candidates that are predicted to be out-of-frame, but it is unclear what the functional significance of these would be.

We have included a panel of boxplots in Figure 1d of the main paper that displays the distribution of the percentage of total transcripts the onco-exaptation candidate is accounting for. In addition, we have included the average fraction of total transcription for every candidate in Supplementary Tables 2 and 3. All candidates are required to account for at least 25% of the transcripts for a particular oncogene. We set a stringent threshold of requiring 10x enrichment in tumor versus normal samples. Thus, for every 1 normal sample, a candidate would need to be present in 126 tumor samples to pass the tumor-enrichment filter. Only 5 of our candidates have expression in any of the normal samples, and there is a total of 8 normal samples with any candidate (Response Table 1). The rest of the candidates were only found in the tumor samples and not in the normal samples.

<table>
<thead>
<tr>
<th>TE</th>
<th>Gene</th>
<th>Number of Tumor Samples</th>
<th>Number of Normal Samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1PA2</td>
<td>SYT1</td>
<td>901</td>
<td>2</td>
</tr>
<tr>
<td>HERVH</td>
<td>SLCO1B3</td>
<td>779</td>
<td>1</td>
</tr>
<tr>
<td>L1PA2</td>
<td>XCL1</td>
<td>559</td>
<td>2</td>
</tr>
<tr>
<td>Tigger3a</td>
<td>ARID3A</td>
<td>362</td>
<td>2</td>
</tr>
<tr>
<td>MER5B</td>
<td>GPM68</td>
<td>126</td>
<td>1</td>
</tr>
</tbody>
</table>

Response Table 1: Candidates with expression in normal samples.

(2.1) TEs are repetitive in nature, and a large number of results are based on next generation sequencing data, whose mapping to TEs could yield ambiguity. Hence, it is important for the authors to provide experimental evidence supporting their RNA-seq findings. It would be great if the authors could pick several top candidates, including AluJb-LIN28, to generate 5’RACE data to validate the prediction of TE as cryptic promoters.

(2.3) We are in complete agreement with the reviewer that the repetitive nature of TEs can lead to mapping issues. For this reason, we performed our RNA-seq analysis with only uniquely mapped reads. We recognize that this analysis biases against young TEs that share high sequence similarity. In addition, we also performed CAGE-seq analysis with only uniquely mapped reads to be confident in our detection. As addressed in Response 1.11, we validated another two candidates out of the top ten with CAGE-seq in H727. Using mate-pair info, we report that reads that initiate from candidate TEs splice into predicted downstream oncogenes, including the AluJb-LIN28B candidate (Response Figure 5a,b). To further verify fidelity of CAGE-seq results, we performed 5’RACE on ARID3A in H727. As shown in Response Figure 5d, we report that the CAGE-seq peak and 5’RACE results are congruent thus we believe that paired-end CAGE-seq can accurately identify onco-exaptation events.

(2.4) We thank the reviewer for bringing up an interesting phenomenon in the transcriptional and translational control of AluJb-LIN28B in cancer. We believe that CAGE-seq provides an unbiased and accurate representation of relative abundance of transcript isoforms. CAGE peak
signals represent number of transcripts that are transcribed from a certain promoter, and thus we can compare CAGE peaks to quantify relative ratio from one transcript to another. Using CAGEr, we report that AluJb-P transcribes 13.7x and 2.4x more LIN28B transcripts than the canonical promoter in H1299 and H838 respectively (Response Table 2). Although the transcript level can be informative, the function of the gene is relevant to the amount of protein that is translated. To our surprise, the LIN28B Western blot revealed that in both H1299 and H838, the dominant species of LIN28B protein is the AluJb-LIN28B isoform. This suggests that there are post-transcriptional and/or post-translation mechanisms favoring the production of the onco-exapted version of LIN28B protein in lung cancer cell lines that have yet to be elucidated.

<table>
<thead>
<tr>
<th></th>
<th>CAGE peak size (tpm)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>H1299</strong></td>
<td></td>
</tr>
<tr>
<td>AluJb-P</td>
<td>370.69 tpm</td>
</tr>
<tr>
<td>Canonical Prom</td>
<td>27.26 tpm</td>
</tr>
<tr>
<td><strong>H838</strong></td>
<td></td>
</tr>
<tr>
<td>AluJb-P</td>
<td>195.26 tpm</td>
</tr>
<tr>
<td>Canonical Prom</td>
<td>81.26 tpm</td>
</tr>
</tbody>
</table>

Response Table 2: CAGE peak tpm of AluJb-P and canonical LIN28B promoter in H1299 and H838.

4. **Does the addition of 22 N-terminal amino acids cause gain-of-function for LIN28?** The author can examine this by overexpressing AluJb-LIN28 or canonical LIN28 in AluJb KO cells to see whether both rescue the tumor phenotype and reduce let-7 expression to a similar extent.

(2.5) This is a great question about potential neo-function of TE-derived peptides! Based on the rescue experiments, AluJb-LIN28B and canonical LIN28B repress let-7 levels to shift the balance of LIN28B/let-7 towards sustaining oncogene expression. Please refer to Response 1.3 and 1.8 for more details. Whether the additional AAs provide alternative function is still unclear. Guo et al. 2018 suggests that the 22 AA might be responsible for higher protein stability, but did not pursue validation experiments. Although the potential neo-function of the 22 N-terminal AAs in AluJb-LIN28B is quite intriguing, we believe this particular question is not within the scope for this paper where we want to focus on the presence and prevalence of onco-exaptation events across numerous cancer types.

5. **In Figure 3f, the interpretation of the results needs more experimental control.** Was the WT sample bulk cells or single clone? Tumor cell lines are largely heterogeneous, with strong oncogenic potential. When single-cell clones from a tumor cell line were established, they sometimes lose the ability to graft and form tumor in vivo as a result of clonal expansion. It’s unclear to this reviewer if the AluJb1 #1 and AluJb2 #1 lines fail to grow tumor in vivo because of decreased LIN28 or because of clonal selection. To make this result truly convincing, the authors need to use multiple WT single-cell clones as control, and demonstrate that LIN28 overexpression could restore the oncogenic potential in vivo using AluJb1 deficient cells.

(2.6) We thank the reviewer for this critical insight! As mentioned in Response 1.3, we agree that clonal expansion could negatively impact the ability of cancer cells to properly graft and grow tumor in vivo. We redesigned the xenograft experiment to include LIN28BP KO clone, which went through same experimental conditions and clonal expansion as AluJb KO clones. LIN28BP KO clone showed slower growth in vivo than WT cells, suggesting that clonal expansion can indeed lead to lower fitness in vivo. However, all LIN28BP KO samples led to properly growing tumors while no visible tumors were formed in the initial month of inspection.
for AluJb KO samples. With this result, we believe that the loss of AluJb-LIN28B could be causal for extremely slow tumor growth in H1299 lung cancer cell lines.

Reviewer #3:

Remarks to the Author:

Several examples have been reported in the literature of human transposable elements (TEs) being transcriptionally activated in cancers and driving ectopic expression of oncogenes by providing alternate promoters and splicing into the gene. This intriguing phenomenon has been termed “oncoexaptation” in a review on the topic. To date, no comprehensive screen of many cancer types has been published to gauge the prevalence of this mechanism of oncogene expression in human malignancy. In the first part of the manuscript, Jang et al perform such a screen on large RNA-seq TCGA datasets from 15 cancer types and, using quite stringent criteria, report 260 TE promoter activations involving 174 known oncogenes in ~3500 tumors. There is an enrichment for LTRs in this dataset and most of the highly prevalent cases preserve the ORF of the gene and are associated with overall higher gene expression. The rest of the paper focuses on an interesting case of an upstream alu element promoting expression of LIN28B (which has known oncogenic properties) in some primary cancers and in some tested lung cancer cell lines. They conduct an analysis of this case by performing promoter assays, identifying potential TFs that bind to the alu, documenting the importance of DNA methylation in controlling activity of the alu, using CRISPR to delete the alu and showing effects on let7 expression and marked effects on tumorigenicity in xenograft models. This second part of the paper is very well done and comprehensive and shows the functional importance of the transcriptional activation of that alu element. The number of oncoexaptation cases that have been well studied and validated is still quite small, so this study is an important contribution, although unfortunately the specific case dissected here is not completely novel (see point 1 below). The first part on the bioinformatics screen is more briefly described and would benefit from additional analyses as outlined below.

Specific comments and concerns with the manuscript are:

1. Many of the findings regarding the alternative isoform of LIN28B driven by the Alu element have been recently reported by Guo et al (PMID: 29466730) “A LIN28B tumor-specific transcript in cancer”, published in Feb. 2018 in Cell Reports. Although Guo et al did not notice (or mention) the alu origin of the alternative transcription start site, they surveyed TCGA data, mapped the TSS, conducted promoter assays, showed a dependence on DNA methylation, examined effects on let7 and reported slower growth in culture and loss of tumorigenicity in cell lines knocked down for expression of this isoform. Hence the novelty of the results on alu-LIN28B is reduced. Although Jang et al are aware of the Guo et al paper (cited as reference 29), it is curiously only briefly mentioned toward the bottom of page 4 on the promoter assays and near the top of page 7, when it is cited regarding the necessity of LIN28B for tumor formation in xenograft models. Jang et al need to more thoroughly discuss and acknowledge this previous paper and make a case for the novelty of their results.

(3.1) We completely agree with the reviewer. In General Response 4 we carefully address the similarity between the two studies as well as our unique contribution to establish the novelty of our results. In our revised manuscript we now designate a significant portion to discuss and acknowledge Guo et al. in our concluding remarks.
2. Please show the actual predicted amino acid sequence of the aluJb-LIN28B isoform. Only a cartoon is shown in figure S4b.

(3.2) As requested, we have included the peptide sequence of AluJb-LIN28B isoform and canonical LIN28B isoform in Supplemental Fig. 6c.

3. Given the above, the primary novelty and excitement of this manuscript is arguably the first part, which is the first large scale comprehensive screen for TE oncoexaptation events and should be of wide interest. This section was covered only briefly and the data could have been analyzed in more depth to make more points. For example, the marked prevalence of the first two cases in figure 1d, which have not been reported before, looks very interesting but the authors do not even mention them in the main text. What is known about these genes? Are these isoforms predicted to make the same or an altered protein? For all the cases shown in figure 1d, please include information on predicted protein sequence – ie, the same protein, novel amino acids added or no protein due to frame shift or no ATG.

(3.3) We thank the reviewer for their suggestions and recommendations on how we could improve and enhance the bioinformatics analysis to make it of greater value to the field. With enhancements that focus on incorporating expression and increased stringency in selecting candidates, the members of our top 10 have changed slightly, but we have incorporated the suggestions of the reviewer to these candidates. First, we have included frame prediction information in Figure 1d. We also include the in-frame, noncoding, or out-of-frame protein prediction in Supplementary Tables 2 and 3. In addition, we have included additional data on two other candidates: L1PA2-SYT1 and Tigger3a-ARID3A that we have CAGE-seq data as proof of the promoter being from the transposable element. In addition, we have included the following discussion highlighting a few of the top 10 candidates:

Based on splicing patterns, eight of these candidates were predicted to form in-frame transcripts that conserve protein sequence, suggesting preservation of canonical oncogene function. Newly discovered onco-exaptation candidates include isoforms of genes such as SALL4 and LIN28B that have recently emerged as potent cancer drivers.21–24 Additionally, the L1PA2-derived isoform of SYT1 occurs in more than 10% of all tumors, suggesting that it could be an important cancer marker. While investigating transcript-level abundance of candidates, we found that many of the onco-exaptation events were driving a significant fraction of expression of the oncogene; some greater than 90% (Fig. 1d & Supplementary Fig. 3). Furthermore, we report that half of our top candidates were associated with worse survival in at least 1 cancer type (Supplementary Fig. 4). As an example, we show that the HERVH-SLCO1B3 transcript, a previously characterized onco-exaptation event, is abundant across various cancer types, highly expressed, and associated with worse prognosis.25 These findings imply that TEs are not only associated with oncogene activation but also contribute significantly to overall oncogene expression and oncogenic potential.

4. In the experimental set-up for detecting “onco-exaptations”, the authors report 174/702 oncogenes meet their criteria but don’t provide a context if this is more or less than expected or what this means biologically. For instance is this more or less than the number of “tumour suppressor-exaptations”, or “mitochondrial exaptations” or any other gene set which may be associated with tumorigenesis? Are oncogenes actually enriched within the set of cancer-specific transcripts initiated by TEs?
We thank the reviewer for this important consideration when evaluating whether this is mechanism specific to genes promoting oncogenesis. Due to the computational burden of attempting to expand the analysis to all genes and the ambiguity of whether certain genes are important in oncogenesis and tumor suppression, we decided to focus on tumor suppressor genes and oncogenes defined in specific databases. We have included a list of 201 tumor suppressor genes identified from COSMIC in our second round of analysis to see if this mechanism does indeed have a preference for oncogenes\(^2\). Based on the filters used in defining our initial candidate list that are highly tumor enriched, we found 129 (106/702 genes 15.1\%) total events for oncogenes and 33 total events (25/201 genes 12.4\%) for tumor suppressor genes. These raw percentages of genes do not show statistically significant difference; however, there are differences that arise when looking at the overall impact of the events on gene expression. First, we have graphed the average expression percentage of each candidate for both the oncogenes and tumor suppressor genes; there is a statistically significant greater fraction expression driven by the oncogene transcripts versus the tumor suppressor gene transcripts (P<.001) (Response Figure 2a). When looking at those that on average drive >50\% of expression, 56 (51/702 genes 7.3\%) total candidates exist for oncogenes, and only 8 (7/201 genes 3.5\%) total events exist for tumor suppressor genes. This is greater than a 2x enrichment and shows that tumor-suppressor gene candidates are less likely to drive a significant percentage of expression. Furthermore, the maximum number of tumors that have a tumor suppressor gene candidate with >50\% of total expression was 94 (mean = 34.4) whereas the maximum tumors for an oncogene candidate was 901 (mean = 68.3); however, the difference in means was not statistically significant (Response Figure 2b). These results indicate that exaptation events are more likely to contribute significantly to the expression of oncogenes rather than tumor suppressor genes.

**Response Figure 2.** Difference in expression contribution and number of samples in between the oncogene and tumor suppressor gene oncoexaptation events. **a,** A boxplot of the average proportion of total expression contributed by each oncogene and tumor suppressor gene candidate in the tumor samples they are present in. A dot plot is overlaid with each dot representing one candidate. There is a statistically significant difference in fraction of expression (Welch’s t-test: P<.001) **b,** A dot plot of the number of tumor samples that each candidate is present in that only includes candidates that on average account for >50\% of total expression of their gene.
Though the evidence is quite compelling, we would like to note that there are limitations in making this conclusion. First, we are not sure what the mechanism of the tumor-suppressor transcripts and potential protein-products are. These could serve as dominant negative isoforms that are actually leading to suppression of activity of the gene; previous examples of splicing variants with shorter isoforms have been found to have dominant negative activity on tumor suppressor genes\(^3\,^4\). Furthermore, there is a grey area between oncogene and tumor suppressor gene definition where some genes can have dual roles; a recent paper characterized 83 such examples\(^5\), 2 of the 8 tumor suppressor exaptation events with >50% expression are on this list (BRCA2 and PPARG). Thus, we elected to not include this enrichment analysis in the final manuscript since we believe we do not have enough evidence with our current study to confidently publish this claim.

5. Related to the above point, after accounting for differences in the number of samples, how often do you see TE transcription start sites for the 702 oncogene list in normal samples?

(3.5) To help answer this and other questions related to alternative transcripts that were filtered out, we have included Supplementary Table 2 that includes all TE-derived transcripts present in at least 1 samples (tumor or normal) before filtering was applied. Before our filtering for highly tumor enriched transcripts, we start with 625 potential TE-derived alternative transcripts, 304 of which were exclusively present in tumor samples; 321 were present in at least 1 normal sample. There are only 5 events that are only present in normal samples, and they detected in only 1 or 2 samples each. Once we apply our 10x tumor-enrichment filter and assure candidates are present in at least 4 samples, only 5 of the remaining examples are present in any normal samples, and those samples are reproduced in the Response Table 1 below.

<table>
<thead>
<tr>
<th>TE</th>
<th>Gene</th>
<th>Number of Tumor Samples</th>
<th>Number of Normal Samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1PA2</td>
<td>SYT1</td>
<td>901</td>
<td>2</td>
</tr>
<tr>
<td>HERVH</td>
<td>SLCO1B3</td>
<td>779</td>
<td>1</td>
</tr>
<tr>
<td>L1PA2</td>
<td>XCL1</td>
<td>559</td>
<td>2</td>
</tr>
<tr>
<td>Tigger3a</td>
<td>ARID3A</td>
<td>362</td>
<td>2</td>
</tr>
<tr>
<td>MER5B</td>
<td>GPM68</td>
<td>126</td>
<td>1</td>
</tr>
</tbody>
</table>

Response Table 1: Candidates with expression in normal samples.

6. The term exaptation suggest that selection is favoring these transcripts over other non-oncogene transcripts. Is there evidence that cancers which express the onco-exaptations reported have better/worse survival in the TCGA cohorts? The Guo et al paper reported worse survival for HCC cases expressing the alu-LIN28B transcript.

(3.6) We thank the reviewer for noting this important consideration, and thus we evaluated our top 10 candidates to see if there was any prognostic value of the presence of these candidates. We took available overall survival data from TCGA and compared patient samples with the candidates present versus those without the candidate present. There were 8 examples of where a candidate was significantly prognostic in a specific cancer (p<.05) and was present in at least 20 samples in that cancer type (Response Figure 6). All 8 indicate a worst progress, including AluJB-LIN28B in LIHC (Liver Hepatocellular Carcinoma) which is consistent with Guo et al. Due to difficulties of detection of 5’ ends with RNA-sequencing data, we do appreciate there are limitations to this analysis; there are a lot fewer samples with the candidate than without for most of these comparisons. Furthermore, we cannot establish that this prognostic effect to this specific isoform rather than just higher expression of the oncogene in general.
However, this data is still quite convincing that the presence of these candidates is prognostic and warrants further in-depth investigation in patient samples to evaluate their impact.
Response Figure 6. Overall survival impact of oncoexaptation candidates. Kaplan-Meier Curves for the 8 examples of where a top 10 candidate was significantly prognostic in a cancer (p<0.05) based on log-rank statistical test. The red line in each graph represents patients where the candidate was found to be present, and in blue line represents all the patients where the candidate was not detected. All were found to negatively impact overall survival.
7. For the table of onco-exaptations, how is error measured? How much confidence can one have in such a table for presenting biologically relevant and real signal over background noise? Related to this, how many of these are initiating transcription as opposed to being artifacts of incomplete 5’ assembly? One metric that can easily be added (as a column in Table S2) to begin to address this important issue is to indicate which TEs in this table overlap a known PHANTOM5 CAGE peak. While this doesn’t prove that a TE is active in exactly the RNAseq sample used, it would demonstrate that these TEs are at least potentially capable of initiating transcription. This is important since the authors perform little validation of their bioinformatics method.

(3.7) We thank the reviewer for their constructive comments that helped inspire many of the changes in our pipeline to attempt to capture the true 5’ end of the transcript as much as possible from the RNA-sequencing and remove noise. Originally, our pipeline was designed to simply screen for events based on initial assembly and read information interrogation. Though this method was able to capture many interesting candidates, it did not allow for more advanced expression analysis. Thus, we made significant changes to incorporate expression information in evaluating presence of candidates which is described in our general response and diagramed in Response Figure 7a. Here, we would like to emphasize the rigorous steps taken to attempt to be as accurate as possible in finding accurate events. First, in our initial screening for candidates (Step 3) we have increased the number of read support needed to 10 uniquely mapping reads needing to be within the TE as well as needing 1 uniquely mapping read to span the distance from the TE to the gene for all candidates. Thus, we are more stringent in our initial call for candidates.

We agree that incomplete 5’ degradation could be leading to some of our candidates, and in fact, we did see patterns of multiple transcripts from adjacent TEs that looked like artifacts of incomplete 5’ assembly. Thus, we incorporated a step (Step 4) that would merge these separate isoforms to choose the most 5’ one and remove the transcripts that might have incomplete 5’ ends. In Response Figure 7b, we showcase an example of Tigger3a-ARID3A and the results of merging using the Interactive Genome Viewer.10

We next use stringtie for transcript-level expression analysis, and again we provide stringent filters of a candidate needing to contribute to at least 25% of total expression of the oncogene and at least 1 FPKM expression of the oncogene to state that a candidate is present. In addition, we look to make sure that in every sample that a candidate is present, there is coverage of the splice junction between the TE exons and the gene exons.

FANTOM5 is an excellent resource to add further evidence of our results, and thus we overlapped the TE initiating our identified candidates with FANTOM5 peaks. We found that 21 of our 129 candidates (16.3%) had a FANTOM5 peak including the candidates. We have labeled the candidates with FANTOM5 peaks in our supplementary tables. We have also done more thorough analysis of this data in Response 3.8. However, we would like to emphasize that FANTOM5 is not comprehensive for TE-promoters, and we still have confidence in our candidates that do not have FANTOM5 peaks. FANTOM5 cancer analysis was limited to a selection of 253 cancer cell-lines and 13 primary cell samples that were all from ovarian cancer patients; the TCGA data is much more diverse and comprehensive. In addition, whereas TCGA has 50 bp and 75 bp reads (most samples were paired-end), FANTOM5 was limited to single-end short reads that could be as small as 20 bp using the single-molecule Helicos sequencer.11 This would severely limit their mappability to TEs. For example, the LTR promoter of FABP7 has been previously published and confirmed experimentally; yet, this is not present as a peak in FANTOM512. Overall, we believe the addition FANTOM5 data provides additional confirmation of a candidate’s validity.
Response Figure 7. RNA-sequencing analysis pipeline. a. Diagram of our bioinformatics pipeline that was used to detect onco-exaptation events using the RNA-sequencing data of 15 cancers from The Cancer Genome Atlas. Further detail about each step is provided in our manuscript Methods section. b. Interactive Genome Viewer visualization of step 4 of our pipeline ("Merge Transcripts"). The first track ("Reference") is the reference sequence of ARID3A, and the second track ("Initial Transfrags") shows the initial candidates. Two seem to have incomplete 5' ends that land in TEs and thus are annotated as being separate oncoexaptation events initially. After merging, the "Merged Transcripts" track shows how the incomplete 5' assembly candidates are removed and the longer exon 1 isoform is kept.
8. Related to the above, while the onco-exaptations reported here were not detected in any of the TCGA datasets from normal tissues, they are not necessarily “cancer-specific” because the number of normal samples was much less than the cancer samples. Also, some of the TEs have substantial CAGE tags from normal tissues, such as the Tigger3a element near ARID3A. The authors should comment on this issue.

(3.8) We agree that the language of cancer-specific can be a bit vague considering the presence of these promoters in other tissue types and the disparity in tumor versus normal tissues. The 10x enrichment requirement in tumor samples is a high-bar for candidates to pass; for every 1 normal sample a candidate is present in it has to be present in 126 tumor samples. However, it is evident that TCGA does not represent every normal tissue type and that some of these candidates might be present in other tissues. To help address this fact, we took the 21 candidates (20 unique TEs) with FANTOM5 peaks and plotted their average expression in cancer, adult, and fetal tissues available from FANTOM5 (Response Figure 8a). Though these promoters can be active in fetal and adult tissues, we see a significant increased expression in the cancer cell line tissues. Next to this plot, we have a table that displays the number of samples and types of tissues incorporated into this analysis; this analysis includes 771 adult samples that represent a diverse array of 38 tissue types. Looking more closely at the few fetal tissues with expression of candidate promoters, we see that the placenta has the highest number of candidates expressed (Response 8b). Endogenous retroviruses are highly active in the placenta and contribute important genes, act as enhancers, and also act as alternative promoters for genes\textsuperscript{13,14}. When we further subset the data by candidate, we see that the majority have higher expression in the cancer cell line tissues rather than the normal adult tissues including Tigger3a-ARID3A (Response 8c). Indeed, a previous review on onco-exaptation noted that some of these TE-promoters were found to be spuriously lowly expressed in normal tissues\textsuperscript{15}. Though these promoters might not be exclusively cancer-specific, we see that indeed they are enriched in tumor samples and that they have consistently higher expression in tumor samples. To address the fact that these promoters can be active in some germ line and adult tissues, we have changed the language in the manuscript to emphasize high enrichment rather than complete cancer-specificity.
Response Figure 8. FANTOM5 expression of candidates. 

a, A boxplot of the average expression per tissue of the 20 TE-promoters present in FANTOM5. They have been separated out by the tissue classes of cancer cell lines, adult tissues and fetal tissues. A table is included next to the plot noting the number of individual samples and the number of tissue types present for each class.  
b, The number of candidates present in the 8 fetal tissue types with the highest expression.  
c, For each candidate, we have a boxplot of the average expression per tissue in cancer cell line and adult tissues. Highly expressed outliers in the adult tissues are labeled. (Welch’s t-test: *=P<.05, **=P<.01, ***P<.001)
9. Perusal of Table S2 reveals a very interesting result that the authors are advised to discuss. In their list of potential oncogenes, they included 4 cases that were covered in a review on oncoexaptation (ref. 4). One of these is SLCO1B3. This gene is listed in Table S2 (row 14) but, remarkably, the putative TE promoter found here is an LTR12C element located upstream of the gene, which is not the same as the intronic LTR promoter reported previously and discussed in reference 4. The case reported previously as being quite common in colon cancer was not detected in this study. Do the authors have an explanation for this? Such usage of different TEs to upregulate or deregulate the same oncogene in different tumors is intriguing and worth discussing. The authors could mention the paper, cited here as reference 15, that showed different LTRs within the ERBB4 gene are activated in different cases of Anaplastic large-cell lymphoma (also discussed in ref 4).

(3.9) We thank the reviewer for this excellent suggestion to increase the thoroughness of our analysis. The use of multiple TEs effecting the same gene is a very interesting phenomenon, and this analysis provides a way to see how often this can occur. To this end, we searched for examples of multiple TEs activating the same oncogene and leading to producing of an in-frame, coding transcript. We found 8 oncogenes with this phenomenon: ALK, CTNND2, ECT2, GRM1, MCF2L, MMS22L, MRE11A, RINT1. We have included the following discussion about this phenomenon in the main text and cited the papers mentioned:

Furthermore, for eight oncogenes, we observed multiple different TEs activating an in-frame isoform of the same gene (Supplementary Table 3), a phenomenon that had only been described for one oncogene\textsuperscript{16}. These additional examples argue for a cancer epigenetic evolution model as previously described\textsuperscript{4}.

We discuss this further in Response 3.10, but many previously published candidates are filtered out due to our stringent thresholds on tumor enrichment. Previously, the SLCO1B3 candidate was filtered out since it was found in 1 normal sample, but now we have featured it as one of our top candidates Figure 1d.

10. On a related point, it would be useful to discuss whether this bioinformatics method actually detected previously known oncoexaptation cases (FABP7, IL33, IRF5, SLO1B3) that the authors included in their screening and if not, why not. Another perusal of Table S2 shows IL33 twice on the list (line 65 and line 243), but again, different TEs were found than the published one reported in colon cancer (ref. 15), which should be commented upon.

(3.10) Our previous analysis focused on candidates that were solely found in tumor samples and in no normal samples, and thus many of the previously published candidates were excluded. We have done a reanalysis, and we now include a supplementary table that contains all candidates in addition to the one that only has the tumor-enriched candidates. Thus, readers will be able to see every TE-derived transcript that was found in this analysis. From our bioinformatics pipeline, we were able to find 6 of the 8 previously published candidates. Below, we have a brief summary of each candidate that was found:

<table>
<thead>
<tr>
<th>Candidate</th>
<th>Number Tumor Samples</th>
<th>Number Normal Samples</th>
<th>Candidate Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>THE1B-CSF1R</td>
<td>472</td>
<td>45</td>
<td>Filtered out due to lack of 10x tumor enrichment</td>
</tr>
<tr>
<td>L1PA2-MET</td>
<td>1018</td>
<td>9</td>
<td>Filtered out due to lack of 10x tumor enrichment</td>
</tr>
</tbody>
</table>
One of the previously published candidates not found, MLT1C/MLTH2-ERBB4, was originally found in cancers not included in our study anaplastic large-cell lymphoma (ALCL)\(^\text{15}\). The other candidate that was not found, LTR16B2-ALK, was confirmed to be present in the TCGA data previously, but that publication had to use manual inspection of samples on a browser to confirm its existence and differentiate it from rearrangements\(^\text{16}\). Though this candidate was initially detected with assembly, our method filtered this candidate out due to a lack of a unique splice junction which can make accurate detection and expression calculation difficult. We have included the following statements in our manuscript to comment on the previously published examples:

In total, we analyzed 7,769 tumor samples and 625 tumor-matched normal samples (Supplementary Fig. 1b), which initially revealed 625 TE-oncogene chimeric transcripts; this list included 5 previously published examples of onco-exaptation (Supplementary Table 2).

11. On a minor point, please include a list of abbreviations for the TCGA tumor types. If this was included somewhere, it was not easy to find.

(3.11) We apologize for this oversight and have included descriptions in the methods section.

Reviewer #4:

Remarks to the Author:

Genome-wide surveys have uncovered transposons as a major source of gene regulatory elements but their functional impact on human health and disease remains poorly understood. Several recent studies suggest that transposons may contribute to gene dysregulation in cancer, by acting as cryptic regulatory promoters that drive oncogene expression in cancer, a process termed “onco-exaptation”. Previous studies have largely focused on specific types of cancer, and the extent of this phenomenon remains unclear. In this manuscript, Jang and colleagues perform a global survey of transposon-initiated gene expression across many cancer types. They perform a comprehensive meta-analysis of TCGA RNA-Seq datasets using a custom pipeline, and identify hundreds of high-confidence transposon-oncogene chimeric transcripts that show recurrent expression across a broad panel of different tumors.

The authors then focus on experimentally characterizing a specific example of an onco-exapted transcript, an AluJb element fused to the LIN28B gene, which is overexpressed in several lung cancer cell lines. Using CAGE-seq, WGBS-Seq, and ATAC-Seq on two lung cancer cell lines, they demonstrate that the AluJb element is normally silenced by methylation in somatic tissues, but becomes demethylated and accessible to transcription factors in cancer cells. The authors further investigate the regulatory activity of the AluJb element using a series of luciferase
reporter experiments, and find that several mutations altering transcription factor binding sites in the ancestral AluJb insertion were likely necessary for its promoter activity in cancer cells.

The authors next use CRISPR to characterize the oncogenic effects of AluJb-LIN28B transcription. They convincingly show that CRISPR-deletion of the AluJb element, but not the ‘native’ promoter of LIN28B, is sufficient to completely ablate LIN28B expression. As anticipated given the experimentally established roles of LIN28B in cancer, the AluJb knockout cells show a variety of phenotypes associated with less aggressive cancer, including reduced migration and growth rates. Intriguingly, AluJb knockout cells completely lose the ability to form tumors in a mouse xenograft model. Finally, the authors use CRISPR epigenome editing to either methylate or demethylate the AluJb locus in various cell lines, and show that this is sufficient to modulate AluJb-LIN28B transcription.

Overall this is a very well-done and comprehensive study that provides the strongest evidence thus far for the widespread occurrence of onco-exaptation of ERV promoters, across many types of cancers. The functional dissection of the AluJb-LIN28B transcript is elegant and convincing, but is missing a few key rescue experiments (see below). Once published, this manuscript is sure to on the top of reading list for scientists from diverse disciplines, including cancer biology, genome regulation, and transposable elements.

(4.1) We want to thank the reviewer for the encouraging comments on our work!

Major comments
1. The oncogenic properties of the AluJb-LIN28B transcript appears to have already been characterized by another recent study, which is cited but rather downplayed in the text (ref 29, Guo et al, “A LIN28B Tumor-Specific Transcript in Cancer” Cell Reports 2018). This reduces the novelty of the present study somewhat, although the previous study was recently published and did not note the transposon-origin of the promoter. The fact that this exact transcript has been previously characterized in the context of liver cancer should be more explicitly referenced and discussed in the text.

(4.2) We completely agree with the reviewer. In General Response 4 we carefully address the similarity between the two studies as well as our unique contribution to establish the novelty of our results. In our revised manuscript we now designate a significant portion to discuss and acknowledge Guo et al. in our concluding remarks.

2. Pg 2, line 63: the authors claim that tumors with “onco-exaptation events were among the samples with the highest associated oncogene expression”. Oncogene expression levels from each tumor sample is presented in the violin plots in Fig 1D and Fig S3. The first issue is that this statement should be statistically supported. Second, it is unclear whether the value associated with each data point is specific to the TE-gene chimeric transcript, or simply represents the total expression of that gene (e.g., counting both the normal and TE-derived promoter), which is likely if the data were extracted from the TCGA GDC portal. An alternative interpretation may be that certain oncogenes are already highly expressed from their native promoter in these samples, and spurious transcription is detected from TE promoters. In this case, it may be misleading to plot the combined expression if the TE-derived transcript represents a minority of the total transcripts.

(4.3) We thank the reviewer for pointing out how our statements on our violin plots could be misleading without more information given to the user. In fact, since our filters require a certain
expression level (>1 FPKM) and at least 25% total expression of the oncogene, it is biased for higher expressing samples and thus we cannot confidently make the claim that these events are in some of the highest expressing samples. We have changed the language in the manuscript to emphasize tumor-enrichment rather than tumor-specificity.

In addition, to clarify this point, we have added the fraction of total oncogene expression that the TE-gene chimeric transcripts account for as a part of our analysis. We have included boxplots in Figure 1d that show the distribution of fraction of total expression for the top candidates, and in Supplementary Table 2 and 3 we have included this value for every candidate so they can be evaluated based on their relative contribution.

3. The phenotypic assays of KO cells presented in Figs 3C-F are compelling, but somewhat noisy, as is expected when generating clones from single cell expansions. These experiments are missing rescue controls which would make the results much more robust. For example, the increased let7 expression (and reduced growth and migration) of KO cells should be rescue-able by overexpressing LIN28BP.

(4.4) We are grateful to the reviewer for this suggestion to make our results more robust. Following the reviewer’s suggestion we have performed rescue experiments. Please refer to General Response 2, Responses 1.3 and 1.8.

4. The complete lack of tumor growth from the AluJb knockouts presented in Fig 3f is dramatic, but results from the control Lin28bp knockout cells are conspicuously missing (as they are shown in all other panels).

(4.5) We apologize for the oversight with the xenograft experiment. We have repeated the xenograft experiment with a LIN28BP KO clone. Please refer to General Response 2, Response 1.3 and 2.6.

Minor comments
1. Fig 2B,C and Fig 3C-F: In these panels where results from H1299 and H383 are displayed side by side, it would be helpful to label the cell lines on the figure (info is presented in legend)

(4.6) We have labeled the figures with appropriate cell lines as requested.

2. There was some analysis comparing the AluJb-P sequence in the reference sequence with the consensus AluJb. Was the sequence region around and including the AluJb-P element Sanger sequenced from the cancer cell lines to verify there were no additional cancer-specific mutations?

(4.7) Since AluJb-P is sufficient and necessary for strong promoter activity based on luciferase reporter assays (Fig. 2c), we checked for additional cancer-specific mutations specifically in the AluJb element in both lung cancer cell lines. As requested by the reviewer, we Sanger sequenced the AluJb element and report no mutations within AluJb element in H1299 and H838.
Response References