INDIAN INSTITUTE OF TECHNOLOGY, DELHI

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

DISCOVERING NOVEL CELL-TYPE SPECIFIC NON-CODING RNA AND DISSECTING THEIR ROLE IN CANCER

PRAKRITHI PAVITHRA

INDIAN INSTITUTE OF TECHNOLOGY DELHI

&

THE UNIVERSITY OF QUEENSLAND

JANUARY 2026

INDIAN INSTITUTE OF TECHNOLOGY, DELHI

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

Discovering Novel Cell-Type Specific Non-Coding RNA and Dissecting their Role in Cancer

by

Prakrithi Pavithra

B.Tech. (Biotechnology)

Submitted

in fulfilment of the requirements for the joint degree of

DOCTOR OF PHILOSOPHY

to the

INDIAN INSTITUTE OF TECHNOLOGY DELHI

&

THE UNIVERSITY OF QUEENSLAND

January 2026

This thesis is dedicated to all cancer victims and survivors

Supervisor Certification

This is to certify that the thesis entitled "**Discovering novel cell-type specific non-coding RNA and dissecting their role in cancer**" being submitted by **Ms. Prakrithi Pavithra** to the Indian Institute of Technology Delhi and The University of Queensland for the award of degree of **Doctor of Philosophy** is a record of *bonafide* research work carried out by her. **Ms. Prakrithi Pavithra** has worked under our guidance and supervision and has fulfilled the requirements for the submission of this thesis, which to our knowledge has reached the requisite standard. The results contained in this thesis are original and have not been submitted, in part or full, to any other University or Institute for the award of any other degree or diploma.

Dr. Ishaan Gupta

Associate Professor,

Department of Biochemical

Engineering and Biotechnology,

Indian Institute of Technology Delhi,

Hauz Khas New Delhi - 110016.

Dr. Quan Nguyen

Associate Professor,

The University of Queensland,

St. Lucia, Australia - 4072.

Abstract

Long non-coding RNAs (lncRNAs), which constitute the majority of the human transcriptome, play pivotal roles in gene regulation and disease mechanisms, particularly in cancer. Despite their biological importance, the understanding of their cell-type and tissue-specific functions remains limited, largely because most available data are from bulk RNA sequencing. Given that lncRNAs are often lowly expressed and highly cell-type specific, their characterisation at single-cell (SC) resolution remains underexplored. This PhD project aims to identify and analyse lncRNAs within a spatial and cell-type context, resulting in the creation of a novel and comprehensive lncRNA pan-cancer atlas. We hypothesised that by integrating high-throughput single-cell RNA sequencing (scRNA-seq) and spatial transcriptomics (ST) data, we could not only detect the majority of known lncRNAs but also uncover novel ones and predict their functions in specific cells and tissues. The overall objective is to build a pan-cancer, single-cell and spatial lncRNA resource that can help uncover the roles of lncRNAs in diagnosis, prognosis, and therapeutic response in cancer.

This thesis comprises three aims: (1) Implementing a bioinformatic pipeline to detect lncRNAs from single-cell and spatial transcriptomics data using 10x Genomics' capture protocols and assessing how off-target measurements of lncRNA can affect gene expression interpretation in probe-based capture methods; (2) Computational and experimental validation of unannotated lncRNAs and new approaches to perform cell-type and location-specific analyses of their potential functions, making this pan-cancer resource of lncRNAs available in the form of an AWS cloud server based website; and (3) Integrating ST with genetic data to comprehensively study lncRNA roles at RNA and DNA level. This aim includes CNV inference to identify malignant regions within tumours, investigating whether novel lncRNAs improve predictive accuracy for malignancy. Additionally, by leveraging genome-wide association studies (GWAS), this aim shows our pioneering research on mapping lncRNAs associated with cancer-relevant traits on specific spatial domains and cell types within a tumour.

During the first year, I applied the TAR-scRNA-seq bioinformatics pipeline to both scRNA-seq and ST datasets to detect lncRNAs in fresh frozen and FFPE cancer tissues (Aim 1). Unexpectedly, we observed that a few novel lncRNAs were detected in 10x Visium datasets using probe-based capture. This led to further analysis of these off-target signals and their impact on interpretation of coding gene expression. The results highlight the importance of

designing ST probes based not only on coding genes but also the entire transcriptome, including lncRNAs, to avoid confounding effects.

In the second year (Aim 2), I applied the same pipeline to fresh frozen tissues from 13 cancer types. A combination of bioinformatics tools was used to assess coding potential, sequence conservation, and overlap with public lncRNA databases like FANTOM-CAT, LncBook, NONCODE and TCGA. I developed new methods to infer potential lncRNA function using spatial autocorrelation, co-expression module networks, eQTL/GWAS SNP enrichment, survival analysis using TCGA bulk data, and differential expression analyses for cell-type and tumour-region specificity, including pre/post-treatment comparisons. For experimental validation, we used qPCR and long-read sequencing technologies like PacBio Iso-Seq and Oxford Nanopore (ONT) to capture longer transcripts from Visium libraries. Since Visium data lacks single-cell resolution, we further analysed datasets from Curio Seeker, a newer spatial method that captures polyadenylated transcripts at single-cell resolution. Using breast cancer and melanoma datasets, we confirmed co-expression of novel lncRNAs with known coding genes in specific cell types and regions. These analyses led to the identification of top candidate functional lncRNAs, which were validated using a custom panel including 76 lncRNAs on the Xenium platform, confirming tissue-specific and cancer-specific expression patterns. For broader usability of our findings, we developed an interactive AWS-hosted database, available at https://spanclnc.click/. This is the largest single-cell and spatial lncRNA resource across multiple cancers, providing annotations, visualisation, and querying tools.

The final aim in this thesis is to further explore the roles of DNA variants in the context of lncRNA expression in cancer, as most of the significant trait-associated genomic variants lie outside the coding regions. This work involved the integration of genetic level information like CNVs and GWAS summary statistics with SC and ST to identify malignant cells/spots and mapped genetic association signals to cells/spots/lncRNAs. First, I evaluated different tools for CNV inference from scRNA-seq/ST data and selected the best one to predict malignancy. Next, I incorporated our unannotated lncRNAs to evaluate if they improve the malignancy prediction and identified those that contribute to malignancy. Finally, I applied an emerging method to map genetic association signals to cells/spots/lncRNAs. This approach leverages GWAS summary statistics to uncover functional insights into lncRNAs by linking them to nearby trait-associated SNPs, enhancing our understanding of their roles.

In summary, this PhD work establishes a robust, scalable framework to detect, validate, and functionally characterise lncRNAs in cancer using single-cell and spatial transcriptomics. The resulting resource and methods significantly advance our ability to study lncRNAs at high resolution and provide a new perspective on cancer biology.

संक्षेप (Abstract in Hindi)

यह पीएचडी शोध कार्य लॉन्ग नॉन-कोडिंग आरएनए (lncRNAs) पर केंद्रित है, जो मानव ट्रांसक्रिप्टोम का एक बड़ा हिस्सा बनाते हैं और जीन विनियमन तथा रोग तंत्रों, विशेष रूप से कैंसर, में महत्वपूर्ण भूमिका निभाते हैं। जैविक रूप से महत्वपूर्ण होने के बावजूद, lncRNAs की कोशिका-प्रकार और ऊतक-विशिष्ट भूमिकाओं की समझ अब भी सीमित है, क्योंकि अधिकांश उपलब्ध डेटा बल्क RNA सीक्वेंसिंग पर आधारित हैं। चूंकि lncRNAs सामान्यतः कम मात्रा में अभिव्यक्त होते हैं और अत्यधिक कोशिका-विशिष्ट होते हैं, इसलिए उनका एकल-कोशिका (single-cell) स्तर पर विश्लेषण अब भी कम शोधित है।
इस पीएचडी परियोजना का उद्देश्य lncRNAs की पहचान करना और उनका विश्लेषण स्थानिक (spatial) और कोशिका-प्रकार के संदर्भ में करना है, जिससे एक नया और व्यापक पैन-कैंसर lncRNA एटलस विकसित किया जा सके। हमारा अनुमान था कि उच्च-गुणवत्ता वाले एकल-कोशिका RNA अनुक्रमण (scRNA-seq) और स्थानिक ट्रांसक्रिप्टोमिक्स (spatial transcriptomics - ST) डेटा को एकीकृत करके, न केवल अधिकांश ज्ञात lncRNAs को पहचाना जा सकता है, बल्कि नए lncRNAs की खोज और विशिष्ट कोशिकाओं/ऊतकों में उनके संभावित कार्यों का भी अनुमान लगाया जा सकता है। इस शोध का मुख्य उद्देश्य एक ऐसा पैन-कैंसर lncRNA संसाधन बनाना है जो निदान, रोग-पूर्वानुमान और उपचार प्रतिक्रिया में lncRNAs की भूमिका को समझने में सहायता कर सके।

यह शोध तीन मुख्य उद्देश्यों में विभाजित है: (1) एक बायोइन्फॉर्मेटिक पाइपलाइन को लागू करना जो single-cell और spatial transcriptomics डेटा से 10x Genomics के कैप्चर प्रोटोकॉल का उपयोग कर lncRNAs का पता लगाए, और यह विश्लेषण करना कि प्रोब-आधारित तरीकों में lncRNA की off-target माप जीन अभिव्यक्ति की व्याख्या को कैसे प्रभावित कर सकती है; (2) अप्रकाशित lncRNAs का कम्प्यूटेशनल और प्रयोगात्मक सत्यापन करना, और उनके संभावित कार्यों का कोशिका-प्रकार और स्थानिक स्तर पर विश्लेषण करने के लिए नई विधियाँ विकसित करना; साथ ही इन लिंकरों पर आधारित पैन-कैंसर डेटाबेस को AWS क्लाउड पर होस्ट की गई वेबसाइट के रूप में उपलब्ध कराना; और (3) ST को आनुवंशिक डेटा के साथ एकीकृत करना, जिससे RNA और DNA स्तर पर lncRNAs की भूमिकाओं का व्यापक अध्ययन किया जा सके। इस उद्देश्य में CNV विश्लेषण शामिल है ताकि ट्यूमर के कैंसरयुक्त क्षेत्रों की पहचान की जा सके और यह जांचा जा सके कि क्या नए lncRNAs कैंसर की भविष्यवाणी की सटीकता को बेहतर बनाते हैं। इसके अतिरिक्त,

GWAS डेटा का उपयोग करते हुए यह अध्ययन यह दिखाता है कि कैसे विशिष्ट स्थानिक डोमेन और कोशिका-प्रकारों में कैंसर-संबंधी गुणों से जुड़े lncRNAs की मैपिंग की जा सकती है।

प्रथम वर्ष में (उद्देश्य1), मैंने TAR-scRNA-seq पाइपलाइन को कई scRNA-seq और ST डेटासेट्स पर लागू किया ताकि ताजे फ्रोजन और FFPE कैंसर ऊतकों में lncRNAs की पहचान की जा सके। आश्चर्यजनक रूप से, हमने देखा कि प्रोब-आधारित कैप्चर द्वारा 10x Visium डेटा में कुछ नए lncRNAs भी पहचाने गए। इसके परिणामस्वरूप इन off-target संकेतों और उनके कोडिंग जीन अभिव्यक्ति पर प्रभाव का गहन विश्लेषण किया गया। यह स्पष्ट हुआ कि ST प्रोब्स को केवल कोडिंग जीन पर आधारित न रखकर पूरे ट्रांसक्रिप्टोम, विशेषकर lncRNAs, को ध्यान में रखकर डिज़ाइन किया जाना चाहिए ताकि ग़लत व्याख्या से बचा जा सके।

दूसरे वर्ष (उद्देश्य 2) में, मैंने इसी पाइपलाइन को 13 अलग-अलग कैंसर प्रकारों के ताजे फ्रोजन ऊतकों पर लागू किया। कोडिंग क्षमता, अनुक्रम संरक्षण, और FANTOM-CAT, LncBook, NONCODE, TCGA जैसे सार्वजनिक lncRNA डेटाबेस के साथ ओवरलैप का मूल्यांकन करने के लिए विभिन्न बायोइन्फॉर्मेटिक टूल्स का उपयोग किया गया। संभावित कार्यों की भविष्यवाणी के लिए स्थानिक स्व-सहसंबंध (spatial autocorrelation), सह-अभिव्यक्ति (co-expression) नेटवर्क, eQTL/GWAS SNP enrichment, TCGA डेटा का उपयोग करके सर्वाइवल विश्लेषण, और कोशिका-प्रकार व ट्यूमर-क्षेत्र विशिष्ट विभेदक अभिव्यक्ति विश्लेषण किए गए — जिनमें उपचार से पहले और बाद की तुलना भी शामिल थी। प्रयोगात्मक सत्यापन के लिए, हमने qPCR और PacBio Iso-Seq तथा Oxford Nanopore (ONT) जैसी लंबी-पढ़ाई तकनीकों का उपयोग किया ताकि Visium लाइब्रेरी से पूर्ण लंबाई वाले ट्रांसक्रिप्ट प्राप्त किए जा सकें। चूंकि Visium डेटा में single-cell रिज़ॉल्यूशन नहीं होता, इसलिए हमने Curio Seeker जैसी नई spatial तकनीकों से प्राप्त डेटा का विश्लेषण किया, जो polyadenylated ट्रांसक्रिप्ट को single-cell स्तर पर पकड़ सकती है। स्तन कैंसर और मेलेनोमा डेटासेट्स का उपयोग करके, हमने विशिष्ट कोशिकाओं और क्षेत्रों में नए lncRNAs की ज्ञात कोडिंग जीनों के साथ सह-अभिव्यक्ति की पुष्टि की। इन विश्लेषणों से शीर्ष कार्यात्मक lncRNAs की पहचान हुई, जिन्हें Xenium प्लेटफॉर्म पर 76 lncRNAs वाले कस्टम पैनल के ज़रिए सत्यापित किया गया। इससे उनके ऊतक-विशिष्ट और कैंसर-विशिष्ट अभिव्यक्ति पैटर्न की पुष्टि हुई। हमारे निष्कर्षों की व्यापक उपयोगिता सुनिश्चित करने के लिए, हमने एक इंटरेक्टिव AWS-होस्टेड डेटाबेस

विकसित किया, जो https://spanclnc.click/ पर उपलब्ध है। यह अब तक का सबसे बड़ा एकल-कोशिका और स्थानिक lncRNA संसाधन है, जो एनोटेशन, विज़ुअलाइज़ेशन और क्वेरी टूल्स प्रदान करता है।

थीसिस का अंतिम उद्देश्य कैंसर में lncRNA अभिव्यक्ति के संदर्भ में DNA वेरिएंट्स की भूमिका का अन्वेषण करना है, क्योंकि अधिकांश रोग-संबंधी एकल न्यूक्लियोटाइड बहुरूपताओं (SNPs) कोडिंग क्षेत्रों के बाहर स्थित होते हैं। इस अध्ययन में एकल-कोशिका और स्थानिक ट्रांसक्रिप्टोमिक्स डेटा को प्रतिलिपि संख्या में परिवर्तन (CNVs) और GWAS सारांश आँकड़ों के साथ एकीकृत किया गया ताकि कैंसरयुक्त कोशिकाओं/स्पॉट्स की पहचान की जा सके और आनुवंशिक संकेतों को lncRNAs और स्थानिक स्तर पर मैप किया जा सके। सबसे पहले, मैंने CNV अनुमान लगाने के लिए विभिन्न टूल्स का मूल्यांकन किया और ट्यूमर की पहचान में सर्वोत्तम प्रदर्शन वाले टूल का चयन किया। इसके बाद, मैंने नए lncRNAs को जोड़ा और यह देखा कि वे कैंसर की भविष्यवाणी में सुधार कर सकते हैं या नहीं। अंत में, मैंने GWAS आँकड़ों को कोशिकाओं/स्पॉट्स/lncRNAs से जोड़ने के लिए एक नई विधि अपनाई, जिससे SNPs के पास स्थित lncRNAs की कार्यात्मक भूमिका के बारे में जानकारी प्राप्त हुई।

सारांशतः, यह पीएचडी कार्य एक मजबूत और स्केलेबल ढांचा प्रस्तुत करता है जो एकल-कोशिका और स्थानिक ट्रांसक्रिप्टोमिक्स के माध्यम से lncRNAs की पहचान, सत्यापन और कार्यात्मक विशेषताओं का अध्ययन करने की सुविधा देता है। इस शोध में विकसित संसाधन और पद्धतियाँ lncRNAs के अध्ययन को उच्च रिज़ॉल्यूशन पर ले जाती हैं और कैंसर जीवविज्ञान को समझने का एक नया दृष्टिकोण प्रदान करती हैं।

Declaration by author

This thesis is composed of my original work, and contains no material previously published or written by another person except where due reference has been made in the text. I have clearly stated the contribution by others to jointly authored works that I have included in my thesis.

I have clearly stated the contribution of others to my thesis as a whole, including statistical assistance, survey design, data analysis, significant technical procedures, professional editorial advice, financial support and any other original research work used or reported in my thesis. The content of my thesis is the result of work I have carried out since the commencement of my higher degree by research candidature and does not include a substantial part of work that has been submitted to qualify for the award of any other degree or diploma in any university or other tertiary institution. I have clearly stated which parts of my thesis, if any, have been submitted to qualify for another award.

I acknowledge that an electronic copy of my thesis must be lodged with the University Library and, subject to the policy and procedures of The University of Queensland, the thesis be made available for research and study in accordance with the Copyright Act 1968 unless a period of embargo has been approved by the Dean of the Graduate School.

I acknowledge that copyright of all material contained in my thesis resides with the copyright holder(s) of that material. Where appropriate I have obtained copyright permission from the copyright holder to reproduce material in this thesis and have sought permission from co-authors for any jointly authored works included in the thesis.

Prakrithi Pavithra

Publications included in this thesis

1. Vo T, **Prakrithi P**, Jones K, Yoon S, Lam PY, Kao YC, Ma N, Tan SX, Jin X, Zhou C, Crawford J, Walters S, Gupta I, Soyer PH, Khosrotehrani K, Stark MS, Nguyen Q. Assessing spatial sequencing and imaging approaches to capture the molecular and pathological heterogeneity of archived cancer tissues. *J Pathol*. 2025 Jan 23. doi: 10.1002/path.6383. PMID: 39846232.

2. **P Prakrithi**, Juwayria , Deepali Jain, Prabhat S Malik, Ishaan Gupta, Caution towards spurious off-target signal in 10x Visium spatial transcriptomics assay from potential lncRNAs, *Briefings in Bioinformatics*, Volume 24, Issue 2, March 2023, bbad031, https://doi.org/10.1093/bib/bbad031.

3. **P. Prakrithi**, Tuan Vo, Hani Vu, Albert Xiong, Loan Nguyen, Andrew Newman, Vicki Whitehall, Jazmina L. Gonzalez Cruz, Ishaan Gupta, Quan Nguyen. Unraveling lncRNA Diversity at a Single Cell Resolution and in a Spatial Context across Different Cancer Types, bioRxiv 2024.08.12.607523; doi: https://doi.org/10.1101/2024.08.12.607523. (Accepted in *Nature Methods*; Impact factor: 32.1 (2024); 5-year Impact Factor: 51.7).

Submitted manuscripts included in this thesis

1. **P. Prakrithi,** Laura Grice, Feng Zhang, et al., Integrating 12 Single-Cell and Spatial Technologies to Characterise Tumour Neighbourhoods and Cellular Interactions in three Skin Cancer Types , bioRxiv 2025.07.25.666708; doi: https://doi.org/10.1101/2025.07.25.666708. (Manuscript under-review in *Nature Cancer*; Impact Factor: 28.5)

Other publications during candidature

1. Sivakumaran H, Nair N., Bitar M., Lu X., Wang L., Liu J., Karunarathne D., **Prakrithi P**., Jacquelin S., Rivera I., Hillman K., Kaufmann S., Ziegman R., Shi W., Veedu R., Nguyen Q., Beesley J., Wykes M.,French J., Edwards S. "BRRIAR lncRNA alters breast cancer risk in by modulating interferon signaling in cis and in trans through BHLHE40 and RIG-I". Mol Cancer. 2026 Jan 7;25(1):5. doi: 10.1186/s12943-025-02510-8. (Impact Factor: 33.9).

2. Wang, Y., Ramnath, D., Das Gupta, K., Miller, G. C., **Prakrithi, P.,** Xiong, Z., Wan, Y., Tejo, E. N., Curson, J. E. B., Abrol, R., Keshvari, S., Gunther, K. S., Loh, Z., Engel, J. A., Engwerda, C. R., Burgener, S. S., Schroder, K., Fairlie, D. P., Clouston, A. D., Powell, E. E., Irvine, K. M., Sullivan, M. A., Nguyen, Q., Sweet, M. J., & Karunakaran, D. (2025). Myeloid HDAC7 drives liver inflammation and systemic glucose dysregulation during diet-induced obesity. bioRxiv 2025. https://doi.org/10.64898/2025.12.04.691405.

Conference Abstracts and Awards

1. Selected for a lightning talk at the Genome Informatics Conference at the Wellcome Sanger Institute, Hinxton, UK in September 2022.
   Abstract title: "Spurious off-target signals from potential lncRNAs by 10X Visium probes"
2. Winner of 'Best Student Talk' (3rd Prize) at the Oz Single Cell Conference, Melbourne, Australia in October 2023.
3. Invited/nominated speaker at the Brisbane Cancer Conference, December 2023.
   Abstract title for 2 and 3: "Unravelling LncRNA Diversity at a Single Cell Resolution and in a Spatial Context Across Different Cancer Types"
4. Winner of a travel grant worth 400€ to present a poster at the Spatial Omics Conference, Ghent, Belgium in June 2024.
5. Selected for a talk at the Keystone Symposia: Single Cell Biology: Unique Cells to Tissue Ecosystems, Whistler, Canada, May 2025.
   Abstract title for 4 and 5: "SPanC-Lnc: Decoding the Spatial and Single Cell Landscape of Long Non-coding RNAs at a Pan-Cancer Level"
6. The Prime Minister's Research Fellowship (PMRF), (December 2022- June 2025).

Contributions by others to the thesis

All chapters

Editorial suggestions were provided by my supervisors, Prof. Quan Nguyen, Prof. Ishaan Gupta and PhD committee, Prof. Mikael Boden, Prof. Sundar D. and Prof. Shilpi Minocha. My supervisors helped conceive the ideas throughout the thesis.

Chapter 2

Dr. Tuan Vo was involved in generating the FFPE sequencing data.

Chapter 3

Dr. Tuan Vo and Dr. Zherui Xiong conducted experiments to generate the in-house spatial data from fresh frozen and FFPE tissues. Dr. Tuan Vo and Ms. Hani Vu assisted in conducting the qPCR experiment. Dr. Zherui Xiong and Ms. Ashandeep Kaur assisted with the Xenium experiment. Dr. Subash Rai, Dr. Loan Nguyen and Dr. Shivangi Wangi were involved in the long read sequencing experiments to generate the raw sequencing data that I further processed and analysed. Andrew Newman contributed to setting up the lncRNA database website and hosting it on the AWS cloud server.

Chapter 4

Thin Melanoma case-control sample-matched dataset, with paired Visium and WGS bulk data were made available from A/Prof Mitchele Starc and Prof Kiarash Khosrotehrani. Dr. Laura Grice provided valuable suggestions in the malignant cell annotation workflow for the skin cancer datasets. No contributions by others were involved in the spatial data analysis.

A detailed breakdown of the author contributions is further specified in the corresponding chapters.

Use of Artificial Intelligence

All the raw text and ideas were originally written by the candidate. The following table summarises the Artificial Intelligence (AI) tools that were used to assist in the thesis writing.

Table I: Use of AI tools to assist thesis writing

AI Tool	Purpose of Use	Sample Prompts Used	Thesis Section
ChatGPT 4.0 Grok	Assisted in improved clarity and conciseness of the abstract, introduction and discussion sections. ChatGPT was used to translate the abstract to Hindi.	"Rewrite the following section(s) of the abstract, introduction and discussion for greater conciseness and clarity."	Abstract, Introduction (Chapter 1, Chapter 4) and Discussion (Chapter 4, Chapter 5)
notebookLM	Curated and summarised relevant research papers, identifying key themes.	"Summarise these recent studies on different sequencing technologies"	Chapter 1, 4 and 5 (Introduction and Discussion sections)

Statement of parts of the thesis submitted to qualify for the award of another degree

No works submitted towards another degree have been included in this thesis.

Research involving human or animal subjects

Ethical Approval and collection of human cancer samples

Included in this study were clinical FFPE biopsies of human dysplastic naevi and melanoma. Institutional approval of experiments involving human tissues was provided by Metro South and The University of Queensland Human Research Ethics Committees HREC/17/QPAH/817, 2018000165, and 2017000318. For the PDOX mouse models for medulloblastoma, formal consent to collect patient samples for the development of PDOXs was reviewed by the Institutional IRB and implemented under the protocol NBTP01. Breast cancer samples used were approved under the project ID 2005/HE000785 by The Royal Women's Hospital Human Research Ethics Committee. The Head and Neck OPSCC samples were approved under the reference number HREC/2019/QMS/49990 by the Metro South Human Research Ethics Committee. The use of colorectal cancer samples was approved by the QIMR Berghofer-HREC (P460). Overall, spatial omics data generation on a pan-cancer level was approved by the QIMR Berghofer-Human Scientific Sub-committee (HSSC) under the reference number P3900. All methods were carried out following relevant guidelines and regulations.

Acknowledgements

First and foremost, I wish to express my sincere gratitude to my supervisors, Dr. Quan Nguyen and Dr. Ishaan Gupta, for their continuous support, invaluable guidance, patience, and immense knowledge.

I cannot thank Dr. Quan enough for funding my extended stay at UQ which has not only helped my career but also helped lay foundation to the life I have always wanted. His expertise in bioinformatics and his mentorship have been instrumental in shaping my PhD research journey. My organisational skills have drastically improved since I started working with him. He has constantly believed in my ability to work on big projects and to learn more. I am very happy and grateful to continue my journey after PhD in the lab and look forwarding to contributing and learning more.

Dr. Ishaan has not only been of great support academically, but also on a personal level. His constant encouragement and belief in me even before my PhD journey, during a brief collaborative project motivated me to pursue a PhD. His expertise in several areas, in-depth knowledge, and his ability to extensively communicate a difficult topic, making it seem so easy and imparting wisdom is truly inspiring. His ability to read his students, especially during difficult times in our personal lives is such a gift. From planning lab outings during stressful times, advising to take breaks when necessary, motivating to travel and attend as many conferences possible to being selfless and letting me pursue most of my PhD at UQ, he keeps proving himself as the best mentor. I owe my growth as a researcher, and much of my personal development, to your exceptional mentorship. For this, I will always be grateful.

Professors, thank you for making my PhD journey not only academically enriching but also joyous, stress-free, and vibrant. Your impact on my life and career is immeasurable, and I am honoured to have had the privilege of learning from and working with you. Without your encouragement and persistent guidance, this thesis would not have been possible.

I would also like to thank the members of my thesis committee, Dr. Mikael Boden, Dr. Sundar D. and Dr. Shilpi Minocha, for their valuable feedback and insightful comments. I would like to extend my deepest gratitude to my lab mates whose diligent efforts in data generation have

been the backbone of this research. I also thank the Professor Incharge of the UQ-IITD Academy, Ms. Ragini Gill and Dr. Nidhi Dubey for their administrative support.

I am profoundly grateful to our collaborators for their expertise, insights, and invaluable contributions which significantly enriched this study. I also thank the patients and their families for their samples, without which this research would have not been possible.

I would also like to extend my gratitude to Dr. Mitali Mukerji who was my first mentor in a research environment. It was under her guidance that I learnt bioinformatics and stepped into the field. Her passion towards science inspired me to pursue research and made me confident as a woman scientist in India. Her appreciation for everything, even the little ones has helped me grow and excel so much personally and professionally.

My sincere appreciation goes to my scholarship providers, The University of Queensland and the PMRF for their financial support. These scholarships were not just a financial relief but a testament to the belief in the potential of my research, which constantly motivated me to push the boundaries. I am also deeply thankful to my Mom, who as a single parent enabled my achievements. I thank her and my brother for their unwavering love, encouragement, and faith in my potential and my Grandfather for supporting my undergraduate education. I am grateful for my fiancé, Sharat, and his parents for their support and for loving me as their own. All their sacrifices, support, and prayers have been my guiding light, and this achievement is as much theirs as it is mine.

I'd like to thank all those who have been a part of my PhD journey, directly or indirectly, for your contributions, support, and goodwill. Lastly, I thank myself for making it through all the hardships over the years of my life without giving up.

Financial support

This work was supported by Australian Research Council (ARC DECRA grant DE190100116), National Health and Medical Research Council (NHMRC Project Grant 2001514) and NHMRC Investigator Grant (GNT2008928) received by Dr. Quan Nguyen, and the UQ-IITD Research Academy (UQIDRA). This research also utilised funds from the Ramalingaswami fellowship No.BT/RLF/Re-entry/19/2018 by Department of Biotechnology (DBT), Government of India and Indian Council for Medical Research (ICMR) research grant IIRP-2023-2341/F1 received by Dr. Ishaan Gupta. This work is also supported by the Ministry of Human resource development, Government of India sponsored Prime Minister's Research Fellowship (PMRF) IITD/Admission/Ph.D./PMRF/2022/36443 awarded to the candidate Prakrithi Pavithra, along with a relocation grant and supervisor funds by the UQ-IITD Research Academy (UQIDRA) and a Student Development bursary by the Institute of Molecular Bioscience.

Keywords

Spatial transcriptomics, lncRNA, uTARs, single-cell, cancer, bioinformatics, long read sequencing.

Australian and New Zealand standard research classifications (ANZSRC)

ANZSRC code: 060102 Bioinformatics, 40%
ANZSRC code: 060407 Genome Structure and Regulation, 40%
ANZSRC code: 060199 Medical Studies, n.e.c., 20%

Fields of research classification

FoR code: 0601, Biochemistry and Cell Biology, 40%
FoR code: 0604, Genetics, 40%
FoR code: 1112 Oncology and Carcinogenesis, 20%

Table of Contents

List of Figures

Figures in Main Chapters

Appendices

*Note: Figures prefixed with **S** refer to Supplementary Figures, and those prefixed with **ED** refer to Extended Data Figures.*

List of Tables

Tables in Main Chapters

Appendices

List of Acronyms/ Abbreviations

Abbreviation	
aTAR	Annotated Transcriptionally Active Region
BAM	Binary Alignment Map
cDNA	Complementary DNA
CNV	Copy Number Variation
CT	Cycle Threshold
cuTAR	Cancer associated Unannotated Transcriptionally Active Region
DE	Differential Expression/ Differentially Expressed
DNA	Deoxyribonucleic Acid
DSP	Digital Spatial Profiler
EMT	Epithelial to Mesenchymal Transition
FF	Fresh Frozen
FFPE	Formalin Fixed Paraffin Embedded
GTF	Gene Transfer Format
GWAS	Genome-wide Association Study
H&E	Hematoxylin and Eosin
H&N OPSCC	Head and Neck Oropharyngeal Squamous Cell Carcinoma
HCP	Human Coding Probability
HD	High Definition
HiFi	High Fidelity
HMM	Hidden Markov Model
MAS-Seq	Multiplexed Array Sequencing
NGS	Next Generation Sequencing
lncRNA	Long non-coding RNA
ONT	Oxford Nanopore Technology
PacBio	Pacific Biosciences
qPCR	Quantitative Polymerase Chain Reaction
RBP	Ribosomal Binding Protein
RNA	Ribonucleic Acid

RNA-seq	RNA sequencing
ROI	Region Of Interest
SC	Single Cell
SCC	Squamous Cell Carcinoma
scRNA-seq	Single-cell RNA sequencing
SMRT	Single Molecule, Real-Time
SNP	Single Nucleotide Polymorphism
SPanC-Lnc	Spatial and Single Cell Pan-Cancer atlas of lncRNAs
ST	Spatial Transcriptomics
TAR	Transcriptionally Active Region
TMM	Trimmed Mean of M-values normalization
uTAR	Unannotated Transcriptionally Active Region
UMI	Unique Molecular Identifier
WGS	Whole Genome Sequencing