

Unraveling lncRNA diversity at a single cell resolution and in a spatial context across different cancer types

Received: 9 July 2024

Accepted: 23 March 2026

Published online: 14 May 2026

 Check for updates

P. Prakrithi^{1,2,3,4}, Tuan Vo^{3,5,6,15}, Zherui Xiong², Hani Vu², Loan T. Nguyen⁷, Andrew Newman^{2,3}, Julie Wilhelmey⁸, Vicki Whitehall^{2,9,10}, Jazmina L. Gonzalez Cruz¹¹, Simon Wood¹², Andrew J. Mallett^{3,13,14}, Ishaan Gupta^{4,16}✉ & Quan Nguyen^{2,3,16}✉

Long noncoding RNAs (lncRNAs) participate in gene regulation underlying development and disease. Overcoming inherent limitations of bulk sequencing lncRNA analysis, we leveraged single-cell and spatial transcriptomics (ST) data to analyze 219,442 potential lncRNAs identified by the TAR-scRNA-seq pipeline across 13 cancer types. The lncRNA functions were assessed by identifying their cell-type specificity and spatial distributions across different tissue regions. Comparing with five existing databases, we confirmed shared lncRNAs and found 94,795 unannotated lncRNAs. The identified lncRNAs were experimentally validated across seven cancer types using three single-cell-resolution ST platforms and two long-read ST sequencing methods, encompassing both targeted and untargeted approaches. Genome-wide colocalization of lncRNAs with regulatory features, disease variants and pairwise spatial autocorrelation with protein-coding genes suggest potential functional roles. We built a free, fast and user-friendly database ‘SPanC-Lnc’ on AWS cloud. SPanC-Lnc will enable discovery of potentially functional lncRNAs specific for cell populations within heterogeneous tissues, providing new biological insights.

Long noncoding RNAs (lncRNAs) constitute the majority of the permissively transcribed genome. Despite this, the majority of our knowledge about transcriptional events is limited to the 1–2% of the genome that encodes proteins¹. With advancements in high-throughput technologies, it has become increasingly apparent that lncRNAs are of functional relevance, contributing to diseases like cancer where they are often misregulated².

The lncRNAs serve as a valuable class of molecular targets for disease identification; however, their detection has been constrained by factors such as their low abundance, cell-type-specific expression patterns and the reliance of existing computational tools on previous annotations³. Previous studies on lncRNAs have predominantly relied on bulk RNA sequencing (RNA-seq) data, which unfortunately results in the loss of cell-type information and tissue spatial context, both of which are

essential for understanding transcriptional regulation in developmental and diseased states. The spatial information provides valuable insights into a transcript’s location within a tissue, its neighboring cells, colocalizing transcripts or proteins and their interacting partners, thereby enhancing our comprehension of biological processes.

Advanced technologies such as single-cell RNA-seq (scRNA-seq) and spatial transcriptomics (ST) sequence transcripts at the single-cell level while preserving spatial information. Up to 40% of lncRNAs are polyadenylated and an additional 40% of them are bimorphic⁴. Thus, using a polyA-capture protocol, such as 10x Visium, these transcripts can be analyzed. Recent studies using scRNA-seq for breast cancer samples^{5,6} evaluated known (annotated) lncRNAs at a single-cell level. Other studies claiming to have analyzed the ‘spatial transcriptome’ of lncRNAs have typically examined lncRNA expression across different

A full list of affiliations appears at the end of the paper. ✉e-mail: Ishaan.Gupta@dbeb.iitd.ac.in; quan.nguyen@qimrb.edu.au

bulk tissues, rather than capturing the nuanced expression patterns of lncRNAs within a tissue section^{7,8}. These studies did not explore novel unannotated lncRNAs. Several methods for novel transcript identification have been established for bulk sequencing data^{9,10}. Applying these methods to ST and scRNA-seq data would not only facilitate the expansion of annotations but also unveil additional layers of information regarding lncRNAs, including cell-type specificity and spatial context to reveal physical manifestation of their crosstalk with protein-coding genes. Although these technologies have not yet addressed the identification of non-polyadenylated lncRNAs or the entire spectrum of rare lncRNAs, creating a repository of potential lncRNAs detected through cutting-edge ST technologies would establish a robust foundation for biological discovery.

In this study, we combined large datasets from the spatially resolved RNA sequencing by 10x Visium and single-cell resolution scRNA-seq by 10x Chromium to discover lncRNAs in tissues from 13 different cancer types, along with cross-validation on samples from the more recent technologies like STOmics, Takara Bio Seeker and Xenium in situ. We demonstrated new types of analyses to identify potential functions of lncRNAs. While previous pan-cancer studies have examined clinically relevant lncRNAs^{11–13}, this study constitutes a large-scale, spatially resolved investigation of uncharacterized or new potential lncRNAs, integrating cell-type-specific information to advance our understanding of their roles in cancer (Supplementary Table 1). The annotations and lncRNA expression data of each tissue are accessible through the interactive website ‘SPanC-Lnc’: Spatial and Single Cell Pan-Cancer Atlas of lncRNAs at <https://spanc-lnc.click/>. Researchers and clinicians can leverage this resource to gain deeper insights into the roles of lncRNAs in cancer biology.

Results

Identification of potential lncRNAs in spatial and single-cell datasets

Samples from 28 previously published in-house ST and 24 scRNA-seq samples across 13 cancer types were analyzed for potential novel lncRNAs (Fig. 1a) using the TAR-scRNA-seq pipeline¹⁰. The cumulative expression of all unannotated transcriptionally active regions (uTARs) detected in each sample was projected on the corresponding tissues and compared to that of the cumulative coding gene expression. Three representative examples for different cancer types are displayed in Fig. 1b. uTAR and coding gene expression for the remaining samples are also shown in Supplementary Figs. 1 and 2, respectively. The expression of several previously identified cancer-specific lncRNAs was also visualized (Supplementary Fig. 3). It is worth noting that while lncRNAs like *MALAT1* and *NORAD* were abundantly expressed across the tissue sections in various samples (Supplementary Fig. 3a), others displayed a certain degree of tissue specificity. For instance, both *PICSA*R and *LINC00520*, known to be squamous cell carcinoma (SCC)-specific¹⁴, were markedly more expressed in cutaneous SCC and head and neck (H&N) oropharyngeal SCC (OPSCC) compared to the other cancer types. In contrast, while *PICSA*R and *LINC00520* were either absent or expressed at lower levels in breast and colorectal cancer (CRC),

TINCR (known to play oncogenic roles in breast cancer^{15,16}) was highly expressed in breast cancer samples (Supplementary Fig. 3b). *TINCR* is also known to be oncogenic in OPSCC and tumor suppressing in SCC and highly expressed in our representative samples. Similarly, *KCNQ1OT1* (commonly implicated in CRC prognosis^{17,18}) showed higher expression in the CRC sample compared to other lncRNAs expressed in this sample.

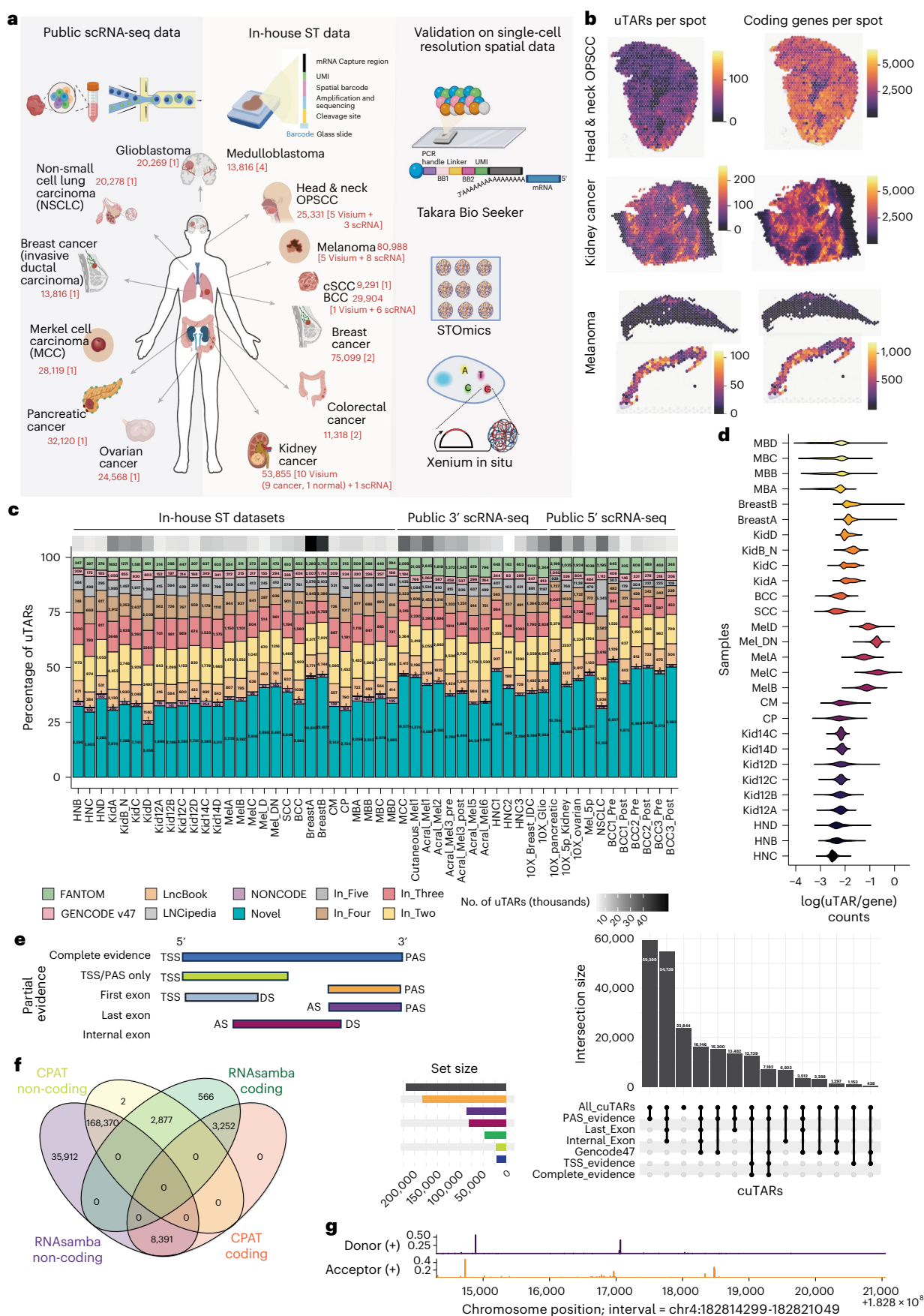
The samples used and the number of uTARs identified are summarized in Fig. 1c, with the numbers varying, depending on the cancer type and coverage. For example, a higher number of uTARs were found in breast cancer samples (~45,000–60,000) and one kidney cancer sample (namely KidA) (~45,000), whereas 4,000–25,000 were detected in other samples. Overall, more uTARs were detected with scRNA-seq compared to ST, as scRNA-seq generally captures a higher number of cells. In most samples, approximately 40–60% of the identified uTARs overlapped with lncRNAs cataloged in lncRNA public databases such as FANTOM, lncBook, LNCipedia and NONCODE (Fig. 1c). Moreover, while our initial analyses used GENCODE v.43 to filter out annotated transcripts, the recent v.47 release has expanded lncRNA annotations¹⁹. To account for this update, we retained uTARs that overlap lncRNA or noncoding pseudogene annotations in v.47 but were unannotated in v.43. Many of these had been identified by our method, demonstrating transcripts that we identified as unannotated lncRNAs in v.43 are now recognized as lncRNAs in GENCODE v.47. The remaining transcripts are likely novel discoveries not yet reported by other databases and merit further investigation (Fig. 1c).

The relative uTAR:gene counts for each tissue at Visium spot-level resolution was calculated as the total number of uTARs divided by the total number of genes detected per spot (Fig. 1d). A higher value indicates a higher number of total uTARs detected per Visium spot with respect to the number of coding genes, as in the melanoma samples (Fig. 1d); however, this could be due to the lower sample quality (<200 median genes per spot). As the uTARs analyzed in this study span various cancer types and tissues, we merged overlapping uTARs across multiple samples into a single uTAR. These were designated as cancer-associated uTARs (cuTARs) (Supplementary Fig. 4) and assigned unique numerical identifiers in our SPanC-Lnc database. The terms uTAR, cuTAR and lncRNA are used interchangeably throughout the paper. A total of 47 cuTARs were expressed on a pan-cancer level (expressed in at least one sample per cancer type in the study) and 1,278 in at least ten cancer types, whereas most of the others exhibited specificity to particular cancer types (Extended Data Fig. 1a,b). The consistency of detection across different sequencing platforms (ST, 3' and 5' scRNA-seq) for melanoma samples was also analyzed. About 1,574 cuTARs were detected in melanoma samples across all the three different platforms (Extended Data Fig. 1c). Following these findings, we further investigated some of these potential disease-associated candidates in specific cancer types.

Characterization of cuTAR transcripts

To further define cuTAR structure, we overlapped cuTAR coordinates with transcription start site (TSS) annotations from FANTOM, poly-A

Fig. 1 | Pan-cancer identification of novel lncRNAs from spatial and single-cell data. **a**, Different cancer types used from public scRNA-seq and previously published in-house ST datasets for lncRNA detection, further confirmed on high resolution spatial technologies. uTARs as of GENCODE v.43 were first identified using the TAR-scRNA-seq pipeline on a per-sample basis; overlapping uTARs across samples were merged to derive a nonredundant set of cuTARs, which form the final lncRNA annotations in our database. Numbers highlighted in red indicate the number of cuTARs detected for each cancer type, followed by the number of samples analyzed within square brackets. **b**, uTAR counts overlaid on the tissue for representative samples as compared to the protein-coding gene counts. **c**, Breakdown of known and novel uTARs identified per cancer sample. The blue region indicates novel uTARs and the other colors indicate an overlap with public datasets (FANTOM, lncBook, NONCODE, GENCODE v.47 novel lncRNA additions and LNCipedia), as well as overlaps in combinations of two, three, four or five of these databases. Bar labels and the gray scale bar on top show the number of uTARs found from each sample. Sample names start with cancer/tissue type abbreviations followed by alphanumeric suffixes. HN, head & neck OPSCC; Kid, kidney; Mel, melanoma; Mel_DN, melanoma dysplastic nevus; MB, medulloblastoma; CP, CRC primary; CM, CRC metastasized. **d**, uTAR:gene detection ratio (number of genes versus number of uTARs), per spot across the ST samples. A higher value indicates a higher number of uTARs detected per spot with respect to the number of genes. **e**, Characterization of cuTARs based on external evidence of (left) and proportions of cuTARs with several layers of evidence (right). **f**, Venn diagram showing the coding-potential-based classification of cuTARs using RNAsamba and CPAT. **g**, Prediction of splice sites using AlphaGenome demonstrated for cuTAR213507.



sites (PAS) and splice sites (Fig. 1e). They were classified as cuTARs with (1) complete evidence (overlapping with a TSS at the 5' end and a PAS at the 3' end within a 1.5-kb window upstream and downstream, respectively); (2) partial evidence comprising of first exons (TSS at the 5' end and a donor site junction at the 3' end), internal exons (acceptor site at the 5' end and a donor site at the 3' end) and last exons (acceptor site at the 5' end and PAS at the 3' end); and (3) incomplete evidence if it overlapped with either TSS or PAS (Fig. 1e). Overall, 9.07% cuTARs had complete evidence (Fig. 1e) and 24.9% cuTARs had PAS, last and internal exon evidence, potentially indicating that these transcripts have at least two exons. Overall, 27% cuTARs had only PAS evidence, which is expected as most cuTARs were biased to the 3' end. To collect further evidence, the cuTARs were overlapped with the new lncRNA annotations from GENCODE v.47 since the release v.43 (Fig. 1e). Overlapping cuTARs with human permissive enhancers from FANTOM, we found that 3.32% were enhancer-derived lncRNAs (e-lncRNAs) with an enhancer overlap at the 5' end and 7.29% with an enhancer spanning anywhere along the cuTAR sequence. Additionally, to ensure cuTARs are independent and not potential read-throughs of nearby genes, we labeled those located within 10 kb downstream of coding genes as 'less confident' (12.5% of cuTARs).

The sequences were analyzed using CPAT and RNAsamba for coding potential. The CPAT algorithm results are affected by the length of the transcript. RNAsamba is an alternative coding potential calculator that works well on partial-length sequences²⁰. RNAsamba identified 97% of all cuTARs to be noncoding. While CPAT filtered out many cuTARs, RNAsamba retained most of them and inferred the coding potential. Overall, 76.6% of cuTARs were predicted to be noncoding using both the tools and 98.3% of the total cuTARs lacked coding potential (Fig. 1f). The lncRNAs with a coding potential might indicate their role in encoding functional micropeptides, as opposed to necessarily being protein-coding transcripts^{21,22}. As the cuTARs represent estimated transcript boundaries using data captured from either the 3' or 5' ends, the calculated coding potentials may be less accurate than those obtained using the full-length sequence, but they can still provide a meaningful reflection of their coding potential. Moreover, 94.75% of them were at least 200 bp, suggesting that the majority of cuTARs were most likely to be 'long' ncRNAs²³.

Additionally, we predict splice sites utilizing AlphaGenome²⁴. To evaluate AlphaGenome predictions, we visualized the *BIRC5* gene (Extended Data Fig. 2a) and compared the predicted splice sites with the RefSeq exon annotations for the cuTARs that have been annotated in GENCODE v.47 but were unannotated as of v.43 (Extended Data Fig. 2b). In most cases, predicted splice site peaks were consistent with the annotated exon–intron junctions. With this serving as a performance reference, we were able to predict and suggest potential splice sites in the unannotated and uncharacterized cuTARs (Fig. 1g).

Conservation and minimum free energy (MFE) are useful metrics for assessing sequence functionality and were calculated for the H&N sample C (HNC). Most top cuTAR candidates had lower conservation than *GAPDH* or the lncRNA *HOTAIR*. They also had a stable secondary structure (Supplementary Fig. 5a).

The genetic-level regulation of the cuTARs through enrichment of eQTL and GWAS SNPs are elaborated in Supplementary Note 1 (Supplementary Figs. 5b and 6–8).

Identification of lncRNAs enriched in cancer tissue regions

Next, we asked whether the cuTAR expression was restricted specifically to the tumor area within the biopsies. Given that lncRNAs can be present in both healthy and tumor tissues, tumor specificity was defined by enrichment within cancerous regions annotated by a pathologist and where histological annotation was unavailable (for example, CP), inferred from tumor-associated gene expression patterns (Fig. 2a). Top cuTAR candidates for each cancer type were shortlisted based on their detection across various datasets and higher expression in the

annotated cancerous regions. The percentage of spots in the cancerous versus normal regions expressing the selected cuTARs in the H&N OPSCC samples were calculated (Fig. 2b).

The expression of these top cuTAR candidates was detected across many cancer types from The Cancer Genome Atlas (TCGA)²⁵ bulk RNA-seq samples. Some cuTARs showed higher expression in the seven TCGA OPSCC samples used, consistent with the observation that all these candidates were tumor-specific in the H&N ST data (Fig. 2b). Among the cuTARs differentially expressed across these two annotated regions (Supplementary Fig. 9), visual examination of gene expression overlaid on the tissues reveal some cuTARs with higher expression in the tumor regions (Fig. 2a,c). Not only cancer region specificity, but also subclone specificity was observed. The tumor subclones were identified based on inferred copy-number variation using CopyKAT²⁶ (Extended Data Fig. 3).

Confirmation of lncRNA detection using spatial long-read sequencing

Since the 10x Visium protocol captures only the 3' ends of fragmented transcripts, the actual transcripts are likely longer than the detected cuTAR boundaries. To increase confidence in transcript identification, we applied long-read sequencing using Oxford Nanopore Technologies (ONT) and PacBio MAS-Seq (Pacific Biosciences) to unfragmented cDNA from 10x Visium libraries. Although these libraries remain 3' biased, long-read sequencing yields longer fragments, allowing more confident mapping of cuTAR signals than using Visium short read data. We found the signals detected to be consistent across different technologies, suggesting that these cuTARs were likely to be genuine signals rather than artifacts from the TAR-scRNA-seq pipeline.

10x Visium unfragmented cDNA libraries from the HNC (OPSCC), SCC and basal cell carcinoma (BCC) samples were used for the ONT experiment. As a representative example, from HNC, we obtained ~16 million ONT reads, with 3,888 spatial barcodes identified by the scNanoGPS pipeline²⁷, allowing us to map to 1,028 Visium spots across the tissue. We successfully recovered ~83.6% reads that were complete (with spatial barcode, unique molecular identifiers (UMIs), 3' adaptors and poly-A/T tails), identifying approximately 985 cuTARs (Supplementary Table 2), 960 of which were present in at least three spots overlapping the tissue. This accounts for 31.7% of those detected using 10x Visium, thus validating the detection of uTAR expression from the short-read data. Similar spatial patterns were observed (Fig. 2d), with up to 38% of spots showing expression in both the technologies and up to 86% for coding genes (Supplementary Note 2 and Supplementary Fig. 10). Notably, when we applied the TAR-scRNA-seq pipeline directly to the Nanopore BAM files, we identified 32 novel uTARs not detected with the short-read data. In addition, we confirmed signals from the transcriptionally active region (TAR) annotations generated using the 10x Visium data, where 857 ONT uTARs (807 cuTARs) were also detected in the HNC Visium sample. Overall, 92 additional cuTARs detected with ONT were present in our entire database but not in the corresponding Visium sample. To evaluate the results, the consistency of expression patterns for coding genes were also verified (Supplementary Fig. 10). There was a substantial overlap of expressing spots across both the platforms for some cuTARs (Extended Data Fig. 4a). The same analysis was performed for cutaneous SCC and BCC using ONT (Extended Data Fig. 4b,c).

For the PacBio MAS-Seq long read sequencing run with complementary DNA of CRC samples from the 10x Visium experiment, we generated approximately 1 million HiFi reads. By aligning the reads and tagging them with spatial barcode information, we found up to 1,500 lncRNAs per sample, which accounted for up to ~20% of those detected with 10x Visium (Extended Data Fig. 4d), thereby also confirming the findings of the 10x Visium Illumina data. We then examined specific cuTARs and found that up to 30% of the spots showed statistically significant consistency in expression for the same set of spots across both the platforms

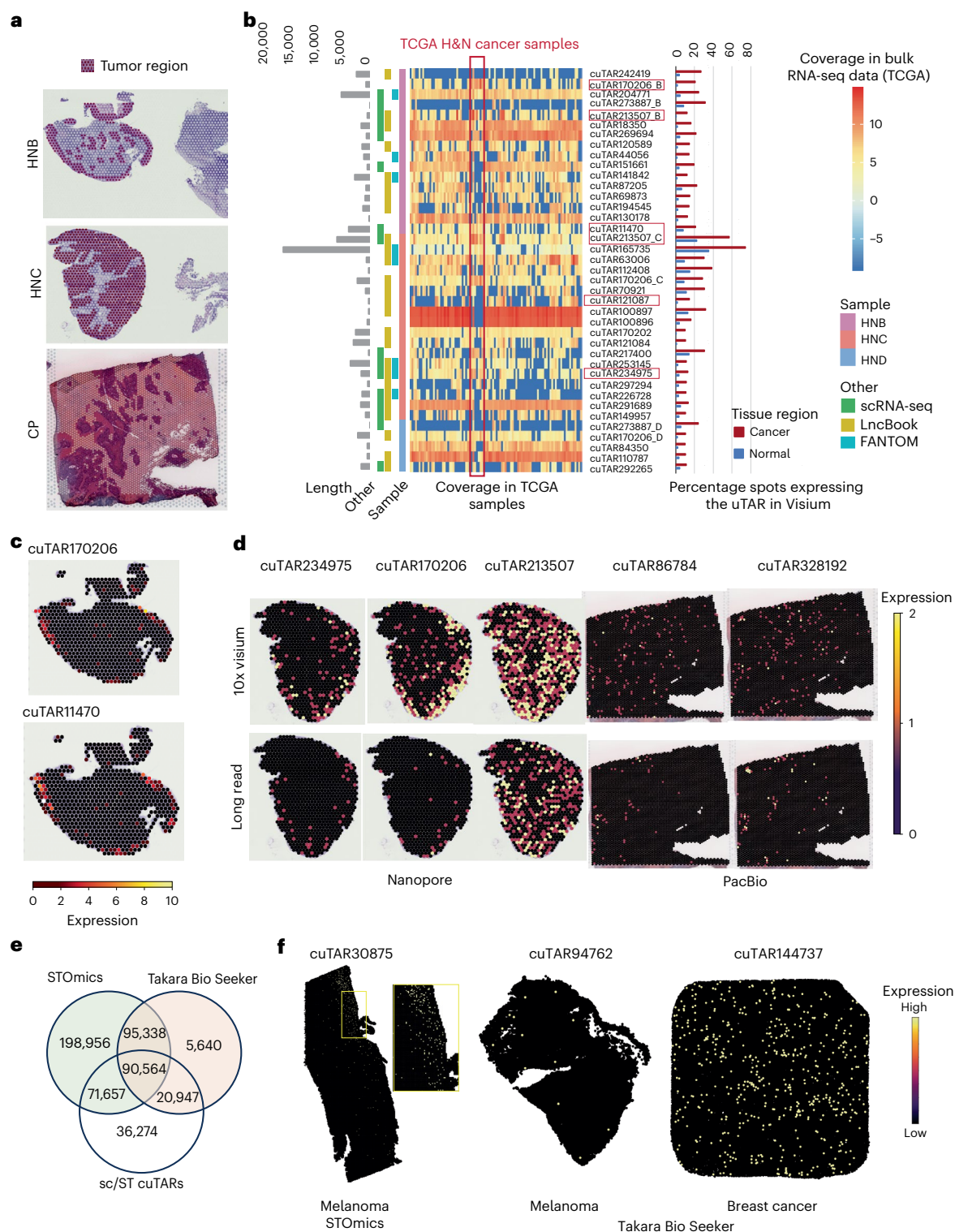


Fig. 2 | Identification of tumor region specific cuTARs and validation with spatial long-read sequencing and single-cell resolved ST. a, Tumor regions (dark shaded) in the H&N cancer samples HNB and HNC and CRC primary tumor sample (CP). **b**, (From left to right) cuTAR lengths, overlap of cuTARs with that from scRNA-seq data and with lncRNAs from public databases, expression of cuTARs in bulk RNA-seq samples from TCGA (normalized bigWig coverage) and quantification of the cuTARs in the cancerous and normal regions of H&N cancer samples. The cuTARs found across multiple samples are indicated by the sample suffix (B, C and D). The highlighted cuTARs with red boxes indicate the ones with higher expression in the tumor region than the normal region, some

of which are shown in c. These show relatively higher expression in the analyzed H&N cancer samples from TCGA highlighted using the vertical red box. **c**, Expression of exemplar cancer region-specific cuTARs projected on the tissues. **d**, Validation of cuTARs with ONT and PacBio MAS-seq long-read sequencing for H&N and CRC samples, respectively. **e**, uTARs shared across single-cell resolution spatial data from STOmics and Takara Bio Seeker and our Visium/scRNA-seq identified cuTARs. Overall, 83.47% of cuTARs identified using 10x Visium ST and scRNA-seq were also detected with STOmics and Takara Bio Seeker. **f**, Expression of cancer-specific cuTARs visualized on STOmics and Takara Bio Seeker samples.

for two CRC-specific cuTARs (Fig. 2d and Extended Data Fig. 4b,c) (for example, cuTAR86784, 19%; $P = 0.011$; cuTAR114310, 29%; $P = 0.014$). Overall, we found the PacBio sequencing mapping efficiency was better than that of ONT considering the number of uniquely mapping cuTARs identified for the number of reads captured and processed. More quality control (QC) statistics for long-read spatial sequencing are described in Supplementary Tables 2 and 3.

Takara Bio Seeker and STOmics spatial data confirmed cuTARs at single-cell resolution

The cuTAR detection at single cell resolution was analyzed with Takara Bio Seeker for one breast cancer and one melanoma sample. We observed consistent expression patterns of coding genes generated using the Takara Bio Seeker pipeline (Extended Data Fig. 5a) and the TAR-scRNA-seq pipeline (Extended Data Fig. 5a) in the breast cancer sample. Some previously identified breast-cancer-specific cuTARs were also visualized (Extended Data Fig. 5b). Similarly, annotated genes and melanoma-specific cuTARs were visualized on the melanoma sample (Extended Data Fig. 5c,d). Overall, 31.4% of the cuTARs identified from Visium and scRNA-seq breast cancer data were also detected in the Seeker breast cancer sample, while 47.56% of those from melanoma data overlapped with the Seeker melanoma sample.

Additionally, we performed STOmics OMNI sequencing for melanoma and CRC samples. cuTAR counts were generated using the SAW pipeline with Visium and scRNA-seq identified cuTAR coordinates as a custom GTF input. 116,862 (52.7%) cuTARs were detected in at least three cells of the STOmics data (Fig. 2e). Total cuTAR counts across the tissues is shown in Supplementary Fig. 11a. Higher expression of some melanoma-specific cuTARs identified using scRNA-seq data was observed in the melanoma sample compared to the CRC samples (Supplementary Fig. 11b). To enable a more unbiased discovery of cuTARs that may be non-polyadenylated, we next applied the TAR-scRNA-seq pipeline directly to the aligned STOmics BAM files at the bulk level. This strategy allowed for de novo identification of uTARs with transcript boundaries independent of Visium, complementing our earlier analysis where transcript counts were restricted to Visium-defined transcripts. The spatial expression patterns of exemplar breast- and melanoma-specific cuTARs are visualized on the Takara Bio Seeker and STOmics datasets (Fig. 2f).

Experimental detection of lncRNA using quantitative reverse transcription PCR

We further experimentally validated and quantified seven cuTARs by qRT-PCR. With specific primers targeting each of the selected cuTARs (Supplementary Table 4), we found cuTARs in the H&N samples (HNC and HNB), SCC, BCC and CRC samples. All the lncRNAs were detected, except cuTAR121087, which showed a low PCR signal, although high expression was detected with Visium for the HNC sample and other samples (Supplementary Figs. 12 and 13). Further results are discussed in Supplementary Note 3.

Inferring potential functions by spatial coexpression analysis of lncRNAs with cancer-associated genes

Transcripts that are expressed together tend to function together. For data-driven prediction of functional cuTARs, we used bivariate Moran's index (I) to identify coexpression patterns between spatially variable cuTARs (Supplementary Fig. 14a) and cancer-hallmark genes, suggesting possible co-regulation gene networks between uncharacterized cuTARs and known cancer-associated genes. Figure 3a shows four top cuTARs with more than 50 high-high (HH) spots showing coexpression with a greater number of cancer-relevant genes compared to the other cuTARs in the H&N sample C (HNC). Some cuTARs were coexpressed with many genes involved in cancer associated pathways, especially the P53 pathway and G2M checkpoint (Fig. 3a), whereas in a kidney cancer sample (KidD), high coexpression was observed with genes involved

in epithelial-mesenchymal transition (EMT) and the P53 pathway (Supplementary Fig. 14b). These cuTARs, although detected in healthy kidney tissue, have a significantly lower expression compared to other kidney cancer samples (Supplementary Fig. 14c). Examples of spatially coexpressed cuTARs and coding genes in the H&N sample are shown in Fig. 3b. For example, cuTAR213507 (chr4:182814299–182820899 (+)) was coexpressed with the messenger RNA of *FGFR3*, a gene upregulated in H&N cancers, also on chromosome 4 (chr4:1793293–1808867 (+)). A total of 24 other cancer-relevant genes are located on the same chromosome with this top spatially autocorrelated cuTAR, suggesting strong evidence for a cancer-associated cuTAR. Moreover, this cuTAR is downstream to a recently annotated lncRNA ENSG00000309834 and the gene *TENM3*. *TENM3* is upregulated in several cancers compared to healthy tissues, particularly in H&N SCCs, pancreatic adenocarcinoma, thymoma and neuroblastoma²⁸. It also serves as an integration site for the human papilloma virus (HPV), causing cervical and a large proportion of oropharyngeal cancers²⁹ and also harbors a genome-wide association study (GWAS) single-nucleotide polymorphism (SNP) rs1516535, an intron variant associated with HPV infection in oropharyngeal cancers. It is worth mentioning that the OPSCC samples utilized in this study were HPV-16⁺ (ref. 30). Fusions of *TENM3* gene have been reported to induce cell proliferation³¹.

Using coexpression network analyses with hdWGCNA, we identified that one of the top candidates cuTAR170206 (lncBook HSALNG0133452) formed part of a coexpression module with genes enriched for various cancer hallmarks. Additionally, it was found to be highly coexpressed with *NDRG1* (Fig. 3c), which is known to be down-regulated in metastatic OPSCC tumors³². In this particular case, *NDRG1* was seen to be expressed only in the periphery and not in the core of the tumor (Fig. 3c). This gene has been shown to have pleiotropic functions, acting as tumor promoters in some cancers, while playing a tumor suppressor role in others³³. This modulation in function could potentially be attributed to its association with interacting lncRNAs.

Inferring functions by machine-learning prediction of interaction with RBPs and colocalization

The lncRNAs interact with RNA binding proteins (RBPs) to regulate mRNA and protein localization and functions³⁴. Several machine-learning models that have been trained on known lncRNA-RBP interactions can help predict new interactions. We applied HLPI-Ensemble³⁵ to predict the interaction of the shortlisted cuTARs with RBPs. A noteworthy association for cuTAR215705 was observed in the ST breast cancer sample (BreastA) (Fig. 3d). cuTAR215705 was predicted to interact with the RBPs, HuR (ELAVL1) and TARDBP (TDP-43). Moreover, these were part of the same coexpression module identified through hdWGCNA, which is enriched for TGF- β negative regulation and positive regulation of protein localization (Supplementary Fig. 14d). Additionally, it shares this module with the gene *BIRC5*, which has recently been reported to interact with ELAVL1. Previous reports indicate that recombinant ELAVL1 is linked to the upregulation of *BIRC5* expression, whereas its silencing correlates with the downregulation of both *BIRC5* mRNA and protein, accompanied by increased apoptosis. Survival analyses from a previous study demonstrated that increased *TTP* (*ZFP36*) and low *BIRC5* expression predicted an overall better prognosis compared to dysregulated *TTP* and high *BIRC5* (ref. 36). cuTAR215705 colocalized with *BIRC5* and *ELAVL1* in the tumor cells (Fig. 3e). The colocalization was significant in the tumor region of the tissue and the absence or reduced expression of both the cuTAR and the two aforementioned genes (*BIRC5* and *ELAVL1*) was observed in other cell types (Supplementary Fig. 14e).

Examples of cell-type-specific lncRNAs (cuTARs) associated with response to cancer therapy

Case study 1: cell-type-specific lncRNAs in acral and cutaneous melanoma and their associations with response to anti-PD-1 immunotherapy. To demonstrate a deeper analysis to explore possible

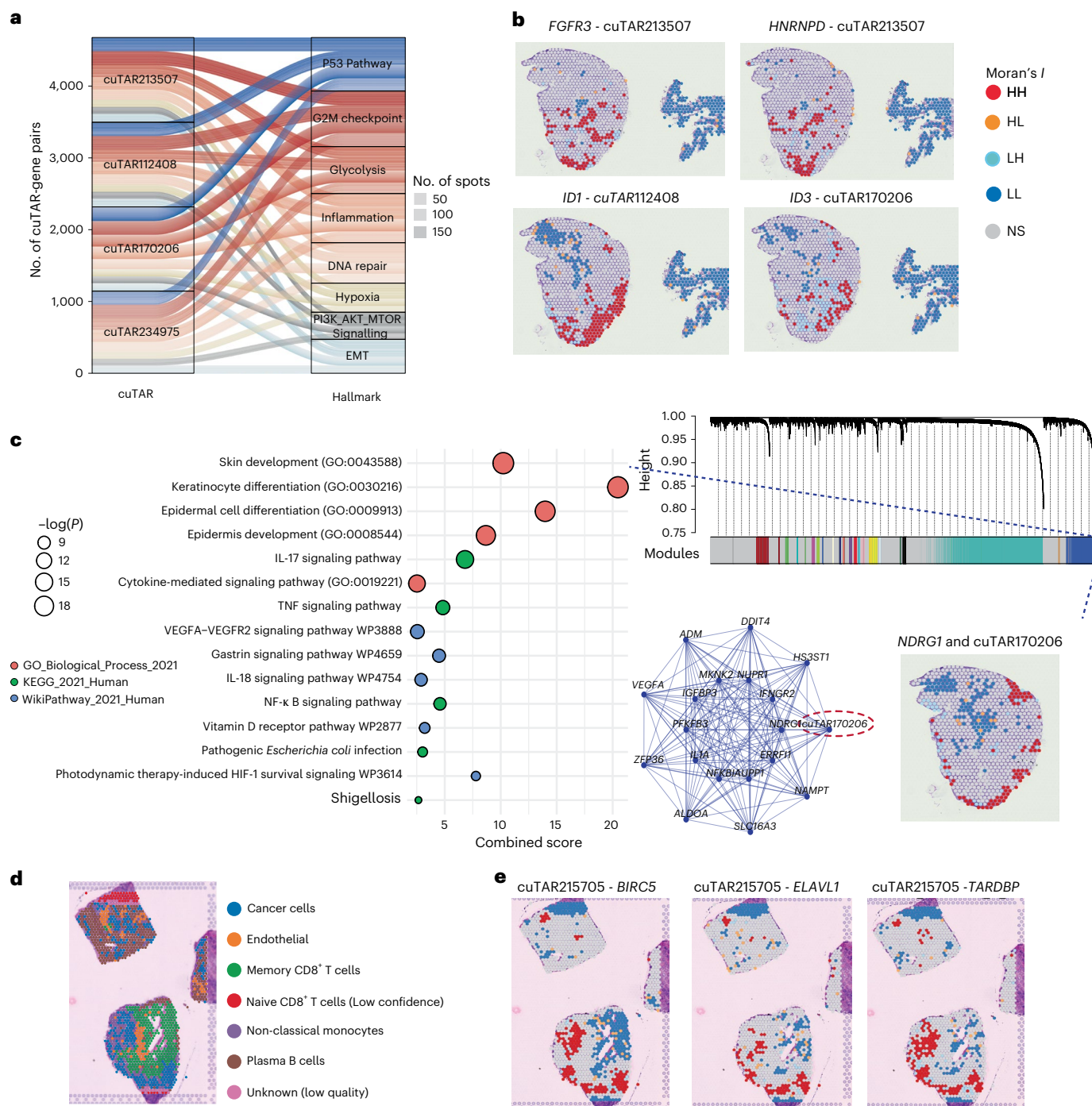


Fig. 3 | Spatial coexpression of cuTARs with cancer-relevant genes. a, Top cuTARs in H&N cancer (sample HNC) showing high spatial autocorrelation with different cancer-relevant gene sets. Here, HH indicates that both the cuTAR and the gene are expressed in the same spot. The y axis shows the number of cuTAR-gene pairs, and the opacity represents the number of spots where spatial autocorrelation (HH) occurs. **b**, Spatial autocorrelation patterns between cuTARs and genes. LH/HL indicates that only one of the features is expressed, LL indicates that neither feature is expressed and NS denotes no significant autocorrelation. **c**, WGCNA analysis shows 32 genes displaying high coexpression with *cuTAR170206* (circled in red) that forms part of the same regulatory module, which includes genes like *NDRG1*, the downregulation of which is associated with metastasis in OPSCC and other genes involved in several cancer hallmark

signaling pathways. Pathway enrichment for gene coexpression modules was performed using Enrichr, accessed through the RunEnrichr() function in R. Enrichment significance was evaluated using a two-sided Fisher's exact test based on the overlap between module genes and pathway-annotated genes. In the dot plot, dot size represents $-\log_{10}$ (adjusted *P* value) and dot color represents the database source (Gene Ontology, Kyoto Encyclopedia of Genes and Genomes (KEGG) and WikiPathways). **d**, Cell-type annotations of a breast cancer tissue (Sample BreastA) inferred from scType. **e**, *cuTAR215705* was predicted to interact with ribosomal binding proteins *ELAVL1* and *TARDBP* using HLPi-Ensemble. The spatial plots show the spatial autocorrelation of *cuTAR215705* with the mRNA encoding the RNA-binding protein *ELAVL1* and its interacting partner *BIRC5* and the mRNA *TARDBP* in breast cancer.

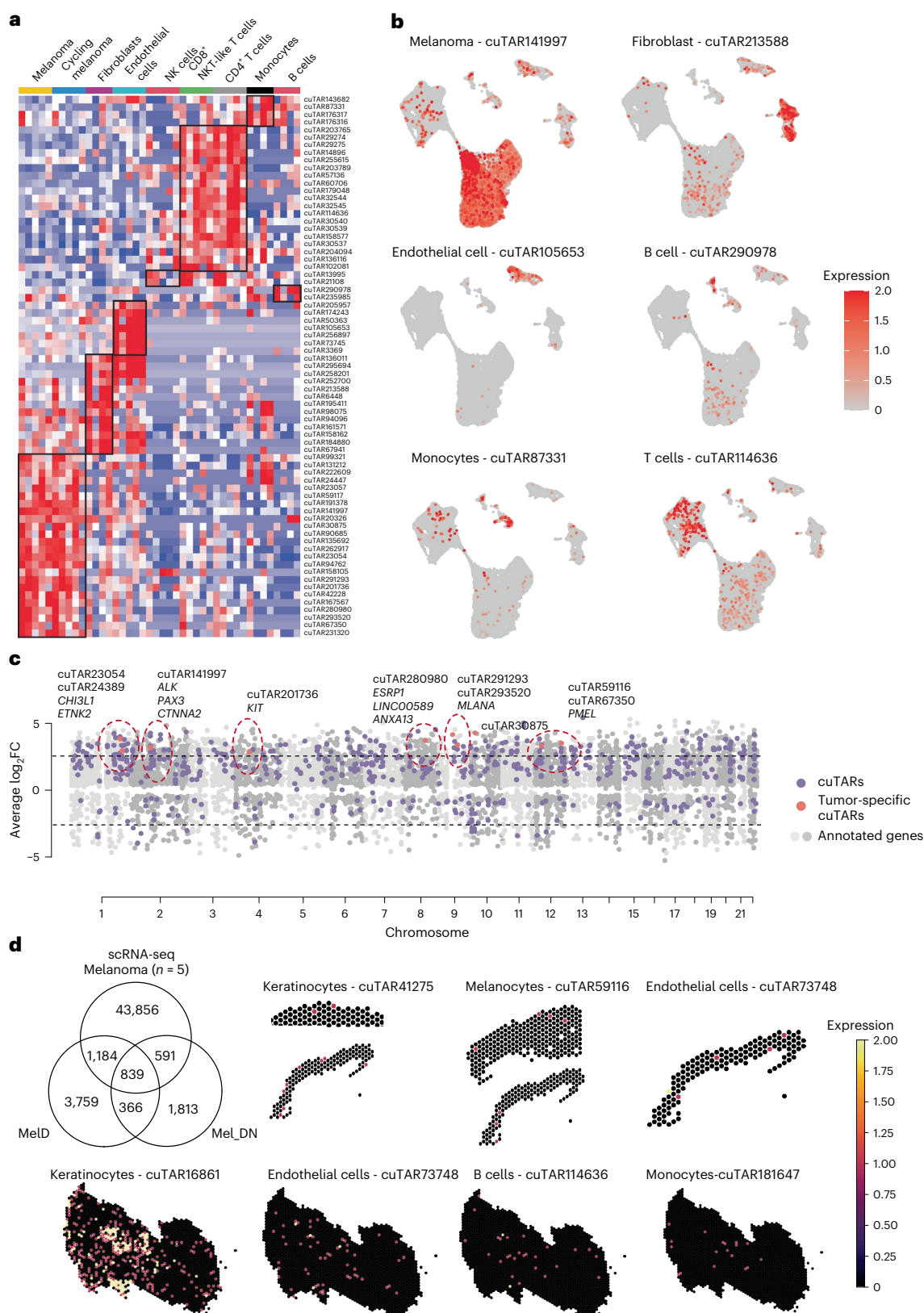


Fig. 4 | Cell-type-specific expression of lncRNAs. a, Cell-type-specific cuTARs (highlighted with black boxes) in scRNA-seq acral and cutaneous melanoma samples. **b**, Expression of some cell-type specific cuTARs projected on the UMAP. **c**, Genes and cuTARs upregulated in tumor cells (melanoma and cycling melanoma) compared to other cells. Tumor cell-specific cuTARs (pink dots) and their proximal cancer-associated coding genes are highlighted in red circles.

Purple dots represent all cuTARs and gray dots represent protein-coding genes. **d**, The venn diagram shows lncRNAs identified in the scRNA-seq data that overlap with Visium ST data from two samples. Spatial tissue plots show the expression of cell-type-specific cuTARs in 10x Visium melanoma samples MelD (top) and Mel_DN (bottom).

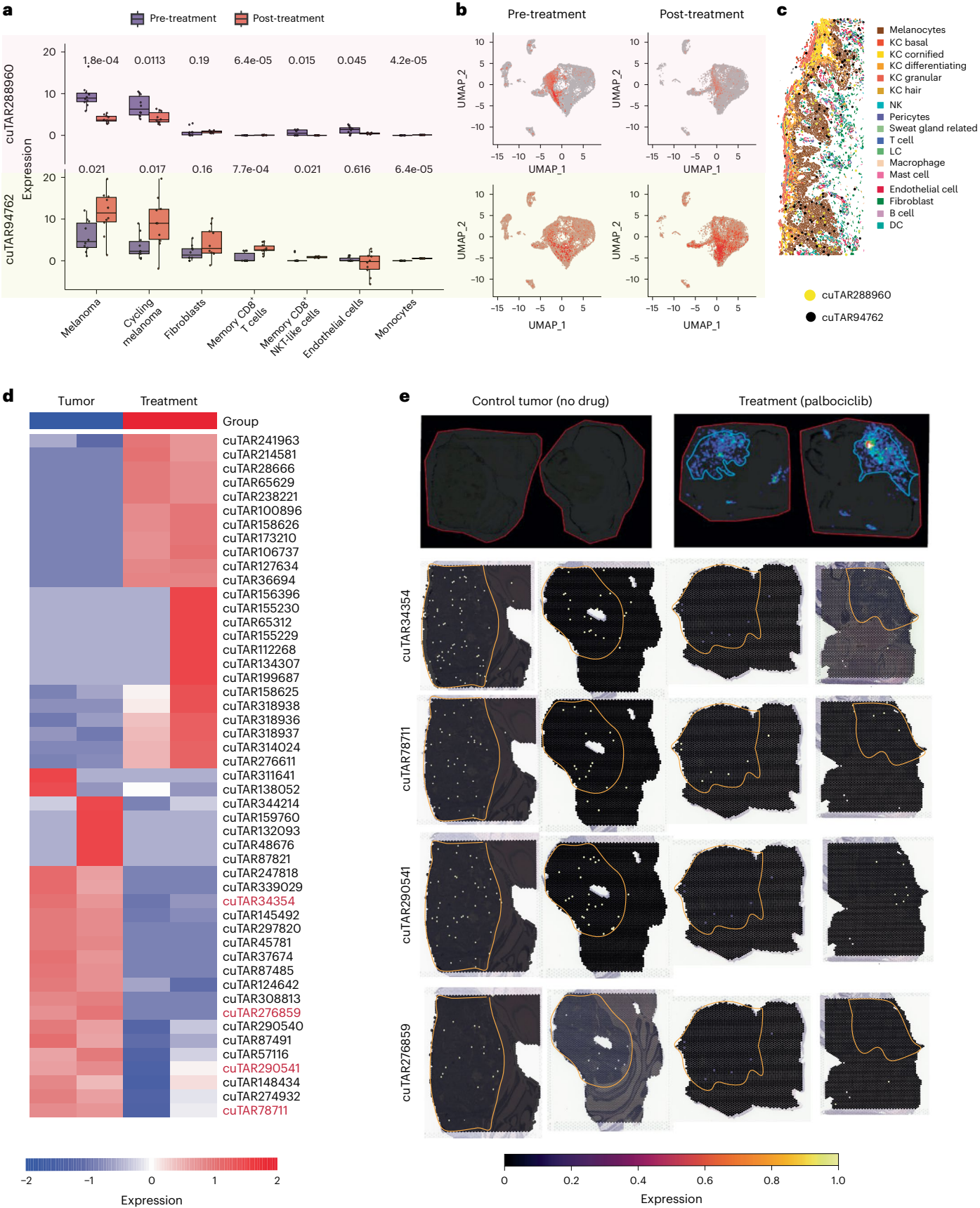


Fig. 5 | Response of lncRNAs to drug treatments. **a**, Expression trends of melanoma-specific cuTARs cuTAR288960 and cuTAR94762 in response to anti-PD-1 therapy in the scRNA-seq dataset. Pseudobulk expression profiles were generated by randomly sampling 50 cells per cell type, repeated ten times and the aggregated expression values across cell types were visualized. Boxplots show these pseudobulk expression values. The center line denotes the median, the box bounds represent the first and third quartiles (25th–75th percentiles) and the whiskers extend to 1.5× the interquartile range (IQR). Statistical significance between conditions was determined using a two-sided Wilcoxon rank-sum test and *P* values are displayed above each comparison. **b**, UMAP projections of

cuTAR expressions pre- and post-therapy. **c**, Expression of cuTAR288960 and cuTAR94762 in Xenium data where most of the transcripts (yellow and black dots, respectively) were seen in melanocytes (brown polygons representing a single-cell). **d**, Differentially expressed cuTARs in medulloblastoma PDOX mouse models in response to the drug palbociclib. Some cuTARs downregulated upon drug treatment are highlighted in red text. **e**, Spatial expression of tumor-specific cuTARs cuTAR78711 (LncBook HSNLNG0100055), cuTAR276859 and cuTAR290541 (LncBook HSNLNG0071008 and LncPedia lnc-PAX5-6) downregulated post-treatment are highlighted in red in **d**. Red outlines in the image define the tissue borders and yellow outlines highlight the human tumor region.

functions of the cuTARs in this atlas, a public scRNA-seq acral and cutaneous melanoma dataset was analyzed to identify cell-type-specific lncRNAs (cuTARs) and to analyze their response to anti-PD-1 immunotherapy (Supplementary Fig. 15). The lncRNAs differentially expressed in each cell-type (Supplementary Fig. 16a,b) were identified with edgeR³⁷ (Fig. 4a) and expression of a subset were visualized on Uniform Manifold Approximation and Projection (UMAP) (Fig. 4b). The cuTARs upregulated in the tumor cells show coexpression with respective proximal protein-coding genes associated with melanoma (Fig. 4c). For example, we observed the co-upregulation of cuTARs with pigmentation genes *PMEL* (cuTAR67350 LncBook HSNLNG0094046) and *MLANA* (cuTAR293520)³⁸. Additionally, cuTAR280980 exhibited coexpression with *ESRP1*, encoding a master splicing regulator in EMT³⁹, an annotated lncRNA, *LINC00589* (also known as *TSNCR8*), implicated in diverse roles across different cancer types⁴⁰, and with *ANXA1*, known to be upregulated in invasive melanomas⁴¹. This suggests that these cuTARs could be part of regulatory modules influencing tumor progression.

Notably, cuTARs from our in-house ST melanoma samples overlapped with the scRNA-seq melanoma data from five samples (Fig. 4d). A subset of these overlapping cuTARs, showing cell-type-specific expression in the scRNA-seq data, were mapped onto the in-house ST melanoma tissues to visualize their spatial distribution (Fig. 4d).

The cuTARs differentially expressed pre and post anti-PD-1 immunotherapy (Fig. 5a–c) were also identified and discussed in Supplementary Note 4 (Supplementary Fig. 16).

Case study 2: response to palbociclib in medulloblastoma PDOX models. Patient-derived orthotopic xenograft (PDOX) mouse models of human tumors are useful preclinical models to study tumors or drug effects. Visium data for treated and untreated samples from human medulloblastoma PDOX mouse models were generated as described previously⁴² (Supplementary Fig. 17a). The lncRNAs differentially expressed between the treated and untreated groups were identified. Some cuTARs that were expressed in the untreated human tumor region were downregulated or unexpressed in the palbociclib-treated human tumor regions (Fig. 5d,e). A similar trend was observed with protein-coding oncogenes like *FOXMI*, *PLK1*, *E2F1* and *GLI2* (ref. 43), which are oncogenes and/or involved in pathways modulated by the drug (Supplementary Fig. 17b).

We demonstrate the prognostic potential of cuTARs and show that, despite weak detection in bulk RNA-seq, their presence is supported in TCGA, consistent with restriction to small, cell-type-specific

populations observed in single-cell data (Supplementary Note 5 and Extended Data Figs. 6 and 7). We further identify cell-type-specific cuTARs in glioblastoma, breast and kidney cancer scRNA-seq datasets (Supplementary Fig. 18).

Validation of lncRNA detection, cancer specificity and coexpression patterns with Xenium customized panel

We performed Xenium in situ hybridization to independently assess findings from Visium and scRNA-seq, enabling us to visualize targets at a cellular level. We designed a custom Xenium panel comprising 76 cuTARs (Supplementary Table 5). Overall, all the cuTARs were found to be expressed, substantially higher than the signals from gDNA negative control probes (Extended Data Fig. 8a). Differential expression analysis confirmed the cancer-type specificity of cuTARs, previously observed with scRNA-seq and Visium (Fig. 6a). The cuTARs were visualized at a single-cell resolution and were seen to be specific to the predicted tumor cells (red polygons) across different cancer types (Fig. 6b). The coexpression patterns with coding genes shown in Fig. 3e were also confirmed. Two particular cuTARs specific to breast cancer samples were analyzed for their coexpression and prognostic value, cuTAR215705 and cuTAR32998 (GENCODE v.47 *LOC107984223*). These were found to be coexpressed with *BIRC5* and associated RBPs (Fig. 3e). The coexpression and localization of one of the cuTARs, cuTAR215705 with *ELAVL1* and *BIRC5* at a subcellular level in one of the breast cancer samples is shown Extended Data Fig. 8b–d; however, these two cuTARs often were not coexpressed.

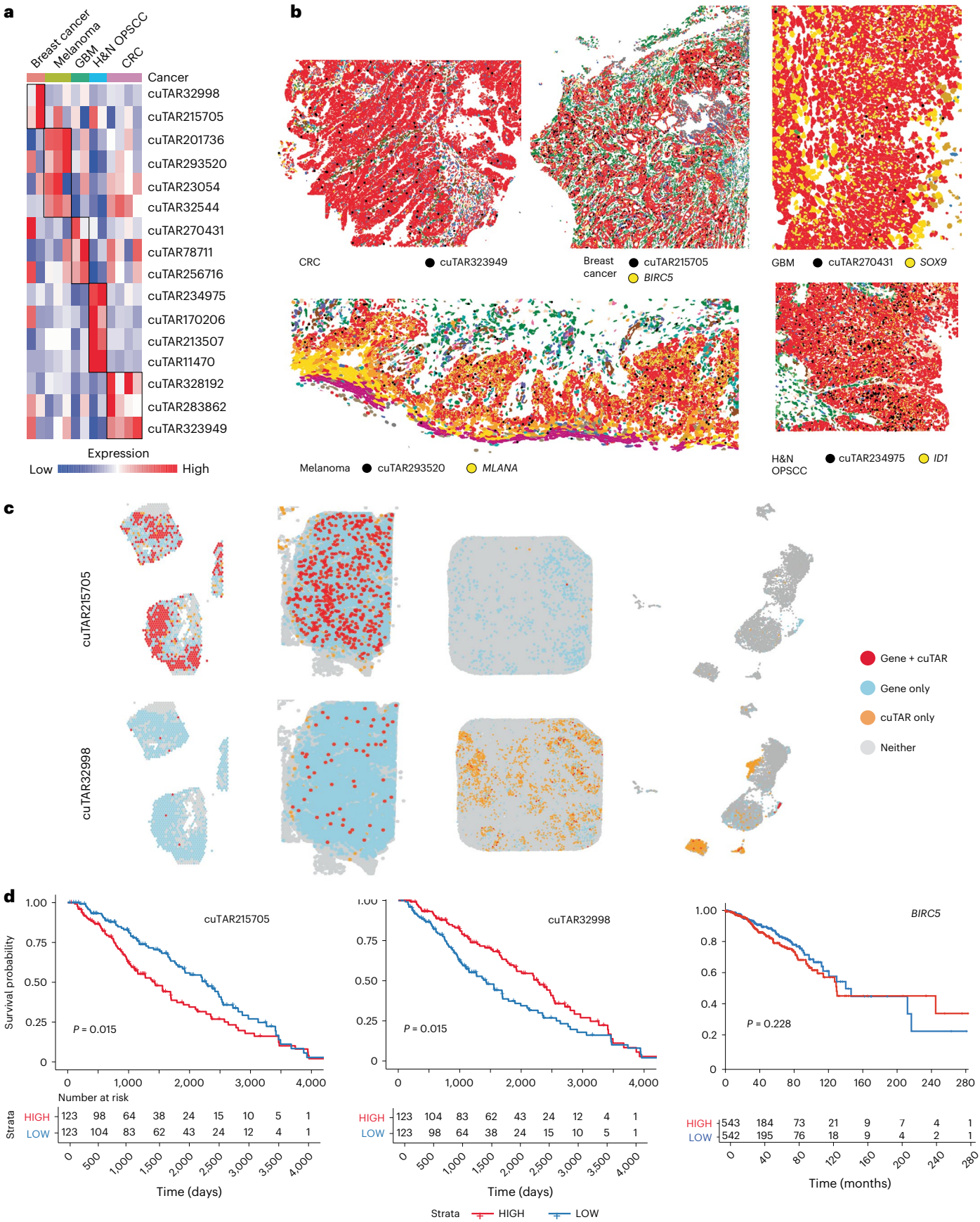
The differences in expression patterns were replicated in samples across 10x Visium, Xenium, Takara Bio Seeker and 10x Chromium scRNA-seq (Fig. 6c). Moreover, in the TCGA breast cancer cohort (246 patients), cuTAR215705 was associated with poor survival, whereas cuTAR32998 correlated with better survival. Notably, these cuTARs stratified patients more efficiently than the coding gene *BIRC5* (Fig. 6d).

A spatial and single cell pan-cancer atlas of lncRNAs

The findings of this study have been made publicly available through our database ‘*SPANC-Lnc*’ on AWS cloud (Extended Data Fig. 9). A total of 219,442 cuTARs across 13 broad cancer types were identified using scRNA-seq and 10x Visium ST, further validated using five ST platforms. This resource allows users to interactively query lncRNA/cuTAR annotations in a user-friendly UCSC-style genome browser, providing visualizations for cross-cancer expression comparisons, cell-type specificity via UMAP overlays and spatial localization within tumor sections. Users

Fig. 6 | Validation of top potential functional cuTAR candidates using Xenium and analysis of prognostic value. **a**, Cancer specificity of cuTARs identified using scRNA and 10x Visium spatial technology were verified in the Xenium experiment where they follow consistent expression patterns. The differential expression of cuTARs were computed using edgeR. The columns in the heatmap under each cancer type represents pseudo bulked expression of cuTARs in the cancer cells of each sample. **b**, Expression of some cancer-specific cuTARs (black dots) observed in the predicted tumor cells (bright red polygons) across different cancer types. Some of their coexpression patterns with coding genes (yellow dots) were also verified. **c**, Co-expression patterns of two cuTARs with *BIRC5* in breast cancer samples. cuTAR215705 and cuTAR32998 show mutually exclusive

expression, with cuTAR32998 absent or low when cuTAR215705 is highly expressed. This pattern is consistently observed across samples from multiple platforms (from left to right): 10x Visium (BreastA, triple-negative breast cancer), Xenium (ER⁺PR⁺HER2⁺), Takara Bio Seeker and 10x Chromium scRNA-seq (Invasive Ductal Carcinoma). **d**, cuTAR215705 is associated with poor prognosis in the TCGA breast cancer cohort, similar to the gene *BIRC5* while cuTAR32998 is associated with better survival. The differences in survival probabilities is more pronounced with the cuTARs than with *BIRC5*. The exact *P* value displayed in the plot was computed using a two-sided log-rank (Mantel–Cox) test implemented via the `survdiff()` function from the survival R package for the cuTARs, and obtained from cBioPortal for *BIRC5*.



can also retrieve splice junction predictions and associated metadata using the table browser. The metadata includes information about cancer/cell-type specificity, overlap with public lncRNA databases, structural characterization and cross-platform validations. SPanC-Lnc is a large pan-cancer resource of known and new lncRNAs with single cell and spatial contexts. Resources like FANTOM-CAT, LncBook and NONCODE provide lncRNA annotations mostly from bulk RNA-seq data of cell lines and tissues (Supplementary Table 1). SPanC-Lnc has lncRNAs with cell type and spatial resolution, lacking from bulk RNA-seq. Recent studies utilizing scRNA-seq and ST data have explored the expression of annotated lncRNAs^{5,6,44,45}; however, information on a pan-cancer scale for potential novel lncRNAs is not available. SPanC-Lnc not only provides annotations and transcript characterization, but also expression comparisons across different cancer types and cell-types and analyzes their potential putative functions. The resource is structured to accommodate incorporation of additional datasets and analytical refinements, and transcript annotations may be further refined as new data become available or as structural and functional validations provide additional evidence. The modular design and documented workflows facilitate transparency, reproducibility and potential future extensions of the atlas.

Discussion

This study utilized advanced transcriptomics technologies for uncovering a diverse repertoire of previously unannotated and/or uncharacterized lncRNAs in cancer. Through integrative analysis across multiple cancer types, we identified lncRNAs exhibiting dynamic expression patterns and tumor-specific spatial distributions, suggesting their potential roles as major regulators in tumorigenesis and cancer progression.

At a structural level, lncRNAs could be derived from enhancers, promoters, pseudogenes or transposons. The lncRNAs derived from the transcription of active enhancers (e-lncRNAs) are known to participate in many cancer-associated biological processes like angiogenesis, proliferation, invasion and metastasis^{46–49}. At the genetic level, disease-associated GWAS SNPs or eQTLs could regulate nearby protein-coding genes through altering the levels of colocalizing lncRNAs⁵⁰. Spatial autocorrelation with coding genes involved in several cancer-related pathways revealed coexpression of these novel lncRNAs in pathways associated with cell cycle, cellular proliferation, immune response, angiogenesis, hypoxia and metastasis, highlighting their putative involvement in critical biological processes underlying cancer development.

In our study, reads within a 500-bp window were merged and annotated as a single uTAR. While this heuristic simplifies annotation, it introduces potential limitations. For example, a long cuTAR (defined from overlapping uTARs across samples) may represent merged signals from independent transcripts, or adjacent cuTARs may be conflated into a single transcript. This was particularly evident in the higher-resolution STOmics/Takara Bio dataset, where longer transcript boundaries led to the merging of multiple Visium-derived cuTARs into a single STOmics cuTAR. Notably, cuTAR discovery and characterization were based primarily on Visium and scRNA-seq data, where transcript boundaries are shorter and more conservative, reducing the risk of over-merging during novel lncRNA identification. STOmics/Takara Bio data were used for validation rather than discovery, and in many cases the extended boundaries reinforced the presence of a single transcript by extending its annotated range. We prioritized expression specificity, reproducibility and biological association (for example, cancer state, treatment response and cell type) rather than precise transcript boundaries. Consequently, the biological signal and specificity of these cuTARs remain robust despite minor annotation discrepancies. This atlas provides a biologically relevant set of cancer-specific, cell-type-specific and treatment-responsive lncRNAs that already serves as a foundation for biological discovery and can be further refined as transcript annotations improve.

While poly-adenylated lncRNAs were the main focus of this study, other classes of ncRNAs like circular RNAs (circRNAs)^{51,52}, pre-miRNAs/miRNAs⁵³, piRNAs^{54,55} and non-polyadenylated lncRNAs⁵⁶ are also aberrantly expressed in cancers and gene regulation. The identification of these lncRNAs in addition to the new lncRNA panel put forth in this study would serve as a valuable resource for future investigations. Notably, we also added STOmics (random hexamer capture) data to complement polyA-capture methods. We detected 93,255 new uTARs and 105,701 uTARs overlapping with 41,573 entries in databases like LncBook, LNCipedia and NONCODE. Of these, 63,584 uTARs overlap with 32,202 lncRNAs on the same strand. This suggests that some of these uTARs may be non-polyadenylated lncRNAs, making them worthwhile candidates for further investigation in future studies.

A large number of lncRNAs identified from scRNA-seq and ST data can be assigned as highly expressed markers specific to subpopulations of cells, including rare cells. SPanC-Lnc constitutes a scalable resource for investigating fundamental aspects of lncRNA biology, offering new approaches to identifying new lncRNAs and inferring their functions at single-cell and spatial resolution. By systematically comparing lncRNA expression across diverse cellular subpopulations, the discovery framework enables interrogation of cell-type specificity, transcriptional heterogeneity and context-dependent expression patterns that are not accessible through bulk RNA-seq. The integration of spatial information further allows examination of how lncRNA expression varies across tissue architecture, such as tumor cores, immune-rich regions and invasive margins. This spatially resolved view provides an opportunity to study microenvironment-associated transcriptional programs and to generate testable hypotheses about lncRNA regulation in situ, for example through the colocalization and co-regulation of lncRNA with coding genes and pathways. Notably, SPanC-Lnc establishes a standardized computational strategy for integrating single-cell and spatial datasets to annotate, prioritize and visualize lncRNAs across different technology platforms and cancer types. As such, it provides a reference database that can be extended to include additional datasets for more tissues, disease states and emerging single-cell and ST technologies. By enabling systematic exploration of the noncoding transcriptome at a cellular resolution, this resource facilitates mechanistic studies aimed at defining the regulatory principles governing lncRNA expression and functions.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41592-026-03071-4>.

References

- Boland, C. R. Non-coding RNA: it's not junk. *Dig. Dis. Sci.* **62**, 1107–1109 (2017).
- Niderla-Bielińska, J., Jankowska-Steifer, E. & Włodarski, P. Non-coding RNAs and human diseases: current status and future perspectives. *Int. J. Mol. Sci.* **24**, 11679 (2023).
- Park, E.-G., Pyo, S.-J., Cui, Y., Yoon, S.-H. & Nam, J.-W. Tumor immune microenvironment lncRNAs. *Brief. Bioinform.* **23**, bbab504 (2021).
- Yang, L., Duff, M. O., Graveley, B. R., Carmichael, G. G. & Chen, L.-L. Genomewide characterization of non-polyadenylated RNAs. *Genome Biol.* **12**, R16 (2011).
- Bitar, M. et al. Redefining normal breast cell populations using long noncoding RNAs. *Nucleic Acids Res.* **51**, 6389–6410 (2023).
- Pinkney, H. R., Black, M. A. & Diermeier, S. D. Single-cell RNA-seq reveals heterogeneous lncRNA expression in xenografted triple-negative breast cancer cells. *Biology* **10**, 987 (2021).

7. Xu, K. et al. Spatial transcriptome analysis of long non-coding RNAs reveals tissue specificity and functional roles in cancer. *J. Zhejiang Univ. Sci. B* **24**, 15–31 (2023).
8. Lv, D. et al. LncSpA: LncRNA spatial atlas of expression across normal and cancer tissues. *Cancer Res.* **80**, 2067–2071 (2020).
9. Weirick, T. et al. The identification and characterization of novel transcripts from RNA-seq data. *Brief. Bioinform.* **17**, 678–685 (2016).
10. Wang, M. F. Z. et al. Uncovering transcriptional dark matter via gene annotation independent single-cell RNA sequencing analysis. *Nat. Commun.* **12**, 2158 (2021).
11. Xu, K. et al. Pan-cancer characterization of expression and clinical relevance of m6A-related tissue-elevated long non-coding RNAs. *Mol. Cancer* **20**, 31 (2021).
12. Li, Y. et al. Pan-cancer characterization of immune-related lncRNAs identifies potential oncogenic biomarkers. *Nat. Commun.* **11**, 1000 (2020).
13. Isaev, K. et al. Pan-cancer analysis of non-coding transcripts reveals the prognostic onco-lncRNA HOXA10-AS in gliomas. *Cell Rep.* **37**, 109873 (2021).
14. Luo, Y., Morgan, S. L. & Wang, K. C. PICSAR: long noncoding RNA in cutaneous squamous cell carcinoma. *J. Invest. Dermatol.* **136**, 1541–1542 (2016).
15. Lu, D. et al. The long noncoding RNA TINCR promotes breast cancer cell proliferation and migration by regulating OAS1. *Cell Death Discov.* **7**, 41 (2021).
16. Wang, Q. et al. LncRNA TINCR impairs the efficacy of immunotherapy against breast cancer by recruiting DNMT1 and downregulating MiR-199a-5p via the STAT1-TINCR-USP20-PD-L1 axis. *Cell Death Dis.* **14**, 76 (2023).
17. Lin, Z.-B. et al. Long noncoding RNA KCNQ1OT1 is a prognostic biomarker and mediates CD8⁺ T cell exhaustion by regulating CD155 expression in colorectal cancer. *Int. J. Biol. Sci.* **17**, 1757–1768 (2021).
18. Mini, E. et al. RNA sequencing reveals PNN and KCNQ1OT1 as predictive biomarkers of clinical outcome in stage III colorectal cancer patients treated with adjuvant chemotherapy. *Int. J. Cancer* **145**, 2580–2593 (2019).
19. Perteghella, T. et al. The GENCODE CLS project: massively expanding the lncRNA catalog through capture long-read RNA sequencing. Preprint at *bioRxiv* <https://doi.org/10.1101/2024.10.29.620654> (2026).
20. Camargo, A. P., Sourkov, V., Pereira, G. A. G. & Carazzolle, M. F. RNAsamba: neural network-based assessment of the protein-coding potential of RNA sequences. *NAR Genom. Bioinform.* **2**, lqz024 (2020).
21. Li, J. & Liu, C. Coding or noncoding, the converging concepts of RNAs. *Front. Genet.* **10**, 496 (2019).
22. Pan, J. et al. Functional micropeptides encoded by long non-coding RNAs: a comprehensive review. *Front. Mol. Biosci.* **9**, 817517 (2022).
23. Ma, L., Bajic, V. B. & Zhang, Z. On the classification of long non-coding RNAs. *RNA Biol.* **10**, 924–933 (2013).
24. Avsec, Ž. et al. Advancing regulatory variant effect prediction with AlphaGenome. *Nature* **649**, 1206–1218 (2026).
25. The Cancer Genome Atlas: citing TCGA. *National Cancer Institute* <https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga/using-tcga/citing-tcga> (2019).
26. Gao, R. et al. Delineating copy number and clonal substructure in human tumors from single-cell transcriptomes. *Nat. Biotechnol.* **39**, 599–608 (2021).
27. Shiau, C.-K. et al. High throughput single cell long-read sequencing analyses of same-cell genotypes and phenotypes in human tumors. *Nat. Commun.* **14**, 4124 (2023).
28. Peppino, G. et al. Teneurins: role in cancer and potential role as diagnostic biomarkers and targets for therapy. *Int. J. Mol. Sci.* **22**, 2321 (2021).
29. Jang, M. K., Shen, K. & McBride, A. A. Papillomavirus genomes associate with BRD4 to replicate at fragile sites in the host genome. *PLoS Pathog.* **10**, e1004117 (2014).
30. Causer, A. et al. Deep spatial-omics analysis of head & neck carcinomas provides alternative therapeutic targets and rationale for treatment failure. *npj Precis. Oncol.* **7**, 89 (2023).
31. Hiwatari, M. et al. Abstract 6734: Identification of the novel TENM3-ALK fusion in an AYA case with ALK rearranged neuroblastoma. *Cancer Res.* **83**, 6734 (2023).
32. de Lima, J. M. et al. NDRG1 deficiency is associated with regional metastasis in oral cancer by inducing epithelial–mesenchymal transition. *Carcinogenesis* **41**, 769–777 (2020).
33. Joshi, V., Lakhani, S. R. & McCart Reed, A. E. NDRG1 in cancer: a suppressor, promoter, or both? *Cancers* **14**, 5739 (2022).
34. Huang, H., Li, L. & Wen, K. Interactions between long non-coding RNAs and RNA-binding proteins in cancer. *Oncol. Rep.* **46**, 256 (2021).
35. Hu, H. et al. HLPI-Ensemble: prediction of human lncRNA–protein interactions based on ensemble strategy. *RNA Biol.* **15**, 797–806 (2018).
36. Al-Yahya, S. et al. Post-transcriptional regulation of BIRC5/survivin expression and induction of apoptosis in breast cancer cells by tristetraprolin. *RNA Biol.* **21**, 1–15 (2024).
37. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
38. Kawakami, A. & Fisher, D. E. The master role of microphthalmia-associated transcription factor in melanocyte and melanoma biology. *Lab. Invest.* **97**, 649–656 (2017).
39. Yao, J. et al. Altered expression and splicing of ESRP1 in malignant melanoma correlates with epithelial–mesenchymal status and tumor-associated immune cytolytic activity. *Cancer Immunol. Res.* **4**, 552–561 (2016).
40. Bai, W. et al. LINC00589-dominated ceRNA networks regulate multiple chemoresistance and cancer stem cell-like properties in HER2⁺ breast cancer. *npj Breast Cancer* **8**, 115 (2022).
41. Delorme, S. et al. New insight into the role of ANXA1 in melanoma progression: involvement of stromal expression in dissemination. *Am. J. Cancer Res.* **11**, 1600–1615 (2021).
42. Vo, T. et al. Spatial transcriptomic analysis of Sonic hedgehog medulloblastoma identifies that the loss of heterogeneity and promotion of differentiation underlies the response to CDK4/6 inhibition. *Genome Med.* **15**, 29 (2023).
43. Huang, D. et al. GLI2 promotes cell proliferation and migration through transcriptional activation of ARHGEF16 in human glioma cells. *J. Exp. Clin. Cancer Res.* **37**, 247 (2018).
44. Ferri-Borgogno, S. et al. Molecular, metabolic, and subcellular mapping of the tumor immune microenvironment via 3d targeted and non-targeted multiplex multi-omics analyses. *Cancers* **16**, 846 (2024).
45. McKellar, D. W. et al. Spatial mapping of the total transcriptome by in situ polyadenylation. *Nat. Biotechnol.* **41**, 513–520 (2023).
46. García-Padilla, C. et al. Molecular mechanisms of lncRNAs in the dependent regulation of cancer and their potential therapeutic use. *Int. J. Mol. Sci.* **23**, 764 (2022).
47. Peng, L. et al. Super-enhancer-associated long noncoding RNA HCCL5 is Activated by ZEB1 and promotes the malignancy of hepatocellular carcinoma. *Cancer Res.* **79**, 572–584 (2019).
48. Yan, L., Chen, H., Tang, L., Jiang, P. & Yan, F. Super-enhancer-associated long noncoding RNA AC005592.2 promotes tumor progression by regulating OLFM4 in colorectal cancer. *BMC Cancer* **21**, 187 (2021).
49. Björklund, S. S. et al. Subtype and cell type specific expression of lncRNAs provide insight into breast cancer. *Commun. Biol.* **5**, 834 (2022).
50. Castellanos-Rubio, A. & Ghosh, S. in *Long Noncoding RNA: Mechanistic Insights and Roles in Inflammation* (ed. Carpenter, S.) 147–160 (Springer, 2022).

51. Kristensen, L. S., Jakobsen, T., Hager, H. & Kjems, J. The emerging roles of circRNAs in cancer and oncology. *Nat. Rev. Clin. Oncol.* **19**, 188–206 (2022).
52. Yarmishyn, A. A. et al. Circular RNAs modulate cancer hallmark and molecular pathways to support cancer progression and metastasis. *Cancers* **14**, 862 (2022).
53. MacFarlane, L.-A. & Murphy, P. R. MicroRNA: biogenesis, function and role in cancer. *Curr. Genomics* **11**, 537–561 (2010).
54. Zhang, Q. et al. The epigenetic regulatory mechanism of PIWI/piRNAs in human cancers. *Mol. Cancer* **22**, 45 (2023).
55. Yao, J. et al. PIWI-interacting RNAs in cancer: biogenesis, function, and clinical significance. *Front. Oncol.* **12**, 965684 (2022).
56. Zhang, Y., Yang, L. & Chen, L.-L. Life without A tail: new formats of long noncoding RNAs. *Int. J. Biochem. Cell Biol.* **54**, 338–349 (2014).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2026

¹The University of Queensland and Indian Institute of Technology Delhi (UQ-IITD) Research Academy, New Delhi, India. ²QIMR Berghofer Medical Research Institute, Brisbane, Queensland, Australia. ³Institute for Molecular Biosciences, The University of Queensland, St. Lucia, Queensland, Australia. ⁴Department of Biochemical Engineering and Biotechnology, Indian Institute of Technology Delhi, New Delhi, India. ⁵School of Biomedical Sciences and Pharmacy, University of Newcastle, Callaghan, New South Wales, Australia. ⁶Precision Medicine Research Program, Hunter Medical Research Institute, New Lambton Heights, New South Wales, Australia. ⁷Queensland Alliance for Agriculture and Food Innovation, The University of Queensland, St. Lucia, Queensland, Australia. ⁸Takara Bio USA, San Jose, CA, USA. ⁹Faculty of Medicine, The University of Queensland, Brisbane, Queensland, Australia. ¹⁰Conjoint Internal Medicine Laboratory, Pathology Queensland, Queensland Health, Brisbane, Queensland, Australia. ¹¹Frazer Institute, The University of Queensland, Brisbane, Queensland, Australia. ¹²Department of Urology, Princess Alexandra Hospital, Brisbane, Queensland, Australia. ¹³Department of Renal Medicine, Townsville University Hospital, Townsville, Queensland, Australia. ¹⁴College of Medicine and Dentistry, James Cook University, Townsville, Queensland, Australia. ¹⁵Present address: Mark Hughes Foundation Centre for Brain Cancer Research, College of Health, Medicine and Wellbeing, University of Newcastle, Callaghan, New South Wales, Australia. ¹⁶These authors jointly supervised this work: Ishaan Gupta, Quan Nguyen. ✉e-mail: Ishaan.Gupta@dbeb.iitd.ac.in; quan.nguyen@qimrb.edu.au

Methods

Primary datasets

Spatial transcriptomics datasets across different cancers. Previously published in-house datasets from fresh frozen tissues of different cancer types (kidney⁵⁷, SCC, BCC⁵⁸, H&N OPSCC³⁰, medulloblastoma⁴², CRC⁵⁹ and breast cancers⁶⁰) and clinical formalin-fixed paraffin-embedded (FFPE) biopsies of melanoma⁶¹ were used in this study (10x Visium poly-A capture). One healthy tissue from kidney was also used (KidB_N). Previously published in-house melanoma datasets⁶¹ and other publicly available datasets of ovarian and prostate cancers from the 10x Visium FFPE probe hybridization based capture were used as the control in our initial phase (published elsewhere)⁶². As this approach relies on predefined gene-targeting probes, novel ncRNAs are generally not captured, although off-target lncRNA detection has been reported⁶². The coverage of the samples used in this study was an average of 100 million reads captured from about 2,000 spots, whereas for the in-house kidney and breast cancer datasets it was 100–300 million reads. Shown in detail are the results for a few cancer types (including H&N cancer, breast cancer and medulloblastoma) of the in-house spatial and public melanoma scRNA-seq datasets.

Single-cell sequencing data

Datasets from 10x Genomics, including glioblastoma, breast cancer (3' scRNA-seq), ovarian, pancreatic, kidney cancer, melanoma and non-small cell lung cancer (NSCLC) (5' scRNA-seq) were used^{63–69}.

Other previously published datasets were also used to analyze lncRNA expressions at a cell-type-specific level. These included 3'-sequencing datasets of acral and cutaneous melanoma from the study [PRJNA862451](#) (ref. 70), peripheral blood mononuclear cells from Merkel cell carcinoma (SRR7722937)⁷¹, patients with H&N cancer (SRR13418965, SRR13419137 and SRR13419168)⁷² and 5'-sequencing datasets of tissue biopsies from patients with BCC. This included data before and after anti-PD-1 therapy from three patients (SRX5128480, SRX5128482, SRX5128486, SRX5128489, SRX5128506 and SRX5128507) out of 12 patients in the original study ([GSE123814](#))⁷³.

Detection of potential novel lncRNAs using an HMM-based approach

The method described by Wang et al.¹⁰ as the TAR-scRNA-seq pipeline was adopted to identify TARs. The pipeline uses an R package GroHMM⁷⁴ that utilizes a two-state hidden Markov model (HMM) to classify regions in an aligned genome as transcriptionally active or not, based on the read coverage in each bin. The position sorted BAM files generated by the 10x SpaceRanger pipeline (SpaceRanger 1.3.0 and 2.0.1 using default parameters) were used as inputs to the pipeline (outputs from 10x Cell Ranger were used for scRNA-seq data). By default, it splits the genome into nonoverlapping bins of 50 bp and is called transcriptionally active if at least three reads were detected in that bin and were labeled as TARs. TARs found within 500 bp apart were merged into one unit. One of the limitations of this approach is that we might wrongly identify two different but adjacent transcripts as a single transcript. The regions identified were then overlapped with reference gene annotations in a strand-specific manner (reference annotations from GENCODE v.43 were used). The TARs overlapping with existing gene annotations (even a few base pairs of overlap) is considered 'annotated' to account for extended gene boundaries) were labeled as TARs (annotated TARs) and the ones falling outside gene boundaries were called uTARs (unannotated TARs). We rule out the unannotated transcripts found on the opposite strand to that of an annotated transcript for this initial phase of the study. A count matrix of the TARs is generated with Drop-Seq tools DigitalExpression function⁷⁵.

Analysis of sequences for coding potential, conservation and stability

To make sure that the identified transcripts were noncoding, the coding potential was analyzed using CPAT⁷⁶ and RNAsamba²⁰. The FASTA

sequences of the cuTARs were extracted using BEDtools⁷⁷ getfasta. CPAT was then run using the inbuilt model for the human genome. A human coding probability (HCP) cutoff of 0.364 was used as described in the tool's documentation. HCP ≥ 0.364 indicates a coding sequence, whereas HCP < 0.364 indicates a noncoding sequence⁷⁶. On the other hand, RNAsamba uses a neural network architecture to classify transcripts as coding or noncoding solely based on sequence information, and is robust to limitations such as truncated open reading frames, and hence more reliable.

Various metrics can help infer the functional significance of a sequence. One such metric is the conservation score, where highly evolutionarily conserved sequences tend to be associated with common essential functions across organisms, while less conserved sequences may exhibit specific functions in an organism. The conservation of the sequences were calculated using the phastCons BigWig files comprising the phastCons scores for multiple alignments of 29 primate/mammalian genome sequences to the human genome GRCh38/hg38 build (<http://hgdownload.cse.ucsc.edu/goldenPath/hg38/phastCons30way/hg38.phastCons30way.bw>)⁷⁸. The BEDtools bigWigAverageOverBed function was used to extract the conservation score for each specified coordinate in a bed file.

MFE is another measure for potential functionality. In general, the lower the MFE, the more stable the secondary structure of the transcript is and hence more likely to be functional. MFE was calculated for each cuTAR in the HNC sample with RNAfold (v.2.6.4)⁷⁹ using default parameters. The most stable secondary structure is predicted.

Characterization of structure and potential functional implications

The identified uTARs were overlapped with TSS and enhancer data from FANTOM for window sizes ranging within 1.5 kb at the 5' end. PAS data from the PolyASite database (<https://polyasite.unibas.ch/>) curated from 3'-end scRNA-seq data was used to support the 3' end of our cuTARs. Splice junction data from Intropolis⁸⁰ (<http://intropolis.rail.bio>), which is a database of splice junctions from analyzing 21,504 human samples sequenced on Illumina was also used to identify splice sites and hence characterize the structure of the identified cuTARs. To further validate our lncRNA identification, the cuTARs were overlapped with the new lncRNA annotations from GENCODE v.47 since release v.43, confirming that these are lncRNAs that were identified by the TAR-scRNA-seq pipeline before GENCODE annotated them as lncRNAs. Additionally, we also predict splice sites utilizing AlphaGenome²⁴, a deep-learning-based tool used to predict several functional genomic features and events. It has been shown that AlphaGenome predicts splicing events with a high accuracy (Pearson correlation of up to 0.82).

The identified cuTAR reads were also overlapped with trait-associated SNPs from the GWAS catalog (all associations v.1.0.2)⁸¹ and with GTEx CAVIAR eQTLs from the UCSC⁸² table browser (downloaded on 23 August 2025)⁸³ with a window size of 100 kb using BEDtools to propose potential genetic-level regulation based on colocalizing features. Three sets of an equal number of random coding regions were considered and the overlap of eQTLs were calculated. There was a markedly higher overlap of cuTARs than with three sets of randomly chosen coding regions. The eQTLs overlapped with both the pan-cancer cuTARs and the cancer-specific cuTARs were identified using direct overlap (BEDtools intersect) and window sizes of 10 kb and 100 kb. The genes associated with these colocalized eQTLs were then queried in enrichR for enriched pathways, GO terms, cell-types and tissues, the significance of which were computed using Fisher's exact test⁸⁴.

The identified cuTARs were overlapped with active enhancers from FANTOM⁸⁵ to identify e-lncRNAs using BEDtools (500-bp window size at the 5' end and also for the entire length of the cuTAR). These can also provide important clues about the regulatory mechanisms and potential functions of these transcripts, helping to shed light on the complex regulatory networks that govern gene expression.

Identification of cancer-specific lncRNAs

Tumor regions were identified using pathologist-provided annotations where available (for example, HNB, HNC), or inferred based on gene expression profiles when such annotations were not available (for example, CP). For analyses requiring cell-type classification (for example, BreastA), *scType*⁸⁶ was employed. Pathologist annotations were exported via Loupe Browser (10x Genomics) and incorporated into downstream analyses. The cuTARs differentially expressed across the annotated cancer and normal regions were identified using Seurat v.4.3 and processed using SCTransform. For FindAllMarkers function, the parameters *min.pct* was set to 0.1 and *logfc.threshold* varied between 0.1–0.25 depending on the sample analyzed. The percentage of spots in the cancerous versus healthy regions expressing the selected cuTARs were also calculated.

Few cuTARs from H&N cancer samples were shortlisted as the cuTARs of interest that could be the top potential functional candidates. They were chosen as they were (1) detected in all H&N OPSCC samples; (2) longer than 1,000 bp; (3) absent from standard annotations (GENCODE v.43) but supported by entries in other lncRNA-specific databases; and (4) differentially expressed in the cancerous region of the tissue. To be more confident about the chosen regions, we also checked their coverages in bulk RNA-seq datasets from TCGA and the scRNA-seq dataset of HNC and checked if they were detected. The bigwig RNA-seq coverage files for 96 random TCGA samples were downloaded using recount. The cuTAR coverages were extracted from these files using *bigWigAverageOverBed* function from UCSC Tools for the specified boundaries in a .bed file. TMM normalization was performed on the raw coverage (counts) using edgeR. By analyzing coverage, gene expression levels can be inferred. Regions with higher coverage indicate higher expression levels of the corresponding cuTARs.

Confirmation of detected signals using long-read sequencing and spatial technologies with single-cell resolution

Oxford Nanopore. Long-read sequencing with ONT was performed using the 10x Visium biotinylated 3' cDNA libraries to validate the detection and expression patterns of the identified potential lncRNAs using PromethION flow cells. SQK-LSK110 and EXP-NBD kits from ONT were used to generate the libraries for the H&N sample C (HNC). Sequence runs were generated using super-accurate basecalling setting with MinKNOW v.22.12.5 and Guppy v.6.4.6 (flow cell type FLO-PRO002). The nanopore run generated ~16 million reads. To demultiplex the reads based on the spatial barcode, *scNanoGPS*²⁷ was used. Some changes were made to the default parameters of the pipeline with respect to the scan region, which was set to 1,000 bp as the 3' adaptors were found that many bases away owing to longer chimeras. The lncRNAs were then identified using a modified version of the TAR-scRNA-seq pipeline, as *scNanoGPS* generates individual alignment files for each spatial barcode identified, rather than an individual BAM file with reads tagged with spatial/cell and UMI barcodes. *Featurecounts*⁸⁷ was used to generate the uTAR expression matrix. GTF annotations of the TARs from the corresponding 10x Visium dataset was also used to generate the count matrix. The identified spatial barcodes were matched with those from the 10x Visium barcode list from the *tissue_positions_list.csv* in the SpaceRanger output of the corresponding Visium data and only those with tissues placed on them were retained. The coordinates of the matched barcodes were also extracted from the aforementioned *csv* file. The expression of some potential candidates were visualized using *stlearn*⁸⁸ and was compared to that of the previously generated short-read data.

Further, the experiment was also performed for skin cancer samples. SQK-NBD-114.24 kit, native barcoding kit 24 V14, was used to prepare libraries for the SCC and BCC samples, which were sequenced on PromethION P24 device using R10 flow cell (FLO-PRO114M). Sequence bases were called during the sequencing run in fast-base-calling mode using PromethION software release 23.11.4 (*minknow-core-promethion*

5.8.3, Dorado v.7.2.13). Then, re-basecalling was performed in super-accuracy mode using Dorado (v.0.5.1). Median read length, read length N50 and median PHRED scores were computed from re-basecalled FASTQ files (*Q* > 10) using NanoStat (v.1.6.0)⁸⁹. The re-basecalled reads were subsequently processed using the same downstream workflow described for HNC, including *scNanoGPS* analysis and cuTAR identification and visualization.

PacBio

Full-length Visium cDNA libraries were used for HiFi sequencing. Multiplexed array sequencing (MAS-Seq), a library preparation method that concatenates cDNA molecules for long-read sequencing, was used following the 10x Single Cell 3' kit protocol (PN: 102-678-600-REV03). After cDNA amplification, cleanup and QC of 10x Visium, 225 pM cDNA per library with concentration of 23 ng μl^{-1} were pooled for sequencing. First, the template switching ligoonucleotide priming artifacts during cDNA synthesis were removed using biotinylated primers. Next, DNA fragments containing orientation-specific MAS segmentation adaptor sequences were generated by performing 16 parallel cDNA amplification reactions. MAS enzyme was used to create single-stranded extensions to enable directional assembly of cDNA segments into a linear array. After DNA damage repair and nuclease treatment, cDNA cleanup was performed with 1.2× SMRTbell beads.

Further, the *pbcrowmwell* workflow of SMRTlink Tools (v.12.0.0)⁹⁰ was used for to process the HiFi BAM files. With the aligned genome file from Minimap2, the TAR-scRNA-seq pipeline was run and the cuTAR expression was overlaid on the tissues and compared to that of Visium.

STOmics

Sample preparation and library construction were performed following the manufacturer's protocol (STOmics Stereo-seq FFPE kit). In brief, stereo-seq chip slides were equilibrated to room temperature and rinsed with nuclease-free water. FFPE tissue blocks were sectioned at 5- μm thickness and mounted onto stereo-seq chip N slides (STOmics Stereo-seq OMNI). Mounted slides were baked at 42 °C for 3 h, followed by overnight incubation at 37 °C. Mounted slides were subjected to deparaffinization, H&E staining and imaging, followed by de-crosslinking to reverse formalin-induced crosslinks. Tissue sections were permeabilized to release RNA molecules, which were hybridized in situ to spatially barcoded free probes embedded on the stereo-seq chip surface, enabling unbiased capture of total RNA. On-chip reverse transcription, ligation and cDNA release were performed, after which cDNA was amplified, purified and converted into sequencing-ready libraries. Libraries were sequenced on the DNBSEQ-T7 platform (MGI Tech) and raw reads were processed using the Stereo-seq Analysis Workflow (SAW; v.8.1.3) for QC, alignment and generation of spatially resolved gene expression matrices integrated with H&E-stained images, enabling downstream analyses such as spatial clustering, cell-type annotation and visualization of coding and noncoding RNA expression at single-cell and subcellular resolution.

Takara Bio Seeker

The Takara Bio Seeker raw and processed data for human breast cancer and melanoma samples were acquired from Curio Bioscience, now part of Takara Bio USA.

RT-qPCR

Six cDNA samples from the 10x Visium libraries, including H&N samples (HNB and HNC), SCC, BCC and CRC (CP and CM) were tested for seven potential lncRNAs (cuTAR100897, cuTAR121087, cuTAR170206, cuTAR17878, cuTAR213507, cuTAR234975 and cuTAR70920). *GAPDH*, a housekeeping gene and *KRT18*, an epithelial marker were used as positive controls. Two negative controls NEG1 and NEG2 were designed to target genomic regions around the centromere that do not code for RNA, hence controls for any genomic DNA contamination.

H₂O was used as a nontemplate negative control. Overall, 12 primers were designed to target seven cuTARs (Supplementary Table 4). The primers were designed in such a way that they did not have any off-targets using the sequence information of the HNC sample. qPCR Master Mix was prepared for each sample following the KAPA SYBR FAST qPCR kit guide to allow each well to contain 1 µl of 10 ng µl⁻¹ cDNA, 5 µl KAPA SYBR FAST qPCR Master Mix (2×) (Roche), 0.2 µl ROX Low Reference Dye (50×) (Roche) and 3.4 µl of RNase-free water. As *GAPDH* was an endogenous control, the template for this primer was further diluted 1:8 to ensure that it was within a comparable quantifiable range. Then, 9.6 µl of the qPCR Master Mix was loaded into each well on the 96-well quantitative PCR with reverse transcription (RT-qPCR) plate before 0.4 µl of 10 nM respective lncRNA primer pairs were added. The qPCR reaction was performed with the following protocol using the ViiA7 96-well Real-Time PCR System with High Resolution Melt (Applied Biosystems); initial denaturation (98 °C for 3 min), two-step amplification (98 °C for 5 s and 63 °C for 30 s) for a total of 35 cycles with the melting curve method. The expression was calculated using the formula $C_{KRT18} \times CT_{KRT18} = C_{cuTAR} \times CT_{cuTAR}$, where CT_{KRT18} and CT_{cuTAR} represent the cycle threshold (CT) values for the positive control gene *KRT18* and the cuTAR target, respectively, and C_{KRT18} and C_{cuTAR} denote their corresponding normalized expression levels, calculated as described below.

The CT was normalized with that of one of the positive controls *KRT18* such that $C_{KRT18} = 1$. The cuTAR expression was calculated as $C_{cuTAR} = 1 \times CT_{KRT18} / CT_{cuTAR}$. The average was used for two primers targeting a same cuTAR. The expression was visualized on a heatmap (Supplementary Fig. 12).

Spatial autocorrelation of lncRNAs with genes relevant to cancer

SpatialDE⁹¹ was used to identify spatially variable features. The significant spatially variable cuTARs and the top chosen candidates chosen for the H&N cancer samples were further used to measure pairwise spatial autocorrelation with coding genes using bivariate Moran's *I* using Python scripts. Several hallmark genes relevant to cancer were retrieved from gene set enrichment analysis (GSEA)-MsigDB⁹² and were used for the analysis. The spots showing (1) high expression of both the gene and the cuTAR were categorized as HH; (2) spots with gene expression and no cuTAR expression as LH; (3) ones with cuTAR expression and lack of gene expression as HL; (4) no expression of both features as LL; and (5) NS for spots with insignificant Moran's *I*. The cuTARs showing high spatial correlation (more than 50 HH spots) with genes of different hallmarks were calculated to identify the more functionally relevant cuTARs and the pathways involved. Moran's *I* was calculated in a simple format using the formula below:

$$I_T = \frac{\sum_i (\sum_j w_{ij} z_{j,t-1} \times z_{i,t})}{\sum_i z_{i,t}^2}$$

Here, I_T is the calculated bivariate Moran's *I*. It measures bivariate spatial correlation for expression values of an lncRNA (cuTAR) (*t*) and a gene (*t* - 1) across all spots with the standardized expression of the cuTAR in the central spots/cells *i*, and the standardized expression of the gene in the neighboring spots/cells *j*. Spatial weights (w_{ij}) represent the proximity between spatial spots/cells *i* and *j*. These weights are defined based on spatial contiguity where $w_{ij} > 0$ if *j* is a neighbor of *i* defined by sharing a border or nearest in proximity, and 0 if otherwise; the total sum of the neighbor weights for each central spot is 1. With this, I_T is 1 if all neighboring spots/cells $z_{j,t-1}$ have the same standardized expression value of the gene with the standardized expression value of the cuTAR at the central spot $z_{i,t}$. Therefore, $I_T = 1$ indicates that the value of a gene has a high spatial autocorrelation with the cuTAR based on the expression of values of neighboring spots, suggesting that the gene and cuTAR are spatially dependent and tend to colocalize across the tissue.

The coexpression network analysis was performed using hdWGCNA⁹³ for merged gene-cuTAR count matrices processed using Seurat v.4.1. The spatial regions of interest were the tumor regions marked by a pathologist for the sample HNC and based on Moran's *I* for BreastA.

Interaction with RBPs and colocalization

Machine-learning models that have been trained using known lncRNA-protein interactions can be used to predict interaction of novel lncRNAs with proteins, adding another layer of information for potential functionality. HLPi-Ensemble, which adopts the ensemble strategy based on three mainstream machine-learning algorithms of support vector machines, random forests and extreme gradient boosting was used to predict the interaction of top cuTARs with RBPs. Further, the spatial colocalization of cuTARs with their interacting RBP-encoding genes and with genes reported to be associated with those RBPs was visualized.

Identification of cell-type-specific lncRNAs and their response to therapy

The melanoma scRNA-seq dataset from the study PRJNA862451 (ref. 70) was used to identify cell-type-specific lncRNAs and their response to therapy. The dataset included acral and cutaneous melanoma samples. The coding gene count matrices of four acral and one cutaneous melanoma samples (10x Chromium V2 chemistry) were integrated using the Seurat workflow and clustering was performed. Cell-type clusters were annotated using canonical markers described in the original study: melanocytes (*MITF*, *PMEL*, *TYR*, *DCT*, *MLANA*, *PMEL*, *APOC1* and *SI00A1*), cycling melanoma cells (*UBE2C*, *NUSAP1*, *MKI67* and *CENPF*), endothelial cells (*VWF* and *PECAM1*), fibroblasts (*COL3A1*, *COL1A2*, *COL1A1* and *LUM*), CD8⁺ T cells (*CD8A*, *HAVCR2*, *LAG3*, *PD1*, *TIGIT*, *CTLA4* and *HOPX*), CD4⁺ T cells (*CD4*, *FOXP3* and *IL2*), monocytes (*CD14*, *LYZ*, *CD74*, *CD68* and *CD79A*), B cells (*MS4A1*), natural killer (NK) cells (*GNLY*, *FGFBP2*, *FCGR3A*, *KLRD1* and *KLRF1*), NK-like T cells (*CD3E*, *CD3D*, *GZMB*, *XCL2*, *IFNG*, *CCL4*, *NKG7*, *GZMA* and *GZMK*) and macrophages (*CD68*, *CD163*, *CD14*, *CD11b*, *CD206*, *CD80*, *CD86*, *CD16*, *CD64*, *CCL18*, *CD115*, *CD11c*, *CD32*, *HLA-DR*, *MRC1*, *MSR1*, *GCA* and *Pf4*).

The cell barcode and its corresponding cell-type identity were used as metadata for the differential expression analysis of the potential lncRNAs. Differential expression analysis was performed by the pseudobulk approach using edgeR. The model was built as design <- model.matrix(~celltype + sample) to identify cell-type-specific lncRNAs accounting for sample batch effects. For the Manhattan plot, the FoldChange function of Seurat was used to find the upregulated genes and cuTARs in melanoma and cycling melanoma cells compared to the other immune cells.

Further there was only one acral melanoma sample for which data were collected before and after anti-PD-1 treatment. The differentially expressed genes and lncRNAs in the tumor cells with respect to the other cells were identified. Ten pseudoreplicates of 50 cells per cell type were created and their aggregated expression across cell types was visualized. As a positive control, marker genes for melanoma were also visualized. Multiple hypothesis testing was controlled using Benjamini-Hochberg false discovery rate (FDR) correction for any DE analysis.

Response to drug in medulloblastoma PDOX models

In-house Visium medulloblastoma data as described by Vo et al.⁴² were also analyzed. In brief, medulloblastoma (SHH MB) tissue from a 4.9-year-old patient was used to generate a PDOX line by implanting tumor cells in the cerebellum of immunocompromised NSG mice within hours of surgical removal from the patient and propagating them from mouse to mouse exclusively without in vitro passaging, as previously described. Two tumor-bearing mice were treated with palbociclib hydrochloride (Pfizer) and the rest were untreated. The raw Visium data were processed using a human-mouse hybrid reference as the tissue comprised the human tumor and the mouse brain tissues.

These treated and untreated samples were analyzed for lncRNAs in this study and those significantly differentially expressed between the tumor regions of the treated and untreated groups were identified. DE analysis was performed with each sample as pseudobulks using edgeR and two treated samples were compared to two untreated tumor samples. In addition to the top statistically significant ($P_{adj} < 0.05$) differentially expressed cuTARs, several medulloblastoma-specific cuTARs (cuTAR290541, cuTAR78711 and cuTAR276859) are also shown, as they exhibited reproducible upregulation in untreated tumors and cancer-specific expression patterns, and had significant FDR-corrected P values when analyzed with increased sampling resolution in the pseudobulk analysis.

Xenium in situ experimental validation of cuTARs

A total of 76 cuTARs were chosen for validation on the Xenium in situ platform to verify their cancer-type specificity as observed using 10x Visium. Five genomic DNA regions around centromeres that do not transcribe into RNA were chosen as a negative control set to prove that the detected cuTAR signals were not noise from genomic DNA but true RNA transcripts. The sequences for probe designing were selected based on performing multiple sequence alignment using ClustalW⁹⁴ across the in-house Visium, spatial long read and scRNA samples, and extracting the most represented sequences. The sequences in FASTA format were sent to the 10x Bioinformatics team, who applied their advanced pipeline to design probes for target domains, screening for off-targets (Supplementary Table 5).

FFPE tissue blocks of two H&N OPSCC, three melanoma, two breast cancer, four CRC and two glioblastoma samples were sectioned at a thickness of 5 μ m and mounted onto Xenium slides, in accordance with the FFPE tissue preparation guide (10x Genomics, CG000578, Rev B). In situ hybridization was carried out overnight using 380 probes from the pre-designed Xenium Immuno Oncology Panel (10x Genomics) and from the add-on custom panel of 76 cuTARs and 19 coding genes not represented in the base panel. 4,6-diamidino-2-phenylindole (DAPI) staining was used to label nuclei, which were used for the estimation of cell boundaries (10x Genomics, CG000582, Rev D). Following completion of the run, H&E staining was conducted on the same tissue region. The Xenium samples from each cancer type were preprocessed individually using Seurat version 5.0. During the QC step, cells with zero expression across all genes were filtered out. Normalization was performed using the SCTransform function. Sample batch correction was performed using RunHarmony(), followed by principal-component analysis using the top 30 principal components. Further, the cancer cells and other cells in the microenvironment were annotated by using the label transfer approach using published and well-annotated single-cell datasets of the respective cancer types for H&N⁹⁵, CRC⁹⁶, breast cancer⁹⁷, GBM⁹⁸ and an in-house annotated scRNA-seq dataset for melanoma⁵⁸. To focus on cancer-specific signal and account for variation across diverse tissue architectures, downstream analyses were restricted to predicted tumor-like cells, enriching for the malignant compartment within the spatial tissues. Cancer-type specificity of candidate cuTARs was assessed using edgeR, with tumor-like cells pseudo-bulked per patient to generate sample-level profiles. Tissue-specific expression patterns were visualized via heatmaps to confirm that cuTARs recapitulate cancer-type specificity in Xenium data, consistent with patterns identified from Visium and single-cell RNA-seq analyses.

Survival analysis with TCGA data

Survival analysis was performed by dividing the samples from TCGA into low- and high-cuTAR-expressing groups to see whether any cuTAR could predict survival and thus be used as a prognostic marker. From the TCGA metadata, 'days_to_last_follow_up' (for alive patients) and 'days_to_death' (for the deceased) were used to define overall survival. Survival analysis was performed using overall survival time and event

status, with cuTAR expression strata as the sole covariate. The statistics were implemented using the survdiff() function from the survival R package (v.3.7-0)⁹⁹. For *BIRC5* in Fig. 6d, the plot and the data were obtained from cBioPortal¹⁰⁰ (on 13 December 2025).

Interactive website for annotations and visualization

The findings of this study have been made available for users to browse through the annotations and to visualize coding gene and cuTAR expression on tissue images. The Angular 13.3 web site is hosted using AWS Amplify. It uses a GitHub repository for version control, AWS Code Build to generate the production artifacts (HTML, CSS and JavaScript) and uses AWS Cloudfront CDN (<https://aws.amazon.com/cloudfront/features/>) to distribute these globally. Libraries used include TypeScript v.4.6 and Node v.10.8.1. The backend web service is running the Django v.5.0 web framework on Python v.3.10, using the slim version of Debian Bookworm (https://hub.docker.com/_/debian). This is deployed on AWS Lambda instances provisioned with 4.5 GB of RAM connected to an AWS elastic file system for accessing HDF5 and SQLite 3 files used during querying. The Python libraries used include AnnData 0.10, h5py 3.11.0, numba 0.59.1, numpy 1.26.4, pandas 2.2.2, scanpy 1.10.1, scipy 1.13.1 and seaborn 0.13.2. For visualization of expressions, Python scripts were used in the backend with preloaded AnnData objects for each sample (.h5ad files). The website will be updated whenever new samples or annotations are generated.

Ethics approval

All samples were approved for research under ethical approval numbers 2018000165 and 2017000318 by the University of Queensland's Human Research Ethics Committees.

Reporting summary

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

Data availability

The count matrices, processed objects of high dimensional data (Xenium, Takara Bio Seeker, STOmics), raw Xenium data and STOmics raw image data are available at UQ eSPACE at <https://doi.org/10.48610/4570a11>. The raw data of the unpublished in-house data including (Visium kidney, head and neck OPSCC and CRC) and STOmics sequencing data (melanoma and CRC) can be requested at <https://doi.org/10.48610/86039c7> and will be accessible upon agreement with the applicable ethics approvals and patient consent policies. The SPanc-Lnc database can be accessed at <https://spanclnc.click/>. Source data are provided with this paper.

Code availability

The annotations generated in this study is available on the website <https://spanclnc.click/>. All the scripts used for the analysis and visualizations can be accessed on GitHub at https://github.com/Genomics-MachineLearning/SPanc_Lnc_PanCancer_LncRNA_Atlas.git.

References

- Raghubar, A. M. et al. High risk clear cell renal cell carcinoma microenvironments contain protumour immunophenotypes lacking specific immune checkpoints. *npj Precis. Oncol.* **7**, 88 (2023).
- Prakrithi, P. et al. Integrating 12 spatial and single cell technologies to characterise tumour neighbourhoods and cellular interactions in three skin cancer types. Preprint at *bioRxiv* <https://doi.org/10.1101/2025.07.25.666708> (2025).
- Kawamata, F. et al. Copy number profiles of paired primary and metastatic colorectal cancers. *Oncotarget* **9**, 3394–3405 (2017).
- Wu, S. Z. et al. A single-cell and spatially resolved atlas of human breast cancers. *Nat. Genet.* **53**, 1334–1347 (2021).

61. Vo, T. et al. Assessing spatial sequencing and imaging approaches to capture the molecular and pathological heterogeneity of archived cancer tissues. *J. Pathol.* **265**, 274–288 (2025).
62. Prakrithi, P., Juwayria, Jain, D., Malik, P. S. & Gupta, I. Caution towards spurious off-target signal in 10x Visium spatial transcriptomics assay from potential lncRNAs. *Brief. Bioinform.* <https://doi.org/10.1093/bib/bbad031> (2023).
63. Human ovarian tumor (FF) (v.2, 150 × 150) single cell immune profiling dataset by Cell Ranger 7.0.0, 10x Genomics (2022).
64. Human invasive ductal carcinoma (3' v.3.1, 150 × 150) single cell gene expression dataset by Cell Ranger 6.0.0, 10x Genomics (2021).
65. Human glioblastoma multiforme (3 v.3.1, 150 × 150) single cell gene expression dataset by Cell Ranger 6.0.0, 10x Genomics (2021).
66. Human kidney tumor (FF) (5 v.2, 150 × 150) single cell immune profiling dataset by Cell Ranger 7.0.0, 10x Genomics (2022).
67. Pancreatic tumor (FF) (5' v.2, 150 × 150) single cell immune profiling dataset by Cell Ranger 7.0.0, 10x Genomics (2022).
68. Dissociated tumor cells from primary human melanoma (Universal 5' Gene Expression + VDJ, 5k cells, Chromium Connect) single cell immune profiling dataset by Cell Ranger v.6.0.1, 10x Genomics (2021).
69. NSCLC tumor (F) (v.2, 150 × 150) 5' single cell immune profiling dataset by Cell Ranger 2.2.0, 10x Genomics (2018).
70. Zhang, C. et al. A single-cell analysis reveals tumor heterogeneity and immune environment of acral melanoma. *Nat. Commun.* **13**, 7250 (2022).
71. Paulson, K. G. et al. Acquired cancer resistance to combination immunotherapy from transcriptional loss of class I HLA. *Nat. Commun.* **9**, 3868 (2018).
72. Kürten, C. H. L. et al. Investigating immune and non-immune cell interactions in head and neck tumors by single-cell RNA sequencing. *Nat. Commun.* **12**, 7338 (2021).
73. Yost, K. E. et al. Clonal replacement of tumor-specific T cells following PD-1 blockade. *Nat. Med.* **25**, 1251–1259 (2019).
74. Chae, M., Danko, C. G. & Kraus, W. L. groHMM: a computational tool for identifying unannotated and cell type-specific transcription units from global run-on sequencing data. *BMC Bioinformatics* **16**, 222 (2015).
75. Bageritz, J. & Raddi, G. Single-cell RNA sequencing with Drop-Seq. *Methods Mol. Biol.* **1979**, 73–85 (2019).
76. Wang, L. et al. CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic Acids Res.* **41**, e74 (2013).
77. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
78. Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* **20**, 110–121 (2010).
79. Lorenz, R. et al. ViennaRNA Package 2.0. *Algorithms Mol. Biol.* **6**, 26 (2011).
80. Nellore, A. et al. Human splicing diversity and the extent of unannotated splice junctions across human RNA-seq samples on the Sequence Read Archive. *Genome Biol.* **17**, 266 (2016).
81. Welter, D. et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* **42**, D1001–D1006 (2014).
82. Casper, J. et al. The UCSC Genome Browser database: 2026 update. *Nucleic Acids Res.* **54**, D1331–D1335 (2026).
83. Lonsdale, J. et al. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).
84. Chen, E. Y. et al. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* **14**, 128 (2013).
85. Andersson, R. et al. An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455–461 (2014).
86. Ianevski, A., Giri, A. K. & Aittokallio, T. Fully-automated and ultra-fast cell-type identification using specific marker combinations from single-cell transcriptomic data. *Nat. Commun.* **13**, 1246 (2022).
87. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
88. Pham, D. et al. Robust mapping of spatiotemporal trajectories and cell–cell interactions in healthy and diseased tissues. *Nat. Commun.* **14**, 7739 (2023).
89. De Coster, W., D'Hert, S., Schultz, D. T., Cruts, M. & Van Broeckhoven, C. NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics* **34**, 2666–2669 (2018).
90. SMRT Link (v.12.0.0). <https://www.pacb.com/smrt-link/> (Pacific Biosciences of California, 2023).
91. Svensson, V., Teichmann, S. A. & Stegle, O. SpatialDE: identification of spatially variable genes. *Nat. Methods* **15**, 343–346 (2018).
92. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* **102**, 15545–15550 (2005).
93. Morabito, S., Reese, F., Rahimzadeh, N., Miyoshi, E. & Swarup, V. hdWGCNA identifies co-expression networks in high-dimensional transcriptomics data. *Cell Rep. Methods* **3**, 100498 (2023).
94. Thompson, J. D., Higgins, D. G. & Gibson, T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**, 4673–4680 (1994).
95. Quah, H. S. et al. Single cell analysis in head and neck cancer reveals potential immune evasion mechanisms during early metastasis. *Nat. Commun.* **14**, 1680 (2023).
96. Pelka, K. et al. Spatially organized multicellular immune hubs in human colorectal cancer. *Cell* **184**, 4734–4752.e20 (2021).
97. Xu, L. et al. A comprehensive single-cell breast tumor atlas defines epithelial and immune heterogeneity and interactions predicting anti-PD-1 therapy response. *Cell Rep. Med.* **5**, 101511 (2024).
98. Wang, X. et al. Decoding key cell sub-populations and molecular alterations in glioblastoma at recurrence by single-cell analysis. *Acta Neuropathol. Commun.* **11**, 125 (2023).
99. Therneau, T. A package for survival analysis in R. R package v.3.8-3 <https://CRAN.R-project.org/package=survival> (2024).
100. Cerami, E. et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* **2**, 401–404 (2012).

Acknowledgements

We acknowledge S. Rai and S. Wani for their contribution to the Nanopore and PacBio experiments respectively. We thank B. Day and R. D'Souza (QIMR Berghofer) for glioblastoma samples; K. Khosrotehrani and S. Tan (Frazer Institute) for melanoma samples and P. Simpson (Centre for Clinical Research, UQ) for breast samples that we used in our Xenium in situ experiment. Raw sequencing data from published work were kindly provided by A. Swarbrick for 10x Visium breast cancer samples. We thank A. Raghobar and T. Humphries for the 10x Visium kidney cancer data. We thank QIMRB National Centre for Spatial Tissue and AI Research and The ACRF Centre for Optimised Cancer Therapies. We acknowledge Q.X.X. Lin from 10x Genomics for designing the lncRNA custom probes. Our sincere gratitude extends to W. Wang from Takara Bio for coordinating with us for sharing their data and D. Kool for processing the datasets through the Seeker pipeline. We also thank B.S. Kumar for initially building the SPanC-Lnc

website and X. Liu from the PacBio team for helping setup SMRTlink. We thank L. Grice and J. French for their critical review of the paper and their valuable suggestions. This work was supported by Australian Research Council (DECRA grant DE190100116), National Health and Medical Research Council (project grant 2001514 and Investigator grant GNT2008928) and the UQ-IITD Research Academy. This work is also supported by the Ministry of Human Resource Development, Government of India sponsored Prime Minister's Research Fellowship IITD/Admission/Ph.D./PMRF/2022/36443 to P.P. This work is supported by funds from the Ramalingaswami fellowship (no. BT/RLF/Re-entry/19/2018) by the Department of Biotechnology, Government of India and Indian Council for Medical Research research (grant IIRP-2023-2341/F1) to I.G.

Author contributions

P.P. conducted formal analysis, methodology, visualization, validation, writing of the original draft and final paper. T.V. conducted data generation and qPCR experiments. Z.X. carried out Xenium experiments. H.V. conducted qPCR experiments. A.N. conducted website building and hosting on AWS cloud. J.W. performed Takara Bio Seeker data generation. V.W. led the CRC data production and writing (review). J.G.C. was responsible for writing (review and editing). A.J.M. and S.W. led the 10x Visium kidney cancer data production. I.G. was responsible for conceptualization, supervision and writing

(review and editing). Q.N. was responsible for conceptualization, data generation, supervision and writing (review and editing).

Competing interests

The authors declare no competing interests.

Additional information

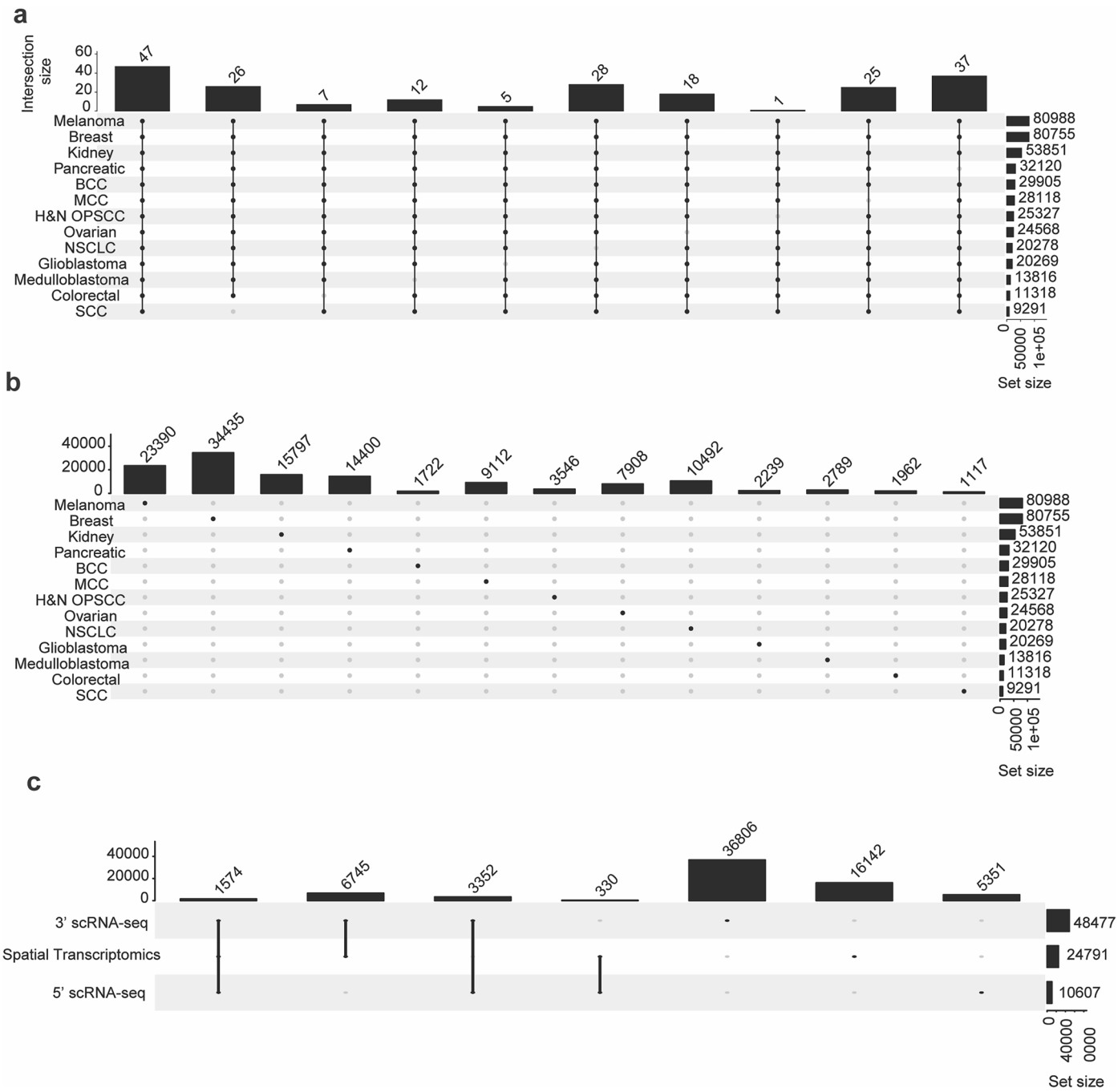
Extended data is available for this paper at <https://doi.org/10.1038/s41592-026-03071-4>.

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41592-026-03071-4>.

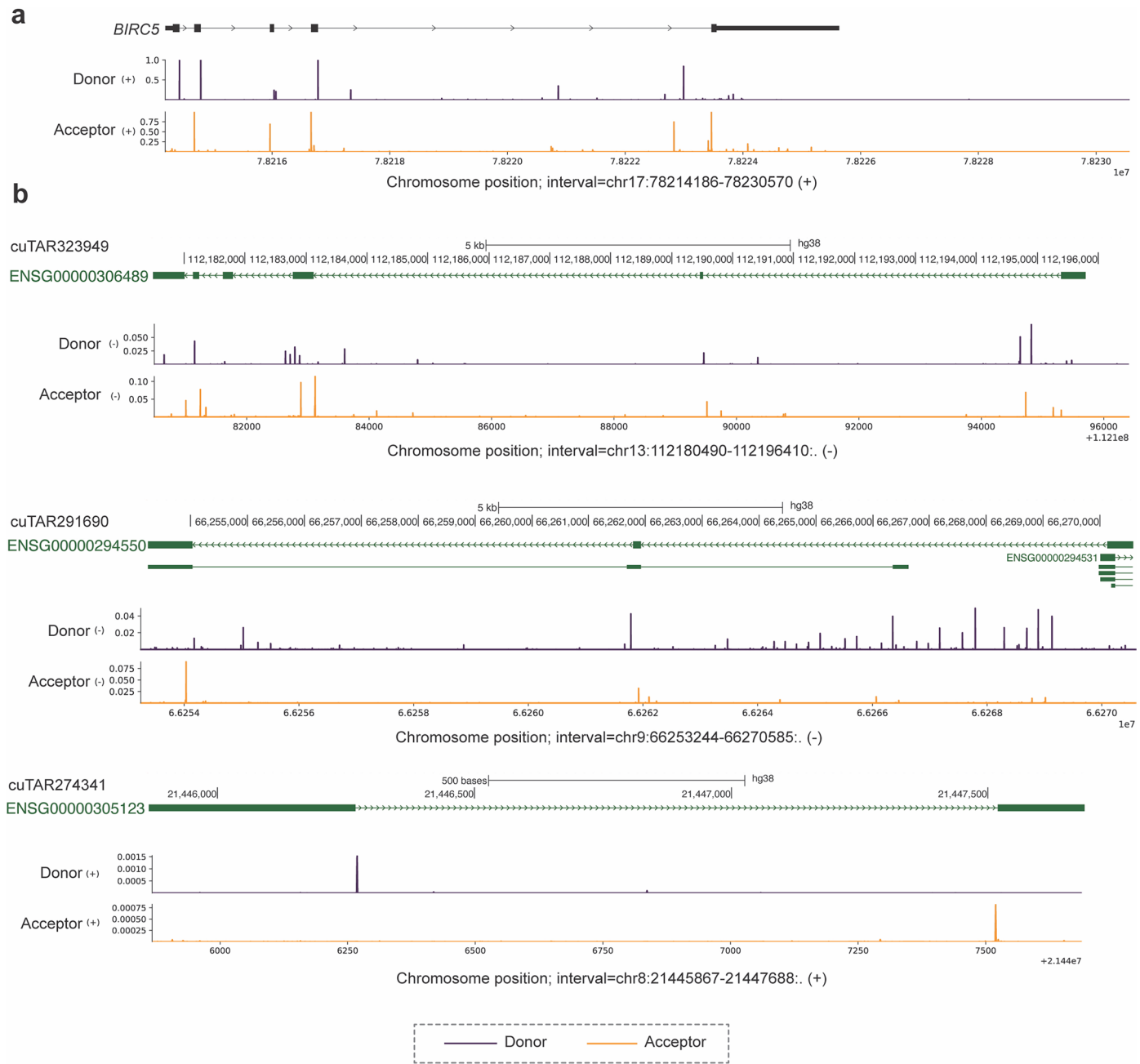
Correspondence and requests for materials should be addressed to Ishaan Gupta or Quan Nguyen.

Peer review information *Nature Methods* thanks RunSheng Chen, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available. Primary Handling Editor: Madhura Mukhopadhyay, in collaboration with the *Nature Methods* team.

Reprints and permissions information is available at www.nature.com/reprints.



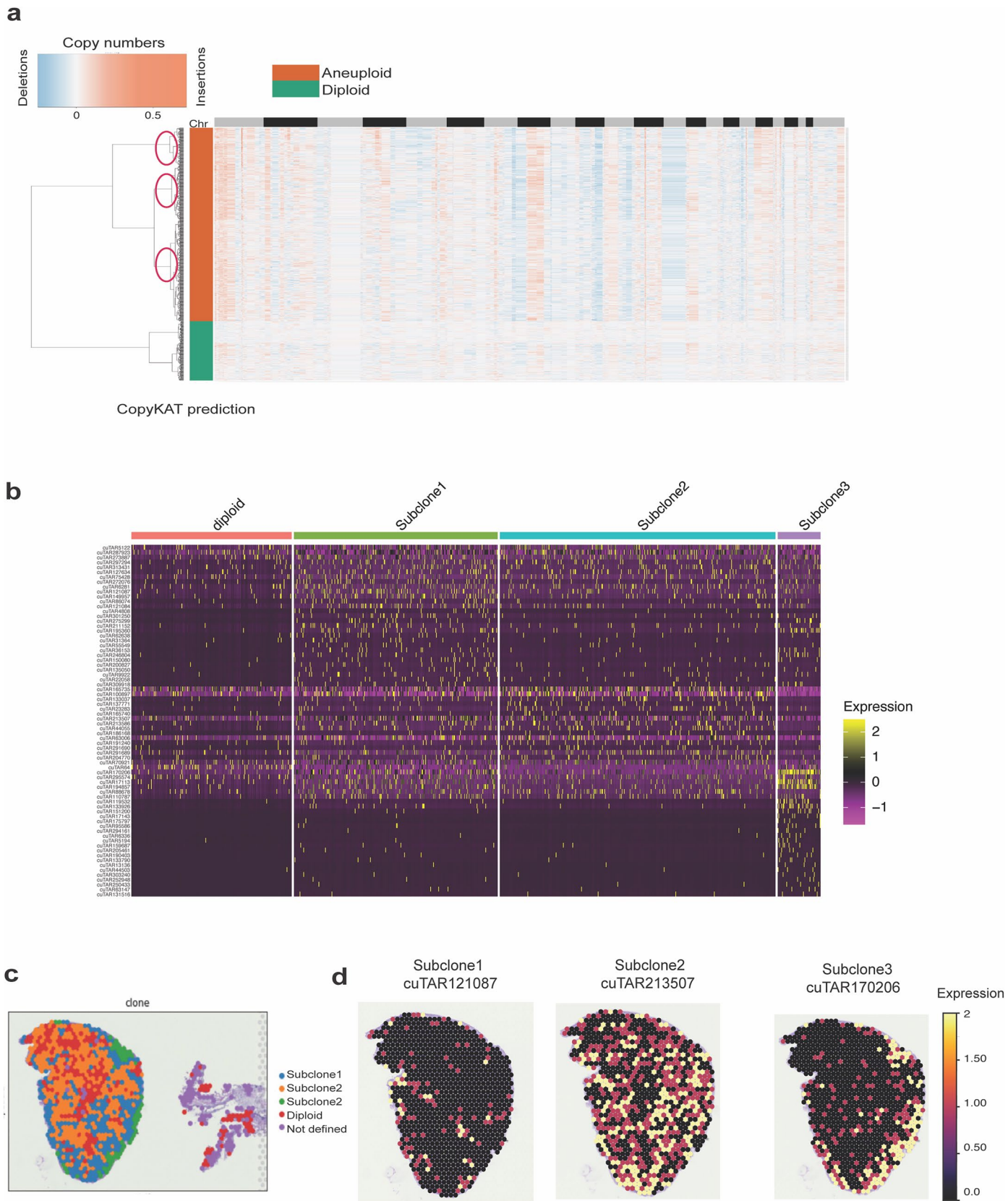
Extended Data Fig. 1 | cuTAR detection across cancer types. The upset plots indicate **a**, The cuTARs shared across 13 cancer types and in at least 12 cancers. **b**, cuTARs unique to each cancer type. **c**, cuTARs in melanoma detected across different sequencing platforms.



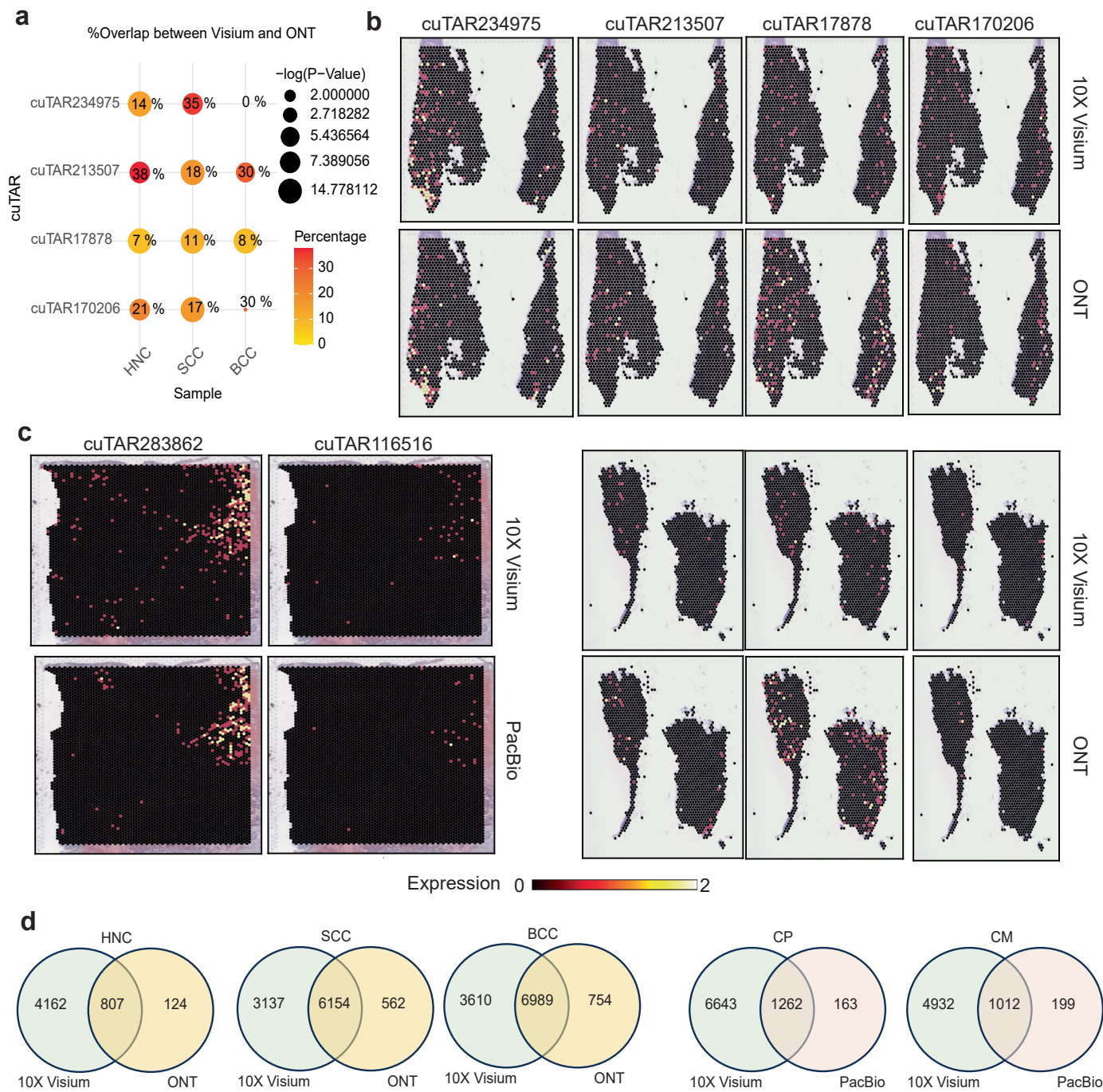
Extended Data Fig. 2 | Splice site prediction using Alphagenome.

a, Comparison of predicted donor and acceptor splice sites by alphagenome with the annotated transcript structure of the protein coding gene *BIRC5* and **b**, cuTARs identified using the TAR-scRNA-seq pipeline with their annotated

counterparts in the GENCODE v.47 that was released after they were identified as cuTARs. The corresponding cuTAR boundaries were extended to match the length of the annotated counterparts for splice site prediction and visualization with AlphaGenome to compare with annotated exonic regions.

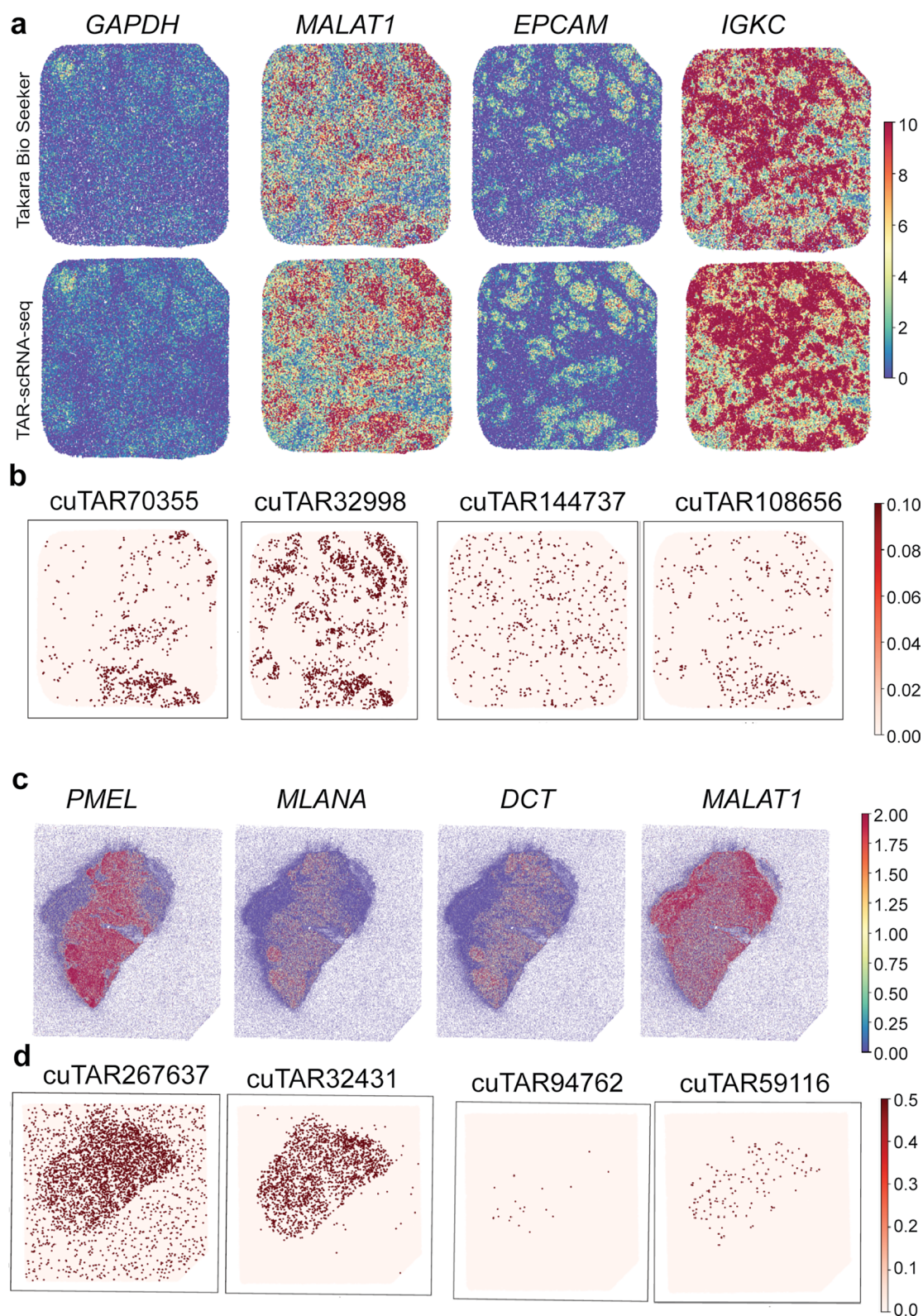


Extended Data Fig. 3 | Sub-clone specific lncRNAs in H&N OPSCC. a, Three major cancer subclones identified using copyKAT. **b**, DE cuTARs across subclones. **c**, Subclones annotated on the Head and Neck OPSCC tissue. **d**, Expression of Subclone-specific cuTARs overlaid on the tissue.



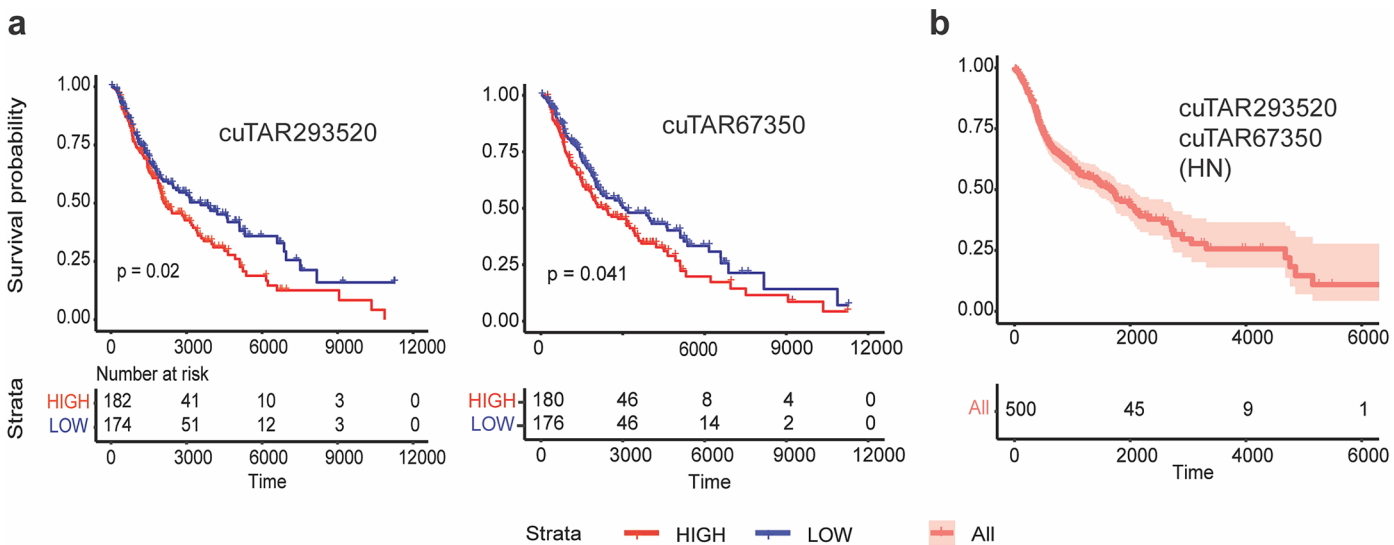
Extended Data Fig. 4 | Cross verification of cuTAR detection with long read sequencing. a, Proportion of spots expressing PCR validated cuTARs across 10x Visium and ONT. Fisher's exact test was used to assess whether ONT and Visium co-detect each cuTAR across spatial spots. For each cuTAR, a 2×2 presence/absence table was generated, and the percentage of spots with non-zero expression in both platforms is shown. Lower p-values indicate stronger co-detection. **b**, Spatial plots for these cuTARs in SCC (top) and BCC

(bottom) across the two platforms. **c**, Spatial plots for cuTARs in colorectal cancer metastasis sample (CM) across 10x Visium and PacBio. **d**, Venn diagrams demonstrating concordance of cuTAR detection between long-read sequencing and 10x Visium. HNC (Head & Neck OPSCC sample C), SCC, BCC samples were used for the ONT experiment and the colorectal cancer primary (CP) and metastasized sample were used for PacBio sequencing.

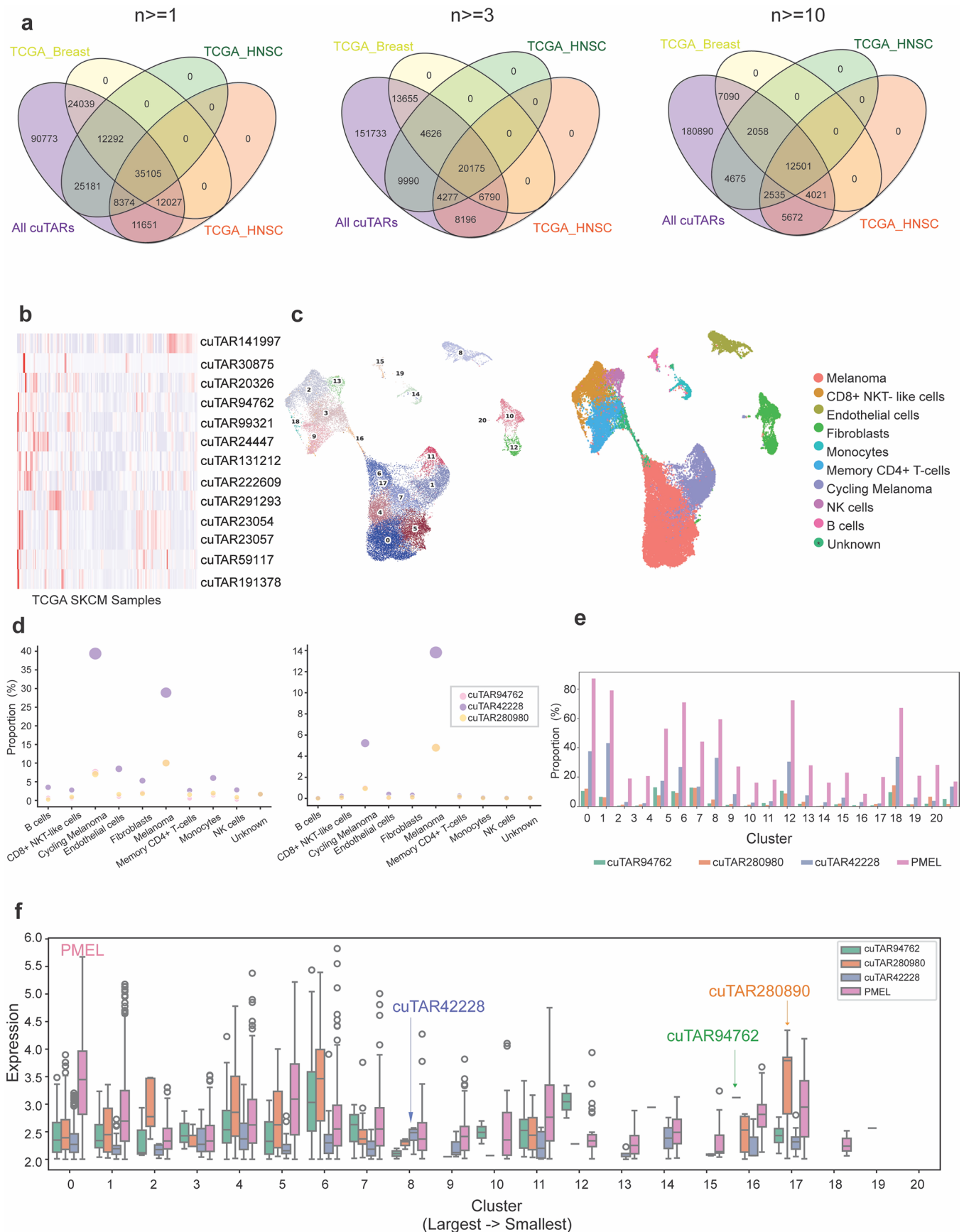


Extended Data Fig. 5 | Cross validation on Curio-Seeker, a spatial platform with single-cell resolution. a, Annotated genes and lncRNA counts obtained using the Curio-Seeker pipeline (top) and the TAR-scRNA-seq pipeline (bottom). **b**, Breast cancer specific cuTARs identified previously were verified on the curio breast cancer sample. **c**, Annotated gene counts for the melanoma sample

obtained using the TAR-scRNA-seq pipeline. **d**, Melanoma specific cuTARs identified previously were verified on the curio melanoma cancer sample. The scale bars represent expression levels of the genes and lncRNAs scaled to different maximum threshold values as displayed.



Extended Data Fig. 6 | Analysis of prognostic value of cuTARs. **a**, Kaplan-Meier survival curves for melanoma-specific cuTARs indicate that the high-expressing group has lower survival probabilities in the SKCM (Skin Cutaneous Melanoma) dataset of TCGA, while **b**, no stratification was observed for these cuTARs in the HN TCGA dataset.

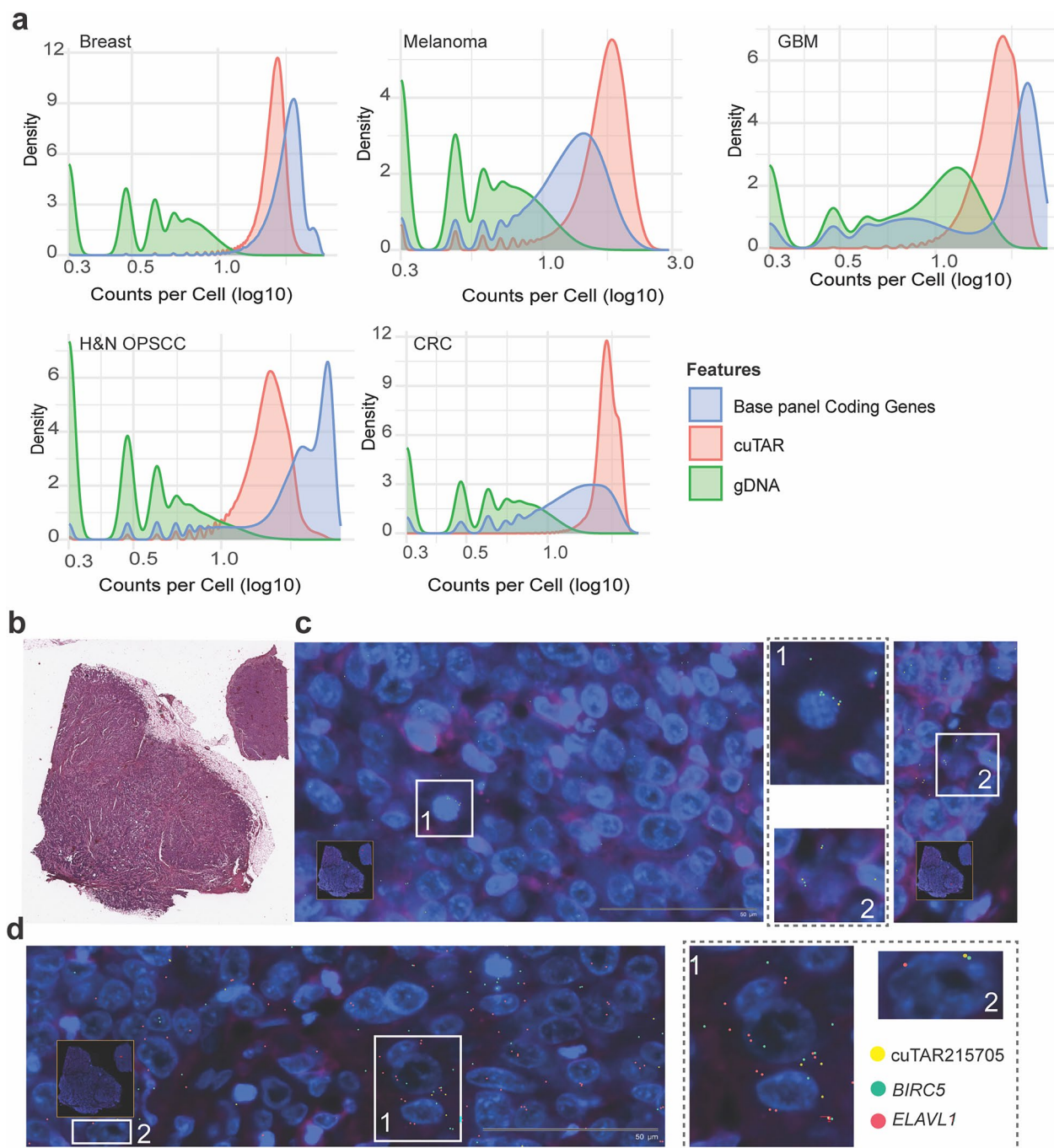


Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | Assessing lncRNA detection and expression specificity.

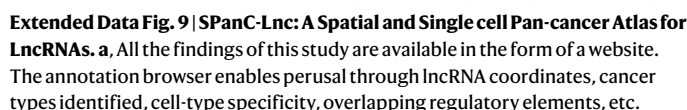
a, Number of cuTARs replicated in TCGA bulk RNA-Seq data in at least one, three and 10 samples. **b**, Expression of some Melanoma specific cuTARs in TCGA Melanoma samples. **c**, Seurat clusters (left) and cell type annotations for the scRNA-Seq Acral and Cutaneous melanoma dataset. **d**, (left) Proportion of cells in each cell type expressing three melanoma tumor-specific cuTARs in the scRNA-seq dataset analyzed. For example, cuTAR94762 is expressed in 7.6% of the Cycling Melanocyte cells. The right panel indicates the proportion of cuTAR expressing cells in the entire dataset. For example, 5.23% of all the cells in the entire dataset are cycling melanocytes expressing cuTAR42228 and 13.82% are

from other non-cycling melanocytes, with an overall representation of 20.6% cells expressing the cuTAR. Similarly, ~6.2% of all cells from the dataset express cuTAR94762 and cuTAR280980 respectively. **e**, Proportion of cells in each Seurat cluster expressing a coding gene and cuTARs. The cuTARs are expressed in very low levels in few cells. **f**, Distribution of expression values (>2) of the coding gene and the cuTARs. cuTARs are expressed in higher levels (>2) with higher median values in smaller clusters, while the coding gene has the highest median in the largest cluster. This suggests that cuTARs are expressed in very low levels in a low proportion in rarer cell clusters and hence more likely be diluted out in bulk RNA-seq data.



Extended Data Fig. 8 | Controls for the Xenium experiment and zoom-in view of Xenium data for coexpression analysis of breast cancer specific cuTAR and coding genes. a, Distribution of counts of the base panel coding genes, cuTARs and gDNA negative controls. **b**, H&E staining image of the breast cancer tissues used for the Xenium experiment **c**, co-expression of cuTAR215705 (yellow) and *BIRC5* (green) **d**, and the gene *ELAVL1* (red) which encodes the RBP HuR that

is predicted to interact with cuTAR215705 and known to interact with *BIRC5*. The three transcripts are represented by individual dots with solid colors and overlap of transcripts does not result in a merged color gradient. DAPI staining (blue) marks the nuclei and ATP1A1/CD45/E-Cadherin (pink) marks cytoplasmic regions. The images within the dotted boxes are zoomed-in views of the areas enclosed in the corresponding numbered white boxes.



b, Visualization of the lncRNA expressions on the spatial tissues across multiple cancer types from Visium and Long Read data. **c**, UCSC genome-style browsing of cuTARs and colocalizing genomic features. **d**, Prediction of splice sites using AlphaGenome.

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

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

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection

10X Visium Spatial transcriptomics data
bcl2fastq_version 2.20.0.422; spaceranger-1.3.0 was used for mapping spatial transcriptomics data of the human kidney cancer (Samples with Prefix Kid) and skin samples (SCC, BCC)
bcl2fastq_version 2.20.0.422; spaceranger-1.2.2 was used for mapping spatial transcriptomics data of the human Melanoma samples
bcl2fastq_version 2.20.0.422; spaceranger-2.0.1 was used for mapping spatial transcriptomics data of the human OPSCC samples
bcl2fastq_version 2.20.0.422; spaceranger-2.1.0 was used for mapping spatial transcriptomics data of the human Medulloblastoma PDOX mouse model samples and human colorectal cancer samples
bcl2fastq_version 2.20.0.422; spaceranger-1.1.0 for mapping spatial transcriptomics data of the human kidney cancer (Samples with Prefix Kid12)
bcl2fastq_version 2.20.0.422; spaceranger-1.2.2 for mapping spatial transcriptomics data of the human kidney cancer (Samples with Prefix Kid14)

The Oxford Nanopore long read sequencing of the Visium cDNA libraries was generated using the following

- SQK-LSK110 and EXP-NBD kits from ONT to generate the libraries. Sequence runs using super-accurate basecalling setting with MinKNOW version 22.12.5 and Guppy version 6.4.6. Flow cell type: FLO-PRO002 for the OPSCC sample.
- SQK-NBD-114.24 kit (Native barcoding kit 14) to prepare libraries for the SCC and BCC samples sequenced on PromethION P24 device using R10 flow cell (FLO-PRO114M), Sequence bases (raw data) were called using PromethION software release 23.11.4 (minknow-core-promethion 5.8.3, dorado version 7.2.13). Re-basecalling was done using dorado version 0.5.1

The PacBio long read sequencing of the Visium cDNA libraries for the colorectal cancer samples was generated using the following
MAS-Seq libraries were prepared using MAS-Seq for 10x Single Cell 3' kit (PN:102-678-600-REV03)

ccs: Version 7.0.0
 lima: Version 2.7.1
 primrose: Version 1.4.0
 samtools: Version 1.12

Data analysis

Several versions of same softwares or packages were used for different analysis based on compatibilities. The full list of softwares and version information for each conda environment used is provided as .yaml files at https://github.com/GenomicsMachineLearning/SPanc_Lnc_PanCancer_LncRNA_Atlas/tree/main/Manuscript_files/Softwares. Information of other stand-alone softwares used are as follows:

- PacBio Analysis - smrtlink_12.0.0.17705
- Coding potential analysis: CPAT 3.0.4
- hlpenseml: <https://github.com/dmitrymyl/hlpenseml.git>

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

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

The cuTAR count matrices generated as part of this study, the associated Visium gene expression and H&E images and processed objects of high dimensional data are available at UQ eSPACE <https://doi.org/10.48610/29c1345>. The raw data (BAM/FASTQ files) will be made available upon request and in agreement with the ethics policy.

- This study also uses raw data from previously published in-house datasets available to access as follows (also described in the manuscript)
 - Visium Melanoma: <https://zenodo.org/records/8053920>
 - Visium Medulloblastoma PDOX Mouse models: PRJEB53588 (ENA)
 - Visium SCC, BCC: PRJEB54964 (ENA)
 - Visium Breast Cancer (<https://www.swarbricklab.com/data-and-code-access>)
 - Public scRNA-Seq datasets: Acral and Cutaneous Melanoma from the study (PRJNA862451), PBMCs from Merkel cell carcinoma (MCC) (SRR7722937), head and neck cancer patients (SRR13418965, SRR13419137, SRR13419168) and 5'- sequencing datasets of tissue biopsies from patients with basal cell carcinoma (BCC). This included data before and after anti-PD-1 therapy from three patients (SRX5128480, SRX5128482, SRX5128486, SRX5128489, SRX5128506, SRX5128507).
 - Curio-Seeker data - Two samples (breast cancer and melanoma) were obtained from Curio/Takara Bio.
- scRNA-seq dataets from 10X Genomics site
 - Breast IDC <https://www.10xgenomics.com/resources/datasets/7-5-k-sorted-cells-from-human-invasive-ductal-carcinoma-3-v-3-1-3-1-standard-6-0-0>
 - Glioblastoma <https://www.10xgenomics.com/resources/datasets/human-glioblastoma-multiforme-3-v-3-whole-transcriptome-analysis-3-standard-4-0-0>
 - Kidney Cancer (5') <https://www.10xgenomics.com/resources/datasets/5k-human-kidney-tumor-isolated-with-chromium-nuclei-isolation-kit-3-1-standard>
 - Lung (NSCLC) <https://www.10xgenomics.com/resources/datasets/nslc-tumor-5-gene-expression-1-standard-2-2-0>
 - Pancreatic cancer <https://www.10xgenomics.com/resources/datasets/5k-human-pancreatic-tumor-isolated-with-chromium-nuclei-isolation-kit-3-1-standard>
 - Ovarian cancer <https://www.10xgenomics.com/resources/datasets/5k-human-ovarian-tumor-isolated-with-chromium-nuclei-isolation-kit-3-1-standard>

Human research participants

Policy information about [studies involving human research participants and Sex and Gender in Research](#).

Reporting on sex and gender

Sex or gender were not considered or reported in any of the methods used in this study.

Population characteristics

The tissues were collected from patients diagnosed with different types of cancers.

Recruitment

- Skin biopsy samples from patients diagnosed with basal cell and squamous carcinoma (BCC and SCC) were collected at the Dermatology Department of the Princess Alexandra Hospital in 2019.
- The fresh frozen Kidney cancer tissue sections collected in the operating theatre from six patients with ccRCC presenting at a tertiary referral hospital from June 2021 to January 2022 (Raghubar et al. 2023).
- The primary colorectal cancer tumours and associated liver metastasis samples were collected from three patients during surgery for CRC carcinoma (Kawamata et al. 2018).
- The breast cancer samples were collected with approval from all relevant human research ethics committees (Sydney Local Health District Ethics Committee, Royal Prince Alfred Hospital zone and the St Vincent's Hospital Ethics Committee) (Wu et al. 2021).
- The Head & Neck OPSCC samples were collected from the Ear, Nose, Throat (ENT) and oncology department at Princess Alexandra Hospital in 2021.
- The FFPE Melanoma biopsies were from archived tissues collected between 2008 and 2018 (Vo et al. 2025.)
- The Medulloblastoma paediatric patient tumour tissues were obtained from Seattle Children's Hospital through the Children's Oncology Group brain tumour biology study ACNS02B3 (NCT00919750). The ACNS02B3 study was approved from the Institutional Review Board (Vo et al. Genome Medicine, 2023).

FFPE samples used for the Xenium in-situ/STOmics experiments

- The breast cancer samples were diagnosed within Pathology Queensland and patients were consented through the Brisbane Breast Bank between 2020-2023.
 - The HPV+ Head & Neck OPSCC FFPE samples were collected from the Ear, Nose, Throat (ENT) and oncology department at Princess Alexandra Hospital in 2021 (approved by Metro South Research Ethics 2019/HE001021).
 - The colorectal cancer samples were collected as part of a previous study by Fennel et al. 2019 between 2008-2012. Written informed consent was obtained from each patient. The study protocol was approved by the Royal Brisbane and Women's Hospital and QIMR Berghofer Medical Research Institute Research Ethics Committees.
 - Patient glioblastoma tumor tissue was collected from patients undergoing surgery at the Royal Brisbane and Women's Hospital (RBWH), with informed consent and human ethics approval from the QIMR Berghofer Medical Research Institute and RBWH human research ethics committees.
- No bias was aware that may impact results.

Ethics oversight

All samples were approved for research under ethical approval numbers 2018000165 and 2017000318 by the University of Queensland's Human Research Ethics Committees.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

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

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

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Life sciences study design

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

Sample size

In this study, we used several independent tissues/ single-cell samples either collected and sequenced in-house or from previously published studies. Our analyses focused on finding spatial patterns within a tissue, which gained statistical power from the large number of spots/cells. From one transcriptomics tissue section, hundreds to thousands of spatial spots are measured. The number of spots in each cell type can be considered as the number of technical replicates of the same cell type within a tissue. Using these replicates, we can account for technical variation, such as sequencing depth and PCR-amplification bias when comparing clusters within the same tissue section. Using simulation data of single cell and spatial data, we had previously calculated that to find a rare population with a frequency of 1% of a sample of spots with 95% detection power we needed 628 spots. All datasets we used had more than >1000 spots, as described above.

Data exclusions

For spatial transcriptomics downstream analysis, we applied a quality control step for excluding lowly expressed uTARs for all samples (features detected in fewer than 3 spots). We did not apply the minimum features per spot filter for the uTAR data, since not all spots were assumed or expected to have more numbers of lncRNAs. For uTAR discovery, we retain those with at least 3 reads mapping to the genome irrespective of spot/cellular level expression.

Replication

For selecting top uTAR candidates we chose the ones that were consistently detected across different samples of the same cancer type or across different cancer types or across sequencing technologies. For PCR validation of selected uTAR regions we also used three technical replicates. 76 uTARs were chosen for validation on Xenium where at least two replicates per cancer type were used for the experiment.

Randomization

We thoroughly assessed false discovery as well as the robustness of null distribution (for example in the case of eQTL co-localization analysis, we performed three sets of random sampling of equal number of genomic regions as the number of uTARs to randomize genomic regions and their co-localization with eQTLs). Our randomization strategy involves random shuffling and pseudobulking of cells for scRNA-Seq DE analysis and random shuffling and choosing genomic regions.

Blinding

All analyses done in this study were data-driven and unsupervised. For example, we applied unsupervised clustering to identify cell types, with no prior information about the expected number of cell types present in the dataset. Only the uTAR-RBP interaction analysis used a pre-trained ML model to predict interactions.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

Methods

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☐	☒ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Animals and other research organisms

Policy information about [studies involving animals](#); [ARRIVE guidelines](#) recommended for reporting animal research, and [Sex and Gender in Research](#)

Laboratory animals

Seven to 9-week-old male NOD.Cg-Prkdcscid Il2rgtm1Wjl/SzJ (NSG) mice (Jackson Laboratory, Bar Harbor, ME) were maintained in a barrier facility on a 12-h light/dark cycle with food and water provided ad libitum, in accordance with the NIH Guide for the Care and Use of Experimental Animals (Vo et al. Genome Medicine, 2023).

Wild animals

The study did not involve wild animals.

Reporting on sex

Male mice were used

Field-collected samples

The study did not involve samples collected from the field.

Ethics oversight

All PDOX mouse experiments were performed with approval from The University of Queensland Molecular Biosciences animal ethics committee (IMB/386/18) and The University of Queensland Institutional Human Research Ethics committee (2015001410) and conducted accordance with the Australian Code for the Care and Use of Animals for Scientific Purposes (Vo et al. Genome Medicine, 2023).

Note that full information on the approval of the study protocol must also be provided in the manuscript.