bioRxiv preprint doi: https://doi.org/10.1101/2021.11.06.467555; this version posted November 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Tumour mutations in long noncoding RNAs that enhance cell fitness

Roberta Esposito*[1,2,3], Andrés Lanzós*[1,2,4], Taisia Polidori [1,2], Hugo Guillen-Ramirez [5,6], Bernard Merlin [1,2], Lia Mela [1,2], Eugenio Zoni [2,9], Isabel Büchi [2,8], Lusine Hovhannisyan [2,7], Finn McCluggage [10,11], Matúš Medo [2,7], Giulia Basile [1,2], Dominik F. Meise [1,2], Sunandini Ramnarayanan [5,6], Sandra Zwyssig [1,2], Corina Wenger [1,2], Kyriakos Schwarz [1,2], Adrienne Vancura [1,2], Nuria Bosch-Guiteras [1,2,4], Marianna Kruithof-de Julio [2,9], Yitzhak Zimmer [2,7], Michaela Medová [2,7], Deborah Stroka [2,8], Archa Fox [10,11], Rory Johnson [1,2,5,6]

1.Department of Medical Oncology, Inselspital, Bern University Hospital, University of Bern, 3010 Bern, Switzerland.

2.Department for BioMedical Research, University of Bern, 3008 Bern, Switzerland

3.Institute of Genetics and Biophysics "Adriano Buzzati-Traverso", CNR, 80131 Naples, Italy.

4.Graduate School of Cellular and Biomedical Sciences, University of Bern, 3012 Bern, Switzerland.

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

6.Conway Institute for Biomolecular and Biomedical Research, University College Dublin, Dublin D04 V1W8, Ireland.

7.Department of Radiation Oncology, Inselspital, Bern University Hospital and University of Bern, Bern, Switzerland

8.University Clinic of Visceral Surgery and Medicine, Bern University Hospital, Inselspital, Department of Biomedical Research, University of Bern, Bern, Switzerland.

9.Department of Urology, Inselspital, Bern University Hospital, Bern, Switzerland.

10.School of Molecular Sciences, University of Western Australia, Crawley, Western Australia, Australia.

11.School of Human Sciences, University of Western Australia, Crawley, Western Australia, Australia.

* Equal contribution

Correspondence: rory.johnson@ucd.ie

Keywords: Cancer; Mutations; Long Non-Coding RNA; LncRNA; Cancer Driver Genes; Pan-Cancer Analysis of Whole Genomes; CRISPR; NEAT1.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.06.467555; this version posted November 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Abstract

Tumour DNA contains thousands of single nucleotide variants (SNVs) in non-protein-coding regions, yet it remains unclear which are "driver mutations" that promote cell fitness. Amongst the most highly mutated non-coding elements are long noncoding RNAs (lncRNAs), which can promote cancer and may be targeted therapeutically. We here searched for evidence that driver mutations may act through alteration of lncRNA function. Using an integrative driver discovery algorithm, we analysed single nucleotide variants (SNVs) from 2583 primary tumours and 3527 metastases to reveal 54 candidate "driver lncRNAs" (FDR<0.1). Their relevance is supported by enrichment for previously-reported cancer genes and by clinical and genomic features. Using knockdown and transgene overexpression, we show that tumour SNVs in two novel lncRNAs can boost cell fitness. Researchers have noted particularly high yet unexplained mutation rates in the iconic cancer lncRNA, NEAT1. We apply i*n cellulo* mutagenesis by CRISPR-Cas9 to identify vulnerable regions of NEAT1 where SNVs reproducibly increase cell fitness in both transformed and normal backgrounds. In particular, mutations in the 5' region of NEAT1 alter ribonucleoprotein assembly and boost the population of subnuclear paraspeckles. Together, this work reveals function-altering somatic lncRNA mutations as a new route to enhanced cell fitness during transformation and metastasis.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.06.467555; this version posted November 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Introduction

Tumours arise and develop via somatic mutations that confer a fitness advantage on cells (Campbell et al., 2020). Such "driver" mutations exert their phenotypic effect by altering the function of genes or genomic elements, and may be identified by signatures of positive evolutionary selection (Rheinbay et al., 2020). Identifying driver mutations, and the "driver genes" through which they act, is a critical step towards understanding and treating cancer (Campbell et al., 2020; Rubio-Perez et al., 2015).

Most tumours are characterised by a limited and recurrent sequence of driver mutations, which are shared by cells within and between tumours (Nowell, 1976; Tomasetti et al., 2015). Evidence from modelling and genetically-modified mouse models (Collins et al., 2012) has cemented the view that such "gatekeeper" mutations are necessary and rate-limiting events for tumorigenesis (Vogelstein and Kinzler, 2015). With the exception of the *TERT* promoter (Vinagre et al., 2013), these almost invariably affect protein-coding sequences (Sondka et al., 2018).

However, the vast majority of single nucleotide variants (SNVs) fall outside protein-coding genes (Khurana et al., 2016). Combined with increasing awareness of the disease roles of noncoding genomic elements (Gloss and Dinger, 2018), this naturally raises the question of whether non-protein coding mutations also contribute to cancer cell fitness (Elliott and Larsson, 2021). Growing numbers of both theoretical (Corona et al., 2020; Hornshøj et al., 2018; Kim et al., 2016; Melton et al., 2015; Puente et al., 2015; Umer et al., 2021) and experimental studies (Cho et al., 2018; Li et al., 2020; Rheinbay et al., 2020; Zhou et al., 2020; Zhu et al., 2020) implicate noncoding SNVs in cell fitness by altering the function of elements such as enhancers, promoters, insulator elements and small RNAs (Shuai et al., 2019). Given the relative robustness of noncoding elements to sequence changes, in contrast to protein-coding open reading frames (ORFs), it is anticipated that noncoding drivers exert weaker phenotypic effects (Elliott and Larsson, 2021). These may help explain the existence of widespread "mini-drivers" predicted by recent modelling (Castro-Giner et al., 2015; Kumar et al., 2020).

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.06.467555; this version posted November 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Surprisingly, one important class of cancer-promoting noncoding gene has been largely overlooked to date: long noncoding RNAs (lncRNAs) (Statello et al., 2021). LncRNA transcripts are modular assemblages of functional elements that can interact with other nucleic acids and proteins via defined sequence or structural elements (Ghandi et al., 2018; Statello et al., 2021). Of the >50,000 mapped in the human genome (Uszczynska-Ratajczak et al., 2018), hundreds of lncRNAs have been demonstrated to act as oncogenes / tumour suppressors (Vancura et al., 2021). LncRNAs are targeted by CNVs (Akrami et al., 2013; Hu et al., 2014; Leucci et al., 2016), targeted by tumour-initiating transposon screens in mouse (Carlevaro-Fita et al., 2020) and by function-altering germline cancer variants (Redis et al., 2016). We and others have previously reported statistical evidence for mutation-enriched candidate driver lncRNAs (Lanzós et al., 2017a), but no fitness-altering somatic lncRNA SNVs have been reported to date.

In the present study, we comprehensively map candidate driver lncRNAs across the largest whole-genome tumour cohort to date, containing both primary and metastatic tumours. The resulting set of dozens of driver lncRNA candidates is enriched for known cancer lncRNAs and carries various independent clinical features of disease genes. We use a variety of experimental perturbations (including *in cellulo* CRISPR mutation) and cell models (including non-transformed backgrounds) to gather the first evidence for fitness-altering mutations acting through lncRNA sequence.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.06.467555; this version posted November 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Results

Integrative driver lncRNA discovery with ExInAtor2

Driver genes can be identified by signals of positive selection acting on their somatic mutations. The two principal signals are *mutational burden* (MB), an elevated mutation rate, and *functional impact* (FI), the degree to which mutations are predicted to alter encoded function. Both signals must be compared to an appropriate background, representing mutations under neutral selection.

To search for lncRNAs with evidence of driver activity, we developed *ExInAtor2*, a driver-discovery pipeline with enhanced sensitivity due to two key innovations: integration of both MB and FI signals, and empirical background estimation (see Methods) (Figure 1A, Supplementary Figure 1). For MB, local background rates are estimated, controlling for covariates of mutational signatures and large-scale effects such as replication timing, which otherwise can confound driver gene discovery (Lawrence et al., 2013). For FI, we adopted functionality scores from the *Combined Annotation Dependent Depletion* (CADD) system, thanks to its wide use and compatibility with a range of gene biotypes (P et al., 2021). Importantly, *ExInAtor2* remains agnostic to the biotype of genes / functional elements, enabling independent benchmarking with established protein-coding gene data.

Accurate discovery of known and novel driver genes

We began by benchmarking ExInAtor2 using the maps of somatic single nucleotide variants (SNVs) from tumour genomes sequenced by the recent PanCancer Analysis of Whole Genomes (PCAWG) project (Campbell et al., 2020), comprising altogether 45,704,055 SNVs from 2,583 donors (Figure 1B, Methods). As it was generated from whole-genome sequencing (WGS), this dataset allows one to search for driver genes amongst both non-protein-coding genes (including lncRNAs) and better-characterised protein-coding genes.

To maximise sensitivity and specificity, we prepared a carefully-filtered annotation of lncRNAs. Beginning with high-quality curations from Gencode (Frankish et al., 2019), we isolated intergenic lncRNAs, from which we removed those with evidence for possible protein-coding capacity. To these, we added the set of confident, literature-curated lncRNAs from Cancer LncRNA Census 2 dataset (Vancura et al., 2021), for a total set of 6982 genes (Figure 1C).

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.06.467555; this version posted November 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

We first evaluated ExInAtor2's ability to identify cancer-related lncRNA genes, compared to ten leading driver discovery methods and PCAWG's consensus measure, which integrates all ten and has better performance than any individual method (Figure 2A) (Rheinbay et al., 2017). Judged by correct identification of known cancer lncRNAs at a false discovery rate (FDR) cutoff of <0.1, ExInAtor2 displayed the best overall accuracy in terms of F_1 measure (Figure 2B). ExInAtor2 displayed good statistical behaviour, since quantile-quantile (QQ) analysis of resulting *p*-values displayed no obvious inflation or deflation and has amongst the lowest Mean Log Fold Change (MLFC) values (Figure 2C), together supporting its low and controlled FDR.

ExInAtor2 is biotype-agnostic, and protein-coding driver datasets are highly refined. To further examine its performance, we evaluated sensitivity for known protein-coding drivers from the benchmark Cancer Gene Census (Sondka et al., 2018). Again, ExInAtor2 displayed competitive performance, particularly with respect to false positive predictions (Supplementary Figure 2A-C).

To test ExInAtor2's FDR estimation, we repeated the lncRNA analysis on a set of carefully-randomised pancancer SNVs (see Methods). Reassuringly, no hits were discovered and QQ plots displayed a neutral behaviour (MLFC 0.08) (data not shown). Similar results were obtained with SNVs from an independent randomisation method (data not shown). Analysing at the level of individual cohorts, ExInAtor2 predicted 3 / 40 lncRNA-cohort associations in the simulated / real datasets, respectively. This corresponds to an empirical FDR rate of 0.075, consistent with the nominal FDR cutoff of 0.1.

We concluded that ExInAtor2 identifies known driver genes with a low and controlled false discovery rate.

The landscape of driver lncRNA in primary human tumours

We next set out to create a genome-wide panorama of mutated lncRNAs across human primary cancers. Tumours were grouped into a total of 37 cohorts, ranging in size from two tumours (Cervix-AdenoCa, Lymph-NOS and Myeloid-MDS tumour types) to 314 tumours (Liver-HCC tumour type), in addition to the entire pancancer set (Figure 3A).

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.06.467555; this version posted November 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

After removing likely false positive associations using the same stringent criteria as PCAWG (Campbell et al., 2020), ExInAtor2 analysis revealed altogether 21 unique cancer-lncRNA associations, involving 17 lncRNAs (Figure 3B) – henceforth considered putative "driver lncRNAs". Of these, 9 are annotated lncRNAs that have not previously been linked to cancer, denoted "novel". The remainder "known" candidates are identified in the literature-curated Cancer LncRNA Census dataset (Vancura et al., 2021). Known lncRNAs are hits in more individual cohorts than novel lncRNAs, with cases like *NEAT1* being detected in four cohorts (Figure 3B). While most driver lncRNAs display exonic mutation rates ~50-fold greater than background (coloured cells, Figure 3B), the number of mutations in such genes is diverse between cohorts, being Pancancer, Lymph-CLL and Skin-Melanoma the biggest contributors of mutations.

Supporting the accuracy of these predictions, the set of driver lncRNAs is highly enriched for known cancer lncRNAs (Vancura et al., 2021)(8/17 or 48%, Fisher test P=2e-6) (Figure 3C). Driver lncRNAs are also significantly enriched in three other independent literature-curated databases (Supplementary Figure 3A).

Driver lncRNAs carry features of functionality and clinical relevance

To further evaluate the quality of driver lncRNA predictions, we tested their association with genomic and clinical features expected of *bona fide* cancer genes. LncRNA catalogues are likely to contain a mixture of both functional and non-functional genes. The former group is characterised by purifying evolutionary selection and high expression in healthy and diseased tissues (Carlevaro-Fita et al., 2020). We found that driver lncRNAs display higher evolutionary sequence conservation and higher steady-state levels in healthy organs (Figure 3D). Their sequence also contains more microRNA binding sites, suggesting integration of gene regulatory networks.

In contrast, we could find no evidence that driver lncRNAs are enriched for genomic covariates and features that could reflect artefactual results. They have earlier replication timing (whereas later replication is associated with greater mutation) (Stamatoyannopoulos et al., 2009), less exonic repetitive sequence (ruling out mappability biases), and similar exonic GC content (ruling out sequencing bias) compared to tested non-candidates (Figure 3D). However, driver lncRNAs tend to have greater processed length, likely reflecting greater statistical power for longer genes that affects all driver methods (Lanzós et al., 2017a).

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.06.467555; this version posted November 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Driver lncRNAs also have clinical features of cancer genes (Figure 3E). They are on average 158-fold higher expressed in tumours (133 vs 0.84 FPKM) (Figure 3E, PCAWG RPKM), are 2.15-fold more likely to contain a germline cancer-associated small nucleotide polymorphism (SNPs) in their gene body (4.7% vs 2.5%) (Figure 3E, SNPs per MB), and are enriched in orthologues of driver lncRNAs discovered by transposon insertional mutagenesis (TIM) screens in mouse (17.6 vs 1.6%) (Supplementary Figure 3A) (Vancura et al., 2021). Finally, driver lncRNAs significantly overlap growth-promoting hits discovered by CRISPR functional screens (11.8 vs 1.3%) (Supplementary Figure 3A). In conclusion, driver lncRNA display evidence for functionality across a wide range of functional and clinical features, suggesting that they are enriched for *bona fide* cancer driver genes.

The landscape of lncRNA drivers in metastatic tumours

We further extended the driver lncRNA landscape to metastatic tumours, using 3,527 genomes from 31 cohorts sequenced by the Hartwig Medical Foundation (Supplementary Figure 3A,C,D) (Priestley et al., 2019). Performing a similar analysis as above, we identified 43 driver lncRNAs in a total of 53 lncRNA-tumour combinations (Supplementary Figure 3B). 8 predicted drivers are previously known cancer lncRNAs (P=0.004) (Figure 3C). Notably, predicted lncRNA candidates from this metastatic tumour cohort significantly overlaps the PCAWG primary tumours and sets of previously-published cancer lncRNAs (Figure 3C).

Mutated novel lncRNAs promote cell fitness

A number of uncharacterised lncRNAs were implicated in the above analyses. We next tested whether they do indeed play any role in cell phenotypes of relevance to cancer. AC078785.2 (RP11-572M11.1) displayed elevated mutation rates in Hepatocellular Carcinoma (HCC) tumours (Figure 4A and Supplementary Figure 4B) and has, to our knowledge, never previously been implicated in cancer. According to the latest Gencode version 38, its single annotated isoform comprises three exons, and displays low expression in normal tissues (Supplementary Figure 4A). We could detect RP11-572M11.1 expression in two HCC cell lines, HuH7 and SNU-475. To perturb RP11-572M11.1 expression, we designed two different antisense oligonucleotides (ASOs) that reduced steady-state levels by >50% in both cell lines (Figure 4B,C and Supplementary Figure 4C). To test the role of RP11-572M11.1 in HCC cell proliferation, we evaluated changes in growth rates following ASO transfection, and observed a significant decrease for both ASOs in both cell backgrounds (Figure 4D and Supplementary Figure 4D).

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.06.467555; this version posted November 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

These results prompted us to ask whether RP11-572M11.1 can also promote cell growth in other cancer types. Thus, we turned to CRISPR-activation, to upregulate the lncRNA from its endogenous locus in HeLa cervical carcinoma cells. Three independent sgRNAs significantly activated gene expression by 4 to ~20-fold (Figure 4E) and two out of three led to significantly increased cell proliferation, while a control sgRNA did not (Figure 4F).

To directly address whether tumour mutations alter the observed effect of RP11-572M11.1 on cell proliferation, we used plasmid constructs to overexpress wild-type or mutated forms of the transcript (Figure 4G). The mutated form contained four SNVs, some of them recurrently observed in independent tumours from both PCAWG and HFM dataset (Supplementary Figure 4B). RP11-572M11.1 carrying tumour mutations had a significantly greater effect on cell proliferation, compared to the wild-type sequence (Figure 4H).

Another lncRNA, AC087463.1, was identified as a potential driver in the Head and Neck (HN) tumour cohort (Figure 4I). AC087463.1 is transcribed from the same locus as the lncRNA PWRN1, previously reported as a tumour suppressor gene in gastric cancer (Chen et al., 2018). It is annotated as a single isoform with three exons (Figure 4I), with the mutation falling in the second exon, unique to this gene (Supplementary Figure 4F). A similar strategy as above showed that expression of a mutated form carrying 5 SNVs (Figure 4K) increased tumorigenicity in HN cells (Figure 4J).

Together, these results show that newly-discovered lncRNAs with positively-selected SNVs are capable of promoting fitness in tumour cell backgrounds, and this activity is enhanced by tumour mutations.

Mutations in NEAT1 promote cell fitness and correlate with survival

To directly test whether fitness-enhancing driver mutations may act through lncRNAs, we turned to a relatively well-understood lncRNA, NEAT1, for which abundant mechanistic and functional data is available. The recent PCAWG project and others have noted particularly elevated mutation rates in the NEAT1 gene, although it remained unclear whether these represent drivers or passengers, possibly linked to the high expression of this gene (Fujimoto et al., 2016; Rheinbay et al., 2020; Wedge et al., 2018). NEAT1 produces short and long isoforms (called NEAT1_1 / NEAT1_2) of 3.7 and 22.7 kb, respectively (Sasaki et al., 2009), which are completely overlapping at the 5' of the gene (Figure 5B).

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.06.467555; this version posted November 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

NEAT1_1 is a ubiquitous, abundant, polyadenylated and highly conserved transcript (Nakagawa et al., 2011). In contrast, NEAT1_2, responsible for formation of membraneless nuclear paraspeckle structures, is not polyadenylated and expressed under specific conditions or in response to various forms of stress (Adriaens et al., 2019; McCluggage and Fox, 2021). Based on ExInAtor2 analysis, NEAT1 mutations, spanning the entire gene length, display evidence for positive selection in altogether 4 and 3 cancer cohorts in PCAWG and Hartwig, respectively.

We hypothesised that driver mutations could be simulated by the small indels created by CRISPR-Cas genome-editing (Cho et al., 2018; Liu et al., 2019). We selected six regions of NEAT1 based on high mutation density, evolutionary conservation and known functions (Yamazaki et al., 2018), hereafter called Reg1, Reg2, etc.., and targeted them with altogether 15 sgRNAs (Figure 5A). To control for the known non-specific fitness effects of double strand breaks (DSBs), we also created two neutral control sgRNAs targeting *AAVS1* locus, and a positive-control paired sgRNA (pgRNA) to delete the entire NEAT1_1 region (Figure 5B and Supplementary Figure 5B). Sequencing of treated cells' gDNA revealed narrowly-focussed substitutions and indels at target regions, similar to that observed in real tumours (Figure 5C and Supplementary Figure 5A).

To quantify mutations' effects on cell fitness, we established a competitive growth assay between mutated mCherry-labelled cells and control GFP-labelled cells (Figure 5D) (Cho et al., 2018). As expected, deletion of entire NEAT1_1 in HeLa cells led to reduced growth (KO), while control sgRNAs did not (Figures 5D). Notably, HeLa cells carrying NEAT1 mutations in defined regions displayed increased fitness: two at the 5' of the gene (Reg2 and Reg3), one in the middle section of NEAT1, close to the alternative polyadenylation site (Reg4) and one at the 3' end (Reg5) (red line, Figure 5D). These findings were supported in 3/4 cases in HCT116 colorectal carcinoma cells (green line, Figure 5D) and showed a similar trend in U2OS osteosarcoma cells (Supplementary Figure 5C).

To corroborate these findings, we repeated HeLa fitness assays in the more complex setup of a "mini" pooled competition screen. Here, mixed populations of mutant cells are quantified by amplicon sequencing of sgRNA barcodes. Consistent with previous results, cells carrying NEAT1 mutations at targeted regions outcompeted control cells over time (Figure 5E).

These results were obtained from monolayer cells, whose relevance to real tumours is disputed. Thus, we performed additional experiments in 3-dimensional spheroids grown from HCT116, and found that Reg2 mutations led to increased growth (Figure 5F).

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.06.467555; this version posted November 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The experiments thus far were performed in transformed cancer cells. To investigate whether NEAT1 mutations also enhance cell fitness in a non-transformed background, we performed similar experiments in MRC5 immortalised foetal lung fibroblasts. Again, NEAT1 mutations were observed to increase fitness, in terms of cell growth (Figure 5G) and, at least for Reg2, in terms of anchorage-independent growth (Figure 5H).

We sought independent evidence for the importance of NEAT1 mutations in patient prognosis. Using data from the PCAWG cohort, we asked whether presence of a NEAT1 mutation correlates with shorter survival. Indeed, in lymphoid cancer patients, NEAT1 mutations correlate with significantly worse prognosis (Figure 5I). This effect remains even after accounting for differences in total mutation rates using the Cox proportional hazards model (P=0.02).

In summary, NEAT1 tumour mutations consistently increase cell fitness *in vitro* independent of genetic background, and are associated with poor prognosis lymphoid cancer patients.

Mutations alter NEAT1 protein interactome and increase paraspeckle formation

NEAT1 is a necessary component of subnuclear paraspeckles (Fox et al., 2002; Hutchinson et al., 2007; McCluggage and Fox, 2021), which assemble when specific architectural proteins bind to nascent NEAT1_2 transcripts (Mao et al., 2011). Paraspeckles are nuclear condensates containing diverse gene regulatory proteins (McCluggage and Fox, 2021). They are very often observed in cancer cells, (Adriaens et al., 2016), and they have been associated with poor prognosis (Li et al., 2018a). Thus we next asked whether NEAT1 mutations might affect cell fitness via alterations in paraspeckle number or structure.

We first evaluated changes in NEAT1 expression and isoform usage in response to mutations. Mutations caused no statistically-significant change in NEAT1_1 expression, while deletion of NEAT1_1 reduced steady-state levels, as expected (Figure 6A). Interestingly, the only mutation to significantly increase NEAT1_2 levels was in Region 4, which is consistent with the fact that it contains the alternative polyadenylation site that mediates switching between the short and long isoforms (Naveed et al., 2021).

Using quantitative analysis of fluorescence in situ hybridisation (FISH) with NEAT1_2 probes, we next asked whether mutations impact on paraspeckle number or structure. Despite changes in isoform expression noted above, mutations in Region 4 resulted in no change in the number or size of paraspeckles, in line with previous findings (Yamazaki et al., 2018) (Figure 6D,E). However, mutations in Region 2 yielded a significant increase in number and size of paraspeckles (Figure 6C-E).

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.06.467555; this version posted November 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

NEAT1 is known to function via a diverse cast of protein partners. Region 2 mutations overlap diverse protein binding sites, and fall in or near to areas of deep evolutionary conservation of sequence and structure (Figure 6F).

To better understand how Region 2 mutations alter NEAT1 function, we compared the protein-interactome of wild-type and mutant RNA by *in vitro* pulldown coupled to mass-spectrometry. We created a 288 nt fragment of NEAT1-Region 2 for wild-type (WT) and mutated sequence, the latter containing two SNVs observed in patient tumours (Figure 6G). We performed RNA pull-down with nuclear lysate from HeLa cells, followed by mass spectrometry. Altogether, 154 interacting nuclear proteins were identified for wild-type sequence. Supporting the usefulness of this approach, interacting proteins highly enriched for both known NEAT1-binders and paraspeckle proteins (see Methods) and contain well known examples like NONO (Simko et al., 2020; Yamazaki et al., 2018) (Figure 6H). Comparing mutant to WT interactomes, we observed widespread changes in NEAT1 complexes: altogether 8 (4.6%) proteins are lost by mutant RNA, and 18 (10.3%) gained (Figure 6I).

We investigated whether mutations create or destroy known binding motifs of changing proteins, but could find no evidence of this. However, we did note that mutations lead to increased binding of previously-discovered interactors, U2SURP and PTBP1 (Figure 6I). Intriguingly, increased binding was also observed for PQBP1 protein, whose disordered domain has been linked to condensate formation, offering a potential mechanism in facilitating paraspeckle formation (Kunde et al., 2011). Conversely, STRING analysis revealed that the proteins lost upon mutation are highly enriched for members of the core RNA Polymerase II complex (strength=2.51, P=0.016; basic list enrichment by STRING, Benjamini-Hochberg corrected) and physically interacting with other proteins of this complex (Figure 6J). In summary, tumour mutations in NEAT1 give rise to reconfiguration of the protein interactome, creating several potential mechanisms by which paraspeckle formation is promoted in transformed cells.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.06.467555; this version posted November 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Discussion

Understanding which tumour mutations give rise to pathogenic cell fitness, and how they do it, are major unsolved challenges of cancer research. Here we have focussed on one particularly mutated class of noncoding elements, the lncRNAs, and provided evidence for the first time that somatic mutations can promote cell fitness by altering lncRNA function.

To identify candidate driver lncRNAs, we developed ExInAtor2, an improved driver discovery method that integrates signals of positive selection from recurrence and functional impact. ExInAtor2 outperforms the present state-of-the-art lncRNA driver prediction methods. This improvement is most likely explained by several conceptual advantages of ExInAtor2. Firstly, it makes minimal assumptions about the nature of lncRNA driver SNVs, by independently estimating the recurrence and functional impact prior to p-value integration, in contrast to some methods that require driver mutations to show both signals (Supplementary Notes). Secondly, ExInAtor2 estimates local background mutation rates in a simple and transparent way, removing the need for numerous covariates. Despite this simplicity, we could find no evidence for false positive calls arising from artefactual covariates such as low expression or late replication (Figure 3D). Finally, ExInAtor2's FI module can accept any base-level functional score, making it versatile and capable in future of employing improved scoring schemes.

We applied ExInAtor2 to a large cohort of 2583 primary and 3527 metastatic tumour genomes. In the primary cohort, we identified 8 known and 9 novel driver lncRNAs. The relevance of these candidates is supported by their association with a range of functional and clinical evidence for cancer roles. This represents the largest available resource of candidate driver lncRNAs to date.

This study has experimentally validated driver lncRNAs for the first time. We used a combination of CRISPR-activation, ASO knockdown and transgene overexpression, to test the effect of wild-type and mutated lncRNAs on cancer-relevant phenotypes in cultured cells. We focussed on two novel lncRNAs, AC087463.1 predicted as driver in Head and Neck cancer and RP11-572M11.1 in Hepatocellular Carcinoma. AC087463.1 is transcribed from the same locus and may be an isoform of *PWRN1* lncRNA, previously reported as a tumour suppressor gene in gastric cancer (Chen et al., 2018). RP11-572M11.1 has, to our knowledge, never previously been implicated in cancer. Using knockdown and overexpression we show that 1) overexpression / knockdown of AC087463.1 and RP11-572M11.1 impact cell proliferation in ways consistent with their being oncogenes; 2) introduction of somatic SNVs further boosts cell fitness compared to the wild-type form. Together, these data support the clinical relevance of novel candidate driver lncRNAs discovered by ExInAtor2, and demonstrate that somatic SNVs can promote cell fitness via lncRNAs.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.06.467555; this version posted November 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Consistent with previous driver discovery analyses from us and others, this project identified NEAT1 as a highly-mutated candidate. Years of evidence indicate that upregulation of NEAT1 supports tumorigenesis in a wide variety of cancers, including prostate cancer, gastric cancer, colon carcinoma, and head and neck squamous cell carcinoma (summarized in (Pisani and Baron, 2020). The NEAT1 locus tends to be highly mutated in tumours, however it remains unclear whether this reflects passenger phenomena (possibly due to the extremely high transcription of the gene) or genuine selected driver mutations (Rheinbay et al., 2020).

Our experimental evidence helps to resolve these questions. We employed Cas9 to simulate tumour-like small indels in chromatin of cancer cells, and observed consistent increases in cell proliferation rates. This was observed in a variety of cell backgrounds, including three-dimensional spheroids and non-cancerous cells. These effects were observed in several regions of NEAT1 that have been implicated in RNA function in previous studies (Adriaens et al., 2019; Yamazaki et al., 2018).

These findings also shed light on the mechanism by which NEAT1 SNVs impact cell phenotype. Mutations in Region 2, located at the 5' end of NEAT1, increase paraspeckle number and size, suggesting that the SNVs in the region could affect the protein binding or the efficiency in assembling the RNP complex into mature paraspeckles. We explored this hypothesis by means of mass spectrometry, observing that SNVs result in a profound shift in NEAT1 RNP composition. Of note, we observe that SNVs result in a significant loss of interaction with the RNA Polymerase II complex mediated by known NEAT1 interactor TAF15. Other known protein interactions are potentiated in mutated RNA, suggesting that changes in paraspeckles may be mediated by both gains and losses of protein interactions.

Overall this work has provided the first experimental evidence that fitness-boosting somatic tumour mutations can act via changes to lncRNA function. We have sketched a first mechanistic outline of how this process occurs via altered protein interaction and changes to membraneless organelles, in this case, paraspeckles. Our catalogue of candidate driver lncRNAs across thousands of primary and metastatic tumours, and improved ExInAtor2 software, provides a foundation for future elucidation of the extent and mechanism of driver lncRNAs.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.06.467555; this version posted November 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Acknowledgements

The results shown here are based upon data generated by the TCGA, PCAWG and GTEx consortia. We thank Iñigo Martincorena (Sanger Institute) for generously providing certain data analysis scripts. We thank Federico Abascal (Sanger Institute) for generously providing cancer cell fraction data. We thank Adrian Ochsenbein, Carsten Riether, Simon Haefliger, Thomas Marti, Renwang Peng, (Inselspital University Hospital of Bern) for many insightful conversations. We thank Basak Ginsbourger (DBMR) for administrative support, and Willy Hofstetter and Patrick Furer (DBMR) for logistical support. All computation was performed on the Bern Interfaculty Bioinformatics Unit computing cluster maintained by Rémy Bruggmann and Pierre Berthier. This publication and the underlying study have been made possible partly on the basis of the data that Hartwig Medical Foundation has made available. Work in the Johnson laboratory is funded by the Medical Faculty of the University of Bern, the University Hospital of Bern, the Helmut Horten Stiftung, Swiss Cancer Research Foundation (4534-08-2018), Science Foundation Ireland through Future Research Leaders award 18/FRL/6194, and the Swiss National Science Foundation through the National Centre of Competence in Research (NCCR) “RNA & Disease”.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.06.467555; this version posted November 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Methods

ExInAtor2 algorithm

ExInAtor2 is composed of two separate modules for detection of positive selection: one for recurrence (RE), comparing the exonic mutation rate to that of the local background; another for functional impact (FI), comparing the estimated functional impact of mutations to background, both estimated in exons.

As an improvement to the first version of ExInAtor (Lanzós et al., 2017), the RE module compares the number of observed exonic mutations against a distribution of simulated exonic counts (Supplementary Figure 1A), obtained by random repositioning of the variants the between the exonic and background regions while maintaining the same trinucleotide spectrum. Background region is defined for each gene as introns plus 10 kb up and downstream, after removing nucleotides overlapping exons from any other gene. Exonic and background regions can be further filtered to remove any additional “masked” regions defined by the user. In this manuscript, this functionality was used to mask with low mappability obtained from the UCSC Genome Browser (Supplementary Files).

The use of local background and controlling for trinucleotide content is intended to avoid known sources of false positives arising from covariates in mutational processes and mutational signatures, such as replication timing, gene expression, chromatin state, etc (Lawrence et al., 2013).

A *p*-value is assigned to each gene, being the fraction of simulations with higher or equal number of mutations compared to the observed number (Formula 6).

$$RE_{p-value} = \frac{\#\ of\ simulated\ exonic\ counts \geq observed\ exonic\ count}{total\ \#\ of\ simulations}$$

Formula 6: p-value calculation for the recurrence (RE) module.

The second FI module compares the mean functional score of the observed exonic mutations to a distribution of simulated values. Simulations are performed by random repositioning of mutations in exonic regions, while maintaining identical trinucleotide content (Supplementary Figure 1B). Similar to the RE model, a *p*-value is obtained by comparing the number of simulations with an exonic mean functional score higher or equal to the observed value (Formula 7). This module work with any base-level scoring method. Given its previous successful use and integrative nature, we selected the Combined Annotation Dependent Depletion (CADD) scoring system (Rentzsch et al., 2021; Rentzsch et al., 2019).

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.06.467555; this version posted November 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

$$FI_{p-value} = \frac{\#\ of\ simulated\ exonic\ means \geq observed\ exonic\ mean}{total\ \#\ of\ simulations}$$

Formula 7: *p*-value calculation for the Functional Impact (FI) module.

In a final step, RE and FI *p*-values are combined using the Fisher method (Formula 8).

$$Combined_{p-value} = -2 * \left[ln\ ln\left(RE_{p-value}\right) + ln\ ln\left(FI_{p-value}\right)\right]$$

Formula 8: Fisher method for *p*-value integration.

Tumour somatic mutations

The principal source of mutations were primary tumours from the Pan-Cancer Analysis of Whole Genomes (PCAWG) project (Campbell et al., 2020). This dataset was created according to a uniform and strict methodology, including collection of samples, DNA sequencing and somatic variant calling, aggressive filtering to remove potential artefacts and false positive mutations (Campbell et al., 2020). For practical reasons, we only considered Single Nucleotide Variants (SNVs) arising from substitutions, insertions and deletions of length 1 bp (indels) (Figure 1B). After this filtering, the PCAWG dataset comprises 37 cancer cohorts, 2,583 samples and 45,703,485 SNVs (Figure 1B). Analyses were performed either on individual cohorts, or on the “Pancancer” union of all cohorts.

Gene annotation and filtering

We employed a filtered lncRNA gene annotation based upon Gencode annotation. Beginning with Gencode v19 annotation, we discarded lncRNA genes overlapping protein-coding genes, or containing at least one transcript predicted to be protein-coding by CPAT (Wang et al., 2013), with default settings of coding potential >=0.364. To the remaining list, we added 521 genes from Cancer LncRNA Census (CLC) (Vancura et al., 2021). The resulting set of 6982 lncRNA genes were used here unless otherwise specified (Figure 1C).

ExInAtor2 benchmarking against other driver discovery methods

We collected driver predictions from 10 methods, in addition to the combined predictions generated by the PCAWG driver group (PCAWG combined, PCAWGc) that displayed best overall performance (Rheinbay et al., 2020). We only selected PCAWG methods that were run in both protein-coding and lncRNAs, and for which predictions were available for individual cohorts (Figure 2A).

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.06.467555; this version posted November 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The original PCAWG publication used carefully filtered annotations for protein-coding and lncRNA genes (Rheinbay et al., 2020). Only coding sequences (CDS) of protein-coding genes were considered, while lncRNAs were strictly filtered by distance to protein coding genes, transcript biotype, gene length, evolutionary conservation and RNA expression. For benchmarking, we ran ExInAtor2 using the same PCAWG annotations.

Evaluation of *p*-value distributions

Under the assumption that most genes are not cancer drivers and follow the null distribution, the collection of p-values should mimic a uniform distribution with deviation of a small number of genes at very low p-values (Tokheim et al., 2016). Quantile-quantile plots (QQ-plot) (Figure 2B and Supplementary Figure 3A) display the observed and expected *p*-values in -log10 scale. In order to generate the theoretical distribution for each driver method across all 37 cohorts and the Pancancer set, we ranked the total list of *n* observed p-values from lowest to highest, then for each *i* observed *p*-value we calculated an expected *p*-value according to the uniform distribution (Formula 1).

$$expected_{\,i} = \frac{i}{n}$$

Formula 1: Expected *p*-value calculation. *i* represents the rank of the corresponding observed *p*-value in the total distribution of *n* observed *p*-values, therefore *i* values range from 1 to *n*.

For each driver method, only genes with a reported *p*-value were included in this analysis, i.e., NA cases were discarded. By visual inspection of the QQ-plots, a correct observed distribution of *p*-values should follow a line with 0 as intercept and 1 as slope, where extreme values beyond approximately 2 in the x-axis should deviate above the diagonal line. We used the Mean Log Fold Change (MLFC) (Formula 2) to numerically estimate such deviation and evaluate the performance of driver gene predictions (Tokheim et al., 2016). The closer to zero the MLFC, the better the statistical modelling of passenger genes following the null distribution (Tokheim et al., 2016).

$$MLFC = \frac{1}{n} * \sum_{i}^{n} \left| \left(\frac{observed_i}{expected_i}\right) \right|$$

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.06.467555; this version posted November 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Formula 2: Mean Log Fold Change (MLFC). *n* represents the total number of *p*-values an *i* the lowest *p*-value.

Gene benchmark sets

We downloaded known driver genes from the Cancer Gene Census (Sondka et al., 2018) (CGC) (www.cancer.sanger.ac.uk/census) on 06/02/2019 as a TSV file. We extracted all Gencode *ENSG* identifiers, resulting in a list of 703 genes. For lncRNAs we used the second version of the Cancer LncRNA Census (Vancura et al., 2021), which contains 521 Gencode lncRNAs.

Precision, sensitivity and F1 comparison

CGC and CLC genes were used as ground truth for driver predictions of protein-coding and lncRNAs, respectively. Three metrics were used to compare driver predictions: Precision, the proportion of predictions that are ground truth genes (Formula 3); Sensitivity, the fraction of ground truth genes that are correctly predicted (Formula 4); F1-score, the harmonic mean of precision and sensitivity (Formula 5).

$$Precision = \frac{TP}{TP + FP} * 100$$

Formula 3: Precision.

$$Sensitivity = \frac{TP}{TP + FN} * 100$$

Formula 4: Sensitivity.

$$F1 - score = 2 * \frac{Precision * Sensitivity}{Precision + Sensitivity}$$

Formula 5: F1-score.

Simulated mutation datasets

To generate realistic simulated data, each mutation was randomly repositioned to another position with identical trinucleotide signature (ATA > ATA, being the central nucleotide the one mutated) within a window of 50 kb on the same chromosome.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.06.467555; this version posted November 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Generation and comparison of genomic features

Evolutionary conservation: We downloaded base-level PhastCons scores for all 46way and 100way alignments (Siepel et al., 2005) from the UCSC Genome Browser (Haeussler et al., 2019). We calculated the average value across all exons of each gene.

Expression in normal samples: We obtained RNA-seq expression estimates in transcripts per million (TPM) units for 53 tissues from GTEx (https://gtexportal.org/home/datasets). For tissue specificity, we calculated *tau* values as previously described (Yanai et al., 2005) (https://github.com/severinEvo/gene_expression/blob/master/tau.R).

Replication timing: We collected replication time data of 16 different cell lines from the UCSC browser (Haeussler et al., 2019) (http://genome.ucsc.edu/cgi-bin/hgFileUi?db=hg19&g=wgEncodeUwRepliSeq).

miRNA binding: We downloaded both bioinformatically predicted (miTG scores) and experimentally validated miRNA binding to lncRNAs from LncBase (Paraskevopoulou et al., 2015) (http://carolina.imis.athena-innovation.gr/diana_tools/web/index.php?r=lncbasev2%2Findex).

Tumour expression: Expression values in units of FPKM-uq (referred as PCAWG RPKM in Figure 4D) were obtained from PCAWG (Campbell et al., 2020).

Drug-expression association: We extracted expression-drug association *p*-values from LncMAP (Li et al., 2018b) (http://bio-bigdata.hrbmu.edu.cn/LncMAP).

Germline cancer small nucleotide polymorphisms (SNPs): We downloaded SNPs from the GWAS Catalogue (Buniello et al., 2019) (https://www.ebi.ac.uk/gwas/).

CIS evidence in mice: We downloaded CIS coordinates from CCGD (Abbott et al., 2015) (http://ccgd-starrlab.oit.umn.edu/download.php) and mapped them to human hg19 with LiftOver (https://genome.ucsc.edu/cgi-bin/hgLiftOver) from the UCSC browser (Haeussler et al., 2019). Then, we calculated the number of CIS intersecting each lncRNA divided by the gene length with a custom script using BEDtools (Quinlan and Hall, 2010). CIS per Mb values are available in Supplementary Files.

Survival analysis

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.06.467555; this version posted November 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Survival plots were constructed using donor-centric whole genome mutations dataset, overall survival data and tumour histology data from UCSC Xena Hub: https://xenabrowser.net/datapages/?cohort=PCAWG%20(donor%20centric)&removeHub=https%3A%2F%2Fxena.treehouse.gi.ucsc.edu%3A443. The whole genome mutations file was intersected with comprehensive gene annotation v37 (https://www.gencodegenes.org/human/release_38lift37.html) using BEDtools intersect to isolate donors with mutations in lncRNA of interest. Survival of donors with mutations in lncRNA of interest was then compared against the group of donors without mutations in lncRNA of interest using R packages "survival" (https://cran.r-project.org/web/packages/survival/index.html) and "survminer" (https://cran.r-project.org/web/packages/survminer/index.html)

NEAT1 structure and element analysis

Elements: The window spanning 300 bp around Mut1a and Mut1b (hg19 chr11:65190589-65190888; hg38 chr11:65423118-65423417) was annotated with the program ezTracks (Guillen-Ramirez and Johnson, 2021) using the following datasets as input: (i) structural features: RNA structures conserved in vertebrates (CRS) (Seemann et al., 2017), DNA:RNA triplex structures (Sentürk Cetin et al., 2019), R-Loops lifted over to hg38 (Sanz et al., 2016); (ii) conservation: phastCons conserved elements in 7, 20, 30 and 100-way multiple alignments (Siepel et al., 2005) retrieved from UCSC genome browser (Kent et al., 2002); (iii) high confidence narrow peaks from eCLIP experiments from ENCODE (Davis et al., 2018) (Complete list of accessions is located at Supplementary Table 2).

RBP motif mapping. The 20 bp-padded sequence around Mut1a and Mut1b (hg19 chr11:65190719-65190775) was extracted and then used to generate the sequence of the three distinct alleles WT, only Mut1a and only Mut1b. The three sequences were used as input for de novo RBP motif matching in the web servers RBPmap (Paz et al., 2014) using the option Genome: other and all Human/Mouse motifs) and RBPDB (Cook et al., 2011) (using the default score threshold, 0.8). Outputs were manually parsed and further processed using an in-house Python script.

SNP structural impact analysis. Sequences for the window spanning 300 bp around each mutation target were extracted. Then, only substitutions were kept and encoded according to their relative position and submitted to the MutaRNA web server (Miladi et al., 2020), which also reports scores from RNAsnp (Sabarinathan et al., 2013).

Cell culture

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.06.467555; this version posted November 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

HeLa, HEK 293T and HCT116 were a kind gift from Roderic Guigo's lab (CRG, Barcelona). The MRC5-SV cells were provided by the group of Ronald Dijkmanthe (Institute of Virology and Immunology, University of Bern) and the HN5 tongue squamous cell carcinoma cells by Jeffrey E. Myers (MD Anderson) to Y. Zimmer. All the cell lines were authenticated using Short Tandem Repeat (STR) profiling (Microsynth Cell Line Typing) and tested negative for mycoplasma contamination.

HeLa, HN5 and HEK 293T cell lines were cultured at 37°C in 5% CO2 in Dulbecco’s Modified Eagle's Medium high-glucose (Sigma) supplemented with: 10% FBS (Gibco), 1% L-Glutamine (ThermoFisher), 100 I.U./mL of Penicillin/Streptomycin (Thermo Fisher).

HCT-116 and MRC5-SV were cultured in McCoy (Sigma) and EMEM (Sigma), respectively, both supplemented with 10% FBS (Gibco), 1% L-Glutamine (ThermoFisher), 100 I.U./mL of Penicillin/Streptomycin (Thermo Fisher). SNU-475 (ATCC) and HuH7 (Cell Line Service) hepatocellular carcinoma cell lines were cultured at 37°C in 5% CO2 in RPMI-1640, GlutaMAX™ (Gibco) supplemented with 10% FBS (Gibco) and 100 I.U./mL of Penicillin/Streptomycin (Thermo Fisher).

Gene overexpression and knockdown experiments

Both the wild-type and mutated lncRNA spliced sequences were synthesized by Gene Universal Inc, into pcDNA3.1 vector backbone. Control pcDNA3.1 plasmids contained the sequence of enhanced green fluorescent protein (EGFP).

Overexpression in HN5 cells: For each transfection 1.6 ug of plasmid DNA has been incubated for 20 minutes with 4 µl of Lipofectamine 2000 transfection reagent (Invitrogen) in 0.2 ml of OptiMEM media (Gibco) and added to the cells cultured in a 6-well plate. As all plasmids contain G418 resistance gene, cells were cultured in 2.5 mg/ml of G418 (Gibco) 48h after transfection.

Overexpression in HuH7 cells: For each transfection, 100 ng of plasmid DNA were incubated for 20 minutes with 0.15 µl Lipofectamine 3000 and 0.2 µl P3000 transfection reagent (Invitrogen) in 10 µl RPMI-1640, GlutaMAX™ (Gibco) and added on top of 2000 HuH7 cells cultured in a 96-well plate. Transfection efficiency was measured with qPCR after 120h.

Knockdown in SNU-475 and HuH7 cells: For the transfections, 10 nM of each ASO were incubated with 0.15 µl Lipofectamine 3000 (Invitrogen) for 20 min in 10 µl RPMI-1640, GlutaMAX™ (Gibco) and added on top of 2000 SNU-475 or HuH7 cells cultured in a 96-well plate. Transfection efficiency was measured with qPCR after 144h.

Crystal violet staining

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.06.467555; this version posted November 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Cells were dissociated with 0.05% trypsin-EDTA (Gibco), resuspended in complete media and counted in Neubauer chamber. Subsequently, 1000 cells per well were plated in a 6-well plate, cultured for one week and stained in a 2% Crystal violet (Sigma) solution. The area percentage covered with cells was analysed using ImageJ (%Area). Data analysis was conducted in Graphpad Prism version 8.0.1. One-way ANOVA was used to determine statistical significance, alpha=0.05.

Proliferation assay – SNU-475 and HuH7

After transfection, the proliferative capacity of SNU-475 and HuH7 was measured every 24h by resazurin assay. Briefly, Resazurin sodium salt (Sigma) was added to each well to reach a final concentration of 3 μM and was incubated at 37°C for 2h. Absorbance was measured with Tecan Spark Plate Reader at 545□nm and 590□nm.

CRISPR sgRNA design and cloning

CRISPR activation in HeLa cells was performed as described by Sanson and colleagues (Sanson et al., 2018). sgRNAs were designed using the GPP sgRNA Designer CRISPRa from the Broad Institute (https://portals.broadinstitute.org/gpp/public/). For each sgRNA, forward and reverse DNA oligos were synthesized introducing the BsmB1 overhangs. The two oligos were phosphorylated with the Anza™ T4 PNK Kit (Thermofisher) according to the manufacturer instructions in a 10 μl final volume. The phosphorylation/annealing reaction was set up in a thermocycler at 20° C for 15 min, followed by 95°C for 5 min and then ramp down to 25° C at 5° C/min rate. For ligation of annealed oligos into the pXPR_502 backbone (Addgene #96923), the plasmid was first digested and dephosphorylated with FastDigest BsmBI and FastAP (Thermofisher) at 37°C for 2 hrs. Ligation reaction was carried out with the Rapid DNA Ligation Kit (Thermo) according to the manufacturer instructions.

sgRNAs targeting NEAT1 were designed using the GPP sgRNA Designer CRISPRKo from the Broad Institute (https://portals.broadinstitute.org/gpp/public/), and cloned into the pDECKO backbone (Addgene #78534) as described above.

Lentivirus production

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.06.467555; this version posted November 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

For lentivirus production, HEK293T cells (2.5 x10^6) were seeded in poly-L-lysine coated 100 mm culture dishes 24 hrs prior to transfection. Cells were then co-transfected in serum-free medium with 12.5 µg of the plasmid of interest (Lenti dCAS-VP64_Blast plasmid or sgRNA-containing pXPR_502 or pDECKO), 4 µg of the envelope-encoding plasmid pVSVg (Addgene 12260) and 7.5 µg of the packaging plasmid psPAX2 (Addgene 8454) with Lipofectamine 2000 (ThermoFisher) according to the manufacturer instructions. After 4-6 hrs the medium was replaced with complete DMEM. Virus-containing supernatant was collected after 24, 48 and 72 hours post-transfection. The three harvests were pooled and centrifuged at 3000 rpm for 15 min to remove cells and debris. The supernatant was collected, and for every four volumes, one volume of cold PEG-it Virus Precipitation Solution was added. The mix was refrigerated overnight at 4ºC and centrifuged at 1500 × g for 30 min at 4ºC.The supernatant was discarded, and the sample centrifuged at 1500 × g for 5 min. The lentiviral pellet was suspended in cold, sterile PBS, aliquoted into cryogenic vials and stored at -70°C.

Lentivirus transduction

CRISPRKo: For the generation and transduction of Cas9-expressing cell lines, HeLa, HCT116 and MRC5-SV Cas9 were incubated for 24 hrs with culture medium containing concentrated viral preparation carrying pLentiCas9-T2A-BFP and 8 µg/ml Polybrene. 24 hrs post-infection, antibiotic selection was induced by supplementing the culturing medium with 4 µg/ml blasticidin (Thermofisher) for 5 days. Blasticidin selected cells were subjected to 3 rounds of fluorescence-activated cell sorting (FACS) to isolate high BFP-expressing cells.

CRISPRa: For the generation and transduction of dCas9-expressing cell lines, HeLa cells were incubated for 24 hrs with culture medium containing concentrated viral preparation carrying pLenti dCas9-T2A-BFP-VP64 and 8 µg/ml Polybrene. Cells underwent FACS sorting to enrich for high BFP expressing cells.

sgRNAs: pLentiCas9-T2A-BFP or dCas9-T2A-BFP-VP64 stable cell line were seeded into 6 well plates at 10^6 cells per well and supplemented with sgRNAs pDECKO or pXPR_502 lentiviral preps, respectively, and spinfected in the presence of polybrene (2 µg/ml) for 95 min at 2000 rpm at 37 °C, followed by medium replacement. 24 hrs post-infection, antibiotic selection was induced by supplementing the culturing medium with 2 µg/ml puromycin (Thermofisher) for at least 3 days.

RT-qPCR gene expression analysis

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.06.467555; this version posted November 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

HeLa cells were lysed, and total RNA was extracted by using the Quick-RNA™ Miniprep Kit (Zymo Research). For each sample, RNA was retro-transcribed into cDNA by using the GoScript™ Reverse Transcription System (Promega) and the expression of the target gene was assessed through Real-Time PCR with the GoTaq® qPCR Master Mix. To this purpose target-specific mostly intron-spanning primers (Supplementary Table 1) were designed by using the online tool Primer 3 version 4.1.0.

Cell viability assay

After puromycin selection, cells expressing controls and candidates' guides were collected and seeded in 96-well plates in at least 3 technical replicates for each time point (3000 cells per well). Proliferation assay was performed using the Cell-Titer Glo 2.0 (Promega) reagent according to the manufacturer instructions. Luminescence was measured with the INFINITE 200 PRO series TECAN reader instrument. Time point 0 (T0) reading was performed 4-5 hours after cell seeding.

1:1 competition assay

HeLa, HCT116 and MRC5-SV cells were infected with pDECKO lentiviruses expressing fluorescent proteins. Control plasmids containing sgRNAs targeting *AAVS1* expressed GFP protein (pgRNAs-AASV1-GFP+), while the sgRNAs targeting the different regions of NEAT1 expressed mCherry. After infection, and seven days of puromycin (2 µg/ml) selection, GFP and mCherry cells were mixed 1:1 in a six-well plate (150,000 cells). Cell counts were analysed by LSR II SORP instrument (BD Biosciences) and analysed by FlowCore software.

Pooled competition assay

Screen: HeLa cells stably expressing sgRNAs targeting NEAT1 Reg2, Reg3, Reg4, Reg5 and KO, and HeLa cells stably expressing sgRNAs Control1 and Control2 were counted and mixed in the following ratio 10:10:10:10:25:25. At Day 0, 2M cells were collected, while 2M were plated and passaged every 2-3 days. Cells were harvested at 7, 14, 21 and 28 days for gDNA extraction. The experiment was conducted in six biological replicates.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.06.467555; this version posted November 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Genomic DNA preparation and sequencing: Genomic DNA (gDNA) was isolated using the Blood & Cell Culture DNA Mini (<5e6 cells) Kits (Qiagen, cat. no. 13323) as per the manufacturer's instructions. The gDNA concentrations were quantified by Nanodrop. For PCR amplification, 1 µg of gDNA was amplified in a 200 µl reaction using Q5® High-Fidelity 2X Master Mix (NEB #M0491). PCR master mix (100 µl Q5, and 10 µl of Forward universal primer, and 10 µl of a uniquely barcoded P7 primer (both stock at 10□µM concentration). PCR cycling conditions: an initial 30 sec at 98□°C; followed by 10□sec at 98□°C, 30□sec at 68□°C, 20 sec at 72□°C, for 22 cycles; and a final 2 min extension at 72□°C. NGS primers are listed in Supplementary Table 1. PCR products were purified with Agencourt AMPure XP SPRI beads according to manufacturer's instructions (Beckman Coulter, cat. no. A63880). Purified PCR products were quantified using the Qubit™ dsDNA HS Assay Kit (ThermoFisher, cat. no. Q32854). Samples were sequenced on a HiSeq2000 (Illumina) with paired-end 150 bp reads. The raw sequencing reads from individual samples were analysed by using a custom shell script to count the number of reads containing each sgRNA. The sgRNA counts were then normalized over the T0 and Control2.

Deep sequencing to determine indel spectrum

Genomic DNA was extracted using the Blood & Cell Culture DNA Mini (<5M cells) Kits (Qiagen, cat. no. 13323) as per the manufacturer's instructions. To prepare samples for Illumina sequencing, a two-step PCR was performed to amplify the different regions of NEAT1. For each sample, we performed two separate 100 ul reactions (25 cycles each) with 250 ng of input gDNA using Q5 MASTER MIX (NEB #M0491) and the resulting products were pooled (PCR reaction: 30 sec at 98□°C; followed by 10□sec at 98□°C, 30□sec at 68□°C, 20 sec at 72□°C, for 22 cycles; and a final 2 min extension at 72□°C). PCR amplicons were purified using solid phase reversible immobilization (SPRI) beads, run on a 1.5% agarose gel to verify size and purity, and quantified by Qubit Fluorometric Quantitation (Thermo Fisher Scientific). The resulting DNA was used for reamplification with primers containing Illumina adaptors using the Q5 master Mix. Illumina adaptors and index sequences were added to 100□ng of purified PCR amplicon (PCR reaction: 30 sec at 98□°C; followed by 10□sec at 98□°C, 30□sec at 68□°C, 20 sec at 72□°C, for 8 cycles; and a final 2 min extension at 72□°C).

RNA-FISH and immunofluorescence

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.06.467555; this version posted November 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

HeLa cells grown on coverslips were fixed using 4% paraformaldehyde and permeabilised by 70% ethanol overnight. For RNA-FISH, Stellaris® FISH Probes, targeting Human NEAT1 Middle Segment, labelled with FAM dye (1:100, Biosearch Technologies) were used and the procedure was carried out according to the manufacturer's instructions. Cells nuclei were counterstained with 1:15,000 DAPI (4′,6-diamidino-2-phenylindole) at room temperature and then mounted onto slides by using the VectaShield (Vector Laboratories) mounting media. Fluorescence signals were imaged at 100× (UPLS Apo 100×/1.40) using the DeltaVision Elite Imaging System and Softworx software (GE Healthcare). Images were acquired as Z-stacks, subjected to deconvolution, and projected with maximum intensity. Images were processed using a custom CellProfiler pipeline to determine paraspeckle number and size.

Soft agar assay

The soft agar colony formation assay was performed as previously described (Borowicz S., et al., 2014). Briefly, the assay was carried out in 6-well plates coated with a bottom layer of 1% noble agar in 2X DMEM (ThermoFisher) supplemented with: sodium bicarbonate, 10% FBS (Gibco), 1% L-Glutamine (ThermoFisher), 100 I.U./ml of Penicillin/Streptomycin (ThermoFisher). Then, 7000 cells were suspended in 2X DMEM and 0.6% noble agar. The suspension mixture was subsequently applied as the top agarose layer. A layer of growth medium was added over the upper layer of agar to prevent desiccation. The plates were incubated at 37 °C in 5% CO2 for 3 weeks until colonies formed. After 20 days the colonies were stained with 200 ml of MTT [(3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide), (5 mg/ml), Sigma] and incubated for 3 hours at 37 °C. Numbers of colonies were counted using the analysis software ImageJ.

3D spheroid assay

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.06.467555; this version posted November 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

HCT116 stably expressing Cas9-BFP and sgRNA-mCherry targeting NEAT1 locus were FACS sorted to enrich the population BFP+/mCherry+. The cells were allowed to grow for 7 days, then detached, counted and seeded onto Corning® 96-well Flat Clear Bottom White (Corning, cat. no. 3610) in 20 µl domes of Matrigel® Matrix GFR, LDEV-free (Corning, cat. no. 356231) and McCoy (Sigma, cat. No. M9309) growth medium (1:1) with a density of 10,000 cells per dome in four technical replicates. Matrigel containing the cells was allowed to solidify for an hour in the incubator at 37 °C before adding 80ul of McCoy growth media on top of the wells. The spheroids were allowed to grow in the incubator at 37°C in a humid atmosphere with 5% CO2. After 4 h the number of viable cells in the 3D cell culture was recorded as time point 0 (T0), CellTiter-Glo® 3D Cell Viability Assay (Promega, cat. no. G9682) was added to the wells, following the manufacturer's instructions for the reading with the Tecan Infinite® 200 Pro. After one week the measurement was repeated.

RNA pull-down and Mass Spectrometry

RNA pull-down analysis was performed as previously described (Marín-Béjar O, Huarte M., 2015). Briefly, wild-type and mutant NEAT1 RNA fragments were transcribed in vitro using HiScribe™ T7 High Yield RNA Synthesis Kit (NEB, #E2040S) and labelled with Biotin using Biotin RNA Labelling Mix (Roche, #11685597910) according to the manufacturers' instructions. Biotinylated RNA (10 pmol) was denatured for 10 min at 65 °C in RNA Structure Buffer (10 mM tris-HCl, 10 mM $MgCl_2$, and 100 mM NH_4C1) and slowly cool down to 4 °C. Nuclear fractions were collected as described previously (Carlevaro-Fita J., et al., 2018) and precleared for 30 min at 4 °C using Streptavidin Mag Sepharose® (Sigma, #GE28-9857-99) and NT2 Buffer [50 mM tris-HCl (pH 7.4), 150 mM NaCl, 1 mM MgCl2, 0.05% NP-40,1 mM DTT, 20 mM EDTA, 400 mM vanadyl-ribonucleoside, RNase inhibitor (0.1 U/µl; Promega), and l× protease inhibitor cocktail (Sigma)]. The precleared nuclear lysates (2 mg) were incubated with purified biotinylated RNA in NT2 buffer along with Yeast tRNA (20 µg/ml; Thermo Fisher Scientific #AM7119) with gentle rotation for 1.5 hours at 4°C. Washed Streptavidin Magnetic Beads were added to each binding reaction and further incubated at 4 °C for 1□h to precipitate the RNA-protein complexes. Beads were washed briefly five times with NT2 Buffer, and the retrieved proteins were then subjected to mass spectrometry analysis, performed by the Proteomics & Mass Spectrometry Core Facility (PMSCF) of the University of Bern, Switzerland, using MaxQuant software for protein identification and quantification.

Mass Spectrometry Data Processing

Intensity Based Absolute Quantification (iBAQ) and label-free quantitation (LFQ) intensities from the MaxQuant output were used for quantitative within-sample comparisons

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.06.467555; this version posted November 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

and fold-enrichment between-sample comparisons respectively. A protein was considered enriched / depleted in a sample condition if its intensity was at least 2-fold greater / lesser than in the reference condition (proteins not detected in one of the conditions are imputed with the lowest value for that sample by MaxQuant). Additionally, the resulting lists of proteins were filtered for nuclear localization (Uhlen et al., 2015) to exclude potential false positives. To calculate the significance of the overlap with known NEAT1 binding proteins (Huang et al., 2020; Spiniello et al., 2018; West et al., 2014) and known paraspeckle proteins (McCluggage and Fox, 2021) a hypergeometric test was applied to the background of all nuclear proteins (n=6758). STRING was used for interaction analysis (physical subnetwork, minimum interaction score=0.4, max number of direct interactors=10) and GO term enrichment analysis (Szklarczyk et al., 2019). Visualization of the results was done with R version 4.1.1 and BioRender.com.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.06.467555; this version posted November 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure Legends

Figure 1

A) ExInAtor2 accepts input in the form of maps of single nucleotide variants (SNVs) from cohorts of tumour genomes. Two signatures of positive selection are evaluated and compared to simulated local background distributions, to evaluate statistical significance. The two significance estimates are combined using Fisher's method.

B) Summary of the primary tumour datasets used here, obtained from Pancancer Analysis of Whole Genomes (PCAWG) project.

C) A filtered lncRNA gene annotation was prepared, and combined with a set of curated cancer lncRNAs from the Cancer LncRNA Census (Vancura et al., 2021).

Figure 2

A) The list of driver discovery methods to which ExInAtor2 was compared. The signatures of positive selection employed by each method are indicated to the right. PCAWGc indicates the combined driver prediction method from Pan-Cancer Analysis of Whole Genomes (PCAWG), which integrates all ten methods.

B) Benchmark gene sets. LncRNAs (blue) were divided in positives and negatives according to their presence or not in the Cancer LncRNA Census (Vancura et al., 2021), respectively, and similarly for protein-coding genes in the Cancer Gene Census (Sondka et al., 2018).

C) Comparing performance in terms of precision in identifying true positive known cancer lncRNAs from the CLC dataset, using PCAWG Pancancer cohort. *x*-axis: genes sorted by increasing *p*-value. *y*-axis: precision, being the percentage of true positives amongst cumulative set of candidates at increasing *p*-value cutoffs. Horizontal black line shows the baseline, being the percentage of positives in the whole list of tested genes. Coloured dots represent the precision at cutoff of $q \leq 0.1$. Inset: Performance statistics for cutoff of $q \leq 0.1$.

D) Driver prediction performance for all methods in all PCAWG cohorts. Cells show the F1-score of each driver method (*x*-axis) in each cohort (*y*-axis). Grey cells correspond to cohorts where the method was not run. The bar plot on the top indicates the total, non-redundant number of True Positives (TP) and False Positives (FP) calls by each method. Driver methods are sorted from left to right according to the F1-score of unique candidates.

E) Evaluation of *p*-value distributions for driver lncRNA predictions. Quantile-quantile plot (QQ-plot) shows the distribution of observed vs expected –log10 *p*-values for each method run on the PCAWG Pancancer cohort. The Mean Log-Fold Change (MLFC) quantifies the difference between observed and expected values (Methods).

Figure 3

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.06.467555; this version posted November 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

A) "Oncoplot" overview of driver lncRNA analysis in PCAWG primary tumours. Rows: 17 candidate driver lncRNAs at cutoff of q ≤ 0.1. Columns: 2580 tumours.

B) LncRNA candidates across all cohorts. Rows: Cohorts where hits were identified. Columns: 17 candidate driver lncRNAs. "Known" lncRNAs are part of the literature-curated Cancer LncRNA Census (CLC2) dataset (Vancura et al., 2021). Functional labels (oncogene / tumour suppressor / both) were also obtained from the same source.

C) Intersection of candidate driver lncRNAs identified in PCAWG primary tumours, Hartwig Medical Foundation (HMF) metastatic tumours and the CLC2 set. Statistical significance was estimated by Fisher's exact test.

D) Genomic features of driver lncRNAs. Each plot displays the values of indicated features for 17 candidate driver lncRNAs (blue) and all remaining tested lncRNAs (non-candidates, grey). Significance was calculated using Wilcoxon test. For each comparison, the ratio of means was calculated as (mean of candidate values / mean of non-candidate values). See Methods for more details.

E) Clinical features of driver lncRNAs. Each point represents the indicated feature. *y*-axis: log2-transformed ratio of the mean candidate value and mean non-candidate value. *x*-axis: The statistical significance of candidate vs non-candidate values, as estimated by Wilcoxon test and corrected for multiple testing. See Methods for more details.

Figure 4

A) The genomic locus of hepatocellular carcinoma (HCC) candidate driver lncRNA RP11-572M11.1. Also shown are SNVs from PCAWG and the three sgRNAs used in CRISPRa experiments.

B) Antisense oligonucleotides (ASOs) were transfected into cells to knock down expression of target lncRNAs.

C) Reverse transcription quantitative polymerase chain reaction (qRT-PCR) measurement of RNA levels in HuH HCC cells after transfection of control ASO, or two different ASOs targeting RP11-572M11.1. Statistical significance was estimated using one-sided Student's *t*-test with n=3 independent replicates.

D) Populations of ASO-transfected cells were measured at indicated time points. Each measurement represents n=3 independent replicates.

E) Overview and performance of CRISPR-activation (CRISPRa) targeting RP11-572M11.1. On the right, qRT-PCR measurements of RNA levels with indicated sgRNAs in HeLa cells. Values were normalised to the housekeeping gene HPRT1 and to a control sgRNA targeting the AAVS1 locus. Values represent n=3 independent replicates.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.06.467555; this version posted November 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

F) The effect of CRISPRa on HeLa cells' viability, as measured by Cell Titre Glo reagent. Values represent n=6 independent replicates, and statistical significance was estimated by comparison to the Control sgRNA by paired *t*-test at the 48 hrs timepoint.

G) Plasmids expressing spliced RP11-572M11.1 sequence, in wild-type (WT) or mutated (Mut) form were transfected into HuH cells. The steady state levels of RNA were measured by qRT-PCR and normalised to cells transfected with similar EGFP-expressing plasmid. Values represent n=3 independent replicates, each one with 6 technical replicates.

H) Populations of plasmid-transfected cells were measured at indicated timepoints. Statistical significance was estimated by one-sided Student's *t*-test based on n=3 independent replicates.

I) The genomic locus of head and neck cancer candidate driver lncRNA AC087463.1. Also shown are SNVs from PCAWG.

K) Plasmids expressing spliced AC087463.1 sequence, in wild-type (WT) or mutated (Mut) form were transfected into HN5 cells. The steady state levels of RNA were measured by qRT-PCR and normalised to cells transfected with similar EGFP-expressing plasmid. Values represent n=3 independent replicates.

J) Results of colony formation assay in HN5 cells. Values indicate the percent of well area covered. Statistical significance was estimated using One-way ANOVA has been used to determine statistical significance, based on 18 culture wells.

Figure 5

A) Overview of the experimental strategy to simulate tumour mutations in the NEAT1 lncRNA gene by wild-type Cas9 protein.

B) A detailed map of the six NEAT1 target regions and 15 sgRNAs. Paired gRNAs used for the deletion of NEAT1_1 are indicated as KO- sgRNA1 and KO- sgRNA2. Previously described functional regions of NEAT1 are indicated below, according to the publication of Yamazaki and colleagues (Yamazaki et al., 2018).

C) Analysis of mutations created by Cas9 recruitment. The target region was amplified by PCR and sequenced. The frequency, size and nature of resulting DNA mutations are plotted.

D) Competition assay to evaluate fitness effects of mutations. Above: Rationale for the assay. Labelled mutated (mCherry, red) and control (GFP, green) cells are mixed in equal proportions at the start of the experiment. At successive timepoints their red/green ratio is measured by flow cytometry, and this value is used to infer fitness effects. Below: Red/green ratios for indicated mutations. "Control1/2" indicate sgRNAs targeting intergenic regions. "KO" indicates paired sgRNAs designed to delete the entire NEAT1_1 region. Separate experiments were performed in HeLa and HCT116 cells. n=4 replicated experiments were performed, and statistical significance was estimated by linear regression model.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.06.467555; this version posted November 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

E) Upper panel: Setup of mini CRISPR fitness screen. HeLa cells are infected with lentivirus carrying defined mixtures of sgRNAs. The sgRNA sequences are amplified and sequenced at defined timepoints. Changes in abundance reflect effects on cell fitness. Lower panel: Abundances of displayed sgRNAs, normalised to the Control 2 negative control. n=4 independent experiments were performed, and statistical significance was estimated by linear regression model.

F) HCT116 cells were cultured as spheroids and their population measured. n=4 replicated experiments were performed, and statistical significance was estimated using Student's one-sided *t*-test.

G) As for Panel D, but with non-transformed MRC5 lung fibroblast cells at timepoint Day 14. Statistical significance was estimated by one-sided Student's t-test based on n=3 independent replicates.

H) MRC5 cells were seeded in soft agar, and the area of colonies at 3 weeks were calculated. The mean of n=2 replicated experiments are shown.

I) The survival time of 184 lymphoid cancer patients from PCAWG is displayed. Patients were stratified according to whether they have ≥1 SNVs in the NEAT1 gene.

Figure 6

A) Normalised steady state RNA levels of NEAT1, as estimated using primers for the NEAT1_1 region. Statistical significance was estimated using Student's one-sided *t*-test. P-values ≥0.05 are not shown.

B) As for Panel A, but using primers for NEAT1_2.

C) Representative images from fluorescence in situ hybridisation (FISH) visualisation of NEAT1 in HeLa cells expressing sgRNAs for Control 2 and NEAT1 Region 2.

D) Counts of paraspeckles in HeLa cells treated with indicated sgRNAs, normalised and compared to Control 2 cells. Values were obtained from 80-100 cells per replicate. N=5 biological replicates. Statistical significance was estimated using paired t-test.

E) As for Panel D, but displaying paraspeckle size.

F) The figure shows various genomic features overlapping NEAT1 Region 2 (highlighted). CRS: Conserved RNA Structure (Seemann et al., 2017). Red bars indicate RNA-binding proteins located by eCLIP (Davis et al., 2018). RNA:DNA indicates experimentally-mapped chromatin interacting domains (Sentürk Cetin et al., 2019). Pink bars indicate conserved regions from PhastCons (Siepel et al., 2005).

G) Sequences of biotinylated probes used for mass-spectrometry analysis of NEAT1-interacting proteins.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.06.467555; this version posted November 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

H) Proteins detected by wild-type (WT) NEAT1 probe, filtered for nuclear proteins only, are ranked by intensity and labelled when intersecting databases of previously-detected NEAT1-interacting proteins (green) and paraspeckle proteins (orange). Statistical significance was calculated by hypergeometric test (to background of all nuclear proteins n=6758).

I) Histogram shows differential detection of proteins comparing mutated (Mut) and wild-type (WT) probes. Dotted lines indicate log2 fold-change cutoffs of -1 / +1.

J) STRING interaction network based on a subset of the proteins lost upon mutation (grey borders) interacting with the RNA polymerase II core complex.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.06.467555; this version posted November 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Supplementary Figure Legends

Supplementary Figure 1

A) Graphic representation of the Mutational Burden module of ExInAtor2. Genes are divided in exons and background regions (including introns and flanking regions defined by the user). The number of real exonic mutations is calculated ('observed'). Mutations in both exons and background are randomly shuffled, maintaining the overall trinucleotide content, a number of times indicated by the user (10,000 as example), and the number of exonic mutations in each iteration is recorded. The *p*-value is calculated by counting how many shuffled values are higher or equal than the real mean score, divided by the total number of shuffles performed.

B) Graphic representation of the Functional Impact module of ExInAtor2. Only exonic regions are considered. The mean functional impact score is calculated for all mutated positions. In a number of simulations defined by the user (10,000 as example), mutations are randomly shuffled along the exons while maintaining the trinucleotide signature, and the mean functional impact score is recorded. The *p*-value is calculated by counting how many shuffled values are higher or equal than the real mean score, divided by the total number of shuffles.

Supplementary Figure 2

A) Evaluation of *p*-value distributions for protein-coding genes. Quantile-quantile plot (QQ-plot) shows the distribution of observed vs expected –log10 *p*-values from each methods across all cohorts. The Mean Log-Fold Change (MLFC) quantifies the difference between observed and expected values (Methods).

B) Benchmark in PCAWG Pancancer dataset for protein-coding genes. *x*-axis represents genes sorted increasingly by *p*-value for each method. *y*-axis shows the percentage of true positives amongst cumulative set of candidates (precision) at each step of the *x*-axis (precision). Black line shows the baseline, being the percentage of positives in the whole list of tested genes, ie the precision expected by random chance. Coloured dots represent the number of candidates for each method with *q*-value<=0.1. Table shows the number of True Positives (TP), False Positives (FP) and F1-score (Methods) for each driver method.

C) Protein-coding benchmark for all PCAWG cohorts. Cells show the F1-score of each driver method (*x*-axis) in each cohort (*y*-axis). Grey cells correspond to cohorts where the method was not run. The bar plot at the top indicates the non-redundant total number of True Positives (TP) and False Positives (FP) unique candidates detected across all cohorts, i.e., if a gene is detected in multiple cohorts it is considered only once. Driver methods on the *x*-axis are sorted from left to right according to the F1-score of unique candidates.

Supplementary Figure 3

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.06.467555; this version posted November 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

A) Clinical / disease properties of lncRNA drivers. Displayed are the percentage of lncRNAs that fulfill indicated criteria. LncRNAs are divided into candidate drivers from PCAWG analysis (red), and all other tested lncRNAs (blue). ‘Growth-promoting’ refer to lncRNAs necessary for proliferation of cancer cell lines discovered by CRISPRi perturbation pooled screens of Liu et al (Liu et al., 2017); ‘Transposon insertional mutagenesis’ refer to lncRNA genes orthologous to sites carrying tumour-initiating *Sleeping Beauty* insertions from mouse genome-wide screens (Vancura et al., 2021); ‘CRlncRNA’ – database of cancer lncrnas (Wang et al., 2018); ‘Lnc2cancer’ – database of cancer lncRNAs (Ning et al., 2016); ‘MiTranscriptome’ – differentially expressed lncRNAs in tumours (Iyer et al., 2015); ‘Cancer LncRNA Census 2’ (Vancura et al., 2021). Statistical significance was calculated by Fisher’s exact test.

B) Statistics of Hartwig Medical Foundation (HMF) tumour SNV data (Priestley et al., 2019).

C) Candidate driver lncRNAs discovered in HMF cohort. ExInAtor2 hits are displayed with a cutoff of FDR<0.1. Numbers inside cells represent the number of exonic SNVs. D) Oncoplot summarising lncRNA SNVs in the HMF cohort.

Supplementary Figure 4

A) Bulk tissue gene expression for RP11-572M11.1. Data was obtained from GTEXportal. **B)** RP11-572M11.1 gene locus, highlighting the region containing mutations tested in experiments (black bar). Below: Somatic mutations from PCAWG and HFM. C) Reverse transcription quantitative polymerase chain reaction (qRT-PCR) measurement of RNA levels in SNU476 HCC cells after transfection of control ASO, or two different ASOs targeting RP11-572M11.1. Statistical significance was estimated using one-sided Student’s t-test with n=3 independent replicates. D) Populations of ASO-transfected cells were measured at indicated time points. Each measurement represents n=3 independent replicates. Statistical significance was estimated using one-sided Student’s t-test with n=3 independent replicates at the last time point. **F)** AC087463.1 gene locus, with a zoom in the region of the mutations selected for experimental validations, indicated in the black bar. Also shown are somatic mutations from PCAWG and HFM.

Supplementary Figure 5

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.06.467555; this version posted November 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

A) We performed deep sequencing to determine the resulting indel distribution arising from targeting Cas9 to Region 2. The reference sequence and the sgRNA are shown on top. **B)** Genomic deletion of NEAT1_1 gene using paired gRNAs. The panel shows agarose gel electrophoresis of PCR product (on the left) and a cartoon showing the deletion strategy (on the right) with primers amplifying the NEAT1 target region. **C)** Competition assay to evaluate fitness effects of mutations. Red/green ratios for indicated mutations. “Control1/2” indicate sgRNAs targeting intergenic regions. “KO” indicates paired sgRNAs designed to delete the entire NEAT1_1 region. Separate experiments were performed in HeLa, HCT116 and U2OS cells. n=4 replicated experiments were performed, and statistical significance was estimated by linear model.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.06.467555; this version posted November 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References

- Abbott, K.L., Nyre, E.T., Abrahante, J., Ho, Y.-Y., Isaksson Vogel, R., and Starr, T.K. (2015). The Candidate Cancer Gene Database: a database of cancer driver genes from forward genetic screens in mice. Nucleic Acids Res. *43*, D844–D848.
- Adriaens, C., Standaert, L., Barra, J., Latil, M., Verfaillie, A., Kalev, P., Boeckx, B., Wijnhoven, P.W.G., Radaelli, E., Vermi, W., et al. (2016). P53 induces formation of NEAT1 lncRNA-containing paraspeckles that modulate replication stress response and chemosensitivity. Nat. Med. *22*, 861–868.
- Adriaens, C., Rambow, F., Bervoets, G., Silla, T., Mito, M., Chiba, T., Asahara, H., Hirose, T., Nakagawa, S., Jensen, T.H., et al. (2019). The long noncoding RNA NEAT1_1 is seemingly dispensable for normal tissue homeostasis and cancer cell growth. Rna *25*, 1681–1695.
- AH, F., YW, L., AK, L., CE, L., J, A., M, M., and AI, L. (2002). Paraspeckles: a novel nuclear domain. Curr. Biol. *12*, 13–25.
- Akrami, R., Jacobsen, A., Hoell, J., Schultz, N., Sander, C., and Larsson, E. (2013). Comprehensive Analysis of Long Non-Coding RNAs in Ovarian Cancer Reveals Global Patterns and Targeted DNA Amplification. PLoS One *8*, e80306.
- Buniello, A., MacArthur, J.A.L., Cerezo, M., Harris, L.W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., et al. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. *47*, D1005–D1012.
- Campbell, P.J., Getz, G., Korbel, J.O., Stuart, J.M., Jennings, J.L., Stein, L.D., Perry, M.D., Nahal-Bose, H.K., Ouellette, B.F.F., Li, C.H., et al. (2020). Pan-cancer analysis of whole genomes. Nature *578*, 82–93.
- Carlevaro-Fita, J., A, L., L, F., C, H., D, M.-P., JS, P., and R, J. (2020). Cancer LncRNA Census reveals evidence for deep functional conservation of long noncoding RNAs in tumorigenesis. Commun. Biol. *3*.
- Castro-Giner, F., Ratcliffe, P., and Tomlinson, I. (2015). The mini-driver model of polygenic cancer evolution. Nat. Rev. Cancer 2015 1511 *15*, 680–685.
- Cho, S.W., Xu, J., Sun, R., Mumbach, M.R., Carter, A.C., Chen, Y.G., Yost, K.E., Kim, J., He, J., Nevins, S.A., et al. (2018). Promoter of lncRNA Gene PVT1 Is a Tumor-Suppressor DNA Boundary Element. Cell *173*, 1398-1412.e22.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.06.467555; this version posted November 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

- Collins, M.A., Bednar, F., Zhang, Y., Brisset, J.-C., Galbán, S., Galbán, C.J., Rakshit, S., Flannagan, K.S., Adsay, N.V., and Magliano, M.P. di (2012). Oncogenic Kras is required for both the initiation and maintenance of pancreatic cancer in mice. J. Clin. Invest. *122*, 639.
- Cook, K.B., Kazan, H., Zuberi, K., Morris, Q., and Hughes, T.R. (2011). RBPDB: a database of RNA-binding specificities. Nucleic Acids Res. *39*, D301–D308.
- Corona, R.I., Seo, J.H., Lin, X., Hazelett, D.J., Reddy, J., Fonseca, M.A.S., Abassi, F., Lin, Y.G., Mhawech-Fauceglia, P.Y., Shah, S.P., et al. (2020). Non-coding somatic mutations converge on the PAX8 pathway in ovarian cancer. *11*.
- Davis, C.A., Hitz, B.C., Sloan, C.A., Chan, E.T., Davidson, J.M., Gabdank, I., Hilton, J.A., Jain, K., Baymuradov, U.K., Narayanan, A.K., et al. (2018). The Encyclopedia of DNA elements (ENCODE): Data portal update. Nucleic Acids Res. *46*, D794–D801.
- Elliott, K., and Larsson, E. (2021). Non-coding driver mutations in human cancer. Nat. Rev. Cancer *21*, 500–509.
- Frankish, A., Diekhans, M., Ferreira, A.M., Johnson, R., Jungreis, I., Loveland, J., Mudge, J.M., Sisu, C., Wright, J., Armstrong, J., et al. (2019). GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res. *47*, D766–D773.
- Fujimoto, A., Furuta, M., Totoki, Y., Tsunoda, T., Kato, M., Shiraishi, Y., Tanaka, H., Taniguchi, H., Kawakami, Y., Ueno, M., et al. (2016). Whole-genome mutational landscape and characterization of noncoding and structural mutations in liver cancer. Nat. Genet. *48*, 500–509.
- Gloss, B.S., and Dinger, M.E. (2018). Realizing the significance of noncoding functionality in clinical genomics. Exp. Mol. Med. 2018 508 *50*, 1–8.
- Guillen-Ramirez, H.A., and Johnson, R. (2021). ezTracks v0.1.0.
- Haeussler, M., Zweig, A.S., Tyner, C., Speir, M.L., Rosenbloom, K.R., Raney, B.J., Lee, C.M., Lee, B.T., Hinrichs, A.S., Gonzalez, J.N., et al. (2019). The UCSC Genome Browser database: 2019 update. Nucleic Acids Res. *47*, D853–D858.
- Hornshøj, H., MM, N., NA, S.-A., MP, Ś., M, J., T, M., R, S., M, K., T, Ø., A, H., et al. (2018). Pan-cancer screen for mutations in non-coding elements with conservation and cancer specificity reveals correlations with expression and survival. NPJ Genomic Med. *3*.
- Hu, X., Y, F., D, Z., SD, Z., Z, H., J, G., Y, Z., L, Y., X, Z., LP, W., et al. (2014). A functional genomic approach identifies FAL1 as an oncogenic long noncoding RNA that associates with BMI1 and represses p21 expression in cancer. Cancer Cell *26*, 344–357.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.06.467555; this version posted November 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

- Huang, J., Sachdeva, M., Xu, E., Robinson, T.J., Luo, L., Ma, Y., Williams, N.T., Lopez, O., Cervia, L.D., Yuan, F., et al. (2020). The long noncoding RNA NEAT1 promotes sarcoma metastasis by regulating RNA splicing pathways. Mol. Cancer Res. *18*, 1534–1544.
- Hutchinson, J.N., Ensminger, A.W., Clemson, C.M., Lynch, C.R., Lawrence, J.B., and Chess, A. (2007). A screen for nuclear transcripts identifies two linked noncoding RNAs associated with SC35 splicing domains. BMC Genomics *8*, 39.
- Iyer, M.K., Niknafs, Y.S., Malik, R., Singhal, U., Sahu, A., Hosono, Y., Barrette, T.R., Prensner, J.R., Evans, J.R., Zhao, S., et al. (2015). The landscape of long noncoding RNAs in the human transcriptome. Nat. Genet. *47*, 199–208.
- Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. (2002). The human genome browser at UCSC. Genome Res. *12*, 996–1006.
- Khurana, E., Fu, Y., Chakravarty, D., Demichelis, F., Rubin, M.A., and Gerstein, M. (2016). Role of non-coding sequence variants in cancer (Nature Publishing Group).
- Kim, K., K, J., W, Y., EY, C., SM, P., M, B., YJ, K., and JK, C. (2016). Chromatin structure-based prediction of recurrent noncoding mutations in cancer. Nat. Genet. *48*, 1321–1326.
- Kumar, S., Warrell, J., Li, S., McGillivray, P.D., Meyerson, W., Salichos, L., Harmanci, A., Martinez-Fundichely, A., Chan, C.W.Y., Nielsen, M.M., et al. (2020). Passenger Mutations in More Than 2,500 Cancer Genomes: Overall Molecular Functional Impact and Consequences. Cell *180*, 915-927.e16.
- Kunde, S.A., Musante, L., Grimme, A., Fischer, U., Müller, E., Wanker, E.E., and Kalscheuer, V.M. (2011). The X-chromosome-linked intellectual disability protein PQBP1 is a component of neuronal RNA granules and regulates the appearance of stress granules. Hum. Mol. Genet. *20*, 4916–4931.
- Lanzós, A., Carlevaro-Fita, J., Palumbo, E., Reverter, F., Mularoni, L., Guigó, R., Johnson, R., Reverter, F., Palumbo, E., Guigó, R., et al. (2017a). Discovery of Cancer Driver Long Noncoding RNAs across 1112 Tumour Genomes: New Candidates and Distinguishing Features. Sci. Rep. *7*, 41544.
- Lanzós, A., Carlevaro-Fita, J., Mularoni, L., Reverter, F., Palumbo, E., Guigó, R., and Johnson, R. (2017b). Discovery of Cancer Driver Long Noncoding RNAs across 1112 Tumour Genomes: New Candidates and Distinguishing Features. Sci. Rep. *7*, 41544.
- Lawrence, M.S., Stojanov, P., Polak, P., Kryukov, G. V., Cibulskis, K., Sivachenko, A., Carter, S.L., Stewart, C., Mermel, C.H., Roberts, S.A., et al. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature *499*, 214–218.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.06.467555; this version posted November 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

- Leucci, E., Vendramin, R., Spinazzi, M., Laurette, P., Fiers, M., Wouters, J., Radaelli, E., Eyckerman, S., Leonelli, C., Vanderheyden, K., et al. (2016). Melanoma addiction to the long non-coding RNA SAMMSON. Nature *531*, 518–522.
- Li, K., Y, Z., X, L., Y, L., Z, G., H, C., KE, D., M, C., W, C., Z, S., et al. (2020). Noncoding Variants Connect Enhancer Dysregulation with Nuclear Receptor Signaling in Hematopoietic Malignancies. Cancer Discov. *10*, 724–745.
- Li, X., Wang, X., Song, W., Xu, H., Huang, R., Wang, Y., Zhao, W., Xiao, Z., and Yang, X. (2018a). Oncogenic Properties of NEAT1 in Prostate Cancer Cells Depend on the CDC5L–AGRN Transcriptional Regulation Circuit. Cancer Res. *78*, 4138–4149.
- Li, Y., Li, L., Wang, Z., Pan, T., Sahni, N., Jin, X., Wang, G., Li, J., Zheng, X., Zhang, Y., et al. (2018b). LncMAP: Pan-cancer atlas of long noncoding RNA-mediated transcriptional network perturbations. Nucleic Acids Res. *46*, 1113–1123.
- Liu, E.M., Martinez-Fundichely, A., Diaz, B.J., Apostolou, E., Sanjana, N.E., and Khurana Correspondence, E. (2019). Identification of Cancer Drivers at CTCF Insulators in 1,962 Whole Genomes.
- Liu, S.J., Horlbeck, M.A., Cho, S.W., Birk, H.S., Malatesta, M., He, D., Attenello, F.J., Villalta, J.E., Cho, M.Y., Chen, Y., et al. (2017). CRISPRi-based genome-scale identification of functional long non-coding RNA loci in human cells. Science (80-.). *06*, 1–19.
- M, G., M, C.-H., S, D., Gandhi, M., Caudron-Herger, M., Diederichs, S., M, G., M, C.-H., and S, D. (2018). RNA motifs and combinatorial prediction of interactions, stability and localization of noncoding RNAs. Nat. Struct. Mol. Biol. *25*, 1070–1076.
- Mao, Y.S., Sunwoo, H., Zhang, B., and Spector, D.L. (2011). Direct visualization of the co-transcriptional assembly of a nuclear body by noncoding RNAs. Nat. Cell Biol. *13*, 95–101.
- McCluggage, F., and Fox, A. (2021). Paraspeckle nuclear condensates: Global sensors of cell stress? Bioessays *43*.
- Melton, C., Reuter, J.A., Spacek, D. V, and Snyder, M. (2015). Recurrent somatic mutations in regulatory regions of human cancer genomes. Nat. Publ. Gr. *47*.
- Miladi, M., Raden, M., Diederichs, S., and Backofen, R. (2020). MutaRNA: analysis and visualization of mutation-induced changes in RNA structure. Nucleic Acids Res. *48*, W287–W291.
- Nakagawa, S., Naganuma, T., Shioi, G., and Hirose, T. (2011). Paraspeckles are subpopulation-specific nuclear bodies that are not essential in mice. J. Cell Biol. *193*, 31–39.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.06.467555; this version posted November 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

- Naveed, A., JA, C., R, L., A, H., J, C., T, L., SD, W., S, F., and AH, F. (2021). NEAT1 polyA-modulating antisense oligonucleotides reveal opposing functions for both long non-coding RNA isoforms in neuroblastoma. Cell. Mol. Life Sci. *78*, 2213–2230.
- Ning, S., Zhang, J., Wang, P., Zhi, H., Wang, J., Liu, Y., Gao, Y., Guo, M., Yue, M., Wang, L., et al. (2016). Lnc2Cancer: a manually curated database of experimentally supported lncRNAs associated with various human cancers. Nucleic Acids Res. *44*, D980–D985.
- Nowell, P.C. (1976). The clonal evolution of tumor cell populations. Science *194*, 23–28.
- P, R., M, S., J, S., and M, K. (2021). CADD-Splice-improving genome-wide variant effect prediction using deep learning-derived splice scores. Genome Med. *13*.
- Paraskevopoulou, M.D., Vlachos, I.S., Karagkouni, D., Georgakilas, G., Kanellos, I., Vergoulis, T., Zagganas, K., Tsanakas, P., Floros, E., Dalamagas, T., et al. (2015). DIANA-LncBase v2: indexing microRNA targets on non-coding transcripts. Nucleic Acids Res. *44*, D231–D238.
- Paz, I., Kosti, I., Ares, M., Cline, M., and Mandel-Gutfreund, Y. (2014). RBPmap: a web server for mapping binding sites of RNA-binding proteins. Nucleic Acids Res. *42*, W361–W367.
- Pisani, G., and Baron, B. (2020). and Chemoresistance. 1–13.
- Priestley, P., Baber, J., Lolkema, M.P., Steeghs, N., de Bruijn, E., Shale, C., Duyvesteyn, K., Haidari, S., van Hoeck, A., Onstenk, W., et al. (2019). Pan-cancer whole-genome analyses of metastatic solid tumours. Nature *575*, 210–216.
- Puente, X., Beà, S., Valdés-Mas, R., Villamor, Gutiérrez-Abril, Martín-Subero, Munar, Rubio-Pérez, R., P, J., Aymerich, et al. (2015). Non-coding recurrent mutations in chronic lymphocytic leukaemia. Nature *526*, 519–524.
- Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.
- Redis, R.S., Vela, L.E., Lu, W., Ambrosio, A.L.B., Gomes Dias, S.M., and Calin, G.A. (2016). Allele-Specific Reprogramming of Cancer Metabolism by the Long Non-coding RNA CCAT2.
- Rentzsch, P., Witten, D., Cooper, G.M., Shendure, J., and Kircher, M. (2019). CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res. *47*.
- Rheinbay, E., Nielsen, M.M., Abascal, F., Tiao, G., Hornshøj, H., Hess, J.M., Pedersen, R.I.I., Feuerbach, L., Sabarinathan, R., Madsen, H.T., et al. (2017). Discovery and characterization of coding and non-coding driver mutations in more than 2,500 whole cancer genomes. BioRxiv 237313.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.06.467555; this version posted November 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

- Rheinbay, E., Nielsen, M.M., Abascal, F., Wala, J.A., Shapira, O., Tiao, G., Hornshøj, H., Hess, J.M., Juul, R.I., Lin, Z., et al. (2020). Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. Nature *578*, 102–111.
- Rubio-Perez, C., Tamborero, D., Schroeder, M.P., Antolín, A.A., Deu-Pons, J., Perez-Llamas, C., Mestres, J., Gonzalez-Perez, A., and Lopez-Bigas, N. (2015). In Silico Prescription of Anticancer Drugs to Cohorts of 28 Tumor Types Reveals Targeting Opportunities. Cancer Cell *27*, 382–396.
- Sabarinathan, R., Tafer, H., Seemann, S.E., Hofacker, I.L., Stadler, P.F., and Gorodkin, J. (2013). RNAsnp: Efficient Detection of Local RNA Secondary Structure Changes Induced by SNPs. Hum. Mutat. *34*, 546–556.
- Sanson, K.R., Hanna, R.E., Hegde, M., Donovan, K.F., Strand, C., Sullender, M.E., Vaimberg, E.W., Goodale, A., Root, D.E., Piccioni, F., et al. (2018). Optimized libraries for CRISPR-Cas9 genetic screens with multiple modalities. Nat. Commun. *9*, 5416.
- Sanz, L.A., Hartono, S.R., Lim, Y.W., Steyaert, S., Rajpurkar, A., Ginno, P.A., Xu, X., and Chédin, F. (2016). Prevalent, Dynamic, and Conserved R-Loop Structures Associate with Specific Epigenomic Signatures in Mammals. Mol. Cell *63*, 167–178.
- Sasaki, Y.T.F., Ideue, T., Sano, M., Mituyama, T., and Hirose, T. (2009). MEN / noncoding RNAs are essential for structural integrity of nuclear paraspeckles. Proc. Natl. Acad. Sci. *106*, 2525–2530.
- Seemann, S.E., Mirza, A.H., Hansen, C., Bang-Berthelsen, C.H., Garde, C., Christensen-Dalsgaard, M., Torarinsson, E., Yao, Z., Workman, C.T., Pociot, F., et al. (2017). The identification and functional annotation of RNA structures conserved in vertebrates. Genome Res. *27*, 1371–1383.
- Sentürk Cetin, N., Kuo, C.-C., Ribarska, T., Li, R., Costa, I.G., and Grummt, I. (2019). Isolation and genome-wide characterization of cellular DNA:RNA triplex structures. Nucleic Acids Res. *47*, 2306–2321.
- Shuai, S., Suzuki, H., Diaz-Navarro, A., Nadeu, F., Kumar, S.A., Gutierrez-Fernandez, A., Delgado, J., Pinyol, M., López-Otín, C., Puente, X.S., et al. (2019). The U1 spliceosomal RNA is recurrently mutated in multiple cancers. Nature *574*, 712–716.
- Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S., et al. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. *15*, 1034–1050.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.06.467555; this version posted November 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

- Simko, E.A.J., Liu, H., Zhang, T., Velasquez, A., Teli, S., Haeusler, A.R., and Wang, J. (2020). G-quadruplexes offer a conserved structural motif for NONO recruitment to NEAT1 architectural lncRNA. Nucleic Acids Res. *48*, 7421–7438.
- Sondka, Z., Bamford, S., Cole, C.G., Ward, S.A., Dunham, I., and Forbes, S.A. (2018). The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. Nat. Rev. Cancer *18*, 696.
- Spiniello, M., Knoener, R.A., Steinbrink, M.I., Yang, B., Cesnik, A.J., Buxton, K.E., Scalf, M., Jarrard, D.F., and Smith, L.M. (2018). HyPR-MS for Multiplexed Discovery of MALAT1, NEAT1, and NORAD lncRNA Protein Interactomes. J Proteome Res *17*, 3022–3038.
- Stamatoyannopoulos, J.A., Adzhubei, I., Thurman, R.E., Kryukov, G. V, Mirkin, S.M., and Sunyaev, S.R. (2009). Human mutation rate associated with DNA replication timing. Nat. Genet. 2009 414 *41*, 393–395.
- Statello, L., CJ, G., LL, C., M, H., Statello, L., Guo, C.J., Chen, L.L., and Huarte, M. (2021). Gene regulation by long non-coding RNAs and its biological functions.
- Szklarczyk, D., Gable, A.L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N.T., Morris, J.H., Bork, P., et al. (2019). STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. *47*, D607–D613.
- Tokheim, C.J., Papadopoulis, N., Kinzler, K.W., Vogelstein, B., and Karchin, R. (2016). Evaluating the Evaluation of Cancer Driver Genes.
- Tomasetti, C., Marchionni, L., Nowak, M.A., Parmigiani, G., and Vogelstein, B. (2015). Only three driver gene mutations are required for the development of lung and colorectal cancers. *112*, 118–123.
- Uhlen, M., Fagerberg, L., Hallstrom, B.M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, A., Kampf, C., Sjostedt, E., Asplund, A., et al. (2015). Tissue-based map of the human proteome. Science (80-.). *347*, 1260419.
- Umer, H., Smolinska, K., Komorowski, J., and Wadelius, C. (2021). Functional annotation of noncoding mutations in cancer. Life Sci. Alliance *4*.
- Uszczynska-Ratajczak, B., Lagarde, J., Frankish, A., Guigó, R., and Johnson, R. (2018). Towards a complete map of the human long non-coding RNA transcriptome. Nat. Rev. Genet. *19*, 535–548.
- Vancura, A., A, L., N, B.-G., MT, E., AH, G., S, H., R, J., Vancura, A., Lanzós, A., Bosch-Guiteras, N., et al. (2021). Cancer LncRNA Census 2 (CLC2): an enhanced resource reveals clinical features of cancer lncRNAs. NAR Cancer *3*.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.06.467555; this version posted November 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

- Vinagre, J., Almeida, A., Pópulo, H., Batista, R., Lyra, J., Pinto, V., Coelho, R., Celestino, R., Prazeres, H., Lima, L., et al. (2013). Frequency of TERT promoter mutations in human cancers. Nat. Commun. 2013 41 *4*, 1–6.
- Vogelstein, B., and Kinzler, K.W. (2015). The Path to Cancer — Three Strikes and You're Out. Http://Dx.Doi.Org/10.1056/NEJMp1508811 *373*, 1895–1898.
- Wang, J., Zhang, X., Chen, W., Li, J., and Liu, C. (2018). CRlncRNA: a manually curated database of cancer-related long non-coding RNAs with experimental proof of functions on clinicopathological and molecular features. BMC Med. Genomics *11*.
- Wang, L., Park, H.J., Dasari, S., Wang, S., Kocher, J.-P., and Li, W. (2013). CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. Nucleic Acids Res. *41*, e74.
- Wedge, D.C., Gundem, G., Mitchell, T., Woodcock, D.J., Martincorena, I., Ghori, M., Zamora, J., Butler, A., Whitaker, H., Kote-Jarai, Z., et al. (2018). Sequencing of prostate cancers identifies new cancer genes, routes of progression and drug targets. Nat. Genet. 2018 505 *50*, 682–692.
- West, J.A., Davis, C.P., Sunwoo, H., Simon, M.D., Sadreyev, R.I., Wang, P.I., Tolstorukov, M.Y., and Kingston, R.E. (2014). The long noncoding RNAs NEAT1 and MALAT1 bind active chromatin sites. Mol. Cell *55*, 791–802.
- Yamazaki, T., Souquere, S., Chujo, T., Kobelke, S., Chong, Y.S., Fox, A.H., Bond, C.S., Nakagawa, S., Pierron, G., and Hirose, T. (2018). Functional Domains of NEAT1 Architectural lncRNA Induce Paraspeckle Assembly through Phase Separation. Mol Cell *70*, 1038-1053 e7.
- Yanai, I., Benjamin, H., Shmoish, M., Chalifa-Caspi, V., Shklar, M., Ophir, R., Bar-Even, A., Horn-Saban, S., Safran, M., Domany, E., et al. (2005). Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. Bioinformatics *21*, 650–659.
- Z, C., H, J., S, Y., T, Z., X, J., P, L., J, J., N, L., B, T., and Y, L. (2018). Prader-Willi region non-protein coding RNA 1 suppressed gastric cancer growth as a competing endogenous RNA of miR-425-5p. Clin. Sci. (Lond). *132*, 1003–1019.
- Zhou, S., JR, H., F, S., G, G., M, T., SA, M.T., JT, H., KJ, K., P, M., M, A., et al. (2020). Noncoding mutations target cis-regulatory elements of the FOXA1 plexus in prostate cancer. Nat. Commun. *11*.
- Zhu, H., Uusküla-Reimand, L., Isaev, K., Wadi, L., Alizada, A., Shuai, S., Huang, V., Aduluso-Nwaobasi, D., Paczkowska, M., Abd-Rabbo, D., et al. (2020). Candidate Cancer Driver Mutations in Distal Regulatory Elements and Long-Range Chromatin Interaction Networks. *77*, 1307-1321.e10.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.06.467555; this version posted November 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 1

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.06.467555; this version posted November 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

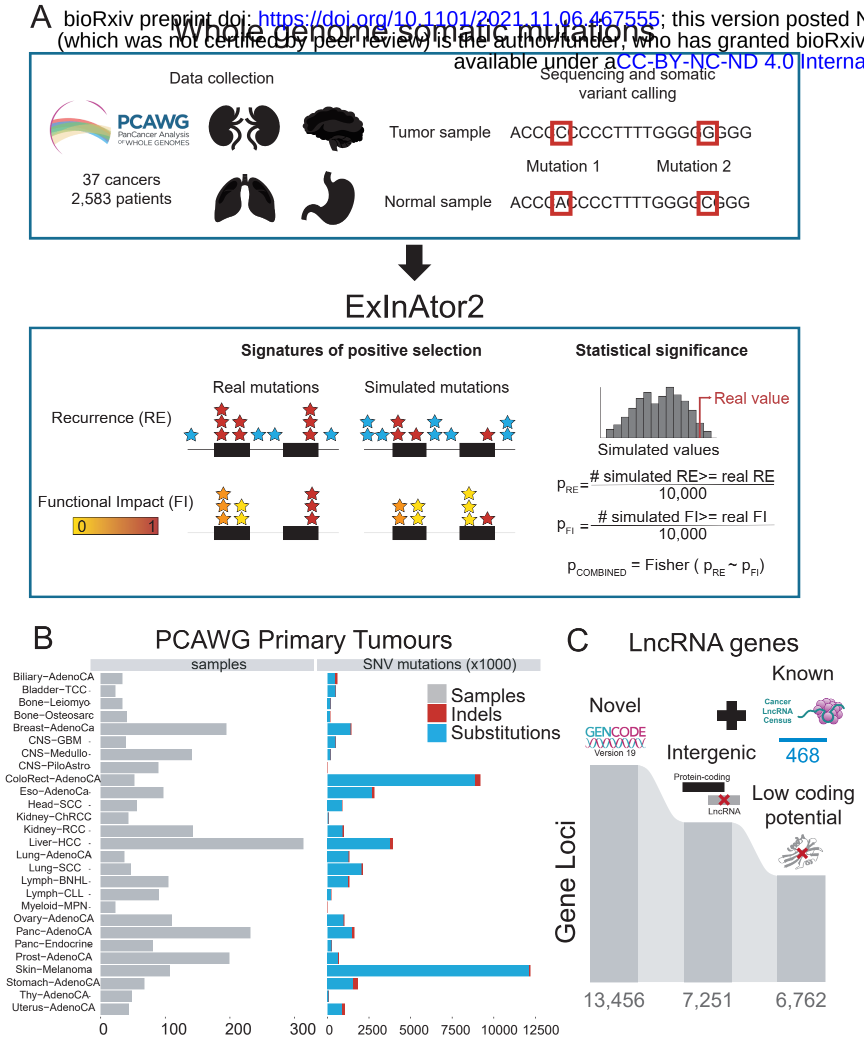


Figure 2

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.06.467555; this version posted November 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

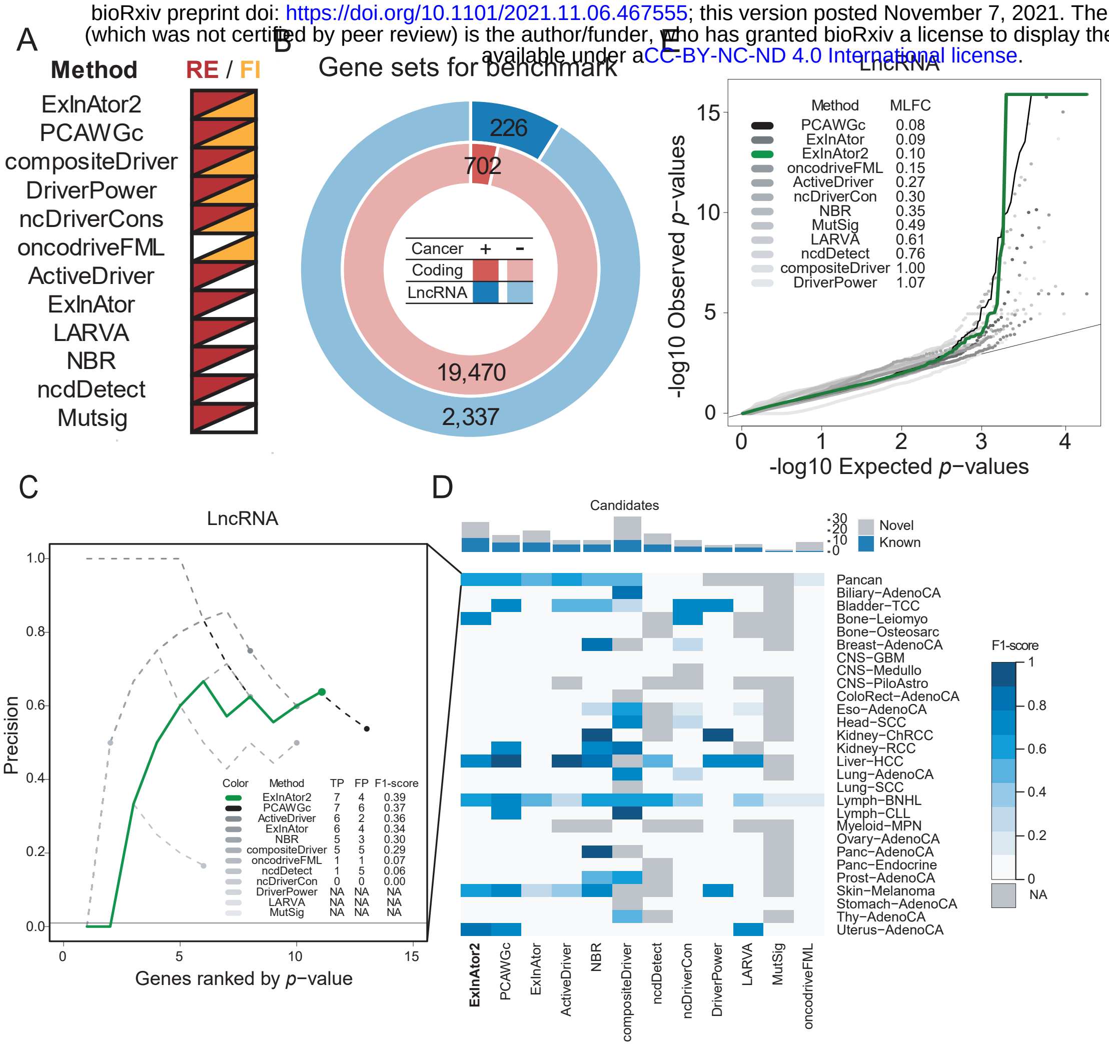


Figure 3

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.06.467555; this version posted November 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

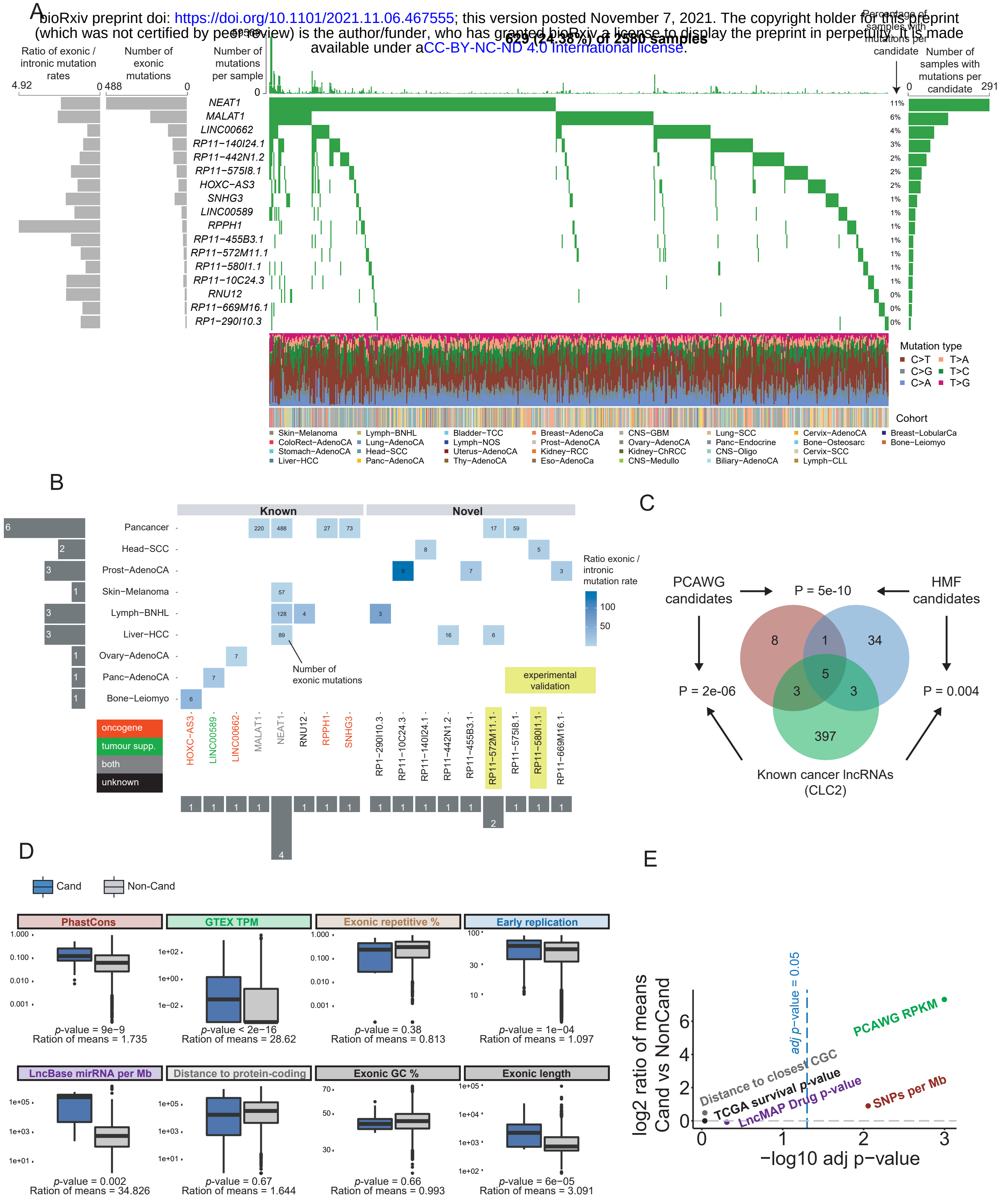


Figure 4

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.06.467555; this version posted November 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

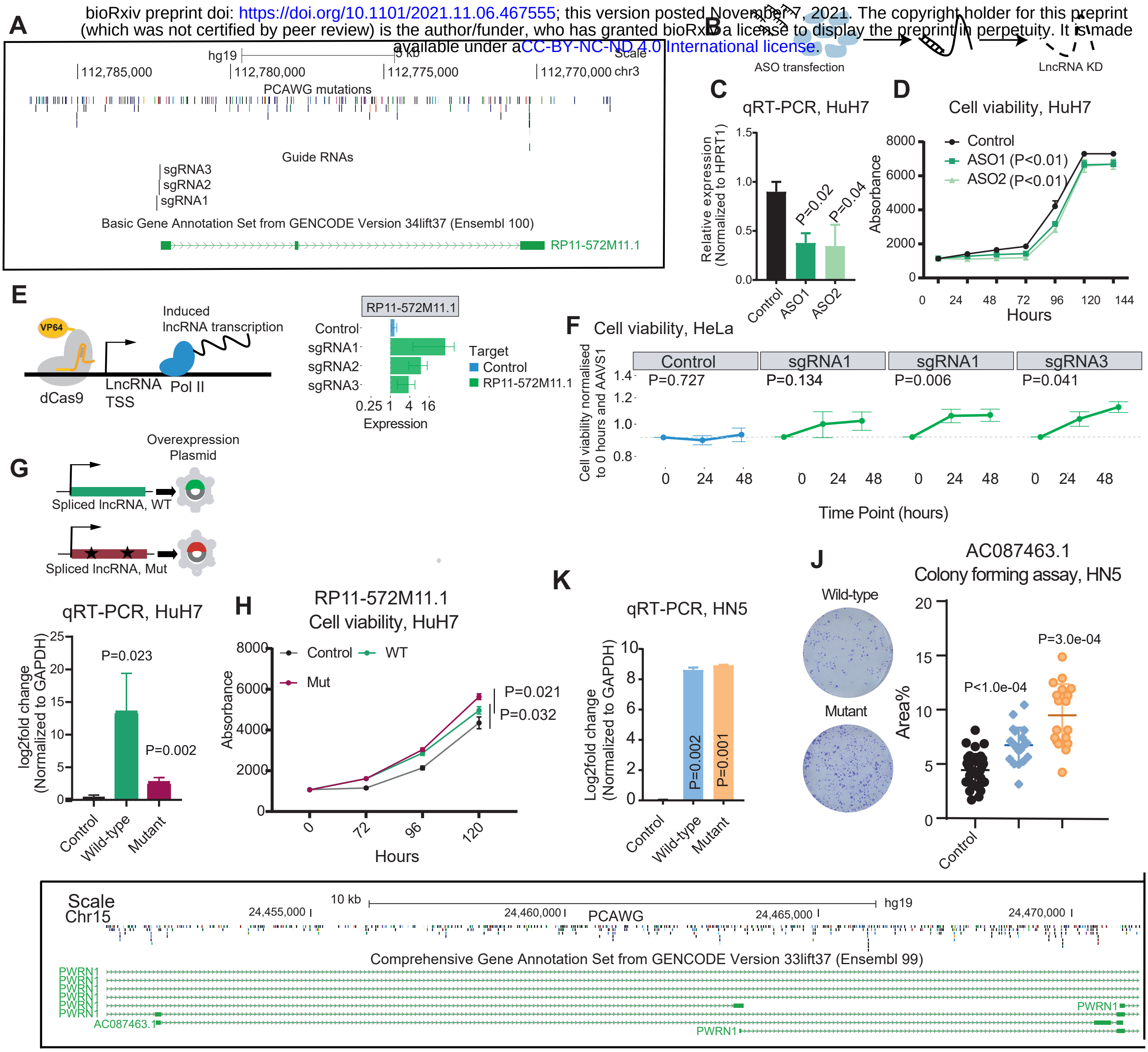


Figure 5

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.06.467555; this version posted November 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

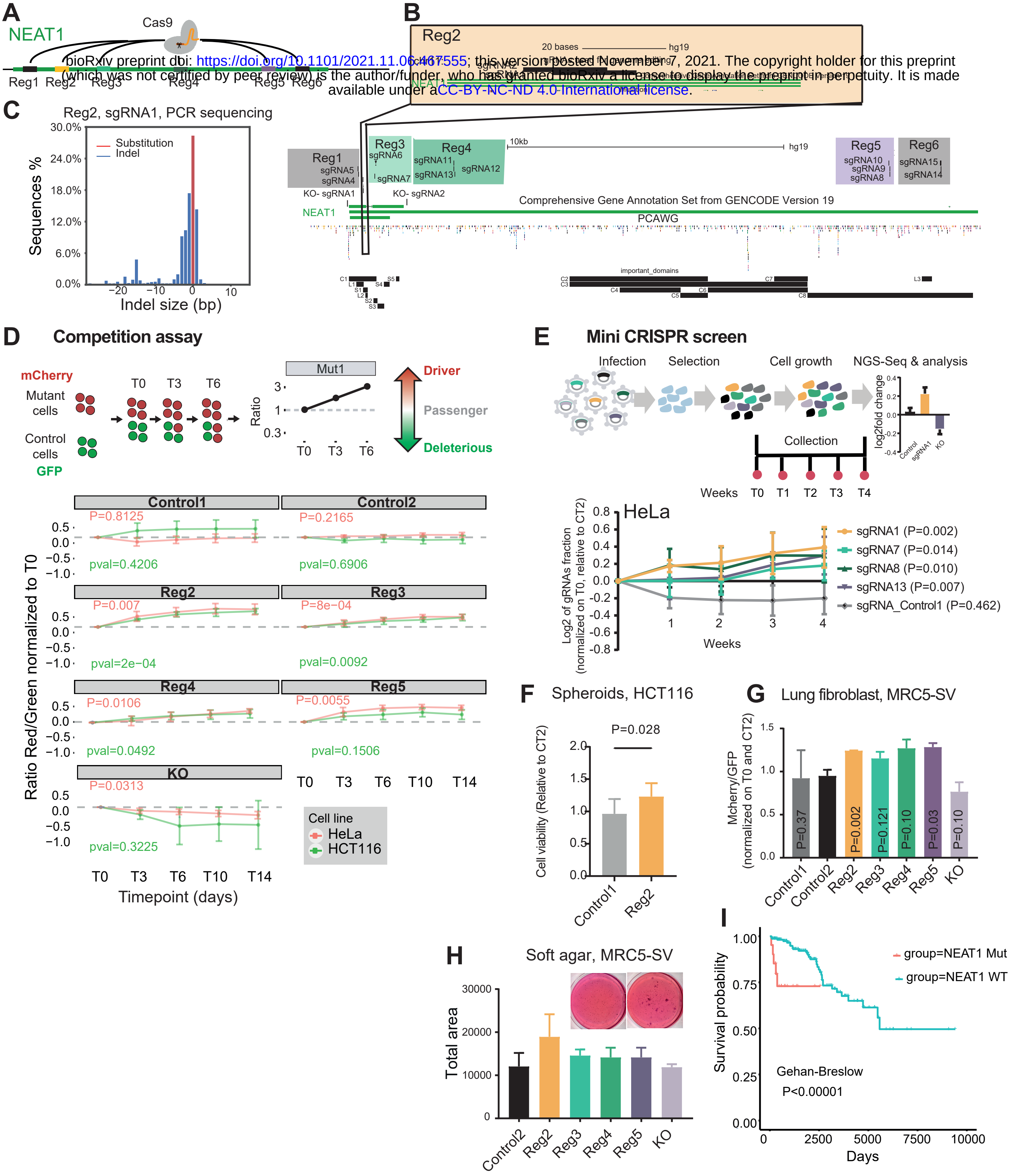
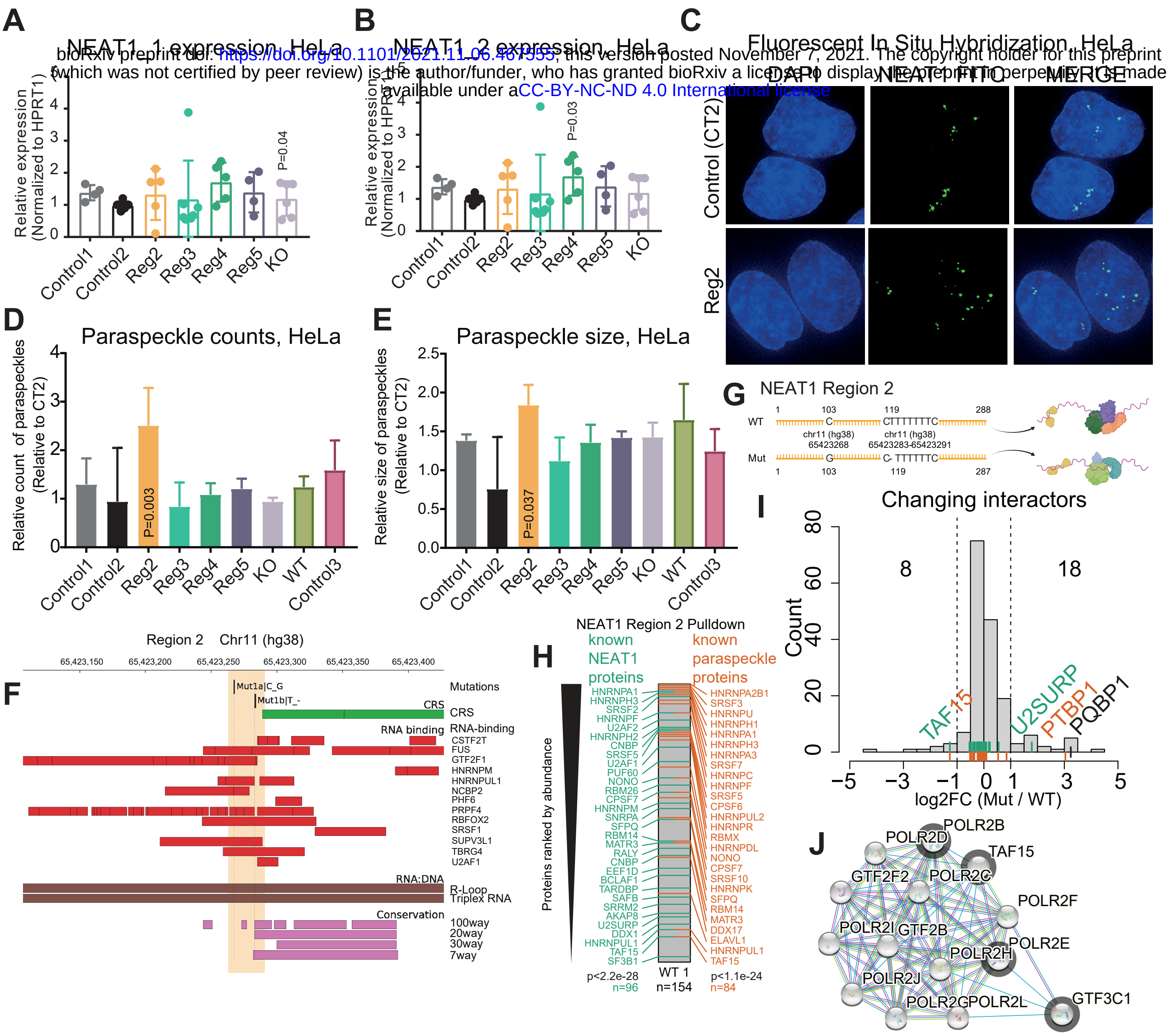


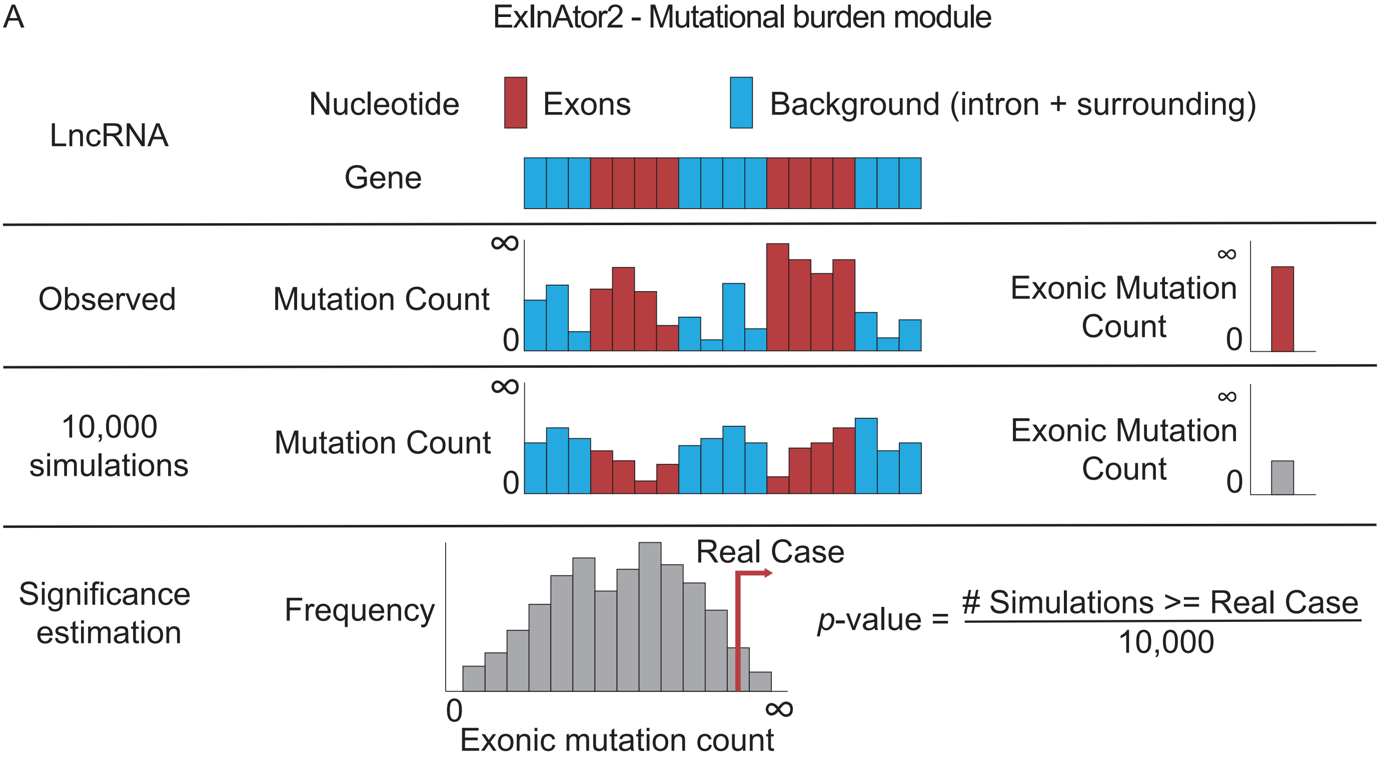
Figure 6

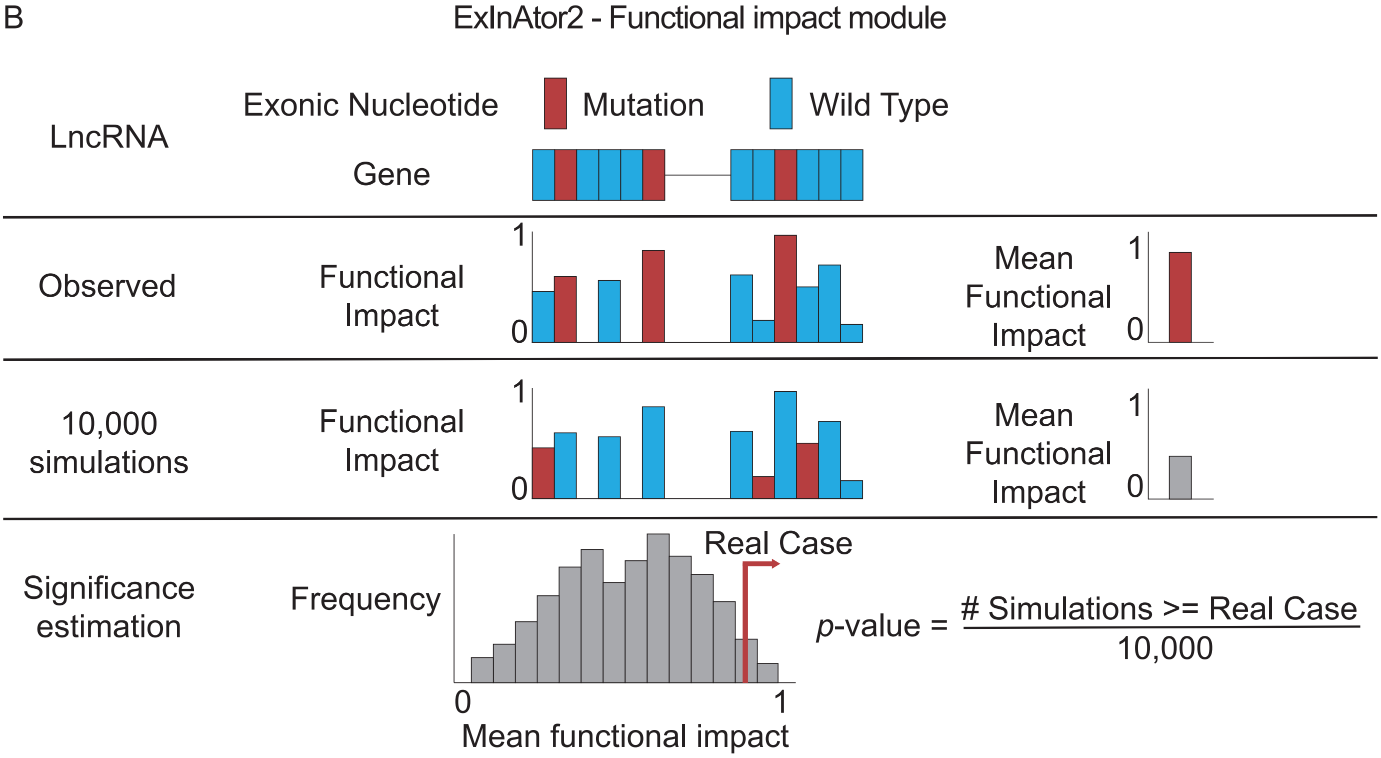
bioRxiv preprint doi: https://doi.org/10.1101/2021.11.06.467555; this version posted November 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



Supplementary Figure 1

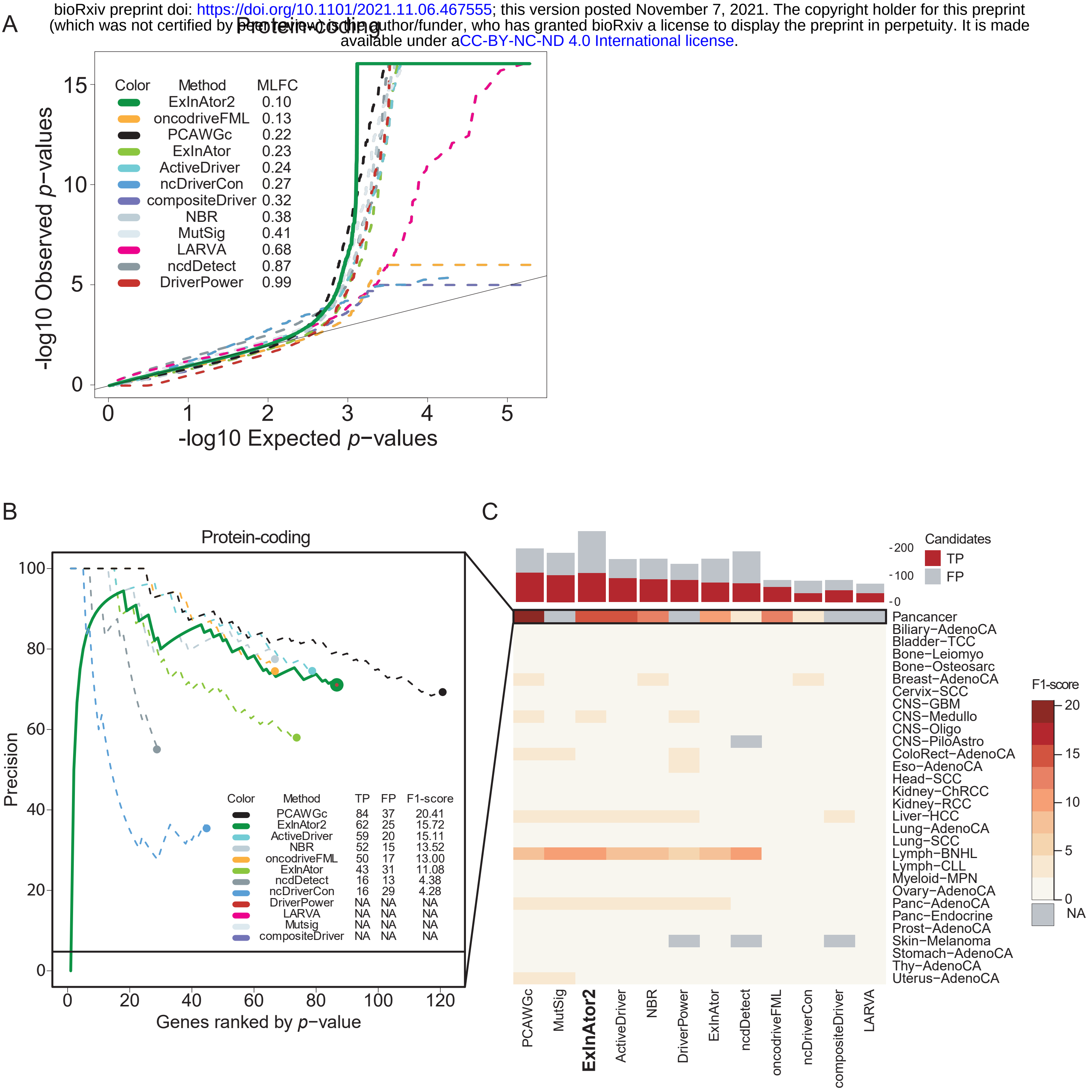
bioRxiv preprint doi: https://doi.org/10.1101/2021.11.06.467555; this version posted November 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.





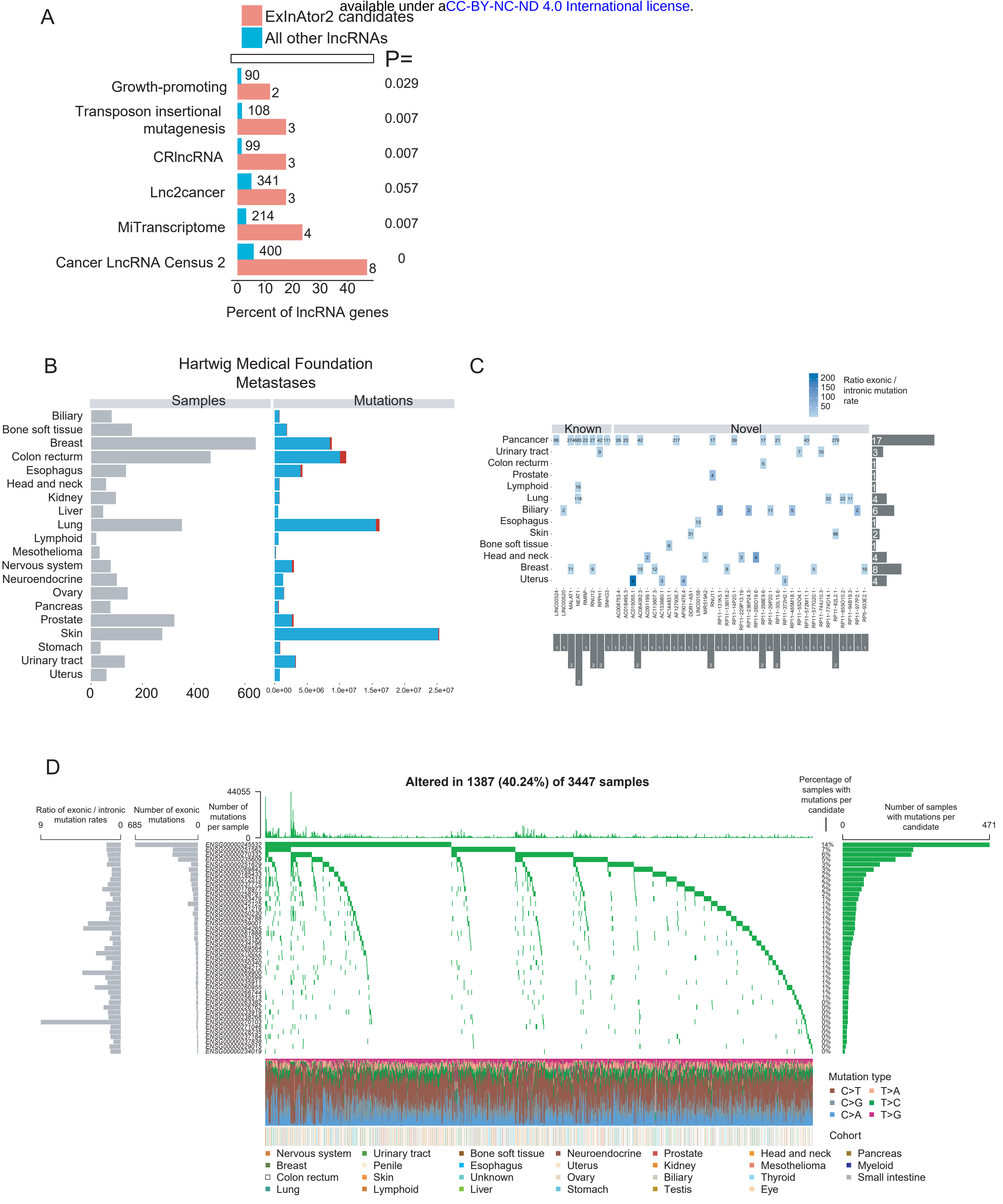
Supplementary Figure 2

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.06.467555; this version posted November 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



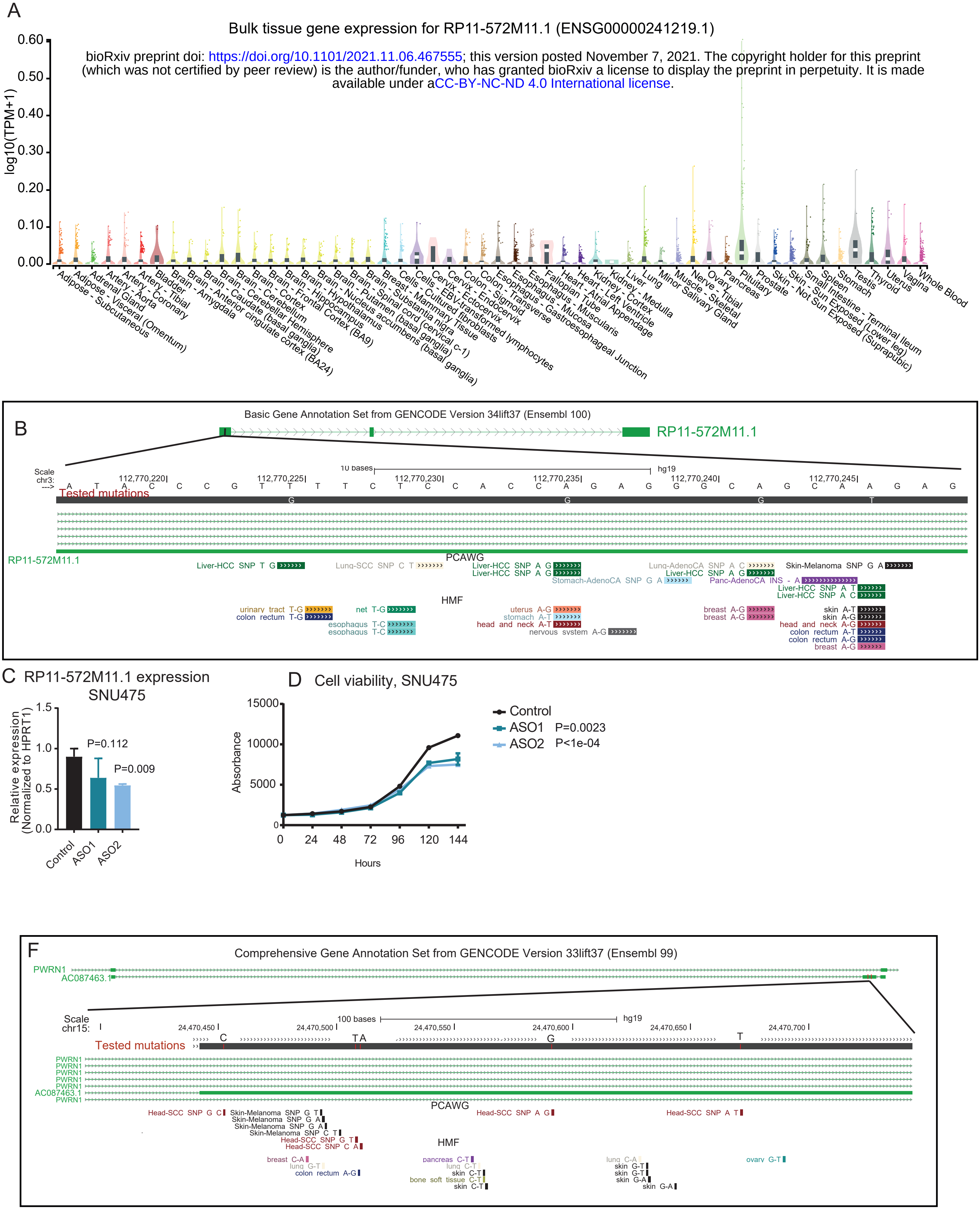
Supplementary Figure 3

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.06.467555; this version posted November 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

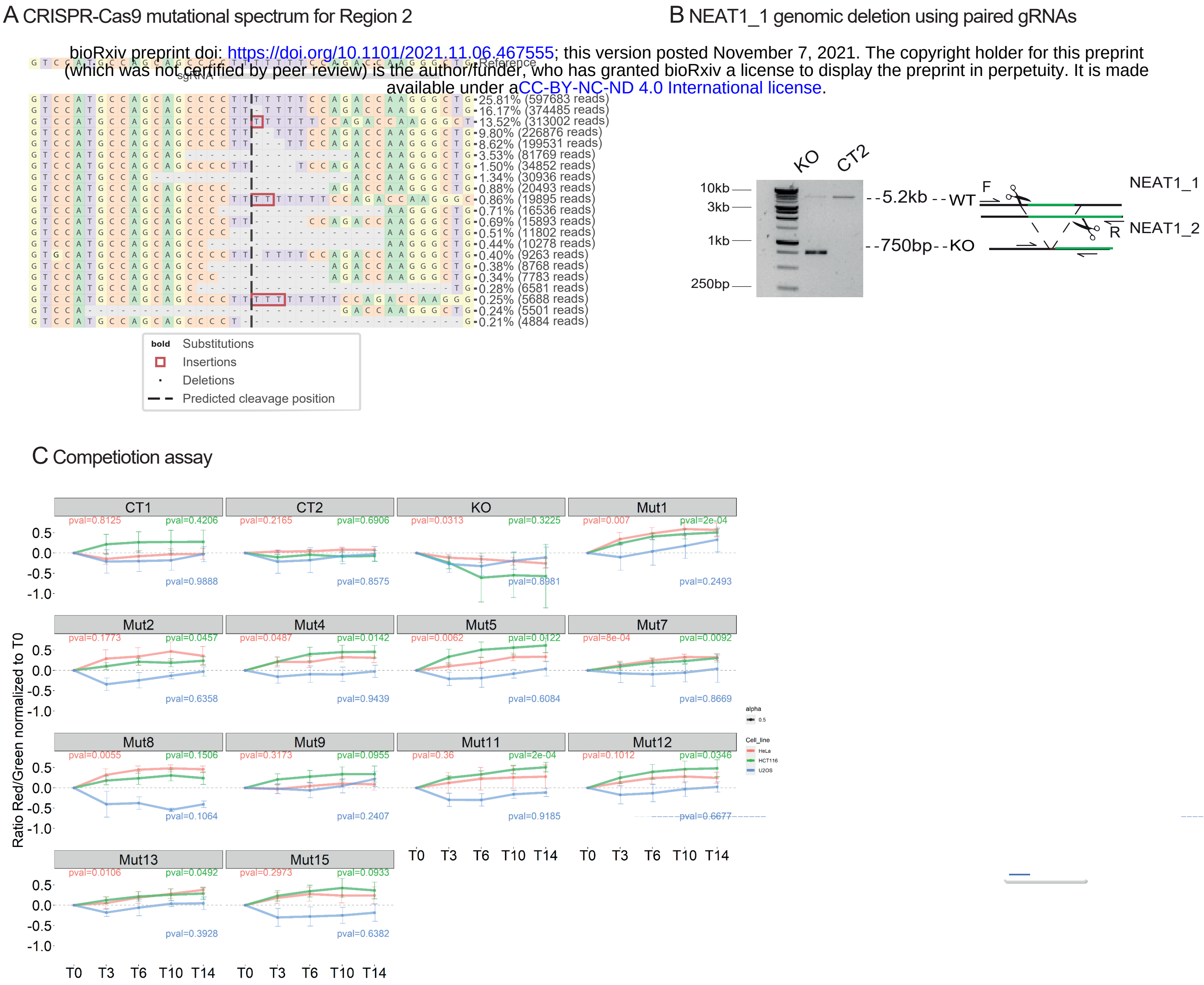


Supplementary Figure 4

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.06.467755; this version posted November 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



Supplementary Figure 5



A CRISPR-Cas9 mutational spectrum for Region 2

B NEAT1_1 genomic deletion using paired gRNAs

C Competiotion assay

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.06.467555; this version posted November 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.