Transposable elements drive widespread expression of oncogenes in human cancers

Hyo Sik Jang1,2,5, Nakul M. Shah1,2,5, Alan Y. Du1,2, Zeea Z. Dailey1,2, Erica C. Pehrsson1,2, Paula M. Godoy1,2, David Zhang1,2, Daofeng Li1,2, Xiaoyun Xing1,2, Sungsu Kim1,3, David O'Donnell1,2,4, Jeffrey I. Gordon1,2,4 and Ting Wang1,2*

Transposable elements (TEs) are an abundant and rich genetic resource of regulatory sequences1,2. Cryptic regulatory elements within TEs can be epigenetically reactivated in cancer to influence oncogenesis in a process termed onco-exaptation3. However, the prevalence and impact of TE onco-exaptation events across cancer types are poorly characterized. Here, we analyzed 7,769 tumors and 625 normal datasets from 15 cancer types, identifying 129 TE cryptic promoter-activation events involving 106 oncogenes across 3,864 tumors. Furthermore, we interrogated the AluJb-LIN28B candidate: the genetic deletion of the TE eliminated oncogene expression, while dynamic DNA methylation modulated promoter activity, illustrating the necessity and sufficiency of a TE for oncogene activation. Collectively, our results characterize the global profile of TE onco-exaptation and highlight this prevalent phenomenon as an important mechanism for promiscuous oncogene activation and ultimately tumorigenesis.

The elucidation of mechanisms behind oncogene activation has been a long-standing goal in cancer biology. Genetic mutation, gene amplification and chromosomal rearrangement are three classic genetic mechanisms that drive cancer progression and identity5,6, but they provide an incomplete explanation for oncogene activation. Recently, a wave of discoveries has demonstrated how TEs change the gene expression landscape during evolution, development and disease7,8,9. Although epigenetically silenced in somatic tissues, TEs can become active in cancer due to DNA hypomethylation, which can expose regulatory sequences and lead to functional consequences10,11. Indeed, some TEs are epigenetically reactivated as cryptic promoters to drive oncogene expression in cancer, a process known as onco-exaptation12,13. To our knowledge, no comprehensive study has investigated whether onco-exaptation is a widespread mechanism for oncogene activation across multiple cancer types.

To globally characterize onco-exaptation events, we canvassed RNA-seq data across 15 cancer types from the TCGA Research Network (http://cancergenome.nih.gov/) (Supplementary Fig. 1a). We constructed a computational pipeline that identifies TE-derived oncogene transcripts that are highly tumor enriched (Supplementary Fig. 1b). A comprehensive list of 702 oncogenes was generated from previously annotated onco-exaptation examples14,15 and ONGene16 (Supplementary Table 1). Considering the technical limitations of RNA-seq data, we set stringent filters (Methods) to maximize the specificity for onco-exaptation events. In total, we analyzed 7,769 tumor samples and 625 tumor-matched-normal samples (Supplementary Fig. 1b), which identified 625 TE-oncogene chimeric transcripts; this list includes five previously published onco-exaptation examples (Supplementary Table 2). After selecting further high tumor-enrichment and expression contribution, we identified 129 high confidence onco-exaptation events across 106 oncogenes (Supplementary Table 3). In addition, we detected at least one onco-exaptation event in 49.7% of all tumors, with prevalence ranging from 12 to 87% across cancer types, indicating that onco-exaptation could be a promiscuous mechanism for oncogene activation (Fig. 1a). On average, each onco-exaptation event was discovered in 51 samples and often distributed across multiple cancer types. We report that the onco-exapted TEs strongly enrich for the long terminal repeat class (Fig. 1b and Supplementary Fig. 2b). Examining the cancer-type distribution of onco-exaptation candidates (Fig. 1c) showed both cancer-type-specific events, such as THE1A-HMGA2 in skin cutaneous melanoma17, and highly prevalent events were present across multiple cancer types. Furthermore, for eight oncogenes, we observed various TEs activating an in-frame isoform of the same gene (Supplementary Table 4), a phenomenon that had only been described for one oncogene18. These additional examples support the cancer epigenetic evolution model as previously described19. In summary, we provide a global profile of onco-exaptation events across 15 cancer types and enumerate TEs’ role in driving oncogene activation and upregulation.

Next, we examined transcript-level information for the top ten most prevalent onco-exaptation candidates that on average accounted for greater than 50% of their target oncogene’s expression (Fig. 1d). Eight of these candidates were predicted to form in-frame transcripts that conserve protein sequence, suggesting preservation of oncogene function. Onco-exaptation candidates include isoforms of genes such as SALL4 and LIN28B that have recently emerged as potent cancer drivers20,21. Additionally, the LIPA2-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 oncogene expression; some greater than 90% (Fig. 1d and Supplementary Fig. 3). Furthermore, we report that half of the top candidates were associated with worse survival in at least one cancer type (Supplementary Fig. 4). For 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 prognosis4,11. These findings...
indicate that TEs not only are associated with oncogene activation but also contribute substantially to overall oncogene expression and oncogenic potential.

For validation, we sought to confirm transcription initiation from a few exapted TEs. We queried the FANTOM5 promoter database25 and discovered five out of the ten most prevalent onco-exaptation candidates show promoter signature. We validated a few FANTOM5 results by mapping transcription start sites (TSS) with cap analysis of gene expression (CAGE)-seq25–27 in the H727 tissues profiled by Roadmap (http://www.roadmapepigenomics.org/) (Supplementary Fig. 5). In H727, the region surrounding the AluJb-LIN28B isoform in lung cancer cell lines, we profiled TSSs that active epigenetic marks encompassed two TEs, a truncated AluJb and MLT1B, upstream of AluJb-P (Fig. 2b). Since various TEs are known to harbor transcription-factor-binding sites that could impact AluJb-P promoter strength. Luciferase assays using various combinations of TEs before a luciferase reporter showed

The AluJb TE is located 20 kilobases (kb) upstream of the canonical promoter of LIN28B and drives the majority of LIN28B’s expression in a substantial number of tumors (Fig. 1d). To verify the existence of the AluJb-LIN28B isoform in lung cancer cell lines, we profiled TSSs in the H1299 and H838 cell lines by using paired-end CAGE-seq. We confirmed a CAGE peak, composed of mate reads that align to LIN28B, which spans ~40 base pairs (bp) in the AluJb element in both lung cancer cell lines, we profiled TSSs from alternative promoters located in TEs (Fig. 2a and Supplementary Fig. 5). In addition, we analyzed 27 RNA-seq datasets from lung cancer cell lines28 and detected five of the ten most prevalent onco-exaptation candidates (Supplementary Table 5). One of the most highly expressed candidates was an AluJb-LIN28B fusion transcript that is present in the H1299, RERF-LC-OK and H838 cell lines. Considering that LIN28B is a well-characterized and potent oncogene22,24,29–31, we pursued this candidate for further functional validation.

Next, we dissected the genetic determinants behind the AluJb-LIN28B onco-exaptation event. In H1299 and H838, we discovered that active epigenetic marks encompassed two TEs, a truncated AluJb and MLT1B, upstream of AluJb-P (Fig. 2b). Since various TEs are known to harbor transcription-factor-binding sites that could have cis-regulatory function2, we tested whether these upstream TEs impact AluJb-P promoter strength. Luciferase assays using various combinations of TEs before a luciferase reporter showed
Fig. 2 | TEs provide bona fide promoters for oncogenes in lung cancer cell lines. a, CAGE-seq profile of H727 across onco-exaptation candidates (ARID3A and SYT1) visualized on WashU Epigenome Browser. Signals in CAGE-seq represent TSS locations. b, CAGE-seq and epigenetic profiles of the AluJb TE in the H1299 and H838. Signal in ATAC-seq represent open chromatin regions. Gray bars in the WGBS-seq track represent CpG locations while the height of blue bars indicates methylation percentage. c, Luciferase assays for transcriptional activity of various TE arrangements in H1299 (left) and H838 (right) (n = 3 independent experiments). d, Luciferase assays for promoter activity in H1299 (left) and H838 (right) with mutagenized transcription factor motifs in AluJb-P (n = 3 independent experiments). c, d, P values were derived from two-tailed Welch t-test. All data are represented as means ± standard error (s.e.m.). WT, wild type.
that vectors without AluJb-P displayed minimal activity (Fig. 2c). Furthermore, the luciferase activity did not diminish in the solo AluJb-P vector relative to other vectors. These results illustrate that AluJb-P contains all the necessary sequences for strong promoter activity, and the upstream TEs have minimal cis-regulatory effect on AluJb-P transcription.

Fig. 3 | AluJb drives LIN28B expression and contributes to oncogenesis in lung cancer cell lines. a, Schematic describing single guide RNA locations and sequence targets within AluJb-P and LIN28BP. b, Cropped western blot for LIN28B protein in H1299 (top) and H838 (bottom) CRISPR clones. This experiment was repeated twice with similar results. c, Relative let-7a, let-7b and let-7g miRNA levels compared to wild type (WT) in CRISPR-knockout clones of H1299 (n = 4 independent experiments) and H838 (n = 3 independent experiments) as measured by quantitative PCR. d, The effect of AluJb-P or LIN28BP deletion on cell growth rate as determined by CCK-8 assay in H1299 and H838 cells (n = 3 independent experiments). e, The effect of AluJb-P or LIN28BP deletion on cell migration in H1299 (top) and H838 (bottom) as measured by scratch migration assay (n = 3 independent experiments). f, Tumor growth of H1299 wild type and H1299 CRISPR-knockout clones injected in nude mouse. Resected tumors of wild type and LIN28BP no. 1 and 2 xenografts. g, Cropped western blot (repeated twice with similar results) of re-expression of human FLAG-LIN28B or AluJb-LIN28B in AluJb knockout clones and its effect on relative let-7 miRNA levels (number of independent experiments indicated in figure with n) and growth rate (n = 3 independent experiments). d.e.g, P values from CCK-8 growth assays and scratch migration assays were derived from comparing to wild type with two-tailed Welch t-test. All data are represented as means ± s.e.m.
AluJb is a primate-specific subfamily in the short interspersed nuclear element (SINE) class of TEs. SINE elements are known to recruit RNA polymerase (RNAP) III to generate short transcripts that can potentially be retrotransposed \(^3\). However, most messenger RNAs are typically transcribed by RNAP II. We hypothesized that AluJb-P accumulated mutations through evolution that generated novel transcription factor binding sites that recruit RNAP II. To explore this hypothesis, we performed pair-wise sequence alignment using EMBOSS Needle \(^3\) between the AluJb-P sequence and the AluJb consensus sequence from Dfam \(^3\). We then identified potential novel transcription factor motifs that were generated by mutations specific to AluJb-P with FIMO \(^3\). Previous work has demonstrated that NFYA binds to AluJb-P and knockdown of NFYA reduces promoter activity in Huh-7 cells \(^3\). However, the degree of NFYA’s impact on AluJb promoter function is still unclear. Our analysis with FIMO detected four other transcription factor motifs that potentially arose from mutations: C/EBPD, SP1, SP4 and YY1 (Fig. 2d).

To interrogate the functional importance of these motifs, we cloned AluJb-P sequences mutagenized for each motif into a luciferase reporter and assessed the change in promoter activity. In both H1299 and H838, mutating SP1, SP4 and YY1 sites significantly diminished relative luciferase expression, which is consistent with previous findings that SP transcription factors cooperate with YY1 to drive strong promoter expression (Fig. 2d)\(^3\). Furthermore, these results were recapitulated in the K562 leukemia cell line (Supplementary Fig. 8a,b), which does not express the AluJb-LIN28B transcript. This finding suggests that K562 cells have all the transcriptional machinery to transcribe from the AluJb-P, but DNA methylation might be suppressing the activity of the promoter (Supplementary Fig. 6a).

To evaluate the functional consequences of the AluJb-LIN28B onco-exaptation event, we first investigated whether the fusion transcript produces a protein product. Within the AluJb-P sequence, we detected a strong start codon 72 bp downstream of the TSS.
This results in the addition of 22 amino acids at the N-terminus of exon 2 of LIN28B (Supplementary Fig. 6c), for a predicted protein size increase of 2.5kDa compared to normal LIN28B. Western blots verified the expected size difference between the onco-exapted AluJb-LIN28B isoform present in H1299 and H838 cells compared to the canonical LIN28B protein present in K562 and HepG2 (Supplementary Fig. 6d). To confirm that the larger protein originated from AluJb-P, we performed CRISPR-Cas9-mediated deletion of AluJb-P in H1299 and H838 (Fig. 3a). In addition, we deleted a 1-kb sequence of the canonical LIN28B promoter (LIN28BP). The deletion of AluJb-P abolished the larger LIN28B protein, while the deletion of LIN28BP did not (Fig. 3b), verifying that AluJb-P produced the larger LIN28B isoform.

Since the AluJb-LIN28B protein is identical to canonical LIN28B, aside from the additional N-terminal amino acids, we examined whether AluJb-LIN28B retained normal LIN28B function. LIN28B represses let-7 miRNAs29,30,38–40, ultimately contributing to oncogenesis through the upregulation of oncogenes such as MYC and RAS46,47. As anticipated, we observed an appreciable increase in the levels of let-7a, let-7b and let-7g in the AluJb-P knockout (KO) cells but not in LIN28BP knockout cells of H1299 and H838 (Fig. 3c).

We further assessed how the deletion of AluJb-P impacts cancerspecific attributes. In both H1299 and H838, AluJb-P knockout cells showed much slower growth (Fig. 3d) and migration (Fig. 3e) relative to the parental cell lines and LIN28BP knockout cells. Also, parental H1299 and LIN28BP knockout clones established rapidly growing tumors in vivo, whereas AluJb-P knockout cells exhibited a marked defect in tumor growth during the time of inspection (Fig. 3f), consistent with the necessity of LIN28B for tumor growth in murine xenograft models41,42. In contrast, the deletion of AluJb-P in K562 cells did not result in elevated let-7 levels (Supplementary Fig. 8e) or loss of proliferation (Supplementary Fig. 8f), suggesting that the loss of AluJb-LIN28B was causal for the decreased oncogenic attributes in H1299 and H838 cells and not due to an off-target effect. Additionally, re-expression of canonical LIN28B and AluJb-LIN28B in H1299 and H838 AluJb-P knockout cells reduced let-7 miRNA levels and modestly rescued proliferation (Fig. 3g).

 Altogether, these results indicate that TE-induced oncogene expression can retain its canonical function, which contributes to cell proliferation, migration and tumor formation.

Most tumors exhibit global DNA hypomethylation, which provides cancer cells with an opportunity to exploit the regulatory potential of TEs. However, whether the loss of DNA methylation is causal for spurring TE’s cryptic promoter activity has been underexplored due to a lack of efficient targeted methylation techniques. To directly assess how DNA methylation regulates AluJb-P activity, we used the CRISPR SuperNova tagging system (SunTag) to recruit either DNM3TA or TET1ICD for targeted methylation or demethylation, respectively (Fig. 4a,b). This system allowed us to modestly increase DNA methylation of the AluJb TE by ~20–30% (Fig. 4c), which led to an ~40% decrease in LIN28B expression in the H1299 (Fig. 4d), suggesting that DNA methylation of the TE is sufficient to decrease oncogene expression. Additionally, demethylation of the AluJb TE in K562 (Fig. 4e) led to the production of AluJb-LIN28B fusion protein (Fig. 4f). These results illustrate that dynamic DNA methylation is a driving epigenetic control that acts as an on-off switch for AluJb-P’s activity and moreover suggests that TE onco-exaptation events arise in tumors due to the unique epigenetic landscape.

Discussion

In conclusion, TEs provide an additional means by which cancer can activate oncogenes. Stochastic, global DNA hypomethylation of cancer cells indiscriminately resurrects TEs of varying regulatory ability that, if they confer a fitness advantage, can be epigenetically inherited and selectively propagated during tumor progression. Here, we present a global profile of tumor-enriched, TE-derived oncogene transcripts across 15 cancer types and show that onco-exaptation is a highly prevalent and promiscuous mechanism that contributes to oncogene activation in close to half of all tumors. By dissecting the mechanisms behind AluJb-derived LIN28B expression, we describe how TEs may be epigenetically and transcriptionally activated to drive oncogene expression. Recently, this tumor-specific LIN28B alternative promoter usage in liver cancer has also been characterized by Guo et al.40, but not in an onco-exaptation context. Our concomitant findings in lung cancer cell lines provide cross-cancer support of the robust oncogenic potential of AluJb-LIN28B. Recognizing onco-exaptation events can provide additional insights into potential genetic and epigenetic mechanisms that drive promoter activity in cancer. For example, we were able to identify additional putative transcription factors that might be controlling AluJb promoter activity by exploring the evolution of the SINE element. Furthermore, we provide evidence that these onco-exaptation events are potentially reversible through targeted epigenetic alterations, which could present a translational avenue for personalized epigenetic oncotherapy. In summary, TEs act as double-edged swords for cancer by offering additional mechanisms for oncogene activation but also providing a potential target for therapeutics.
17. Wiesner, T. et al. Alternative transcription initiation leads to expression of a novel ALK isoform in cancer. Nature 526, 453–457 (2015).
18. Wolff, E. M. et al. Hypomethylation of a LINE-1 promoter activates an alternate transcript of the MET oncogene in bladders with cancer. PLoS Genet. 6, e1000917 (2010).
19. Liu, Y., Sun, J. & Zhao, M. ONGene: a literature-based database for human oncogenes. J. Genet. Genomics 44, 119–121 (2017).
20. Raskin, L. et al. Transcriptome identifies HMG2 as a biomarker of melanoma progression and prognosis. J. Invest. Dermatol. 133, 2585–2592 (2013).
21. Zhang, X. et al. SALL4: An emerging cancer biomarker and target. Cancer Lett. 357, 55–62 (2015).
22. Wang, T. et al. Aberrant regulation of the LIN28A/LIN28B and let-7 loop in human malignant tumors and its effects on the hallmarks of cancer. Mol. Cancer 14, 125 (2015).
23. Nguyen, L. H. et al. LIN28b is sufficient to drive liver cancer and necessary for its maintenance in murine models. Cancer Cell 26, 248–261 (2014).
24. Viswanathan, S. R. et al. Lin28 promotes transformation and is associated with advanced human malignancies. Nat. Genet. 41, 843–848 (2009).
25. Forrest, A. R. R. et al. A promoter-level mammalian expression atlas. Nature 507, 462–470 (2014).
26. Brocks, D. et al. DNMT and HDAC inhibitors induce cryptic transcription start sites encoded in long terminal repeats. Nat. Genet. 49, 1052–1060 (2017).
27. Shiraki, T. et al. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. Proc. Natl Acad. Sci. USA 100, 15776–15781 (2003).
28. Suzuki, A. et al. Aberrant transcriptional regulations in cancers: genome, transcriptome and epigenome analysis of lung adenocarcinoma cell lines. Nucleic Acids Res. 42, 13557–13572 (2014).
29. Johnson, C. D. et al. The let-7 microRNA represses cell proliferation pathways in human cells. Cancer Res. 67, 7713–7722 (2007).
30. Newman, M. A., Thomson, J. M. & Hammond, S. M. Lin-28 interaction with the Let-7 precursor loop mediates regulated microRNA processing. RNA 14, 1539–1549 (2008).
31. Zhou, J., Ng, S. B. & Chng, W. J. LIN28/LIN28B: an emerging oncogenic driver in cancer stem cells. Int. J. Biochem. Cell Biol. 45, 973–978 (2013).
32. Moqtaderi, Z. et al. Genomic binding profiles of functionally distinct RNA polymerase III transcription complexes in human cells. Nat. Struct. Mol. Biol. 17, 635–640 (2010).
33. Rice, P., Longden, I. & Bleasby, A. EMBOSS: The European Molecular Biology Open Software Suite. Trends Genet. 16, 276–277 (2000).
34. Hubley, R. et al. The Dfam database of repetitive DNA families. Nucleic Acids Res. 44, D81–D89 (2016).
35. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: Scanning for occurrences of a given motif. Bioinformatics 27, 1017–1018 (2011).
36. Guo, W. et al. A LIN28B tumor-specific transcript in cancer. Cell Rep. 22, 2094–2106 (2018).
37. Beketaev, I. et al. cis-regulatory control of Mesp1 expression by YY1 and SPI1 during mouse embryogenesis. Dev. Dyn. 245, 379–387 (2016).
38. Hao, I. et al. Lin28 mediates the terminal uridylation of let-7 precursor microRNA. Mol. Cell 32, 276–284 (2008).
39. Viswanathan, S. R., Daley, G. Q. & Gregory, R. I. Selective blockade of microRNA processing by Lin28. Science 320, 97–100 (2008).
40. Rybak, A. et al. A feedback loop comprising lin-28 and let-7 controls pre-let-7 maturation during neural stem-cell commitment. Nat. Cell Biol. 10, 987–993 (2008).
41. Tanenbaum, M. E., Gilbert, L. A., Qi, L. S., Weissman, J. S. & Vale, R. D. A protein-tagging system for signal amplification in gene expression and fluorescence imaging. Cell 159, 635–646 (2014).
42. Morita, S. et al. Targeted DNA demethylation in vivo using dCas9-peptide repeat and scFv-TET1 catalytic domain fusions. Nat. Biotechnol. 34, 1055–1056 (2016).
43. Huang, Y. H. et al. DNA epigene editing using CRISPR-Cas SunTag-directed DNMT3A. Genome Biol. 18, 1–11 (2017).

Acknowledgements
We would like to thank J. Hossington-López and M.L. Jaeger from The Edison Family Center for Genome Sciences & Systems Biology (CGSSB) for assistance with sequencing; B. Kooebbe and E. Martin from CGSSB for data processing; M. Savio, M. Patana and D. Schweppe from the Siteman Flow Cytometry core for FACS-related expertise; M. Goodell and Y. Huang for valuable expertise with the SunTag-DNMT3A system and L. Maggi (Washington University School of Medicine) for generously gifting the H838 cell line. This work was funded by NIH grant numbers 5R01HG007175, U24ES026699 and U01HG009391 and the American Cancer Society Research Scholar grant number RSG-14-049-01-DMC. H.S.J. was supported by a grant from NIGMS (no. T32 GM007067). A.Y.D. is supported by a grant from NHGRI (no. T32 HG000045). N.M.S. is a Howard Hughes Medical Institute (H.H.M.I.) Medical Research Fellow, http://www.hhmi.org/. The xenograft work cited in this publication was performed in a facility supported by NCRB grant number C06 RR015502. E.C.P. was supported by Postdoctoral Fellowship PF-17-201-01-TBG from the American Cancer Society.

Author contributions
H.S.J., N.M.S. and T.W. conceived and implemented the study. N.M.S., H.S.J., E.C.P., D.L. and T.W. contributed to the computational analysis. H.S.J. generated transcriptomic and epigenomic profiles of cell lines. H.S.J., X.X. and D.Z. performed the CRISPR-mediated deletion experiments. H.S.J. and Z.Z.D. performed the promoter-luciferase, motif mutagenesis and let-7 qPCR experiments. H.S.J. and A.Y.D. performed the growth and migration assays. H.S.J., A.Y.D., D.O. and J.I.G. performed the xenograft experiments. H.S.J., P.M.G. and S.K. conceived and implemented the study. N.M.S., H.S.J., E.C.P. and D.L. performed the growth and deletion experiments. H.S.J. and Z.Z.D. performed the promoter-luciferase, motif mutagenesis and let-7 qPCR experiments. H.S.J. and A.Y.D. performed the growth and migration assays. H.S.J., A.Y.D., D.O. and J.I.G. performed the xenograft experiments. H.S.J., P.M.G. and S.K. performed the targeted methylation experiments. H.S.J. performed the rescue experiments. The manuscript was prepared and revised by H.S.J., N.M.S. and T.W. with input from all authors.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41588-019-0373-3.
Reprints and permissions information is available at www.nature.com/reprints.
Correspondence and requests for materials should be addressed to T.W.
Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2019.
Methods

Data download. All patient sample RNA-seq data analysis was done on the GDC Data Release 9.0 of TCGA data (24 October 2017). Normal and tumor RNA-seq BAM files for the following 15 cancers were downloaded using the gdc-client v.1.3.0: bladder urothelial carcinoma, colon adenocarcinoma, head and neck squamous cell carcinoma, kidney renal cell carcinoma, low grade glioma, liver hepatocellular carcinoma, lung adenocarcinoma, lung squamous cell carcinoma, ovarian serous cystadenocarcinoma, prostate adenocarcinoma, skin cutaneous melanoma, stomach adenocarcinoma, thyroid carcinoma and uterine endometrial carcinoma. In addition, normalized gene expression data (HTSeq-FPKM-Uq) and clinical metadata for all samples were downloaded using the gdc-client v.1.3.0. A total of 7,769 tumor samples and 625 matched-normal samples were used for analysis. A total of 26 lung adenocarcinoma cancer cell line RNA-seq files were downloaded using sratools with the following accession: DRA001846. We included RNA-seq of the H384 lung cancer cell line, which had been previously generated in our laboratory and will be publicly available. GENCODE v.25 was used as the transcript reference39. The gene transfer format file of consensuses transcripts was downloaded from https://www.gencodegenes.org/releases/25.html. Repeatmasker annotations were downloaded from the UCSC table browser for hg38 (refs. 40–41). FANTOM5 hg38-aligned peaks used for annotating the supplementary tables were downloaded from http://fantom5.gsc.riken.jp/%5Fdatafiles/reprocessed/hg38_latest/. Then, 698 protein-coding oncogenes were obtained from the ONGene database42. Another four genes from previous publications noting ‘onco-exaptation’ were included in the list: IRE5, FABP7, SLC18B3 and IL33 (ref. 43). More details about the software used in our analysis can be found in the Life Sciences Reporting Summary.

Assembly and annotation of transcripts. BAM files were sorted and indexed and chrl–22, X and Y were extracted. Stringtie v.1.3.3 was used to assemble the BAM files for all the RNA-seq samples (stringtie –m 100 –c 1)44. These transcripts were then annotated with features from GENCODE v.25 with a custom script. Briefly, GENCODE v.25 was first processed into a coordinate dictionary on the basis of chromosome, start and end location. Only the transcripts that were considered ‘appris_principal’ were used so that alternative transcripts of the gene would not be excluded as potential TE-derived candidates. This set of principal transcripts as well as the Repeatmasker TE coordinates were used to annotate the transcripts generated from the stringtie assembly for each sample. The starting position of the transcript was annotated using the Repeatmasker table to find TE-derived TSSs. Then, the first exon of the transcript was annotated on the basis of overlap with exonic or intronic features of GENCODE v.25. If the exon overlapped both an exon and intron, then the exon was selected as the annotation for that element. Then, all subsequent exons in the transcript were annotated until one overlapped with a TE. This exon of the protein-coding gene was selected as the ‘splice target’ of that transcript. After all transcripts were annotated, candidate transcripts were selected based on the following criteria: the start site of the transcript being within a TE, the TE being intergenic or intronic, the starting exon not overlapping with exon 1 of the canonical gene, and the transcript splicing into a protein-coding gene. We further limited our analysis to only include a list of 702 oncogenes to increase likelihood of finding candidates with tumorigenic impact.

Generating a reference transcriptome including onco-exaptation candidates. Aggregating annotation data across all tumor and normal RNA-seq datasets, we constructed a list of unique onco-exaptation candidates based on the subfamily of the TE, the genomic coordinates of the TE and the exon of the protein-coding gene where the transcript was spliced into. To remove potential assembly artifacts and genomic contamination, we removed candidates that had an average exon one length greater than the ninety-ninth percentile of all GENCODE v.25 transcript first exons (2,588 bp). Furthermore, transcripts with first exons that retained an intron were also removed. Finally, we only included candidates that were present in at least two samples.

To further increase confidence of promoter activity, we interrogated all reads that uniquely mapped to each candidate TE. We subsequently annotated the mate pair of those reads to see if any overlapped directly with oncogene exons. For single-end reads, we annotated the portion of the read mapping outside the TE to see if it overlapped with an oncogene exon. First, we removed candidates that had zero files where there were at least ten uniquely mapped reads that started in the TE. In addition, these events were required to have at least one sample with uniquely mapped paired-end reads where one of the pairs mapped to the TE and the other to the splice target of the candidate. For intronic onco-exaptation events, we also removed candidates that had evidence of exomization (there were reads mapping both to the exon and the TE in more than 15% of cases). Finally, candidates that were exclusively in single-end RNA-seq files were removed. The remaining candidate TE-derived transcripts were then merged with the reference GENCODE v.25 annotation file using Cuffmerge to create a reference transcriptome inclusive of potential onco-exaptation events that have not been previously annotated.

Transcript-level quantification and candidate selection. To determine the contribution of candidates to overall gene expression, we used stringtie (+e -b) with the merged transcriptome as the reference. For each sample, we labeled a candidate as being present if it met the following criteria: (1) the transcript accounted for at least 25% of total gene expression, (2) there was at least one read covering the splice junction between the TE and the splice target (candidates without unique splice junctions were removed) and (3) the target gene had at least one FPKM expression. Next, we filtered for candidates that were highly tumor enriched (>10x enrichment in the tumor samples) and present in at least four tumor samples. For the two cancers where there were no normal samples (ovarian serous cystadenocarcinoma and low grade glioma), we removed candidates that were present in >75% of their samples in these tumor types to avoid simply enriching for tissue-specific alternative promoters. This gave us a master list of 129 tumor-enriched onco-exaptation candidates involving 106 oncogenes. We then explored the abundance of these 129 candidates across the various cancer types to determine the prevalence of this phenomenon.

Open-reading-frame prediction and FANTOM5 annotation. After determining the predicted transcript sequences of our candidates, we used CPC2 that predicted whether candidates were coding or non-coding45. For coding transcripts, we subsequently used the start codon identified by CPC2 for the longest open-reading frame and evaluated if it was in-frame or out-of-frame in relation to the canonical isoform. For FANTOM5 promoter annotation, we used the FANTOM5 peaks in hg38 for samples that were not part of exposure or time-course experiments. Subsequently, we evaluated whether there were any peaks that overlapped with the onco-exapted TE that were on the same strand as our candidate transcript.

Cell culture methods. All cell lines were grown in a humidified incubator with 95% CO2 at 37°C. H1299, H384, H727 and K562 cell lines were cultured in RPMI1640 medium (Gibco, 11875–085) supplemented with 10% fetal bovine serum (Corning, 35–011-CV) and 100 U ml−1 penicillin-streptomycin (Gibco, 1510–122). HEK295 cell line was cultured in DMEM (Gibco, 11965–084) supplemented with 10% fetal bovine serum and 100 U ml−1 penicillin-streptomycin. Adherent cell lines were passaged at 70–90% confluency with 0.05% Trypsin-EDTA (Gibco, 25300–54).

Epigenome and transcriptome profiling. H1299 and K562 whole-genome bisulfite (WGBS)-seq and CAGE-seq were obtained from previously published resources. To generate WGBS-seq of H384 cell lines, we extracted genomic DNA with Quick-DNA Miniprep Kit (Zymo, D3024) and bisulfite-converted 200 ng of DNA using EZ DNA Methylation-Direct kit (Zymo, D3020). For WGBS-seq, we processed the bisulfite-converted DNA with TruSeq DNA Methylation Kit (Illumina, 15066014). To evaluate DNA methylation of targeted regions, we performed BS–PCR using Zymo/1aq PreMix (Zymo, E2003) following the manufacturer’s protocol. Illumina adapters for sequencing were ligated to the BS–PCR product and amplified for sequencing. WGBS-seq and targeted BS–PCR libraries were sequenced on Illumina NextSeq and MiSeq platforms, respectively. The sequencing reads were aligned to hg19 genome with Bismark and CpG methylation values were calculated using the bismark_methylation_extractor function.

To generate chromatin accessibility profiles for H1299 and H384, we followed the published Omni-ATAC-seq protocol46. Omni-ATAC-seq libraries were sequenced on Illumina NextSeq platform and reads were mapped to hg19 genome using bwa-mem.

Total RNA was extracted using TRIzol Reagent (ThermoFisher Scientific, 15596026) following the manufacturer’s protocol with few modifications. We performed an extra chloroform wash after transferring the aqueous phase. Furthermore, we added 5 µg of glycerol and 750 µl of isopropanol to the aqueous phase and incubated the solution overnight at −20°C to precipitate the RNA. Total RNA was treated with TURBO DNase (ThermoFisher Scientific, AM2238). H384 RNA-seq library was generated using TruSeq RNA Library Prep Kit v2 (Illumina, RS-122–2001).

To annotate TSS locations, we generated CAGE-seq libraries using CAGE Preparation Kit (DNAFORM). In brief, 1 µg of total RNA was reverse transcribed using SuperScript III (ThermoFisher Scientific, 18080093) and 5 µl of mRNA was biotinylated. Biotinylated RNA/complimentary DNA hybrid was purified using Dynabeads M-280 Streptavidin beads (ThermoFisher Scientific, 112050D) and processed to be sequenced on the Illumina sequencing platforms. For H727, we generated nanoCAGE-seq libraries47. In summary, poly(A) mRNA was extracted using Dynabeads mRNA DIRECT Purification Kit (ThermoFisher Scientific, 61011). The mRNA was enriched for 5′-capped mRNA via Terminator exonuclease (Lucigen, TERR51026) digestion. Then we followed standard nanoCAGE protocol to generate the DNA via template-switching technology. H1299 and H384 CAGE-seq reads were aligned to the hg19 genome while H727 nanoCAGE-seq was aligned to hg38 genome with HISAT and processed using CAGE package in R statistics48. All browser tracks are visualized with the WashU Epigenome Browser.

qPCR of let-7 micro RNA and LIN28B. Let-7 miRNA levels were profiled using a published real-time PCR-based platform49. To summarize, 500 ng of total RNA was reverse transcribed using SuperScript IV First-Strand Synthesis System.
CRISPR-SunTag vector construction. We obtained the scFv-sgFP-DNMT3A1 vector (Addgene, 102278) for the targeted methylator vector. We purchased pHRD5V4-dCas9-10xGCN4_v4-P2A-BFP plasmid (Addgene, 60903) and pLKO.5.sgRNA.EFS.IRFP675 plasmid (Addgene, 57824). For targeted demethylation, we replaced the DNMT3A sequence with a TET1 catalytic domain sequence, which was amplified from the pPlaTET-grNA2 plasmid (Addgene, 82559). Recent work revealed that dCas9-SunTag with 22aa linkers between GCN4 had higher demethylation efficiency14. In pHRD5V4-dCas9-10xGCN4_v4-P2A-BFP plasmid, we excised the 10x GCN4 sequence and cloned in GCN4–22aa sequence from pPlaTET-grNA2 as Gibson Assembly. sgRNA was cloned into pLKO.5.sgRNA.EFS.IRFP675 plasmid.

Lentivirus production and transduction of CRISPR-SunTag vectors. HEK293T cells were seeded in 2 ml of DMEM complete medium and grown to 50% confluence. We co-transfected CRISPR-SunTag plasmids with pMD2G and psPAX2 following polyethylenimine (PEI) transfection protocol. In brief, 6µg of PEI and 2µg of combined plasmids was added to 200µl of OptiMEM (ThermoFisher Scientific, 31985062) and incubated at room temperature for 30 min. The incubated PEI–vector mixture was added directly to HEK293T cells. After 48h, the viral supernatant was collected and filtered through 0.45-µm polyethersulfone filter (Sigma-Aldrich, SLHV0335R). Then, polybrene (Sigma-Aldrich, TR-1003-G) was supplemented to the viral supernatant to a concentration of 5µg µl−1. The polybrene-viral supernatants of dCas9-SunTag-BFP, scFv-sgFP-DNMT3A1/TET1CD and sgRNA/IRFP675 were added directly on top of H1299 and H838 cells in six-well plates. The transfected cells were rinsed with PBS and analyzed by flow-cytometry (Beckman Coulter MoFlo) for BFP and GFP and farIRFP675 fluorescence. Individual triple-positive fluorescent cells were sorted into 96-well plates and expanded. Once sufficiently expanded, the CRISPR-SunTag clones are resorted on the MoFlo for strong fluorescence and collected for downstream analysis of DNA methylation, gene expression and peptide expression.

Human LIN28B and AluJb-LIN28B rescue. We purchased pBABE-H-Lin28B plasmid (Addgene, 26358) that expresses FLAG-tagged human LIN28B protein24. We generated AluJb-LIN28B coding sequence from H1299 mRNA and cloned AluJb-LIN28B coding sequence in lieu of FLAG-hLIN28B peptide expression. We generated AluJb-LIN28B rescue vectors by co-transfected AluJb-LIN28B plasmids with pMD2G and pUMVC following PEI transfection protocol into HEK293T cells. AluJb knockout clones were transduced with viral supernant supplemented with polybrene (5µg µl−1) for 2d. Successfully infected cells were selected by 2µg µl−1 puromycin (A.G. Scientific, P-1033-sol) treatment for 5d before subsequent analysis.

Statistical analysis. Kaplan–Meier distributions between samples with or without candidate expression were compared using the log-rank test. All statistics for in vitro experiments were performed using a two-tailed Welch’s t-test. Enrichment for the TE class was calculated with this formula: (number of TE family onco-expanded/number of total TE onco-expanded)/(number of total TE family/number of all TEs)).

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Code availability
All custom scripts are available from the authors upon request.

Data availability
Datasets generated and analyzed in this study are available on Gene Expression Omnibus under accession code GSE113946.

References
44. Harrow, J. et al. GENCODE: the reference human genome annotation for the ENCODE project. Genome Res. 22, 1760–1774 (2012).
45. Taira-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. Curr. Protoc. Bioinforma. 25, 1–14 (2009).
46. Karolchik, D. The UCSC Table Browser data retrieval tool. Nucleic Acids Res. 32, D493–D496 (2004).
47. Pertea, M. et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat. Biotechnol. 33, 290–295 (2015).
48. Van Ommen, G. J., van Boeijen, A. & van Driessche, E. A fast and accurate coding potential calculator based on sequence intrinsic features. Nucleic Acids Res. 45, W12–W16 (2017).
49. Dunham, I. et al. An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57–74 (2012).
50. Krueger, F. & Andrews, S. R. Bismark: a flexible aligner and methylation caller for Bisulfite-seq applications. Bioinformatics 27, 1571–1572 (2011).
51. Corces, M. R. et al. An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. Nat. Meth. 14, 959–962 (2017).
52. Bayat, A., Gaëta, B., Ignjatovic, A. & Parameswaran, S. Improved VCF normalization for accurate VCF comparison. *Bioinformatics* 33, 964–970 (2017).
53. Salimullah, M., Mizuho, S., Plessy, C. & Carninci, P. NanoCAGE: a high-resolution technique to discover and interrogate cell transcriptomes. *Cold Spring Harb. Protoc.* 6, 96–111 (2011).
54. Haberle, V. et al. CAGEr: precise TSS data retrieval and high-resolution promoterome mining for integrative analyses. *Nucleic Acids Res.* 43, e51(2015).
55. Zhou, X. et al. The human epigenome browser at Washington University. *Nat. Meth.* 8, 989–990 (2011).
56. Wang, X. Primer sequences for 96 cancer-related miRNA assays. *RNA* 15, 716–723 (2009).
57. Haeussler, M. et al. Evaluation of off-target and on-target scoring algorithms and integration into the guide RNA selection tool CRISPOR. *Genome Biol.* 17, 1–12 (2016).
58. Moreno-Mateos, M. A. et al. CRISPRscan: designing highly efficient sgRNAs for CRISPR-Cas9 targeting in vivo. *Nat. Meth.* 12, 982–988 (2015).
Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

- The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided
- Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted. Give P values as exact values whenever suitable.
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen's d, Pearson's r), indicating how they were calculated
- Clearly defined error bars
- State explicitly what error bars represent (e.g. SD, SE, CI)

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection

- gdc-client version 1.3.0 (https://gdc.cancer.gov/access-data/gdc-data-transfer-tool) was used to download indexed RNA-seq BAM files and FPKM-UQ Expression files for 15 cancers (BLCA, BRCA, COAD, HNSC, KIRC, LGG, LIHC, LUAD, LUSC, OV, PRAD, SKCM, STAD, THCA, UCEC).
- sra-tools v2.8.2: Used to collect lung cancer cell line RNA-sequencing data
- bismark (v0.181): Aligning WGBS-seq and BSPCR-seq reads
- bwa mem (v0.7.15): Aligning ATAC-seq reads
- HISAT2 (v2.1.0): Aligning CAGE-seq reads
- Leica Application Suite X (2.0.0.14332): Imaging scratch migration assays

Data analysis

- samtools version 1.6: Used to extract chr1-22, X, and Y from TCGA RNA-seq BAM files. tabix from samtools was used to serially index transposable element locations from repeatmasker downloaded from UCSC table browser for hg38 as well as Gencode v25 exon and intron locations.
- stringtie version 1.3.3: Used to assemble BAM files to generate GTF files with transcript predictions. Also used for transcript-level
quantification.
cufflinks version v2.2.1: Used for the cuffmerge function to create a consensus reference transcriptome with gencode transcript as well as onco-exaptation candidate.

bedtools v2.27.1: bamtobed function used to generate a bed files for single end and paired end reads in region of interest from BAM files.

CPC2-beta: Used for predicting if transcripts are non-coding or coding. In addition, the start codon used for CPC2 calculations was used to predict the frame of the transcripts.
genecode_to_dic.py: Custom script used to generate a cPickle dictionary of all exons and introns in Genecode v25. This dictionary is used for annotation in later scripts.

rmsk_annotate.py: Custom script used to annotate paired-end BAM file reads in region(s) of interest
rmsk_annotate_bedpe.py: Custom script used to annotate single-end BAM file reads in region(s) of interest
WashU EpiGenome Browser (legacy v.46.1)
R Statistics (v3.4.1)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data
All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:
- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

TCGA data was accessed from https://portal.gdc.cancer.gov/.
ONGene data was accessed from http://ongene.bioinfo-minzhao.org/.
FANTOM5 data was accessed from http://fantom.gsc.riken.jp/5/datafiles/reprocessed/hg38_latest/
26 lung cancer cell line RNA-sequencing data was accessed from the following accession: DRA001846
Data sets generated and analyzed in this study are available on Gene Expression Omnibus (GEO): GSE113946.

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☑️ Life sciences     ☐ Behavioural & social sciences     ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size

No statistical methods were used to predetermine sample size. The sample size for each experiment is defined in figure or figure legend. For CRISPR experiments, we chose two independent clones for each gRNA pair to minimize interpretation of data from potential off-target effects. For in-vitro experiments, we chose to perform 3 to 4 independent experiments (as detailed in figure legend) to identify statistically significant trends using two-tailed Welch's t-test. For xenograft experiments, we chose a sample size of 4 for WT and 6 for CRISPR KO clones to account for potential biological variation between individual animals.

Data exclusions

The following TCGA samples were excluded:
TCGA-HT-8106-01A-11R-2404-07- Did not have corresponding FPKM-UQ expression file available
TCGA-V1-A8MJ-01A-11R-A36G-07- Did not have corresponding FPKM-UQ expression file available
TCGA-44-2665-01B-06R-A277-07- Did not have corresponding FPKM-UQ expression file available
TCGA-EM-A3AN-01A-11R-A206-07- Did not have corresponding .bai file even when downloaded with multiple attempts. Indexing the file with samtools also failed.

We pre-established that only samples with usable transcriptomic data would be selected for analysis.
Replication: All attempts at replication were successful in this study. The number of independent experiments performed for each assay is defined in the figure and figure legend.

Randomization: No randomization was performed for this study. However, all results were replicated through numerous independent experiments.

Blinding: No blinding was performed for this study for the xenograft experiment. To minimize potential bias, two investigators were present to measure tumor size and provided a consensus measurement.

Reporting for specific materials, systems and methods

### Materials & experimental systems

| n/a | Involved in the study |
|-----|------------------------|
| ☑️  | Unique biological materials |
| ☑️  | Antibodies |
| ☑️  | Eukaryotic cell lines |
| ☑️  | Palaeontology |
| ☑️  | Animals and other organisms |
| ☑️  | Human research participants |

### Methods

| n/a | Involved in the study |
|-----|------------------------|
| ☑️  | ChIP-seq |
| ☑️  | Flow cytometry |
| ☑️  | MRI-based neuroimaging |

### Antibodies

- **Lin28B Antibody** #4196 (Cell Signaling Technology, Cat #4196S, Lot #5) 1:1000 dilution
- **Anti-ACTB Mouse Monoclonal Antibody** (GenScript, Cat #A00702, Lot #15E000905) 1:5000 dilution of 0.5mg/ml antibody

**Validation**: Both antibodies are commercially available and validated by manufacturers and utilized in numerous publications.

- LIN28B publications provided by manufacturer (species reactivity = human):
  - Balzer, E. and Moss, E.G. (2007) RNA Biol 4, 16–25.
  - Piskounova, E. et al. (2008) J Biol Chem 283, 21310–4.
  - Cho, W.C. (2007) Mol Cancer 6, 60.
  - Viswanathan, S.R. et al. (2009) Nat Genet 41, 843–8.

- ACTB publications provided by manufacturer (species reactivity = mouse, rabbit, chicken, human, hamster, cow, goat, fish and pig):
  - Gruenwald, K. et al. (2013) J Bone Miner Res. 3,666-75.
  - Au, KY. et al. (2013) Biochim Biophys Acta. 1840(3), 958-63.
  - Wang, H. et al. (2013) J Biol Chem. 288(37), 26668-77.

### Eukaryotic cell lines

**Policy information about** [cell lines](#)

- **Cell line source(s)**
  - NCI-H838 (H838) ATCC CRL-5844
  - NCI-H1299 (H1299) ATCC CRL-5803
  - K-562 (K562) ATCC CCL-243
  - 293T (HEK293T) ATCC CRL-3216
  - NCI-H727 (H727) ATCC CRL-5815

- **Authentication**: None of the cell lines were authenticated after purchase from ATCC.

- **Mycoplasma contamination**: Cell lines were not tested for mycoplasma contamination since purchase from ATCC.

- **Commonly misidentified lines** ([See ICCLAC register](#))
  - No commonly misidentified lines were utilized in this study.

### Animals and other organisms

**Policy information about** [studies involving animals; ARRIVE guidelines](#) recommended for reporting animal research

- **Laboratory animals**
  - Jackson Lab, 002019, 4 weeks old homozygous NU/J females
| Category                | Description                                      |
|------------------------|--------------------------------------------------|
| Wild animals           | The study did not involve wild animals.          |
| Field-collected samples| The study did not involve field-collected samples. |