

Bioinformatics Approaches For Lung Cancer Early Detection And Diagnosis Based On Liquid Biopsy Data

by

Stavros P. Giannoukakos



**UNIVERSIDAD
DE GRANADA**

PhD Program in Fundamental and Systems Biology

Department of Genetics

University of Granada

Supervisors

Professor Michael Hackenberg

Professor Alberto Luis Fernández Hilario

March 2024

Editor: Universidad de Granada. Tesis Doctorales
Autor: Stavros Panagiotis Giannoukakos
ISBN: 978-84-1195-567-6
URI: <https://hdl.handle.net/10481/97444>

Financial support

This Doctor of Philosophy thesis was conducted at the Faculty of Science, Department of Genetics, University of Granada.

The research was fully supported by the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement for the *European Liquid Biopsy Academy (ELBA)* with Grant No 765492.



European Union



Marie Curie
Actions



ELBA

Acknowledgements

I am profoundly grateful for the unwavering support and encouragement that has illuminated my journey throughout my doctoral pursuit.

To my family, whose love and support have been the bedrock of my strength, I extend my deepest gratitude. Your unwavering belief in me has been my guiding light, and I am endlessly thankful for the sacrifices you've made to see me through this challenging yet rewarding endeavour. To my beautiful niece, born amidst the chapters of my PhD, you are a symbol of joy and new beginnings, and your presence has added immeasurable warmth to this academic odyssey.

To my friends, your companionship and understanding during the peaks and valleys of this journey have been a source of solace. Special thanks to Asli, Kiveli, Konstantinos, Aggelos, Maro, Despo, Thomas, Vaso, Christos, and Margarita for their steadfast friendship and unwavering support.

Within this consortium, I extend my sincere appreciation to Silvia D'Ambrosi, Carlos Pedraz-Valdunciel, Diogo Fortunato, and Martyna Filipka. Your individual and collective contributions have been instrumental, turning challenges into triumphs and significantly enhancing the richness of this academic journey.

In addition to the ELBA consortium, my heartfelt gratitude extends to my team at Granada and CIBM, where the unwavering support of Ernesto, Cristina, and Angel has been instrumental throughout my academic journey. Their tireless efforts and technical expertise have been a constant source of assistance, guiding me through various challenges with unwavering dedication.

Heartfelt appreciation goes to my supervisors, Michael Hackenberg, Alberto Fernandez, and Danijela Koppers-Lalic. Your patience, guidance, and belief in my potential have been the compass that guided me through the labyrinth of academia. Without your mentorship, my PhD would have remained a distant dream.

This achievement is not mine alone; it is a testament to the collective strength of those who have surrounded me with love, wisdom, and support. Thank you, from the depths of my heart, for being the pillars of my academic success.

“Nothing in life is to be feared, it is only to be understood. Now is the time to understand more, so that we may fear less.”

- Marie Curie

Summary

Cancer, especially lung cancer—the leading cause of deaths related to this disease—highlights the limitations of traditional tissue biopsies which are invasive and cannot continuously monitor tumour evolution. This emphasises the need for less invasive and more dynamic diagnostic methods. Liquid biopsy, using bodily fluids to detect molecular changes, offers a less invasive alternative, enabling the exploration of diverse biosources for early detection, diagnosis, and disease monitoring. While advances in gene expression technologies like Second and Third Generation Sequencing and NanoString have enhanced our understanding of tumour biology, the full potential of liquid biopsy has yet to be realised. Further development in bioinformatics and machine learning is necessary to harness liquid biopsy’s capabilities for personalised cancer management, bridging the gap between innovative gene expression technologies and clinical application.

As part of the European Liquid Biopsy Academy (ELBA) Innovative Training Network consortium, this thesis contributes to the collective mission of enhancing cancer diagnostics. Various projects within ELBA employ diverse technologies with the shared goal of improving cancer diagnosis through liquid or tissue biopsy. Specifically, this thesis aims to thoroughly address three primary goals: first, developing a novel method called Ensemble Learning for Liquid Biopsy Analysis (ELLBA) for analysing liquid biopsy RNA sequencing data; second, using Oxford Nanopore Technologies’ direct RNA sequencing for tissue biopsy of non-small cell lung cancer to identify prognostic biomarkers, by means of a new pipeline named ‘DRseeker’; and third, creating ‘NanoInsights’, a comprehensive solution for NanoString nCounter technology that integrates advanced bioinformatics and machine learning to improve data analysis.

In pursuit of our first aim to advance liquid biopsy-based transcriptomics, we introduced a new methodology called Ensemble Learning for Liquid Biopsy Analysis (ELLBA). Our hypothesis posited that extracting discriminative molecular features from Second Generation Sequencing liquid biopsy-based RNA-Seq data could improve cancer predictions. ELLBA integrates six biofeature types—gene expression, isoform expression, Fraction of Canonical Transcript, gene fusion, RNA editing, and Single Nucleotide Variants—enabling comprehensive molecular characteristic capture in cancer diagnostics. Utilising intra-sample CPM normalisation and standard ensemble classification methods, ELLBA outperforms traditional gene expression analysis in predictive accuracy. Rigorously assessed across diverse datasets and biosources, ELLBA consistently showed superior performance through integrated biofeature data analysis via ensemble classification.

Regarding the second aim, the emergence of Third Generation Sequencing, particularly Oxford Nanopore Technologies' direct RNA sequencing (DRS) protocol, presents notable strides in cancer transcriptomics. DRS facilitates the capture of complete transcript lengths in their native state, offering insights into various aspects of the transcriptome. To fully harness the potential of DRS research, a tailored bioinformatics pipeline named 'DRseeker' was created for comprehensive transcriptomic analysis. In its application to a lung cancer dataset, DRseeker facilitated the identification of significant shifts in transcript expression, the discovery of novel transcripts, and the detection of alterations in crucial genes. Moreover, the examination of polyadenylation variations and epitranscriptomic changes, such as methylation, illuminated intricate regulatory mechanisms within cancer cells.

Finally, the NanoString nCounter system represents a significant technological advancement in transcriptomics, particularly useful in translational research and clinical applications. This system offers numerous advantages and can play an important role in various applications, including liquid biopsy. However, it introduces new challenges in data analysis, such as normalisation and interpretation. To tackle these challenges, the 'NanoInsights' web service was developed, integrating bioinformatics and machine learning to enhance NanoString data analysis. Featuring a user-friendly interface, extensive quality control, multiple normalisation methods, gene enrichment analysis, and diverse machine learning approaches, NanoInsights caters to researchers of all expertise levels, offering a comprehensive solution for interpreting complex datasets.

The integration of liquid biopsy, cutting-edge gene expression technologies, innovative bioinformatics, and state-of-the-art machine learning algorithms represents a transformative leap in cancer diagnostics. This convergence not only enriches our comprehension of cancer's molecular intricacies but also lays the groundwork for early detection and diagnosis. Ultimately, it opens new avenues for personalised medicine and targeted therapies, promising more effective treatments and improved patient outcomes.

Key Words: bioinformatics, machine learning, lung cancer, liquid biopsy, transcriptomics, liquid biopsy-based RNA-Seq, direct RNA sequencing, NanoString nCounter

Resumen

El cáncer, especialmente el cáncer de pulmón, que es la principal causa de muertes relacionadas con esta enfermedad, resalta las limitaciones de las biopsias de tejido tradicionales, pues resultan invasivas y no pueden monitorear continuamente la evolución del tumor. Así, es evidente la necesidad de métodos diagnósticos menos invasivos y más dinámicos. La biopsia líquida, que utiliza fluidos corporales para detectar cambios moleculares, ofrece una alternativa que incluye las características anteriores, permitiendo la exploración de diversas fuentes biológicas para la detección temprana, diagnóstico y seguimiento de la enfermedad. Aunque los avances en tecnologías de expresión génica, como la Secuenciación de Segunda y Tercera Generación y NanoString, han mejorado nuestra comprensión de la biología tumoral, aún no hemos alcanzado el potencial completo de la biopsia líquida. Se requiere un mayor desarrollo en bioinformática y aprendizaje automático para aprovechar plenamente las capacidades de la biopsia líquida en la gestión personalizada del cáncer, cerrando la brecha entre las tecnologías innovadoras de expresión génica y su aplicación clínica.

Como parte del consorcio de la Academia Europea de Biopsia Líquida (en inglés European Liquid Biopsy Academy o ELBA), esta tesis contribuye a la misión colectiva de mejorar el diagnóstico del cáncer. Varios proyectos dentro de ELBA emplean diversas tecnologías con el objetivo compartido de mejorar el diagnóstico del cáncer mediante biopsias líquidas o de tejido. Específicamente, esta tesis tiene como objetivo abordar de manera exhaustiva tres metas principales: primero, desarrollar un nuevo método llamado Aprendizaje del Conjunto para el Análisis de Biopsias Líquidas de Biopsias Líquidas para analizar datos de secuenciación de ARN de biopsias líquidas; segundo, utilizar la secuenciación directa de ARN de Oxford Nanopore Technologies para la biopsia de tejido del cáncer de pulmón de células no pequeñas para identificar biomarcadores pronósticos, mediante un nuevo

pipeline llamado 'DRseeker'; y tercero, crear 'NanoInsights', una solución integral para la tecnología NanoString nCounter que integra bioinformática avanzada y aprendizaje automático para mejorar el análisis de datos.

En busca de nuestro primer objetivo de avanzar en la transcriptómica basada en biopsia líquida, introdujimos una nueva metodología denominada Aprendizaje del Conjunto para el Análisis de Biopsias Líquidas de Biopsia Líquida (Ensemble Learning for Liquid Biopsy Analysis o ELLBA, por sus siglas en inglés). Nuestra hipótesis postulaba que la extracción de características moleculares discriminativas de los datos de RNA-Seq de biopsia líquida de secuenciación de segunda generación podría mejorar las predicciones de cáncer. ELLBA integra seis tipos de biomarcadores: expresión génica, expresión de isoformas, Fracción de Transcrito Canónico, fusión de genes, edición de ARN y Variantes de Nucleótido Único, permitiendo la captura integral de características moleculares en el diagnóstico del cáncer. Utilizando la normalización CPM intra-muestra y métodos estándar de clasificación de conjuntos, ELLBA supera el análisis tradicional de expresión génica en precisión predictiva. Evaluado rigurosamente en diversos conjuntos de datos y biofuentes, ELLBA mostró un rendimiento consistentemente superior a través del análisis integrado de datos de biomarcadores mediante clasificación combinada (tipo ensemble).

En cuanto al segundo objetivo, la aparición de la secuenciación de tercera generación, en particular el protocolo de secuenciación directa de ARN (en inglés direct RNA sequencing o DRS) de Oxford Nanopore Technologies, representa avances notables en la transcriptómica del cáncer. El DRS facilita la captura de longitudes completas de transcritos en su estado nativo, ofreciendo información sobre diversos aspectos del transcriptoma. Para aprovechar al máximo el potencial de los avances que permite el uso de DRS, se creó una pipeline bioinformática específica denominada "DRseeker" para el análisis transcriptómico integral. En su aplicación a un conjunto de datos de cáncer de pulmón, DRseeker facilitó la identificación de cambios significativos en la expresión de transcritos, el descubrimiento de nuevos transcritos y la detección de alteraciones en genes cruciales. Además, el examen de las variaciones de poliadenilación y los cambios eptranscriptómicos, como la metilación, permitió elucidar mecanismos regulatorios intrincados dentro de las células cancerosas.

Finalmente, el sistema NanoString nCounter representa un avance tecnológico significativo en transcriptómica, particularmente útil en la investigación traslacional

y las aplicaciones clínicas. Este sistema ofrece numerosas ventajas y puede desempeñar un papel importante en diversas aplicaciones, incluida la biopsia líquida. Sin embargo, introduce nuevos desafíos en el análisis de datos, como la normalización e interpretación. Para abordar estos desafíos, se desarrolló el servicio web "NanoInsights", que integra bioinformática y aprendizaje automático para mejorar el análisis de datos de NanoString. Con una interfaz de usuario amigable, control de calidad exhaustivo, múltiples métodos de normalización, análisis de enriquecimiento de genes y diversos enfoques de aprendizaje automático, NanoInsights atiende a investigadores de todos los niveles de experiencia, ofreciendo una solución integral para interpretar conjuntos de datos complejos.

La integración de la biopsia líquida, las tecnologías de expresión génica de vanguardia, la bioinformática innovadora y los algoritmos de aprendizaje automático de última generación representa un salto transformador en el diagnóstico del cáncer. Esta convergencia no solo enriquece nuestra comprensión de las complejidades moleculares del cáncer, sino que también sienta las bases para la detección y el diagnóstico tempranos. En última instancia, abre nuevas vías para la medicina personalizada y las terapias dirigidas, lo que promete tratamientos más efectivos y mejores resultados para los pacientes.

Palabras clave: bioinformática, aprendizaje automático, cáncer de pulmón, biopsia líquida, transcriptómica, secuenciación de ARN basada en biopsia líquida, secuenciación directa de ARN, NanoString nCounter

Contents

Copyright Commitment	i
Financial support	ii
Acknowledgements	iii
Summary	iv
Resumen	vii
List of Figures	xv
List of Tables	xxvi
Abbreviations	xxvii
1 Introduction	1
1.1 The Global Cancer Landscape	1
1.1.1 Understanding Cancer	3
1.1.2 Cancer MicroEnvironment	4
1.1.3 Overview of Lung Cancer	5
1.1.4 Diagnostic Challenges of Lung Cancer	6
1.2 Tissue Biopsy: Unraveling its Process and Challenges	7
1.3 Liquid Biopsy: Unveiling New Frontiers	9
1.3.1 Circulatory Biomarkers in Blood-Based Liquid Biopsy	10
1.3.2 Tumour-Educated Platelets	11
1.3.3 Extracellular Vesicles	12
1.3.4 Circulating Epithelial Cells	14
1.4 Sequencing Technologies	15
1.4.1 Second Generation Sequencing	17
1.4.2 Third Generation Sequencing	18
1.4.3 NanoString nCounter Technology	21
1.5 Sequencing Technology: Paving the Way for Liquid Biopsy	23

1.6	Bioinformatics Data Analysis	24
1.7	The Emergence of Liquid Biopsy in Bioinformatics Research	27
1.8	Machine Learning in Genomics Research	28
1.9	Machine Learning and Liquid Biopsy	30
1.10	Thesis Overview	31
2	Objectives	34
3	Methods	37
3.1	Molecular Methods and Sequencing	37
3.1.1	Lung Tissue Collection and RNA Isolation in NSCLC Patients	37
3.1.2	RNA Isolation from Lung Tissue Samples	38
3.1.3	MinION Nanopore Direct RNA sequencing	38
3.2	Bioinformatics Methods	38
3.2.1	Preprocessing and Quality Control Workflow for NGS Data .	39
3.2.2	Mapping and Quality Assessment for NGS Data	39
3.2.3	Transcriptome Quantification for NGS Data	40
3.2.4	Identification of RNA Editing and SNVs in NGS Data . . .	40
3.2.5	Re-Basecalling and Initial Filtering Workflow for TGS Data	41
3.2.6	Mapping and Quality Assessment for TGS Data	42
3.2.7	Gene and Transcript Identification and Quantification Using TALON for TGS Data	42
3.2.8	CAGE Analysis for TSS Verification for TGS Data	43
3.3	Statistical Analysis	43
3.3.1	Exploratory Analysis for NGS Data	43
3.3.2	Normalisation of the NGS Data	44
3.3.3	Gene Set Enrichment Analysis for the NGS Data	44
3.3.4	Exploratory Analysis for TGS Data	45
3.3.5	Differential Expression and Usage Analysis for TGS Data . .	45
3.3.6	Novel Transcript Characterisation for TGS Data	46
3.3.7	PolyA-Tail Length, DPA, and APA Analysis for TGS Data .	47
3.3.8	Methylation Detection for TGS Data	47
3.3.9	Initial Quality Control and Exploratory Analysis for NanoS- tring nCounter Data	47
3.3.10	Normalisation and Differential Expression for NanoString nCounter Data	48
3.4	Machine Learning Methods	51
3.4.1	Filtering Process	52
3.4.2	Feature Selection Techniques	52
3.4.3	Classification Algorithms	54
3.4.4	Ensemble Learning Techniques	56
3.4.5	Evaluation Metrics	57
4	Assessing the complementary information of biologically relevant features in lbrNA-Seq data	61

4.1	Introduction	62
4.2	Materials and Methods	64
4.2.1	Workflow and Implementation	64
4.2.2	Data Preprocessing and Biofeature Extraction	65
4.2.3	Gene Expression	65
4.2.4	Isoform Expression	67
4.2.5	Fraction of Canonical Transcript	67
4.2.6	Gene Fusion	67
4.2.7	RNA Editing Events	68
4.2.8	Single Nucleotide Variants	68
4.2.9	Machine Learning and Ensemble Learning Implementation .	69
4.2.10	Functional Enrichment Analysis	71
4.3	Results	71
4.3.1	Data Collection and Description	71
4.3.2	Overview of the ELLBA Methodology	73
4.3.3	Comparative Analysis of Normalisation Methods for Gene and Isoform Expression Data	75
4.3.4	Optimal Classification Models for the Different Biofeature Types	77
4.3.5	Enhancing Predictive Output through Ensemble Learning .	79
4.3.6	Biological Feature Selected Interpretation	82
4.4	Discussion	84
4.5	Conclusion	87
5	Expanding the landscape of cancer transcriptome by native RNA sequencing of NSCLC tissue samples	88
5.1	Introduction	89
5.2	Materials and Methods	91
5.2.1	Methodology for Lung Tissue Collection and RNA Isolation in NSCLC Patients	91
5.2.2	RNA Isolation from Lung Tissue Samples	91
5.2.3	Direct RNA Sequencing of NSCLC Samples Using Oxford Nanopore MinION Technology	92
5.2.4	Re-Basecalling and Initial Filtering Process	92
5.2.5	Initial Quality Control and Preprocessing	93
5.2.6	Genome Alignment and Post-Alignment Quality Control . .	93
5.2.7	Gene and Transcript Identification and Abundance Estimation	94
5.2.8	Exploratory Analysis	94
5.2.9	DGE, DTE, and DTU	95
5.2.10	Gene Functional Enrichment Analysis	96
5.2.11	Transcript Functional Annotation and Functional Enrich- ment Analysis	96
5.2.12	CAGE Analysis	96
5.2.13	PolyA-Tail Length Estimation, DPA and APA Analysis . . .	97
5.2.14	Methylation Detection	97

5.3	Results	98
5.3.1	Comprehensive Analytical Pipeline for Direct Long-Read RNA-Seq Data Investigation	98
5.3.2	Transcriptome Profiling of Lung Cancer Tissue and Adjacent Non-Transformed Lung Tissue Using ONT Direct RNA Sequencing Protocol	100
5.3.3	TALON: Detecting and Measuring Both Known and Novel Transcripts	102
5.3.4	Exploratory and DGE Analysis in NSCLC and Adjacent Non-Transformed Tissues via Long-Read Sequencing	104
5.3.5	Assessment of Differential Transcript Expression in NSCLC Using Long-Read Sequencing	105
5.3.6	Transcriptomic Diversity in Cancer: Investigating Differential Transcript Usage	109
5.3.7	Unraveling the Complexity of NSCLC: Insights from PolyA Tail Length Variations	111
5.3.8	Deciphering RNA Methylation Patterns in the NSCLC Transcriptome: A New Dimension of Analysis	114
5.4	Discussion	119
5.5	Conclusion	122
6	NanoInsights: A Web Platform for Advanced NanoString nCounter Data Analysis	123
6.1	Introduction	123
6.2	Materials and Methods	126
6.2.1	Web-Service Development and Hosting Details	126
6.2.2	Quality Control and Preliminary Data Exploration	126
6.2.3	Filtering Process for Genes and Samples	127
6.2.4	Normalisation and Differential Expression Analysis	127
6.2.5	Optional Preprocessing and Feature Selection	129
6.2.6	Selection of Classification Algorithm	129
6.2.7	Selection of Test Set for final classifier evaluation	130
6.2.8	Classification Output	130
6.2.9	Analysis of Gene Set Enrichment	131
6.3	Results	131
6.3.1	Overview of the NanoInsights Platform Workflow	131
6.3.2	User Interface Overview	134
6.3.3	Case Study	137
6.4	Discussion	140
6.5	Conclusion	142
7	Brief Discussion and Future Perspectives	144
8	Conclusions	147
9	Conclusiones	149

Bibliography	151
Appendix	193
A.1 Assessing the complementary information of biologically relevant features in lbRNA-Seq data (Supplementary Materials)	193
A.1.1 Detailed analysis applied to the individual datasets	193
A.1.2 Comprehensive elucidation of the Decision Output module	195
A.1.3 Normalisation and Benchmarking of Gene and Isoform Expression Matrices	196
A.1.4 Implementation of different ensemble learning techniques	197
A.1.5 Calculation of Accuracy and Misclassification rate	197
A.1.6 Interpretation of the Biological Significance of Selected Features in the NSCLC Dataset	198
A.1.7 Supplementary Figures	200
A.1.8 Supplementary Tables	203
A.2 Expanding the landscape of cancer transcriptome by native RNA sequencing of NSCLC tissue samples (Supplementary Materials)	220
A.2.1 TALON transcript classification categories.	220
A.2.2 Supplementary Figures	221
A.2.3 Supplementary Tables	228
A.3 NanoInsights: A Web Platform for Advanced NanoString nCounter Data Analysis (Supplementary Materials)	239
A.3.1 NanoString nCounter Technology: A Multiplexed Approach for RNA Target Analysis	239
A.3.2 Supplementary Figures	240
A.3.3 Supplementary Tables	243

List of Figures

1.1	Worldwide Cancer Case Estimates in 2020 by Globocan. Estimated global cancer occurrences for 2020, covering individuals of every gender and age group across all continents.	2
1.2	Illustration of the Tumour Microenvironment. The figure provides a comprehensive depiction of the TME, capturing both the transformed tumour cells and the non-transformed cells that constitute the intricate tumour landscape. Adapted from the work of Belli, Antonarelli, Repetto, <i>et al.</i> 2022, originally published in <i>Cancers</i> [17].	4
1.3	Lung Cancer Subtypes. Overview of lung cancer’s main histological types: SCLC and NSCLC. SCLC includes small cell carcinoma and mixed small cell/large cell cancer, typically associated with rapid growth and strong correlation with cigarette smoking. NSCLC encompasses adenocarcinoma (predominantly found in the lung’s outer regions), squamous cell carcinoma (typically located near bronchial tubes), and large cell carcinoma (with a tendency for rapid growth).	5
1.4	Circulating Biomarkers Revealed Through Blood-Based Liquid Biopsy. It highlights circulating components in plasma or serum (top), including EVs like Exosomes, Proteins, circulating cell-free RNA (encompassing both noncoding and messenger RNA), ctDNA, and TEPs. Simultaneously, the cellular fraction (bottom) unveils: i. tumour cells (CTCs, either singular or clustered), and ii. the non-tumour cell fraction, such as immune cells, CECs, and Cancer-Associated Fibroblasts.	10
1.5	Reciprocal Influences Between Cancer and Platelets. This illustration, based on research by Ding, Dong, and Song 2023 published in <i>Cancer Cell International</i> [75], shows the complex interactions between cancer and platelets. It focuses on how tumours educate platelets, leading to their activation, aggregation, and release of substances, which promotes thrombocytosis by affecting bone marrow megakaryopoiesis. In turn, platelets support tumour growth and spread through angiogenesis, vascular remodelling, protecting CTCs, evading immune detection, and recruiting stromal cells. Critical elements in these processes are megakaryocyte progenitor, MKs, and hematopoietic stem cells.	12

-
- 1.6 **Representation of distinct extracellular vesicles subtypes.** Modified from the work of Fox, Kim, Le, *et al.* 2015, originally published in the Journal of Controlled Release [79]. The figure highlights exosomes, microvesicles, and apoptotic bodies, each presenting distinct characteristics and roles. 13
- 1.7 **Overview of the RNA Sequencing Workflow.** This schematic depicts RNA sequencing via NGS technology in six steps: 1) RNA isolation from samples; 2) fragmentation into short pieces; 3) conversion into cDNA; 4) ligation of sequencing adapters and amplification; 5) sequencing of amplified cDNA fragments; and 6) mapping reads to the genome or transcriptome to identify exon and intron regions. Adapted from Prashant Dahal’s work on RNA sequencing [104]. 18
- 1.8 **RNA sequencing using nanopore technology.** This illustration, edited by ONT [117], outlines the direct RNA Sequencing Workflow: a) Library Preparation details full-length poly-A RNA preparation for nanopore sequencing, beginning with ligation of a reverse transcription splint (red) to RNA (blue), followed by sequencing adapter attachment (blue with brown tips) for nanopore entry. b) Ionic Current Trace shows the ionic current changes as the RNA-transcript passes through the nanopore, highlighting the baseline open-pore current, adapter and poly-A tail disruptions, a longer disruption for the 1,500 nt transcript, and a return to baseline after transcript passage. 20
- 1.9 **nCounter Digital Nucleic Acid Counting.** Overview of the nCounter Digital Nucleic Acid Counting system workflow, illustrating a three-step process: 1) hybridisation of mRNA with CodeSet probes, 2) purification of the hybridised complexes, and 3) counting and identification of target nucleic acids using the Digital Analyser for high-multiplex quantification of nucleic acid molecules without cDNA synthesis or amplification [122]. 22
- 4.1 **Comprehensive Overview of Datasets in the Study.** A detailed overview of the datasets employed in this study, showcasing six distinct datasets: NSCLC, GBM, CRC, ESCC, PDAC, and HCC, as depicted in the outer donut plot. Light blue colouring (NSCLC, CRC, and PDAC) signifies datasets with independent external validation sets, while grey shading (GBM, ESCC, HCC) represents datasets without external validation. The inner circle categorises the biosource origin of each dataset: dark yellow for TEPs in NSCLC, GBM, CRC, and ESCC; cinnamon red for EVs in PDAC; and brown for CECs in HCC. 72

- 4.2 **Overview of the ELLBA Workflow.** ELLBA methodology features two main components: bioinformatics analysis (light green) and ML (light blue). The workflow includes seven modules. Bioinformatics analysis involves Input, Mapping, Biofeature Extraction, and Biofeature Processing. ML includes Feature Selection, Classification, and Decision Output. The process starts with data Input and progresses through Mapping, Biofeature Extraction, and Biofeature Processing for bioinformatics analysis. Then, Feature Selection, Classification, and Decision Output handle ML analysis. Biofeatures are individually processed in Biofeature Processing, involving data cleaning and normalisation or discretisation. In ML, Feature Selection and Classification are applied to each biofeature. The Decision Output combines individual classification outputs using ensemble learning (soft voting) for the final decision. 74
- 4.3 **Summary of Various Normalisation Methods.** A total of eight normalisation techniques were assessed across all six datasets. Each column corresponds to a distinct normalisation method, while each row represents a different dataset employed. The x-axis illustrates the various models utilised within each normalisation. and the y-axis depicts the mean AUC score achieved through 5-fold CV. Each model is represented by dots indicating the AUC for each CV fold. Additionally, a dashed line indicates the mean AUC across the 5 folds. while a solid line represents the mean AUC across all models. 77
- 4.4 **Classifier Selection Overview.** Each column in the plot represents a different feature type, while each row corresponds to a different dataset. On the x-axis, six classifiers are evaluated, and the y-axis displays the mean AUC score achieved through 5-fold CV. The dashed lines represent the average AUC score from the 5-fold CV evaluation. For the first two feature types (gene and isoform expression), AdaBoost with ExtraTrees, is highlighted in red, as it is better suited for these feature types. For the remaining feature types, Logistic Regression, is highlighted in blue, as it is better suited for these types of features. 79
- 4.5 **Performance Overview of Feature Types and Ensemble Learning Methods Across Datasets.** Each column in this figure corresponds to a specific dataset, with dataset information and the presence of an independent validation set indicated above. (A) The plot displays AUC scores for the test sets across all feature types. (B) Dot plots illustrate the percentage of misclassifications for each feature type across different datasets. The colourful squares and dots in both figures A and B represent the ensemble learning techniques, while the grey squares and dots represent the remaining feature types. 80

- 4.6 **Average Misclassification Percentage Across All Datasets.** This graph displays the percentage of average misclassifications for each feature type on the x-axis, with feature types arranged from the least to the most misclassifications from left to right. Each dot represents the percentage of average misclassifications, providing an overview of the classification performance across all datasets. 82
- 4.7 **Enrichment Analysis results.** Presented is a Manhattan plot illustrating the results of an enrichment analysis performed on the genes selected for ensemble learning within the NSCLC dataset. The x-axis is dedicated to functional terms, which have been meticulously organised and colour-coded based on their respective data sources. Simultaneously, the y-axis represents the adjusted enrichment p-values, thoughtfully presented in a logarithmic negative scale. Terms of lesser significance are discreetly depicted as faint circles, while encircled numbers within the figure denote statistically significant enriched GO terms. 83
- 5.1 **Schematic representation of the data analysis workflow.** This figure presents a structured overview of the data analysis workflow divided into four main stages: Data Preprocessing, Alignment, Core Analyses, and Complementary Analyses. Each stage outlines the key steps taken, from initial raw data re-basercalling and QC, through genomic alignment and genes/isoforms quantification, to the core examination of transcriptomic features including differential expression and alternative splicing, and the complementary analyses such as polyadenylation assessment, functional annotation, and methylation detection. 98
- 5.2 **Overview of RNA sequencing workflow and data characteristics.** (A) Illustrating the process flow from tissue collection in NSCLC patients, including tumour and non-transformed tissue, through RNA extraction, library preparation, and sequencing using the MinION platform. The inset graph shows a representative sample of sequencing current over time. (B) Displaying a histogram of read lengths for the combined set of raw, mapped, and assigned reads from both tumour and non-transformed samples. (C) Depicting the gene body coverage profiles of all six samples. 101
- 5.3 **Overview of RNA sequencing workflow and data characteristics.** (A) Depicting the overall filtering statistics, contrasting prefiltered (light green) and filtered (dark blue) counts for unique genes, unique transcripts, known genes, known transcripts, novel genes, and novel transcripts. (B) Illustrating the overall gene-level novelty, depicting counts for known genes, intergenic regions, and antisense transcripts. (C) Presenting the overall transcript-level novelty, with counts for known transcripts, ISMs, NICs, NNCs, Intergenic and Antisense transcripts. Please note that both gene and transcript level figures are based on the filtered data. 103

- 5.4 **Composite Overview of Gene-Level Analysis in cancer versus non-transformed samples.** (A) Shows a PCA at the gene level that distinguishes cancerous samples from non-transformed tissues, noting the percentage of variance explained by the first principal component. (B) Displays a Volcano Plot illustrating differences in gene expression between the two sample types, with genes categorised as up-regulated (in green), down-regulated (in red), or unchanged (in grey), based on a log₂ fold-change threshold of 2 and an adjusted p-value cutoff of 0.05. (C) Details a Gene Ontology Enrichment Analysis, with bars representing the frequency of genes involved in various biological processes, coloured by significance of enrichment. (D) Presents a comparative Venn diagram that quantifies and contrasts genes identified in this study with those found in prior research by Bang et al. in 2017, emphasising the unique and overlapping findings relevant to NSCLC genomics. 104
- 5.5 **Transcriptomic Landscape of Differential Expression in Cancer.** (A) MA plot illustrating the log₂ fold change against log Counts Per Million, distinguishing between known and novel transcripts. The significant differentially expressed transcripts are marked as up-regulated (green) and down-regulated (red), with novel transcripts denoted by triangles. (B) Bar graph summarising the number of known and novel transcripts that are up-regulated and down-regulated, providing a clear distribution of expression changes in the cancer samples. 106
- 5.6 **Differential Expression Analysis of AGER in non-transformed and cancer States.** This composite bar chart illustrates the gene and isoform expression levels of AGER between non-transformed and cancer conditions. The bars indicate the mean expression levels, and the error bars represent the variability within the sample conditions, reflecting the standard deviation of the measurements. The condition of the samples is colour-coded for clarity. The left panel shows the aggregated gene expression, indicating a statistically significant decrease (* $p < 0.05$) in cancer compared to non-transformed samples. The right panel details the expression of individual isoforms, with the y-axis representing normalised expression levels. The various isoforms are distinguished by their transcript types (Known, NNC, ISM Suffix). Notable variations in expression are marked with significance levels, where a single asterisk (*) denotes $p < 0.05$ and double asterisks (**) denote $p < 0.01$ 108

- 5.7 **Isoform Usage Shift in MEST Gene between Non-transformed and Cancer Conditions.** The figure depicts the isoform switch in the MEST gene between conditions. The top panel illustrates exon-intron structures of three distinct MEST isoforms, with coding regions in grey and the Abhydrolase_1 domain in orange. The lower panels show quantified gene and isoform expressions, and isoform usage fractions (IF), contrasting non-transformed with cancer conditions. Bars represent mean expression levels, and error bars indicate variability. Significant changes in isoform usage are marked by asterisks, underscoring the potential regulatory implications of isoform switching in oncogenesis. 111
- 5.8 **Comprehensive Analysis of Polyadenylation Variability in Cancer.** (A) Bubble plot displaying the differential polyA tail length (DPA) across all transcripts, with bubble size corresponding to the median cancer PolyA length and colour intensity indicating significance level. (B) Boxplot of global comparison of polyA tail lengths (DPA) between non-transformed and cancer groups, with median values annotated, showcasing a statistically significant difference as determined by the Wilcoxon test. (C) Volcano plot highlighting genes with alternative polyadenylation (APA) patterns, where the vertical axis represents the negative log₁₀ adjusted p-value and the horizontal axis shows the delta usage. Points are coloured based on regulatory direction, with labels identifying notable genes with significant APA events. 113
- 5.9 **Comparative Heatmap of Methylation Rates in Cancer vs Non-Transformed samples.** This heatmap displays varying methylation rates across a spectrum of tumour-suppressive and oncogenic genes related to lung cancer, comparing cancerous to non-transformed samples. Each horizontal band represents a sample, with upper bands for cancer and lower for non-transformed samples; vertical bands correspond to different genes. Colour gradients from dark green to bright orange reflect methylation rates, with darker tones indicating higher methylation. Genes marked with asterisks below the heatmap show significant methylation differences. 119

- 6.1 **Detailed Flowchart of the NanoInsights Analytical Process.** The figure presents an eight-stage progression through NanoInsights' analytical platform. The initial phase (Stage 1) involves the collection of .RCC files and the associated clinical data. In Stage 2, this data is uploaded for processing, and the analysis pipeline is initiated, with an option to customise or utilise default parameters. Pre-processing, quality control, and exploratory analysis define Stage 3, setting the groundwork for data analysis. Stage 4 applies selective gene or sample filters based on earlier parameter settings. Normalisation is the focus of Stage 5, where users can select from a suite of algorithms or rely on the system's automatic algorithm selection. Machine Learning takes centre stage in Stage 6, employing data to train and evaluate a selected classifier. Stage 7 integrates the outcomes of differential expression and Machine Learning for gene set enrichment analysis. The final phase, Stage 8, offers data visualisation and insight extraction, providing users with interactive tools to explore the analysis findings and facilitating the download of the fully processed results. This overview encapsulates NanoInsights' methodical process for thorough RNA analysis, spanning from data intake to the derivation of insightful conclusions. 134
- 6.2 **Screenshot of the NanoInsights Platform Interface.** Screenshot of the NanoInsights web service main page interface, providing an overview of the user experience upon entering the website. . . . 135
- 6.3 **Visual Analytics of RNA Expression Data in Response to Chemoradiotherapy.** This compilation of visualisations showcases the data analysis conducted with NanoInsights for patients with LARC, segmented by their response to Preoperative Chemoradiotherapy. **(A)** A box plot visualises the unnormalised data, delineating expression levels across the samples, with colours corresponding to the CartridgeIDs denoting different Runs. **(B)** A box plot derived from IQR analysis contrasts the expression profiles between good responders and non-responders. **(C)** Post-Loess normalisation, this box plot depicts RLE levels, with good responders in blue and non-responders in red. **(D)** A volcano plot captures DE genes post-Loess normalisation; red points indicate significantly down-regulated genes, and green points represent up-regulated ones, based on an adjusted p-value of 0.05 and an absolute log2 fold change of 0.5. **(E)** The feature selection process is illustrated using RFECV with a Random Forest Classifier; it plots cross-validated scores against the number of features, using 'balanced accuracy' as the scoring metric. The red point marks the optimal number of selected features. **(F)** A ROC plot provides a comparative overview of different classifiers' abilities to differentiate between good and non-responders, with varying colours representing the different classifiers and datasets, such as the training or test sets. 139

- 1 **ELBA Methodology Integration into Clinical Screening.** This schematic illustrates the integration of the ELBA methodology into the routine clinical screening process. When an individual undergoes regular screening at a healthcare facility, a blood sample is collected, and one of the blood-based biosources is extracted. Subsequently, the content of this biosource can be sequenced, and the resulting raw data serves as input for the ELBA methodology. The ELBA pipeline efficiently processes the raw data and provides a final prediction regarding whether the individual may potentially develop the disease or not, aiding in early disease detection during clinical screening. 200
- 2 **Overview of Different Normalisation Methods Applied to Isoform Expression.** A comprehensive evaluation of eight normalisation techniques was conducted across all six datasets. Each column corresponds to a specific normalisation method, and each row pertains to a different dataset utilised. The x-axis illustrates the diverse models employed within each normalisation method, while the y-axis portrays the mean AUC score achieved through 5-fold CV. Each model is denoted by dots representing the AUC for each CV fold. Furthermore, a dashed line signifies the mean AUC across the 5 folds, while a solid line represents the mean AUC across all models. 201
- 3 **Confusion Matrices for the Six Analysed Datasets.** Each confusion matrix illustrates the prediction outcomes for the respective test set of each dataset. Dataset names are specified in the title of each matrix. The x-axis denotes the predicted classes, and the y-axis represents the actual classes. The colour intensity within each quadrant of the confusion matrix corresponds to the percentage of instances for each class. Darker colours indicate a higher percentage, while lighter colours denote a lower class percentage, offering a visual representation of the soft voting classification performance. 202
- 4 **Classification of Transcripts Based on TALON Analysis.** The schematic illustrates various categories of transcripts identified: Known isoforms (blue) represent exact matches to established transcript models; Incomplete splice match (ISM) transcripts (green) are partially matching sequences, further divided into prefix (aligning at the start), suffix (aligning at the end), or both; Novel in catalog (NIC) transcripts (yellow) connect known splice sites in new configurations; Novel not in catalog (NNC) transcripts (red) contain at least one novel splice junction; Genomic transcripts (light blue) are typically excluded due to partial overlap or potential DNA contamination; Antisense transcripts (grey patterned) are transcribed in the opposite direction to known genes. 221

- 5 **Transcript Type Composition Across Samples.** The bar plot summarises the transcript composition for non-transformed and cancer samples based on the reference genome annotation and TALON analysis. It displays the proportion of reads categorised as protein-coding, non-coding RNAs (ncRNAs), long non-coding RNAs (lncRNAs), pseudogenes, novel transcripts, artifacts, and those requiring confirmation. Each bar represents a different sample, allowing for a comparative view of transcript types between non-transformed and cancerous tissues. Please note that the statistics are derived from the filtered expression matrix. 222
- 6 **Distribution of Transcript Categories Across Samples.** The circular bar chart depicts the proportion of known and novel transcript categories for each sample, including cancerous and non-transformed tissues. Each segment represents a category such as known transcripts, ISM Prefix, ISM Suffix, ISM Both, NIC, NNC, Antisense, and Intergenic, with the length of the bar corresponding to the percentage composition of that category within the sample. The chart provides an overview of the transcriptomic landscape between the different sample types. 223
- 7 **Heatmap of Gene Expression Variability in NSCLC.** This heatmap showcases the top 100 most variable genes across non-transformed and cancerous lung tissue samples. Each row represents a gene, and each column corresponds to a sample. The colour gradient from blue to red indicates the expression level from low to high, standardised across samples. Hierarchical clustering on both genes and samples illustrates the relative similarity of expression patterns, with the dendrogram on the top and left reflecting the clustering results. 224
- 8 **Heatplot Visualisation.** The heatplot presents a visual mapping between genes and biological concepts through a heatmap representation. In instances where the network of gene-concept connections is exceedingly intricate, particularly when numerous significant terms are involved, the heatplot provides a streamlined perspective. This streamlining aids in the clearer identification of expression trends. 225
- 9 **Gene Ontology and Pathway Enrichment Analysis for DE Novel Transcripts.** (A) The top panel shows the enrichment analysis of up-regulated novel transcripts, identifying significant biological processes, molecular functions, and cellular components, with the degree of enrichment indicated by the negative log₁₀ p-value. (B) The bottom panel presents a similar enrichment analysis for down-regulated novel transcripts, highlighting differentially involved pathways and biological terms. The color coding corresponds to various categories of gene ontology and pathways, while the size of the dots represents the magnitude of enrichment. 226

- 10 **Detailed visualisation of all captured isoforms of the AGER gene.** The figure provides a comprehensive genomic landscape of the AGER gene, outlining both novel and known isoforms. On the left, novel isoforms are cataloged with their respective names, detailing the type of isoform (NNC or ISM Suffix), and an indication of CAGE support for their 5' ends (Yes or No). The central part of the figure graphically maps out the structure of these novel isoforms, juxtaposed with the lower-positioned known isoforms of AGER. Grey circles adjacent to four of these isoforms signify those that were quantified and met the quality and filtering standards of the study, indicating their presence at detectable levels. Conversely, grey squares next to other isoforms represent those that, despite being identified, did not pass the filtering step due to insufficient read counts. The absence of markers alongside the rest of the isoforms suggests that they were not detected in the analysis. The figure, taken from Ensembl Genome Browser, is colour-coded to distinguish between different genomic elements: protein-coding sequences, RNA genes, and processed transcripts are each depicted with a distinct colour as explained in the gene legend. 227
- 11 **NanoInsights Web Service Interface and How to Use Guide.** Screenshot of the NanoInsights web service interface showcasing the 'How to use?' tab. This tab provides comprehensive guidance on navigating and utilising the website efficiently. It includes detailed explanations of each accessible variable that users can adjust to tailor their experience. The guide offers insights into the functionality of the website, enabling users to make informed decisions and optimise their interactions with the platform. 240
- 12 **Standard NanoString QC Metrics for Gene Expression Analysis (Training Set).** This multi-panel figure showcases the standard QC metrics for a NanoString gene expression analysis. **(A)** The Imaging QC boxplot displays the field of view uniformity across different cartridges. The red line in 0.75 represents the minimum limit. Sample below this limit will be potential outliers. **(B)** The Binding Density QC boxplot shows the level of image saturation. Samples outside the red lines (indicating the low and upper limit) would be potential outliers. **(C)** The Positive Control Linearity QC boxplot shows the linearity of the assay across a range of positive control concentrations. These positive controls are typically used to measure the efficiency of the hybridization reaction. Sample with a linearity below 0.95 (red line), would be indicated as potential outliers. **(D)** The Limit of Detection QC boxplot indicates the sensitivity of the assay in detecting low-abundance targets. Each panel delineates the QC metrics across various cartridge IDs, providing a comprehensive overview of the assay's performance and reliability. . 241

-
- 13 **PCA Plot Highlighting Batch Variation in the Training Set.** This Principal Component Analysis plot visualises the variance in gene expression data across different batches, represented by CartridgeIDs. Each point corresponds to a sample, with its position reflecting the sample's score on the first two principal components that together explain 45% of the variance (PC1: 37%, PC2: 8%). The colour coding indicates the batch each sample belongs to, providing a clear illustration of batch effects within the dataset. . . . 242

List of Tables

4.1	Implementation details on the six biofeature types utilised in this study.	66
4.2	Concise summary of genes recurring in more than three, up to six out of six datasets. The table includes the following attributes: GeneID, Gene Name, Gene Type, Occurrence (indicating in how many datasets the gene was selected as a feature), and a brief description of each gene’s function.	83
5.1	Direct RNA sequencing general overview.	102
6.1	Comprehensive Overview of Normalisation Methods and Their Explanations in Data Analysis	128
6.2	Detailed Clinical Information of Patients in Training and Test Cohorts from the Study.	137
1	List of software and packages used in the ELLBA software.	203
2	Detailed machine learning configuration.	204
3	Detailed overview of the publicly available datasets utilised in the study	204
4	Number of extracted features per dataset prior any filtering steps .	204
5	Summary of features selected for each dataset using the Genetic Algorithm.	205
6	Machine learning output stats per dataset	208
7	Table of genes exhibiting recurring appearances exceeding a threshold of two occurrences across all datasets.	209
8	Significantly Differentially Expressed Genes in Non-Transformed vs. Cancer samples.	228
9	Significantly Differentially Expressed Transcripts in Non-Transformed vs. Cancer.	232
10	Significantly Differentially Used Transcripts in Non-Transformed vs. Cancer samples.	235
11	Significantly Alternative Polyadenylated Genes in Non-Transformed vs. Cancer samples.	236
12	Comprehensive Inventory of Packages Employed in NanoInsights. .	243
13	Gene Features Selected for Predictive Modelling in Chemoradiotherapy Response.	244
14	Comparative Performance Metrics of the Utilised Classifier Models.	244

Abbreviations

A3	Alternative 3'
A5	Alternative 5'
AF	Alternative F irst Exon
AL	Alternative L ast Exon
APA	Alternative P olyadenylation A nalysis
AS	Alternative Splicing
AUC	Area Under the ROC C urve
ccRCC	clear c ell R enal C ell C arcinoma
CEC	Circulating E pithelial C ell
circRNA	c ircular R NA
CM	Confusion M atrix
CPM	Count P er M illion
CRC	Colorectal C ancer
CSF	Cerebro S pinal F luid
CSS	Cascading S tyle S heets
CT	Computed T omography
CTC	Circulating T umour C ell
ctDNA	circulating tumour D NA
ctRNA	circulating tumour R NA
CV	Cross- V alidation
DAPA	Differential Alternative P olyadenylation A nalysis
DE	Differential E xpression
DGE	Differential G ene E xpression
DPA	Differential P olyadenylation A nalysis

DRS	D irect R N A S equencing
DTE	D ifferential T ranscript E xpression
DTU	D ifferential T ranscript U usage
ECM	E xtra C ellular M atrix
ELLBA	E nsemble L earning for L iquid B iopsy A nalysis
ESCC	E sophageal S quamous- C ell C arcinoma
ETC	E xtra T rees C lassifier
EV	E xtracellular V esicle
exLR-Seq	E xtracellular V esicles L ong R N A S equencing
FoCT	F raction of C anonical T ranscript
FQ	F ull Q uantile
GB	G radient B oosting
GBM	G lio B lasto M a
GCO	G lobal C ancer O bservatory
GO	G ene O ntology
GSEA	G ene S et E nrichment A nalysis
HCA	H ierarchical C lustering A nalysis
HCC	H epatoc C ellular C arcinoma
IARC	I nternational A gency for R esearch on C ancer
IHC	I mmuno H isto C hemistry
IQR	I nter Q uartile R ange
KNN	K - N earest N eighbours
LARC	L ocally A dvanced R ectal C ancer
LB	L iquid B iopsy
lbRNA-Seq	liquid biopsy-derived R N A - S eq
lncRNA	long non-coding R N A
LOOCV	L ease- O ne- O ut C ross- V alidation
LR	L ogistic R egression
MDS	M ulti D imensional S caling
miRNA	m icro R N A
MK	M ega K aryocytes

ML	Machine Learning
MRI	Magnetic Resonance Imaging
MRLE	mean Relative Log Expression
mRNA	messenger RNA
mRNA-Seq	mRNA Sequencing
MX	Mutually Exclusive Exons
NA	Not Applicable
ncRNA	non-coding RNA
NGS	Next Generation Sequencing
NSCLC	Non-Small Cell Lung Cancer
ONT	Oxford Nanopore Technologies
PCA	Principal Component Analysis
PDAC	Pancreatic Ductal AdenoCarcinoma
PET	Positron Emission Tomography
PFI	Permutation Feature Importance
PSI	Percent Spliced In
QC	Quality Control
RCS	RNA Calibration Strand
RF	Random Forest
RFECV	Recursive Feature Elimination with Cross-Validation
RI	Retained Intron
RLE	Relative Log Expression
RNA-Seq	RNA Sequencing
ROC	Receiver Operating Characteristic
RPKM	Reads Per Kilobase per Million mapped reads
SBS	Sequencing By Synthesis
SCLC	Small Cell Lung Cancer
scRNA-Seq	single-cell RNA Sequencing
SE	Skipped Exon
SGS	Second Generation Sequencing
SNP	Single Nucleotide Polymorphisms

SNV	S ingle N ucleotide V ariant
TBLB	T rans B ronchial L ung B iopsy
TEP	T umour- E ducated P latelet
TGS	T hird G eneration S equencing
TME	T umour M icro E nvironment
TMM	T rimmed M ean of M -values
TNM	T umour, N ode, and M etastasis
TPR	T rue P ositive R ate
TSS	T ranscription S tart S ite
UQ	U pper Q uartile

Chapter 1

Introduction

This Chapter sets the groundwork for the thesis by underscoring the critical necessity of progressing in cancer diagnostic techniques, emphasising lung cancer due to its high mortality rate. It navigates the reader from the existing diagnostic methodologies to the emergent realms of biopsy and the integration of sequencing technologies into oncological research. Additionally, it acquaints the reader with key concepts in bioinformatics and machine learning, essential for understanding the forthcoming content of the thesis. Concluding the chapter is a synopsis of each subsequent chapter.

1.1 The Global Cancer Landscape

Cancer presents a significant health challenge, manifesting as a complex disease with a diverse array of pathologies that disrupt the normal function of cells. In 2020 alone, an estimated 19.3 million new cases were diagnosed globally, marking a 2.9% increase from 2018 and underscoring the persistent and escalating burden of cancer. The Global Cancer Observatory (GCO) further reports a distressing 10.0 million cancer-related deaths in 2020, reflecting a somber 10.6% mortality rate [1].

As per the Global Cancer Statistics (GLOBOCAN 2020) estimates, female breast cancer has surpassed lung cancer and now holds the top position as the most commonly diagnosed cancer, accounting for an estimated 2.3 million new cases (11.7%). Lung cancer closely follows with 11.4%, while colorectal, prostate, and

stomach cancers constitute 10.0%, 7.3%, and 5.6%, respectively (Figure 1.1). Despite this shift, lung cancer maintains its unfortunate status as the leading cause of cancer-related deaths, claiming approximately 1.8 million lives (18%). Colorectal, liver, stomach, and female breast cancers follow, with mortality rates of 9.4%, 8.3%, 7.7%, and 6.9%, respectively [2].

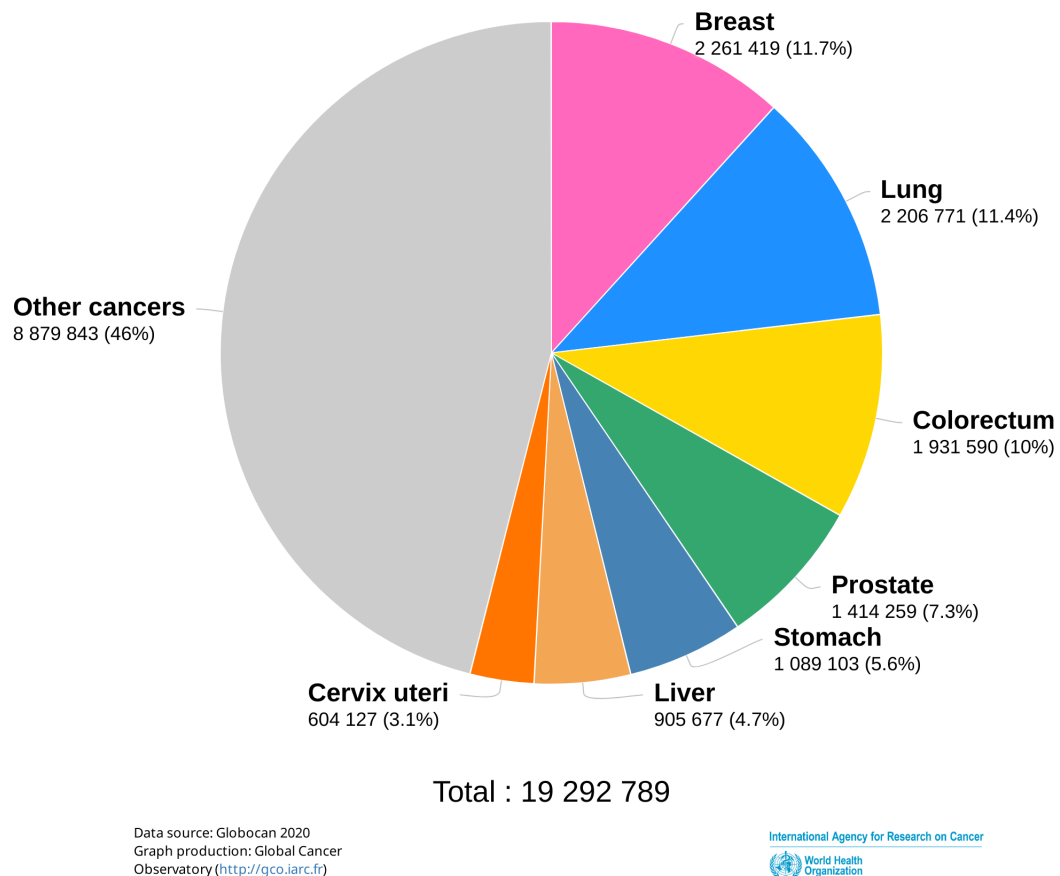


FIGURE 1.1: **Worldwide Cancer Case Estimates in 2020 by Globocan.** Estimated global cancer occurrences for 2020, covering individuals of every gender and age group across all continents.

Globally, it is anticipated that there will be approximately 28.4 million new cases of cancer in 2040, marking a substantial 47% surge compared to the recorded 19.3 million cases in 2020 [1]. Addressing these imbalances and enhancing cancer outcomes necessitates substantial research efforts. There is a crucial requirement for the advancement of methodologies focusing on early diagnosis, prevention, and treatment.

1.1.1 Understanding Cancer

Cancer, a diverse and complex group of diseases, manifests in various forms, impacting nearly every organ and tissue in the human body. It is a genetic disease and arises from the uncontrolled growth and division of cells, resulting in the formation of tumours that can invade surrounding tissues and metastasise to other parts of the body [3]. This intricate process involves a disruption in the normal regulatory mechanisms governing cell growth and apoptosis (programmed cell death) [4].

Despite considerable progress in cancer research, the exact causes of many cancers are still not fully understood. Factors such as genetic predisposition, exposure to environmental carcinogens, and lifestyle choices collectively influence cancer development, highlighting the disease's complexity [5]. This intricate interplay of factors poses significant challenges to both the understanding and treatment of cancer. These diverse elements interfere with the cellular signalling pathways essential for maintaining cellular balance, further complicating cancer's nature [6].

Tumour outcomes exhibit significant variation, ranging from benign growths to aggressive and life-threatening malignancies. Precancerous and early-stage lesions also demonstrate diverse biological features, spanning from favourable to unfavourable outcomes, with potential for malignant transformation. Conversely, advanced cancers often present with unfavourable (sub)clones, posing greater challenges for treatment. Early detection plays a crucial role in improving outcomes by enabling timely intervention. Nevertheless, certain cancers may progress silently, underscoring the importance of regular screenings and heightened awareness [7].

The complexity of cancer stems not only from diverse origins involving different cell types, but also from distinct molecular and cellular processes driving its progression [8]. Different cancer types exhibit unique characteristics, response patterns, and diagnostic and treatment challenges [9], [10]. This intricate understanding sets the stage for exploring the subsequent aspect of cancer care, delving into the classification and description of cancer stages, a crucial step in guiding treatment decisions and predicting patient prognosis [11].

1.1.2 Cancer MicroEnvironment

The tumour microenvironment (TME) is a dynamic and intricate ecosystem surrounding cancer cells, comprising cellular and non-cellular components (Figure 1.2). This intricate network not only supports tumour growth and invasion but also facilitates immune evasion, angiogenesis, and resistance to treatments, contributing to a poor prognosis. The TME's complexity is further amplified by the ongoing interactions within it; cancer cells manipulate the microenvironment to promote angiogenesis and immune tolerance, while immune cells can influence the tumour's growth dynamics [12]–[15]. This constant evolution underscores the challenges and opportunities in targeting the TME for cancer therapy [16].

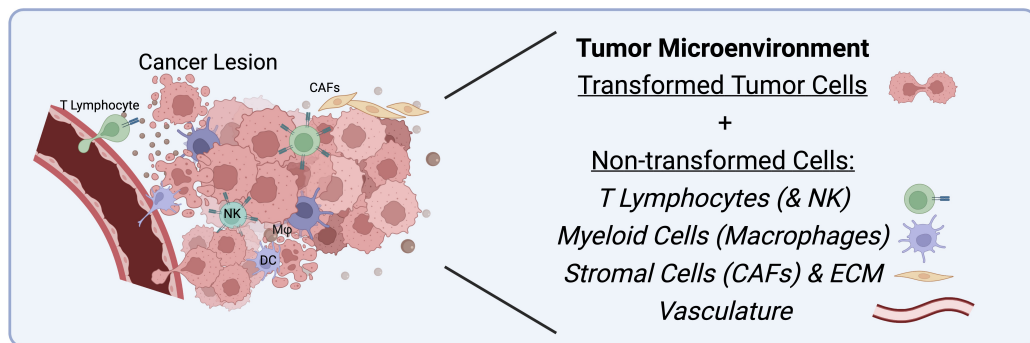


FIGURE 1.2: **Illustration of the Tumour Microenvironment.** The figure provides a comprehensive depiction of the TME, capturing both the transformed tumour cells and the non-transformed cells that constitute the intricate tumour landscape. Adapted from the work of Belli, Antonarelli, Repetto, *et al.* 2022, originally published in *Cancers* [17].

At the core of the tumour microenvironment lies the transformed tissue, where cancer cells undergo uncontrolled proliferation and evade the normal regulatory mechanisms that govern cellular behaviour. Genetic mutations and epigenetic changes drive the transformation, leading to the hallmark characteristics of cancer, such as sustained angiogenesis, resistance to cell death, and the ability to invade surrounding tissues. The transformed tissue becomes a nexus for intricate signalling pathways and interactions that shape the TME [13]. Moreover, the development of the tumour entails sophisticated interactions with various components of the blood microenvironment.

Surrounding the transformed tissue, non-transformed stromal cells, including fibroblasts, immune cells, and endothelial cells, actively participate in the tumour microenvironment. Fibroblasts contribute to the formation of the extracellular

matrix (ECM) and facilitate tumour cell invasion [18]. Immune cells, both infiltrating immune cells and those residing in lymphoid organs, engage in a dynamic interplay with cancer cells, impacting the balance between tumour promotion and suppression [19]. Endothelial cells play a crucial role in angiogenesis, ensuring the tumour receives a sufficient blood supply for sustained growth [20].

1.1.3 Overview of Lung Cancer

Lung cancer is marked by uncontrolled cell growth in the lungs, making it the primary contributor to global cancer-related deaths. The histological examination's morphological subclassification of lung cancer historically recognised two distinct subtypes: small cell lung cancer (SCLC) and non-small-cell lung cancer (NSCLC), each exhibiting markedly different clinical characteristics [21]. NSCLC further undergoes subdivision based on pathological criteria, encompassing broad categories such as adenocarcinoma, squamous cell carcinoma, and other histological subtypes like large cell carcinoma (Figure 1.3). According to the American Cancer Society, NSCLC is the predominant type, constituting approximately 80% to 85% of all lung cancers. In contrast, SCLC, although less common, tends to be more aggressive and has a higher likelihood of metastasis to other parts of the body [22].

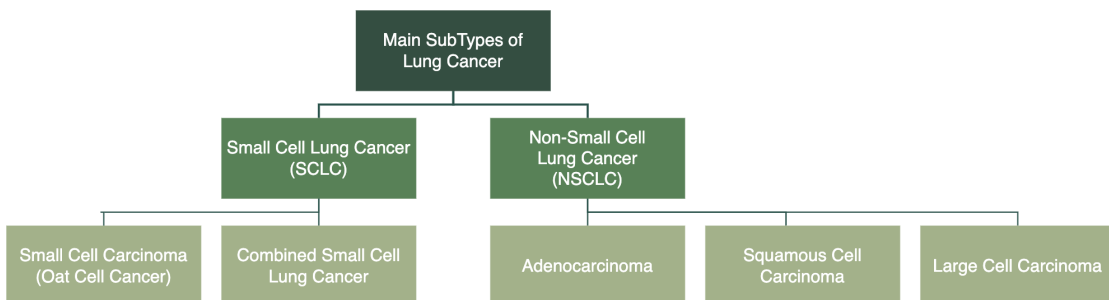


FIGURE 1.3: **Lung Cancer Subtypes.** Overview of lung cancer's main histological types: SCLC and NSCLC. SCLC includes small cell carcinoma and mixed small cell/large cell cancer, typically associated with rapid growth and strong correlation with cigarette smoking. NSCLC encompasses adenocarcinoma (predominantly found in the lung's outer regions), squamous cell carcinoma (typically located near bronchial tubes), and large cell carcinoma (with a tendency for rapid growth).

The risk factors for lung cancer include smoking, exposure to secondhand smoke, radon gas, asbestos, and certain occupations, such as mining and construction.

People who have a family history of lung cancer are also at increased risk of developing the disease [23].

The diagnosis of lung cancer is made through a variety of tests, including a chest X-ray, computed tomography (CT) scan, positron emission tomography (PET) scan, magnetic resonance imaging (MRI), bronchoscopy, and biopsy [24].

The treatment for lung cancer depends on the type, stage, and patient's overall health. Treatment options include surgery, chemotherapy, radiation therapy, immunotherapy, and targeted therapies [25]–[27].

1.1.4 Diagnostic Challenges of Lung Cancer

Navigating the intricate landscape of lung cancer diagnosis, from its silent inception to diverse histological variations, necessitates a comprehensive approach [28], [29]. This involves leveraging advanced imaging technologies, molecular profiling, and a nuanced understanding of clinical presentations [24].

Lung cancer poses a formidable challenge in early detection due to its asymptomatic nature during the initial stages. Patients may remain oblivious to the disease until it advances to a less manageable state [30], underscoring the need for innovative screening strategies and increased awareness among healthcare professionals and the general populace.

The emergence of symptoms, when they do occur, often mirrors those of other respiratory conditions, such as chronic obstructive pulmonary disease (COPD) or pneumonia [31], [32]. Persistent cough [33], shortness of breath [34], and chest pain [35], while indicative, overlap with various respiratory ailments, complicating the identification of lung cancer without thorough diagnostic scrutiny.

Even though current diagnostic tools are crucial, they come with their own set of challenges. Chest X-rays [36], once primary, lack sensitivity in early-stage lung cancer, while CT scans, though more sensitive, present radiation exposure and a risk of false positives [37]. PET scans, providing functional insights, are limited by cost and availability [38].

Although tissue biopsy is still considered the gold standard, it comes with several limitations. Transbronchial biopsies (TBLB) entail the risk of sampling error,

along with significant inconsistencies or disagreement among various pathologists. This variability can result in false negatives, compromising the reliability of the diagnostic process [39]. Moreover, thoracoscopic biopsies, while yielding larger samples, are more invasive with increased risks [40].

Innovations in diagnostics are addressing these challenges. Emerging imaging modalities, like low-dose CT scans and AI-driven analysis, enhance sensitivity in early detection [41]. Evolving biopsy techniques, including minimally invasive bronchoscopic procedures [42] offer less invasive approaches for obtaining cancer cells from the bloodstream [43].

In the realm of novel diagnostic approaches, biomarkers, and liquid biopsies stand out. Analysing blood-based biomarkers offers non-invasive avenues for detecting genetic alterations associated with lung cancer, promising early detection and treatment response monitoring. Ongoing research focuses on standardisation, sensitivity, and specificity of these techniques [43].

1.2 Tissue Biopsy: Unraveling its Process and Challenges

Tissue biopsy plays a pivotal role in the diagnosis and characterisation of cancer, providing essential information for treatment planning and prognosis. While imaging tests, such as CT scans or MRIs, are valuable in identifying masses or irregular tissue, they fall short in distinguishing between cancerous cells and non-cancerous ones. In cancer diagnosis, a tissue biopsy involves the removal of a small sample of abnormal tissue or cells from the body, typically from a tumour or a suspicious lesion. This procedure allows pathologists to examine the tissue under a microscope, identifying cellular abnormalities, and determining the cancer type, grade, and extent of spread.

Tissue biopsy in cancer is differentiating between benign and malignant tumours, and understanding the specific genetic and molecular characteristics of the tumour. Additionally, tissue biopsy aids in the assessment of the tumour's aggressiveness, guiding oncologists in tailoring treatment strategies. Through techniques such

as immunohistochemistry and molecular profiling, tissue biopsy enables the identification of specific biomarkers, paving the way for targeted therapies that can address the unique features of an individual's cancer [44].

However, the application of tissue biopsy is not without challenges. In their work, Ilić and Hofman [45] have highlighted some key points:

- **Invasiveness of the Procedure:** Tissue biopsy is inherently invasive, carrying potential risks for patients, especially when tumours are located in sensitive or challenging anatomical areas. The invasiveness of the procedure can lead to complications, discomfort, or additional health risks for individuals undergoing biopsy, making it a critical consideration in the overall diagnostic process.
- **Difficulty in Obtaining Adequate Tissue Samples:** Obtaining a sufficiently large and representative tissue sample is essential for accurate diagnosis. However, challenges arise, particularly in cases where tumours are small or exhibit heterogeneity. The heterogeneous nature of tumours, where different parts show distinct genetic features, complicates the sampling process and may result in an incomplete representation of the cancer's characteristics. This limitation underscores the need for precision in sample collection.
- **Tumour Heterogeneity:** Tumour heterogeneity adds a layer of complexity to the analysis of biopsy samples. Different regions within a tumour may display diverse genetic features, making it challenging to capture the full spectrum of the cancer's genetic makeup in a single biopsy. This diversity can impact treatment decisions, as a biopsy from one part of the tumour may not accurately represent the entire disease, potentially leading to suboptimal therapeutic strategies.
- **Evolution of Tumour Characteristics Over Time:** Studies have revealed the dynamic nature of tumour characteristics, demonstrating the emergence of treatment-resistant subclones. These subclones, present at a minimal frequency in the primary tumour, may become more predominant over time. This evolving nature underscores the importance of considering temporal changes in tumour features, emphasising the need for timely and repeated biopsies to inform treatment decisions effectively.

- **Timing and Impact on Treatment Decisions:** The timing of tissue biopsy is crucial, and delays in obtaining results can have a significant impact on treatment decisions. In some cases, patients may need repeat biopsies to capture changes in the tumour over time or in response to treatment. This iterative process can be physically and emotionally challenging for individuals, highlighting the importance of streamlining the diagnostic timeline to minimise the burden on patients facing a cancer diagnosis.

1.3 Liquid Biopsy: Unveiling New Frontiers

In the expansive domain of diagnostics and treatment, a groundbreaking innovation has surfaced – liquid biopsy. Departing from the invasive nature of traditional tissue biopsies, liquid biopsy leverages various bodily fluids to unlock critical insights into a spectrum of health conditions, notably cancer. Liquid biopsy seek to complement traditional tissue biopsy methods, providing clinicians with a broader toolkit for comprehensive cancer care while minimising the burden on patients.

This technique signifies a paradigm shift, offering a minimally-invasive, and sometimes non-invasive method for detecting molecular alterations associated with diverse diseases. Unlike its conventional counterpart, liquid biopsy presents an innovative approach that aligns with the principles of precision medicine.

Several biofluids exhibit potential utility for liquid biopsy, including saliva [46], urine [47], peritoneal fluid [48], cerebrospinal fluid (CSF) [49], seminal fluid [50], and more. However, among these options, blood stands out as the most commonly used biofluid for liquid biopsies. Its prevalence stems from the ease of collection and the abundance of biosources within its dynamic composition.

Liquid biopsy emerges as a valuable and versatile tool, holding the promise of transforming diagnostic landscapes and overcome the limitations of tissue biopsy. Its applications extend beyond traditional boundaries, offering unprecedented opportunities for early cancer detection [51], tailoring personalised treatment options [52], assessment of treatment response and resistance [53], [54], and monitoring cancer progression [55].

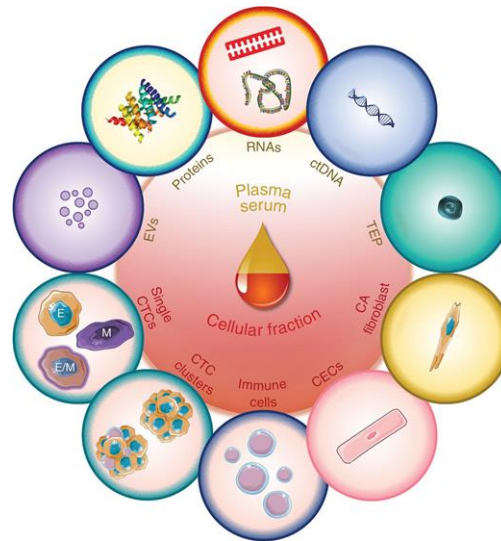


FIGURE 1.4: **Circulating Biomarkers Revealed Through Blood-Based Liquid Biopsy.** It highlights circulating components in plasma or serum (top), including EVs like Exosomes, Proteins, circulating cell-free RNA (encompassing both noncoding and messenger RNA), ctDNA, and TEPs. Simultaneously, the cellular fraction (bottom) unveils: i. tumour cells (CTCs, either singular or clustered), and ii. the non-tumour cell fraction, such as immune cells, CECs, and Cancer-Associated Fibroblasts.

1.3.1 Circulatory Biomarkers in Blood-Based Liquid Biopsy

A key aspect of blood-based liquid biopsy is the analysis of genetic materials circulating in the bloodstream to detect cancer. These biomarkers include whole cells, cellular fragments, and tumour-released molecules. Figure 1.4, published in *Cancer Discovery* by Alix-Panabières and Pantel in 2021 [56], showcases these circulating biomarkers, encompassing cells from both primary and metastatic tumours, as well as molecules and metabolites emitted by tumour cells. Such biomarkers yield insights into tumour heterogeneity, mechanisms of metastasis, and resistance to treatments [57], [58], and they mirror the tumour’s genomic and proteomic profiles. Analysing these markers provides a real-time snapshot of the tumour’s genetic makeup, unveiling crucial genomic alterations and deepening our understanding of cancer’s biology, including its progression, behaviour, and the dynamics of cellular communication within the tumour microenvironment [59]. This comprehensive insight facilitates the development of targeted therapies, significantly advancing cancer treatment strategies [60].

Since the inception of liquid biopsy development, a multitude of biosources has been harnessed, falling within the categories mentioned earlier. Among these,

the most prevalent include circulating tumour DNA (ctDNA), circulating tumour RNA (ctRNA), circulating tumour cells (CTCs), extracellular vesicles (EVs), Tumour-Educated Platelets (TEPs), Circulating Epithelial Cells (CECs), and others [56], [61]. However, this thesis will revolve around biosources endowed with a transcriptome, specifically TEPs, EVs, and CECs.

1.3.2 Tumour-Educated Platelets

In recent years, the quest for innovative biomarkers in cancer diagnosis and monitoring has spotlighted a promising entity: platelets (Figure 1.5). These minute anucleate cells originate from megakaryocytes (MKs) within the bone marrow and lung niches [62].

Platelets, the second most abundant anucleate cells in circulation, after red blood cells, boasting an average lifespan of merely 7 days [63]. Beyond their primary role in homeostasis, platelets assume significance in tumourigenesis and tumour progression [64]. They orchestrate tumour angiogenesis and vascular remodelling, shield CTCs from shear forces, elude immune surveillance, and orchestrate the recruitment of stromal cells to foster the establishment of metastatic niches and advance metastasis. Conversely, tumours can also influence platelets, triggering their activation, aggregation, and release of platelet-derived substances into circulation. This bidirectional interaction between tumours and platelets results in the systematic and local responses of platelets to cancer. Simultaneously, platelets absorb free proteins, nucleic acids, vesicles, and particles [65]–[68], leading to alterations in their RNA, DNA (comprising genomic DNA fragments [69]) and proteomic expression profiles [70]–[74]. This phenomenon is termed "tumour-educated platelets" [64].

The alterations in the profile of TEPs constitute a concentrated biorepository abundant with tumour-derived and bioactive molecules, suggesting the potential of TEPs as cancer-specific biomarkers. Consequently, their content mirrors the tumour's current state and bioactivity, making TEPs crucial for detecting and tracking cancer progression across various types, including colorectal carcinoma, glioblastoma, and NSCLC [75].

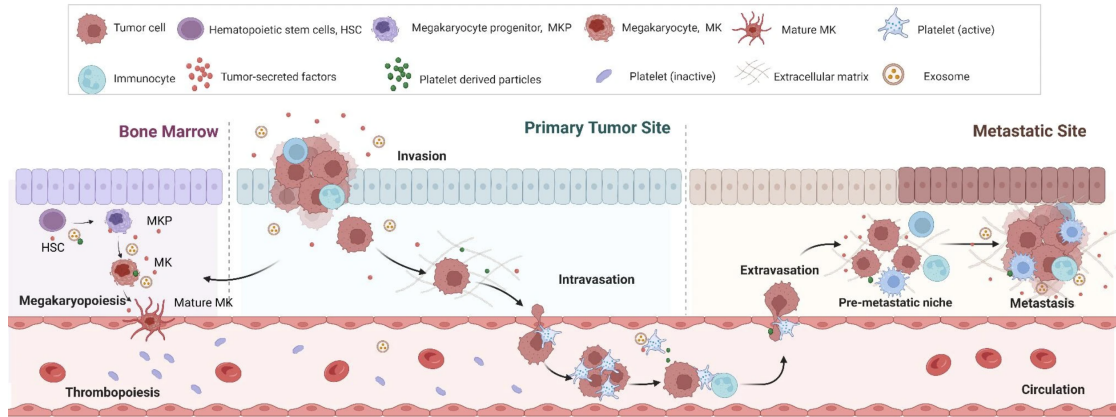


FIGURE 1.5: **Reciprocal Influences Between Cancer and Platelets.** This illustration, based on research by Ding, Dong, and Song 2023 published in *Cancer Cell International* [75], shows the complex interactions between cancer and platelets. It focuses on how tumours educate platelets, leading to their activation, aggregation, and release of substances, which promotes thrombocytosis by affecting bone marrow megakaryopoiesis. In turn, platelets support tumour growth and spread through angiogenesis, vascular remodelling, protecting CTCs, evading immune detection, and recruiting stromal cells. Critical elements in these processes are megakaryocyte progenitor, MKs, and hematopoietic stem cells.

Utilising TEPs in liquid biopsies promises advancements in early cancer detection and subtyping, offering a window into tumour heterogeneity for a more precise disease characterisation [74], [76]. As non-invasive biomarkers, TEPs could significantly improve cancer diagnosis, especially when traditional biopsies pose challenges.

Moreover, TEPs serve as a dynamic tool for monitoring treatment responses and forecasting prognosis. Changes in TEP profiles during treatment reveal insights into therapeutic success and potential resistance [77], supporting personalised treatment adjustments for better patient outcomes.

1.3.3 Extracellular Vesicles

Extracellular Vesicles (EVs) have emerged as significant players in liquid biopsy, offering a wealth of information for cancer diagnosis and monitoring. EVs encompass a diverse array of membranous structures, including exosomes, microvesicles, and apoptotic bodies (Figure 1.6), released by various cells into the extracellular space. These minute particles, enclosed by a lipid bilayer, have been identified in

numerous biological fluids, with blood being particularly abundant in these entities [78].

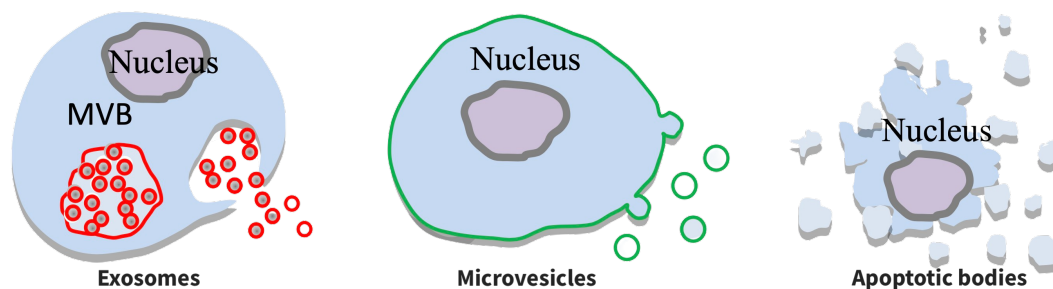


FIGURE 1.6: **Representation of distinct extracellular vesicles subtypes.** Modified from the work of Fox, Kim, Le, *et al.* 2015, originally published in the *Journal of Controlled Release* [79]. The figure highlights exosomes, microvesicles, and apoptotic bodies, each presenting distinct characteristics and roles.

One of the remarkable features of EVs is their stability, both in terms of morphology and chemical properties. The lipid bilayer enveloping EVs provides protection for their cargo, shielding it from extracellular proteases and enzymes. This stability ensures that the molecules trapped within EVs remain intact, making them reliable carriers of biomarkers indicative of the originating cells' health or disease condition [80].

EVs play pivotal roles in intercellular communication, acting as messengers that transport a diverse cargo, including proteins, lipids, nucleic acids (RNA and DNA), and various bioactive molecules. Originating from different cellular sources, including tumour cells, EVs can sensitively reflect an individual's health status. Notably, tumour cells release EVs, and accumulating evidence suggests that these tumour-derived EVs contribute to processes such as metastasis, angiogenesis, and chemotherapy resistance [81].

The contents carried by EVs, spanning proteins, microRNAs (miRNAs), messenger RNAs (mRNAs), circular RNAs (circRNAs), DNA, and lipids, offer a comprehensive range of biomolecules for assessing the health or disease state of the originating cells. Nucleic acids and proteins encapsulated within EVs provide a snapshot of the genetic and proteomic signatures of cells, contributing to a deeper understanding of underlying biological processes [80].

In the context of liquid biopsy, EV-based approaches present a promising avenue for gaining valuable insights into tumour progression and the characteristics of

the tumour itself. By analysing the molecular cargo carried by EVs in bodily fluids, such as blood, researchers and clinicians can potentially detect cancer at early stages, tailor personalised treatment options, monitor treatment response, and track cancer progression with a level of precision previously unimaginable [82].

Moreover, EVs' involvement in various biological cancer processes, including cell growth, proliferation, and migration, through the transfer of cargos between different cells, underscores their significance in cancer research and diagnostics [83]. Leveraging EVs for liquid biopsy holds great potential for advancing precision medicine and transforming the landscape of cancer diagnostics and monitoring. As research in this field progresses, the integration of EV-based liquid biopsy into clinical practice could usher in a new era of non-invasive and highly informative cancer diagnostics.

1.3.4 Circulating Epithelial Cells

Circulating epithelial cells (CECs) have emerged as a compelling focus within the realm of liquid biopsy, offering a unique opportunity to study the dynamic aspects of cancer biology.

CECs refer to epithelial cells that circulate in the bloodstream. Epithelial cells are abundant cells that line the skin, body cavities, and blood vessels. CECs can originate from various parts of the body, with the intestines and lungs being common sources. Unlike white blood cells and other circulating components, CECs carry unique molecular signatures that reflect their tissue of origin. This characteristic makes them valuable for studying the genetic and phenotypic diversity of associated tumours [84]. Understanding the biology of CECs is essential for harnessing their potential as informative biomarkers.

The molecular cargo carried by CECs includes DNA, RNA, and proteins, offering an opportunity to study the genetic and proteomic makeup of tumours through a non-invasive approach [85]. The analysis of CECs allows for the identification of key mutations, gene expression patterns, and alterations reflective of tumour heterogeneity. This biomarker potential positions CECs as valuable targets for liquid biopsy applications.

Isolating CECs from the peripheral blood poses technical challenges due to their rarity and the presence of other blood components. Various isolation methods,

including immunomagnetic separation, microfluidic technologies, and density gradient centrifugation, have been developed to enrich and capture CECs [86]–[88]. These methods aim to maximise purity and viability while minimising contamination from other blood cells.

CECs hold significant promise in the field of cancer diagnosis and monitoring. The detection and molecular characterisation of CECs provide valuable information for early cancer diagnosis, stratification of patients, and monitoring treatment response. CEC-based liquid biopsy offers a real-time and non-invasive approach to understanding the evolving genetic landscape of tumours, enabling clinicians to tailor treatment strategies for optimal outcomes.

In various cancer types, including liver, breast, lung, and colorectal cancers, the presence and characteristics of CECs have been correlated with disease progression, metastasis, and overall survival [89]–[91]. The clinical significance of CECs extends beyond diagnostics to include prognostication, guiding treatment decisions, and assessing the risk of recurrence.

1.4 Sequencing Technologies

The advent of sequencing technologies has precipitated a paradigm shift in the landscape of oncological research. Initially, the field witnessed the emergence of first-generation sequencing modalities, also known as Sanger sequencing, which marked the inception of high-throughput methodologies. This was succeeded by the evolution to second-generation sequencing (SGS), also known as next-generation sequencing (NGS). NGS, particularly epitomised by Illumina’s sequencing technology, marked a revolutionary stride, delivering unprecedented throughput and accuracy. This facilitated a granular exploration of cancer genomics, unveiling critical mutations and genomic variations integral to oncogenesis [92].

NGS has been indispensable in the realm of cancer transcriptomics, facilitating the simultaneous analysis of a multitude of RNA molecules in a cost-effective and efficient manner. This technology has unraveled the convoluted gene expression patterns intrinsic to tumourigenesis, progression, and therapeutic response. A notable aspect of NGS in the study of cancer transcriptomics is its ability to simultaneously profile coding and non-coding RNA species. This enabled researchers to explore and define the functions of various RNA types, including mRNAs, microRNAs

and long non-coding RNAs, in the context of cancer biology. This holistic examination of the transcriptome enables scientists to decipher the complex regulatory networks that govern gene expression in cancer cells. The heightened sensitivity and resolution of NGS are pivotal in detecting rare and low-abundance transcripts, thereby illuminating hitherto obscured facets of cancer transcriptomics [93], [94].

The contemporary era is witnessing the ascendance of third-generation sequencing (TGS), a technology that is poised to further revolutionise our understanding of cancer genomics. TGS introduces long-read sequencing capabilities, with innovations led by entities such as Oxford Nanopore Technologies. This advancement enables the exploration of intricate genomic structures, encompassing repetitive regions, structural variations, and epigenetic markers, which were previously elusive due to the limitations of earlier sequencing technologies [95].

In the context of cancer transcriptomics, TGS is anticipated to provide a profound impact. Such technology allows for the comprehensive examination of extended RNA molecules, thereby enhancing our understanding of transcriptional regulation, RNA modifications, and the interplay among various RNA species. This ability to capture longer RNA sequences is instrumental in elucidating complex aspects of gene expression, alternative splicing, post-transcriptional processing steps, such as polyadenylation, and non-coding RNA dynamics, which are crucial in the pathogenesis of cancer. Consequently, TGS is set to unveil novel layers of genomic information, thereby enriching our understanding of the molecular landscape of cancer and paving the way for more targeted and effective therapeutic interventions [95], [96].

Complementing these sequencing technologies, platforms such as NanoString provide an additional dimension to cancer research by extending their application to clinical settings. NanoString's capability for direct quantification of RNA molecules without amplification offers a precise and robust method for gene expression profiling. This is particularly valuable in validating sequencing results and advancing translational research, thereby contributing significantly to the personalised medicine paradigm in oncology. The integration of NanoString technology alongside advanced sequencing methods represents a significant stride in the quest to decode the complexities of cancer at a molecular level [97].

1.4.1 Second Generation Sequencing

Illumina's technology, known for its innovative sequencing by synthesis (SBS) method, has become a pivotal technology in genomic research. This method has not only transformed the way genomics, transcriptomics, epigenomics, and epitranscriptomics are approached, but also established Illumina's technology as a powerful and widely recognised tool in the scientific community. This innovative method combines the amplification and sequencing of DNA fragments on a solid surface, starting with the preparation of a DNA library. Here, genomic DNA is fragmented, and adapters are ligated to each end, serving both as primers for the sequencing process and facilitators for the subsequent clonal amplification of DNA fragments.

The library is then bound to a flow cell surface where bridge amplification leads to the formation of clusters of identical DNA molecules. In the sequencing phase, fluorescently labeled nucleotides are incorporated into the growing DNA strands. A camera captures the emitted fluorescent signals upon each nucleotide's incorporation, enabling the precise identification of each base and thus, the determination of the DNA sequence. This high-throughput sequencing process, capable of handling millions of fragments simultaneously, significantly enhances the speed, efficiency, and accuracy of the sequencing data [98], [99].

Illumina NGS stands out for its high throughput, allowing the parallel sequencing of many DNA strands, which markedly reduces time and cost per sequence. It is known for its high accuracy, particularly in detecting single-nucleotide polymorphisms (SNPs). This technology is also versatile and scalable, suitable for a broad spectrum of applications, ranging from whole-genome sequencing to targeted re-sequencing [100].

In the realm of transcriptomics, Illumina's RNA sequencing (RNA-Seq) technology is of particular note. It begins with the extraction of RNA from a sample, followed by its conversion into a complementary DNA (cDNA) library. This cDNA is then sequenced using Illumina's SBS method, as shown in Figure 1.7. RNA-Seq is instrumental in unraveling transcriptomic landscapes, uncovering gene expression patterns, splicing variants, and post-transcriptional modifications. The process not only provides insights into gene regulation and function in various biological contexts, but also in disease states [101]–[103].

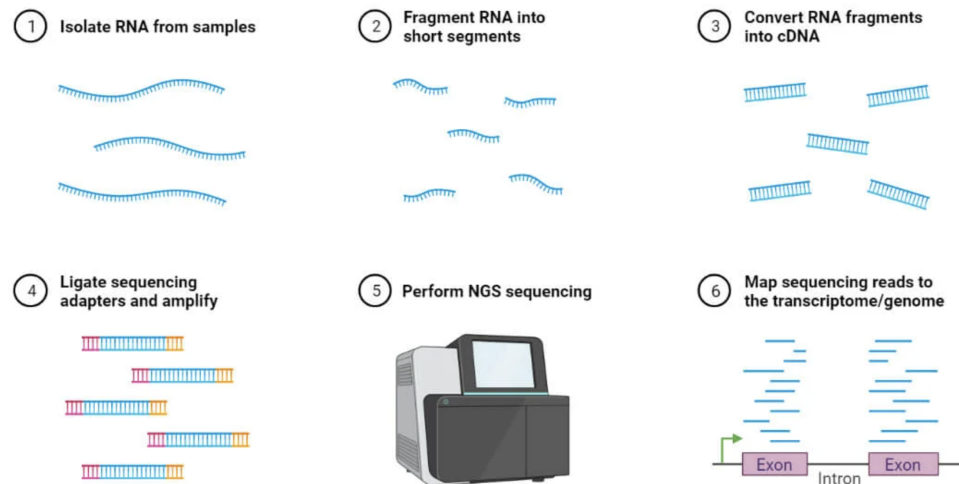


FIGURE 1.7: **Overview of the RNA Sequencing Workflow.** This schematic depicts RNA sequencing via NGS technology in six steps: 1) RNA isolation from samples; 2) fragmentation into short pieces; 3) conversion into cDNA; 4) ligation of sequencing adapters and amplification; 5) sequencing of amplified cDNA fragments; and 6) mapping reads to the genome or transcriptome to identify exon and intron regions. Adapted from Prashant Dahal’s work on RNA sequencing [104].

Although Illumina’s NGS offers numerous benefits, it is not without its challenges. One significant limitation is the generation of short reads, usually between 100 to 300 base pairs, which can complicate the reconstruction of complex genomic regions such as repetitive regions [105]. Moreover, it can exhibit GC bias, leading to the underrepresentation of regions with high or low GC content [106]. While it is more cost-effective compared to earlier technologies, the initial setup and operational expenses of Illumina sequencing are still substantial, especially for large-scale projects.

Nevertheless, Illumina’s NGS technology, despite its drawbacks, remains a vital tool in genomic research. Its contributions to understanding the genetic and transcriptomic make-up of organisms are vast, with a broad application range that continues to evolve, significantly impacting both research and clinical fields.

1.4.2 Third Generation Sequencing

TGS technologies, exemplified by Oxford Nanopore Technologies (ONT), also known as nanopore sequencing, represent a significant advancement in genomics,

offering unique advantages compared to previous sequencing generations. Nanopore sequencing operates on the principle of passing a DNA or RNA molecule through a nanoscale pore, allowing real-time, single-molecule sequencing with long read lengths.

In nanopore sequencing, a biological sample is prepared into a library, and a voltage is applied across the nanopore. As the DNA or RNA molecule traverses the pore, the changes in electrical current are recorded. Each nucleotide induces a characteristic disruption in the current, enabling base identification [107], [108]. Nevertheless, ONT systems do not recognise single nucleotides, as the measured current reflects the properties of short nucleotide sequences. These sequences, typically consisting of about five bases known as k-mers, result in a diverse array of signals – with over 1000 distinct signals corresponding to each unique micropolymer (k-mer) [109]. This real-time, single-molecule approach provides long reads, facilitating the sequencing of complex genomic regions [110], detection of structural variations [111], detection of epigenetic modifications [112], attain complete chromosome assemblies from one telomere end to the other [113], and direct RNA sequencing for comprehensive transcriptomic analysis [108].

The impact of nanopore sequencing in the field of transcriptomics is constantly growing, with ONT's Direct RNA Sequencing (DRS) emerging as a key player. This method directly sequences RNA molecules, eliminating the need for reverse transcription into cDNA (Figure 1.8). DRS maintains the integrity of RNA modifications and captures the complete structure of transcript isoforms. Its effectiveness in simultaneously identifying new transcripts, isoforms, and various RNA modifications at a single-molecule level significantly enhances our understanding of alternative splicing mechanisms [114]. Additionally, DRS excels in accurately quantifying isoforms, and identifying polyadenylation sites [115], promoters, and splice sites [116], offering a more true-to-nature view of the transcriptome by avoiding biases inherent in cDNA synthesis.

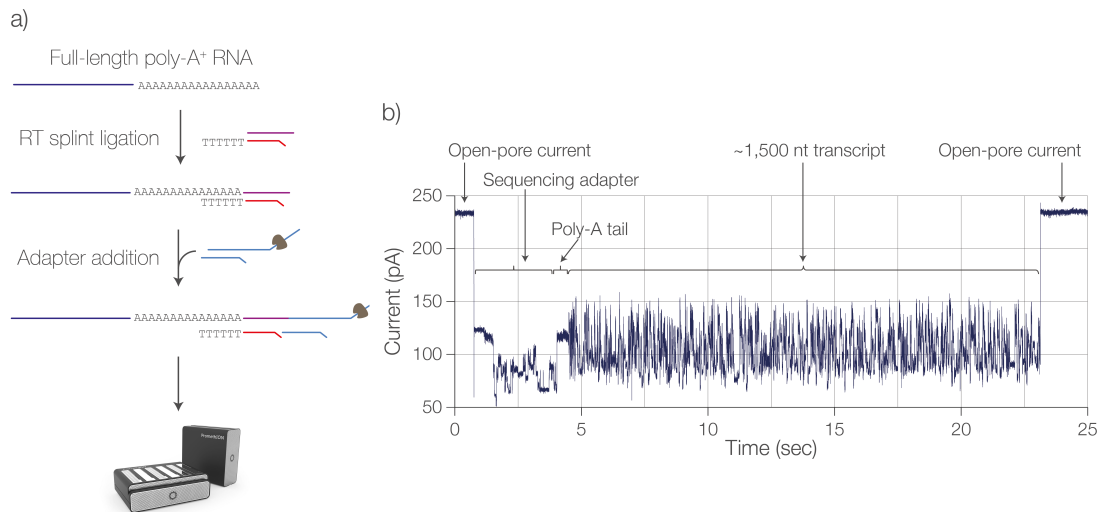


FIGURE 1.8: RNA sequencing using nanopore technology. This illustration, edited by ONT [117], outlines the direct RNA Sequencing Workflow: a) Library Preparation details full-length poly-A RNA preparation for nanopore sequencing, beginning with ligation of an reverse transcription splint (red) to RNA (blue), followed by sequencing adapter attachment (blue with brown tips) for nanopore entry. b) Ionic Current Trace shows the ionic current changes as the RNA-transcript passes through the nanopore, highlighting the baseline open-pore current, adapter and poly-A tail disruptions, a longer disruption for the 1,500 nt transcript, and a return to baseline after transcript passage.

This approach paves the way for major breakthroughs in diverse areas such as personalised medicine, gene therapy, pharmacogenomics, medical sequencing for viruses and microbes, as well as in epigenetics and cancer research. DRS's potential to reshape RNA research is further enhanced by its capacity to probe the dynamic world of RNA modifications, thereby shedding light on post-transcriptional regulation.

In the process of nanopore sequencing, RNA is converted into a sequenceable library and analysed using the nanopore platform. Bioinformatics tools are then used for data analysis, including mapping reads to a reference genome and quantifying transcript and gene expression. The longer read lengths of nanopore sequencing improve the resolution of complex transcript structures and aid in identifying isoforms and alternative splicing events.

While nanopore sequencing brings several benefits, it also faces challenges such as higher error rates compared to short-read sequencing technologies. The direct RNA method requires a substantial initial amount of poly A+ RNA and typically

yields less output than direct cDNA and PCR-cDNA methods. Ongoing enhancements in base-calling algorithms are aimed at improving the accuracy of nanopore sequencing [95]. Moreover, comprehensive bioinformatics solutions to fully exploit this technology's data layers are still developing. The scalability and portability of this technology make it versatile for a wide range of applications, from in-lab research to field studies.

1.4.3 NanoString nCounter Technology

NanoString Technologies has become a prominent player in the realms of genomics, proteomics, and especially in quantifying transcriptomes, thanks to its nCounter Analysis System. This system, known for its high precision and sensitivity, provides unique benefits for gene expression profiling among other applications. It has been extensively used in various research areas, including the discovery of gene expression signatures, biomarker identification, and the molecular profiling of diseases. Specifically, in cancer research, the nCounter System has played a pivotal role in delineating gene expression patterns that correlate with tumour subtypes, prognosis, and response to treatment, as evidenced by numerous studies [118]–[121].

The nCounter System operates on a barcode technology that utilises colour-coded molecular barcodes to quantify target molecules, such as RNA transcripts or microRNAs, in a highly multiplexed manner. The workflow begins with the hybridisation of target nucleic acids to specific capture and reporter probes. Each probe set consists of a pair of oligonucleotides, with one end attached to a solid support (capture probe) and the other carrying a unique colour-coded molecular barcode (reporter probe). After hybridisation, the samples are loaded into the nCounter cartridge, where they undergo a series of purification steps. The cartridge is then placed into the nCounter Digital Analyser, where the colour-coded barcodes are counted and tabulated for each target. This digital counting approach provides a direct and quantitative measure of the target molecules without the need for amplification, preserving the integrity of the input RNA (Figure 1.9) [123].

The nCounter System is capable of analysing as many as 800 targets in a single reaction, which positions it as a medium-throughput, multiplexed tool. This system is distinguished by its sensitivity, detecting even low-abundance transcripts within complex biological samples efficiently, such as rare transcripts in liquid

regardless of their technical background. By forgoing the need for cDNA synthesis and PCR amplification, it simplifies the experimental process and reduces the amount of active laboratory time required [124].

Within clinical and translational research, the nCounter System has become highly valued across various domains, notably in oncology, immunology, and infectious diseases. This broad applicability is partly due to its proficiency in analysing both limited and degraded RNA samples, including those derived from formalin-fixed, paraffin-embedded (FFPE) tissues. This feature significantly enhances its utility in clinical environments where obtaining ample, high-quality samples can often be challenging [125], [126]. For instance, in the realm of cancer diagnostics, the platform's ability to reliably assess RNA from years-old FFPE tissue sections opens doors to retrospective analyses, enabling researchers to uncover potential biomarkers from stored samples. Additionally, the system's efficiency in processing samples of minimal quantity makes it particularly suited for liquid biopsies, offering comprehensive molecular insights even from limited sample volumes.

1.5 Sequencing Technology: Paving the Way for Liquid Biopsy

The use of sequencing technologies has become a cornerstone in the realm of precision medicine, especially their application in liquid biopsies, which shows immense promise. NGS has greatly enhanced the capabilities of liquid biopsies by providing sensitive and specific analysis of various biological sources in the bloodstream, such as TEPs, EVs, CECs, and CTCs. This innovation is reshaping cancer care by enabling minimally invasive tumour analysis, real-time monitoring of disease progression, and pinpointing therapeutic targets and mechanisms of drug resistance, which are crucial for crafting individualised treatment plans.

NGS provides the benefit of simultaneously analysing the expression of thousands of genes [127] and transcripts, as well as scrutinising numerous genes for mutations [128]. This comprehensive examination reveals the distinct molecular profile of a patient's cancer, encompassing the identification of gene fusions and epitranscriptional markers, like RNA editing events [129].

Yet, adopting sequencing technologies for liquid biopsy comes with its set of obstacles. The often-dilute nature of these biological sources in the blood, particularly in the early stages of cancer, calls for sequencing approaches that are both highly sensitive and accurate. The task is further complicated by the fragmented state of nucleic acids and the genetic noise created by the normal apoptosis of cells, which can obscure the analysis [130].

The demand for high-quality sequencing to accurately pinpoint and quantify rare genetic variants at low frequencies is critical. This demands advanced bioinformatics tools that are adept at navigating the complexities of the data, distinguishing genuine signals from potential sequencing errors. As sequencing technologies progress, they amplify the utility of liquid biopsies, enabling more profound insights into tumour genetics. However, this also complicates the data analysis and interpretation process. Bioinformatics becomes indispensable in this context, transforming intricate genetic data into meaningful, actionable insights for clinical application. By addressing these analytical challenges, bioinformatics not only facilitates a more nuanced understanding of cancer but also guides the shift towards more individualised diagnosis and treatment strategies.

1.6 Bioinformatics Data Analysis

Bioinformatics, a rapidly progressing discipline, merges computer science with biology to tackle the complex task of analysing and interpreting biological data. This interdisciplinary field is especially pivotal in genomics, where the creation of large-scale datasets is routine. This field becomes increasingly indispensable as sequencing technologies advance, producing larger and more complex datasets [131], [132]. The surge in data necessitates bioinformatics to not only manage and analyse these datasets but also to interpret them meaningfully. The expansion of sequencing capabilities has propelled bioinformatics forward, fuelled by improvements in computational hardware and the development of sophisticated algorithms and software. This enables researchers to perform a wide range of tasks, from initial data processing to advanced analysis for clinical applications.

The rapid evolution of bioinformatics is essential for navigating the complexities of genomic data, playing a key role in modern biology. It employs state-of-the-art

computational strategies to translate vast genomic datasets into valuable scientific knowledge, aiding in the discovery of disease mechanisms, identification of therapeutic targets, and unraveling the complex processes of life.

Despite varying research objectives, bioinformatics analyses of sequencing data share a foundational workflow, standardised across diverse platforms despite their unique challenges and requirements [133]. This structured workflow is delineated into three essential stages, integral for deciphering the complex data generated by sequencing technologies. The process initiates with raw data preprocessing, where data is refined to ensure the highest possible quality. Subsequent alignment of sequences to a reference genome pinpointing their origins. The workflow culminates in a tailored analysis and interpretation phase, directly addressing the research question and yielding insights that enrich our comprehension of genetics and molecular biology. This efficient approach facilitates the methodical examination of sequencing data, driving significant advances in the field. The four stages integral to the bioinformatics analysis of sequencing data are:

1. **Preprocessing:** Sequencing processes are prone to errors, which are reflected in the quality scores attached to sequence reads within the raw fastq files. These files are pivotal for initial quality assessments, as they contain all raw sequencing reads along with their corresponding quality scores and identifiers. Two essential preprocessing tasks are filtering reads based on their quality scores and trimming adapters. Filtering removes reads (or part of reads) that fall below a certain quality threshold, ensuring that only high-quality, reliable data proceeds to further analysis stages. Adapter trimming involves removing residual adapter sequences from the ends of reads. This step is crucial, as leftover adapter sequences can disrupt subsequent mapping. By filtering out low-quality reads and trimming adapters, researchers can enhance the accuracy of sequence alignment and the overall reliability of the analysis [134].
2. **Sequence Alignment:** After preprocessing, the subsequent pivotal phase involves aligning sequencing reads with a reference genome. This alignment is crucial for identifying the specific genomic origins of each read within the genetic context provided by the reference sequence. This process is a foundational element of bioinformatics, illuminating the functions, structural characteristics, and evolutionary paths of the sequences. Through such

comparative analysis, a base is established for examining genetic variations among sequences, thereby deepening our understanding of the intricate connections and evolutionary journeys of these genetic sequences [135].

- 3. Downstream Analysis and Interpretation:** The final stage of transcriptomic analysis, a pivotal component of bioinformatics, unfolds through a comprehensive suite of analytical techniques, each crafted to dissect different facets of transcriptomic data. Gene Expression Profiling measures the concurrent expression levels of numerous genes, offering a comprehensive view of cellular activity under differing conditions or within distinct cell types [136]. Differential Gene Expression Analysis contrasts these expression levels between samples, identifying genes with significant changes to highlight potential biological pathways affected by a condition or treatment [137]. The Discovery of Novel Transcripts seeks out previously unknown RNA sequences, expanding our understanding of the genome's complexity [138]. Alternative Splicing Analysis reveals the process by which a single gene can give rise to multiple mRNA variants, shedding light on the intricate regulation of gene expression [139]. Gene Fusion Detection focuses on identifying hybrid genes from the fusion of two separate genes, crucial in cancer research for diagnostic and therapeutic implications [140]. Quantification of Isoform Expression measures the expression levels of different gene variants produced through alternative splicing, providing insights into their distinct functional roles [141]. Identification of Single Nucleotide Polymorphisms (SNPs) within RNA sequences can indicate genetic variations affecting gene function or disease susceptibility [142]. Lastly, the Examination of Post-transcriptional Modifications, such as RNA editing or methylation, investigates alterations made to RNA after synthesis, which can significantly influence RNA function and stability, showcasing the dynamic complexity of gene regulation [143].

Each of these components plays a pivotal role in providing a comprehensive understanding of the genome's function, structure, and evolution. This phase is where the data begin to reveal insights into genetic function and regulation, disease mechanisms, and potential therapeutic targets.

The evolution of bioinformatics in response to the deluge of data from sequencing platforms underscores a dynamic field that continuously adapts and innovates. Through its multifaceted analysis pipelines, bioinformatics not only manages the

technical challenges of massive data volumes but also unlocks the potential for groundbreaking discoveries in genetics, medicine, and beyond.

1.7 The Emergence of Liquid Biopsy in Bioinformatics Research

The transformative integration of sequencing technologies with liquid biopsy represents a significant advancement in cancer research and diagnostics. In this context, the role of bioinformatics emerges as crucial. It serves to bridge the gap between the innovative capabilities of sequencing technologies and the practical application of liquid biopsy. By analysing sequencing data, bioinformatics facilitates early cancer detection, monitoring, and informed treatment decision-making. Employing advanced tools and methodologies, bioinformatics navigates the challenges posed by the small quantities of DNA and RNA in body fluids. It enables the detection and quantification of genetic mutations, epigenetic alterations, and expression patterns that serve as biomarkers for cancer. The synergy between bioinformatics and liquid biopsy technologies is pivotal, enhancing the sensitivity and specificity of biomarker identification and paving the way for the discovery of novel therapeutic targets [144].

Moreover, bioinformatics tools are indispensable for identifying low-frequency variants, copy number alterations, and structural rearrangements critical for understanding the complex nature of tumours [145]. The continuous monitoring capability of liquid biopsies, powered by bioinformatics, opens new avenues for real-time disease monitoring and assessing treatment efficacy [146]. This integration underscores the potential of liquid biopsy to transform cancer care into a more precise, minimally invasive, and individualised approach.

The necessity for bioinformatics extends beyond handling technical challenges; it is fundamental in developing predictive models and identifying novel biomarkers. As sequencing technologies evolve with higher resolution and throughput, the demand for sophisticated bioinformatics expertise in liquid biopsy research escalates. This demands a profound understanding of both biological and computational aspects of liquid biopsy analysis and the development of algorithms capable of differentiating between tumour-derived and normal DNA/RNA amidst high genetic noise.

At the frontier of this convergence between bioinformatics and liquid biopsy is the exploration of machine learning (ML) and artificial intelligence (AI). Leveraging ML and AI promises to refine data analysis significantly, enhancing the accuracy and efficiency of identifying actionable insights from complex and large-scale datasets [147]. These sophisticated computational methods for data mining, pattern recognition, and multi-omic data integration highlight the critical role of bioinformatics in translating the wealth of data generated by sequencing technologies into actionable clinical insights.

As we stand on the brink of a paradigm shift in cancer diagnosis and treatment, the integration of bioinformatics and liquid biopsy research symbolises a crucial confluence of disciplines. It not only advances liquid biopsy as a tool for precision oncology but also demonstrates the broader role of bioinformatics in utilising large-scale datasets to revolutionise healthcare. The necessity of bioinformatics in bridging the innovative leap in sequencing technologies to clinical practice emphasises its pivotal role in the era of precision oncology. The exploration of methodologies, computational challenges, and future directions of bioinformatics will further enhance the utility of liquid biopsy, underscoring its significance in the ongoing quest to understand and combat cancer.

1.8 Machine Learning in Genomics Research

The burgeoning field of bioinformatics has been instrumental in managing, processing, and analysing the immense volumes of data produced by modern sequencing technologies. However, the increasing complexity and scale of this data have begun to stretch the capabilities of traditional bioinformatics approaches. ML, a dynamic branch of AI, steps in at this juncture, enhancing the predictive capabilities, classification problems, and pattern recognition necessary to navigate the complexities of large-scale datasets in bioinformatics [148].

ML refers to a collection of algorithms that allow computers to learn from data without explicit programming. In contrast to traditional programming, where every step is predefined, ML algorithms can identify patterns and relationships in data, and then use those insights to make predictions on new, unseen data [149]. This makes them particularly well-suited for analysing the massive datasets generated by sequencing technologies.

Types of ML commonly Used in Genomics Data Analysis:

- **Supervised Learning:** This approach trains an algorithm on a dataset where each entry is clearly labeled, presenting both the input variables and their anticipated outcomes. In this structured training process, the algorithm learns to associate inputs with outputs, constructing a model from existing data to identify patterns that can accurately predict outcomes for new, unseen data. The ultimate goal is for the algorithm to generalise from its training, enabling it to offer reliable forecasts for unencountered data. In the realm of genomics, supervised learning is especially critical for prognosis and customising treatment plans. For example, it is often used to assess disease risks associated with specific genes or genetic variants that the model has been trained to recognise, which are linked to particular health conditions [150]. This application of supervised learning is essential in advancing personalised medicine strategies, allowing for treatments that are tailored to the genetic makeup of individual patients.
- **Unsupervised Learning:** Unlike supervised learning, unsupervised learning offers a distinct advantage by analysing data without pre-labeled outcomes, uncovering hidden patterns and structures within genetic information that may not be immediately apparent. This type of ML algorithm is particularly useful for clustering genetic data based on similarities, identifying novel genetic markers without prior knowledge, and discovering new subtypes of diseases. By grouping genes or variants with similar expression profiles or mutations, unsupervised learning facilitates a deeper understanding of genetic relationships and their implications for health and disease. Such insights are invaluable for segmenting patient populations into more precise categories, enhancing the specificity of research studies, and paving the way for more targeted therapeutic approaches [151].

Data preparation is a fundamental stage in the ML pipeline and is crucial for the successful training and performance of ML models. This stage involves cleaning and filtering the data (handling missing values, removing outliers, eliminating non-informative data), transforming variables (normalisation, scaling), and feature selection or extraction to best represent the problem at hand. Proper data preparation ensures that the ML model has the highest quality and most relevant

information available, free from skew or bias that could adversely affect its ability to learn and generalise.

Effective data preparation reflects a deep understanding of the problem domain and directly influences the accuracy and efficiency of the resulting model. For instance, in a healthcare setting, properly prepared data helps in accurately predicting patient outcomes or diagnosing diseases, thereby directly impacting treatment decisions and health outcomes.

ML's predictive modelling capability, which integrates genomic data with clinical and environmental factors, offers a comprehensive approach to healthcare. It can predict an individual's disease susceptibility, treatment response, and potential side effects, underscoring ML's potential to revolutionise personalised medicine.

1.9 Machine Learning and Liquid Biopsy

The integration of ML with liquid biopsy represents a groundbreaking novelty in the early detection, prognosis, and ongoing monitoring of several diseases including cancers. ML has emerged as a transformative technology in this domain, significantly enhancing the precision, efficiency, and predictive capabilities of liquid biopsy analyses [152]. This synergy between ML and liquid biopsy improves the detection, analysis, and tracking of diverse cancer biomarkers in bodily fluids, offering new avenues for personalised cancer care.

ML is particularly effective in addressing key oncological challenges such as the precise detection and quantification of rare biomarkers, which are often present only in minute quantities against a complex background of normal cellular material. ML algorithms excel in analysing large datasets to identify disease-specific patterns and markers that are too subtle for traditional methods due to the complexity and volume of the data.

In practical terms, ML algorithms facilitate nuanced distinctions in cancer diagnostics. For example, with ctDNA, these algorithms can differentiate between cancerous and healthy genetic profiles, pinpointing mutations linked to cancer [153]. When evaluating CTCs, machine learning plays a vital role in distinguishing cancerous from non-cancerous cells, which is essential for correct disease staging

and prognosis. Additionally, for exosomes and other vesicles laden with molecular data, machine learning is instrumental in unlocking their payload, offering a glimpse into the tumours condition and activity [154].

Implementing ML in liquid biopsy entails several key stages: Initially, data pre-processing and filtering refine and standardise the data. Feature selection follows, identifying significant indicators that inform subsequent analyses. Model training is conducted to recognise and learn patterns within the data. Optionally, more advanced machine learning techniques, such as ensemble learning which combines predictions from multiple models, can be applied to enhance predictive accuracy. The process concludes with rigorous testing and validation to verify the model's reliability and effectiveness. These steps are essential to ensure the successful application of machine learning in liquid biopsy analyses.

1.10 Thesis Overview

Chapter 1 lays the essential groundwork of this thesis, offering a foundational overview that is crucial for understanding the key components utilised throughout the study. Its goal is to offer an in-depth explanation of the topics covered within the thesis. More than setting the thematic stage, the introduction ensures that readers are equipped with the fundamental concepts, theories, and terminology necessary to grasp the nuances of the study. This approach facilitates a clear understanding of the research's significance and the analytical findings discussed in subsequent sections, catering to both experts and those new to the topic.

Chapter 2 meticulously outlines the aims and objectives guiding the research. This chapter introduces the specific goals the thesis aims to achieve, offering a clear roadmap for the investigation. It details the research questions to be addressed, the hypotheses to be tested, and the contributions it seeks to make to the existing body of knowledge.

Chapter 3 is dedicated to a comprehensive exposition of the methodologies employed across the entirety of this thesis. Within this chapter, an extensive dive into the analytical tools and statistical techniques utilised in the research is undertaken. It details the selection and application of specific materials, the experimental setups, and the procedural approaches adopted throughout the following chapters.

Furthermore, it provides an in-depth explanation of the software tools and statistical frameworks that underpin the data analysis, ensuring the reader has a thorough understanding of how the research findings were derived.

Chapter 4 unveils "ELLBA", a novel methodology designed for the analysis of RNA sequencing data derived from liquid biopsies (lbRNA-seq) in the context of cancer diagnostics. ELLBA addresses existing challenges by establishing an extensive Machine Learning-based Ensemble Classification framework that assimilates a wide array of molecular data and employs effective normalisation techniques for clinical application. The methodology has undergone validation across several independent datasets, affirming its efficacy and dependability. The proposed workflow of the study demonstrates improved predictive accuracy, indicating its suitability for clinical implementation owing to its bespoke approach. Related publications include: [155].

Chapter 5 introduces "DRseeker," a pipeline for analysing DRS protocol sequenced samples from ONT. It supports comprehensive downstream analysis, extracting multiple information layers. Key functionalities include Differential Gene Expression and Isoform Expression analysis for insights into gene regulation and expression patterns, alongside novel isoform detection to explore genetic novelties. It assesses polyA tail lengths and conducts Alternative Polyadenylation Analysis, crucial for mRNA stability and translation insights. The pipeline also facilitates Differential Polyadenylation Analysis and Differential Transcript Usage, illuminating post-transcriptional regulation intricacies. DRseeker serves as a pivotal resource for deciphering gene expression and regulation via nanopore sequencing. This project stands out as a robust resource for researchers seeking to unravel the intricate details of gene expression and regulation via nanopore sequencing technology.

Chapter 6 introduces "NanoInsights," a versatile web service for analysing NanoString nCounter data, crucial in genomics and clinical applications. The platform features eight unique normalisation methods, enabling customised data processing for diverse research requirements. NanoInsights employs machine learning algorithms for sample classification, offering insights to support clinical decisions and genomics advancements. This service simplifies data analysis, enhancing accessibility and usability for both researchers and clinicians. Related publications include: [156]–[160].

Chapter 7 offers an in-depth analysis of the insights and outcomes presented in Chapters 4, 6, and 5. It underscores the pivotal discoveries and contributions of each segment, examining the broader impact of these findings and their implications for the field. The chapter also identifies key areas ripe for further exploration, proposing targeted research directions that promise to enrich understanding and foster advancements. Moreover, it recommends strategic enhancements and innovative methodologies to bolster the effectiveness and breadth of subsequent studies. Through this discourse, the chapter lays the foundation for future endeavours, building on the solid base established by this research to propel the discipline forward.

Chapters 8 and 9 provide summaries of the tools developed and the research findings discussed in Chapters 4, 6, and 5, with conclusions offered in both English and Spanish. From these outcomes, recommendations are made for researchers, bioinformaticians, and research organisations. The thesis concludes by identifying potential directions for future research as recognised by the author.

Chapter 2

Objectives

This thesis was guided by key hypotheses aimed at advancing cancer diagnostics through the integration of innovative bioinformatics and ML techniques applied to various gene expression technologies within the ELBA consortium. Focusing particularly on lung cancer, which often remains asymptomatic in its early stages, this research emphasised improving detection methods using blood-based liquid biopsies and tissue biopsies. While holding significant promise, they also face considerable challenges, especially in bioinformatics processing, which must be addressed to fully realise their potential.

The initial hypothesis posited that by enhancing well-established high-throughput technologies such as NGS and RNA-Seq with previously underutilised layers of information through advanced ML, significant improvements could be achieved in cancer detection and diagnosis. A secondary hypothesis suggested that exploring emerging gene expression technologies would allow for a deeper understanding of cancer's molecular mechanisms, potentially leading to the development of more effective diagnostic tools. These hypotheses stem from the recognition of liquid biopsies' potential alongside existing challenges in bioinformatics processing that limit their efficacy.

Ultimately, the overarching goal was to develop more accurate and accessible diagnostic tools that would not only advance technological progress but also provide real-world benefits in clinical settings, thereby improving patient care and treatment outcomes. To achieve this goal, the hypotheses outlined above guided the formulation of the following focused objectives:

1. Develop a clinical-ready methodology for lbRNA-Seq analysis

- **Quantify Molecular Properties of Coding Transcripts:** Develop a method that quantifies different molecular properties (biofeatures) of coding transcripts from lbRNA-Seq data, such as RNA editing events and SNVs. Each biofeature may contain unique information that, when considered in conjunction with others, could improve the discriminative power for identifying various disease states.
- **Optimise Feature Outputs:** Enhance the values of the quantified molecular biofeatures to increase their analytical accuracy. Specifically, improve data processing by implementing inter-sample and intra-sample normalisation techniques where applicable, as these are crucial for ensuring data consistency. These optimisations are vital for clinical applications that require precise and standardised methods.
- **Combine Biofeatures with ML:** Leverage advanced ML and classification techniques, such as Ensemble Classification, to create a single ML classifier that integrates all the extracted biofeatures from the data. This approach aims to improve the analysis process by allowing the ML model to identify complex patterns within the combined data, potentially leading to higher prediction accuracy.
- **Develop a User-Friendly Pipeline:** Construct a user-friendly pipeline that is designed for straightforward adoption by the broader scientific community. This pipeline will facilitate access to advanced bioinformatics tools, ensuring that users can easily implement and benefit from the developed methodology.

2. Design and create a methodology for DRS data processing

- **Enhanced Analysis of Sequencing Data:** Integrate a range of analyses to maximise the use of sequencing output, ensuring a comprehensive extraction of all potential biological insights unique to DRS technology. This includes studying gene expression at the transcriptome level and exploring consequential features such as Differential Isoform Usage. Additionally, simultaneously investigate complementary features like polyA tail length and post-transcriptional modifications, which can all be extracted in a single run without the need for supplementary techniques.

- **Developing an Optimised DRS Data Analysis Pipeline:** Construct a comprehensive and easy-to-use pipeline tailored for DRS data from ONT. Ensure the software is continuously updated and optimised for real-world data to consistently deliver the best results, keeping pace with rapid advancements in technology and bioinformatics tools.

3. Develop a web-based platform for NanoString nCounter analysis

- **Creating an Application for Data Analysis:** Design a pipeline specifically for analysing NanoString nCounter data, demystifying bioinformatics and ML complexities for users across varying levels of expertise.
- **Streamlined NanoString Analysis with a Web Platform:** Incorporate the pipeline into a user-friendly web-based platform designed specifically for individuals with limited expertise in bioinformatics. This platform streamlines the analysis process while maintaining depth and quality. This includes intuitive interfaces, guided workflows, and comprehensive tutorials, making advanced analytical tools accessible to a broader scientific community.
- **Enhancing NanoString Data Analysis:** Develop a robust, easy-to-navigate platform for conducting complex analyses of NanoString nCounter data. Features should include diverse normalisation techniques, real-time interactive data visualisation, automated reporting for streamlined workflows, and the ability to download high-resolution figures for easy data presentation.

Chapter 3

Methods

3.1 Molecular Methods and Sequencing

3.1.1 Lung Tissue Collection and RNA Isolation in NSCLC Patients

In our second study, we focused on a cohort of three non-small cell lung cancer (NSCLC) patients, with an average age of 57.3 years, each diagnosed with KRAS-mutant adenocarcinomas. The collection of tissue samples occurred during tumour resection surgeries, where we obtained both the malignant tissues and the adjacent non-transformed lung tissues for each patient. Immediate snap-freezing of these samples in liquid nitrogen followed, securing their preservation at -80°C for future analyses.

Preparatory steps for histological evaluation included the creation of cryosections from both peripheral and central tissue regions, which were then subjected to hematoxylin and eosin staining. Microscopic examination of these stained sections verified the samples' minimal necrotic content, deeming them suitable for RNA isolation. To mitigate RNA degradation, we implemented stringent protocols, including the use of RNase-free water and thorough decontamination of equipment with RNA zap. Tissue sections designated for RNA isolation were processed with a lysis buffer from the MirVana kit, immediately stored at -20°C , continuously kept on ice during processing, and then placed at -80°C until RNA isolation commenced.

3.1.2 RNA Isolation from Lung Tissue Samples

The isolation of RNA from the collected tissue samples was conducted utilising the MirVana Total RNA Isolation Kit, adhering to the provided manufacturer's guidelines. The integrity and quantity of the extracted RNA were meticulously assessed using the Agilent Bioanalyzer RNA 6000 Picochip and quantified with the NanoDrop 2000 spectrometer by Thermo Scientific, ensuring a precise evaluation of the RNA samples.

3.1.3 MinION Nanopore Direct RNA sequencing

The sequencing project focused on six samples, including two matched pairs of NSCLC adenocarcinoma tissues alongside their non-cancerous counterparts. However, due to insufficient reads from the non-cancerous sample of the third pair, a non-transformed sample from an additional fourth subject was sequenced instead. Direct RNA sequencing of each sample involved the use of 500 ng of Poly-A RNA, enriched via magnetic beads in accordance with the Direct RNA Sequencing Kit (RNA Kit SQK-RNA002) instructions from Oxford Nanopore Technologies Ltd. A pivotal step in our protocol was the integration of 0.25 μ l of the RNA Calibration Strand (RCS) for Enolase II (ENO2), serving as a calibration standard. The sequencing workflow began with the priming of the MinION flow cell using the Flow Cell Priming Kit (EXP-FLP002) from the same company. Subsequent to this priming stage, 75 μ l of the RNA library was loaded into the SpotON sample port of the MinION device, facilitating the commencement of the sequencing process.

3.2 Bioinformatics Methods

The computational analyses presented in this thesis primarily utilised Python for the development of the majority of scripts, supplemented by additional scripts written in R for specific tasks. The complete suite of projects, which will be accessible post-publication, is hosted on the GitHub repository at <https://github.com/sgiannouk>.

The research detailed in Chapter 4 and Chapter 6 within the thesis was based on datasets publicly available through the National Center for Biotechnology Information (NCBI). For ease of reference and reproducibility, each of these chapters meticulously lists the accession numbers associated with the datasets employed in our analyses.

This section also serves to introduce a concise overview of the key bioinformatics methodologies and resources that were employed across the various segments of this thesis. These methodologies encompass a range of computational tools and analytical techniques pivotal for the data processing and analysis undertaken in our research. For a detailed exploration of the specific materials and methods utilised in individual chapters, where these aspects are further elaborated upon.

3.2.1 Preprocessing and Quality Control Workflow for NGS Data

The preprocessing of initial raw fastq files from Illumina sequencing (NGS) begins with FastQC [161], which facilitates the automated identification of potential adapter sequences within the data. Upon detecting these adapters, the BBDuk [162] tool is then utilised to carry out adapter removal, alongside quality trimming and filtering processes. These steps involve trimming of poly-A or poly-T sequences longer than 20 nucleotides, removal of reads containing ‘N’ bases after trimming, cutting off read ends to remove bases scoring below 20 on the Phred scale, discarding reads shortened to less than 40 nucleotides after trimming, and eliminating reads whose average post-trim quality score falls below 20 Phred. Following the completion of these preprocessing measures, FastQC performs a Quality Control check and an initial statistical analysis of the data. To provide a comprehensive overview of the dataset, MultiQC [163] compiles a final report, offering an aggregated view of the data.

3.2.2 Mapping and Quality Assessment for NGS Data

The STAR aligner [164] is employed for mapping the high-quality sequencing reads to the human reference genome GRCh38.p13 primary assembly, alongside the GENCODE v35 reference gene annotation [165]. This precise alignment process utilises the ‘GeneCounts’ and ‘TranscriptomeSAM’ parameters to accurately

quantify both gene and transcript levels. Moreover, STAR is fine-tuned to work in synergy with the Arriba software [166], significantly boosting its capacity to identify fusion genes, while also performing gene and transcriptome quantification. Quality metrics for the alignment are derived using RSeQC and Picard tools [167], [168], essential for assessing the alignment process and further evaluate the quality of the data. The entire post-alignment review process is streamlined through the use of MultiQC, resulting in a detailed report that encapsulates the results. This final document provides a thorough overview of the data quality ensuring a solid foundation for further genomic analysis.

3.2.3 Transcriptome Quantification for NGS Data

Utilising the ‘TranscriptomeSAM’ output from STAR aligner, the Salmon software [169] operates in an alignment-based mode to process this data. Default settings are applied, with the enhancement of specific options such as ‘seqBias’ and ‘gcBias’ for the correction of sequence and GC content biases, respectively. The ‘libType U’ setting is chosen to accommodate unstranded data, and the analysis includes 100 bootstrap iterations to ensure accurate and robust quantification. The final transcript-level expression matrix is derived by amalgamating the quantification results through Salmon’s ‘quantmerge’ script, focusing on the ‘numreads’ column to construct the expression matrix.

3.2.4 Identification of RNA Editing and SNVs in NGS Data

De-duplication of the genome-aligned BAM files was conducted using samtools [170] through its rmdup function. To adjust the base quality scores, the GATK4 BaseRecalibrator tool [171] was applied. For variant calling preparations, the BAM files underwent further processing with BCFtools mpileup [172], applying parameters such as minimum mapping quality (min-MQ) and base quality (min-BQ) both set to 15, alongside other adjustments for improved accuracy. Variant calling was executed with BCFtools call, setting specific conditions for ploidy, variant selection, and calling methods, while excluding common variants listed in the dbSNP database [173] to refine the dataset.

For the identification of RNA editing events, the REDIttools software [174] was configured to recognise edits with stringent quality and frequency criteria. These

identified events were then consolidated into a comprehensive matrix, applying filters to only include events with more than 20% occurrence across the dataset.

Following a similar initial approach for SNP analysis, the protocol diverged post common variant filtering to exclude low-quality positions (SNPs with quality greater than Phred score of 20, depth greater than 5 reads and mapping quality greater than 20 Phred score) and previously noted RNA editing sites. The consolidation of variant files was performed using BCFtools merge, leading to the creation of a unified variant matrix.

3.2.5 Re-Basecalling and Initial Filtering Workflow for TGS Data

Basecalling of the MinION output fast5 files was executed locally using ONT's Guppy software [175], version 6.3.7, to ensure high-quality read inclusion based on a minimum Phred score of 7, as recommended by ONT. We customised the basecalling process by converting U bases to T and producing reads in reverse orientation, with settings adjusted to specific flowcell and kit types, calibration detection, RNA trimming strategies, sequence reversal, quality score filtering, and enabling U substitution.

For a thorough assessment of our sequencing data's quality and the overall success of the sequencing effort, we employed NanoPlot [176]. This dedicated tool provides detailed statistical analyses and a range of quality control graphics showcasing read length and quality variability across the sequencing timeline. Bivariate plots helped explore the correlation between read length and quality scores for a deeper insight into the data's fidelity.

In the subsequent stages, we excluded reads shorter than 50 nucleotides and undertook meticulous error correction with IsONcorrect [177]. This approach capitalised on the diverse gene isoform information to correct errors efficiently, proving particularly beneficial for enhancing read accuracy, even with minimal sequencing coverage.

3.2.6 Mapping and Quality Assessment for TGS Data

The fastq files, designated as ‘pass’ (with a minimum Phred score of 7), were mapped to the GRCh38 primary assembly of the human genome using the Minimap2 aligner [178] in a configuration that supported splice-aware alignment. The alignment process utilised 14-mer k-mers excluding secondary alignments.

To organise and condense the resulting SAM files, and to generate detailed metrics for assessing the quality of the sequencing experiment, we employed a suite of tools including Samtools [170] for file sorting and compression, PycoQC [179] for quality control analysis, and RSeQC [180] for comprehensive evaluation of read mapping and alignment quality. These tools facilitated a thorough examination of the sequencing output, ensuring a high standard of data integrity and alignment accuracy.

3.2.7 Gene and Transcript Identification and Quantification Using TALON for TGS Data

For the purpose of identifying existing and predicting new genes and transcripts within long-read transcriptomic data, we utilised the TALON software [181]. As a preliminary step, and following the guidance provided by the developers, we applied TranscriptClean [182] to enhance the quality of aligned reads. This step focused on correcting any deviations from canonical splice junctions. The preprocessed BAM files from our six samples, in conjunction with the GENCODE v35 human reference annotation, served as the input for our analyses with TALON.

TALON’s workflow facilitated the quantification of transcript abundance across our samples, yielding two separate matrices. The first matrix offered a detailed view of expression levels for both known and newly discovered transcripts, mapped to established genomic locations, without imposing any filtering based on counts or other criteria. The second matrix applied a more rigorous approach, focusing on newly identified transcripts, necessitating their presence in at least one sample group with a minimum of five counts for inclusion.

An additional custom script was implemented to refine our selection of novel transcripts further. This script mandated that all novel transcripts demonstrate a significant polyA tail length across a minimum of ten transcripts. Furthermore,

we set a criteria of at least 20 counts per isoform for novel transcripts to be included under the ISM Prefix, Suffix, None, and Both categories. For transcripts classified under the novel ISM Suffix category, we ensured their completeness through an exhaustive comparison of their 5' ends against CAGE assay data, as elaborated in a later section. Lastly, we merged the ISM None and Both categories into a consolidated ISM Both category, streamlining our categorisation of novel transcript types.

3.2.8 CAGE Analysis for TSS Verification for TGS Data

Utilising human CAGE peak data from FANTOM5 database [183], updated to hg38 using the LiftOver tool, we identified transcription start sites (TSS) for our transcript models, checking for CAGE peaks within a 100 bp range of these TSS locations with Bedtools.

3.3 Statistical Analysis

3.3.1 Exploratory Analysis for NGS Data

Before delving into statistical analysis, it's crucial to undertake an exploratory analysis to uncover essential insights such as batch effects and the presence of lowly expressed data. To this end, multiple preliminary analyses are performed to ensure the data's readiness for further examination. One of the primary steps involves conducting a Principal Component Analysis (PCA) using the DESeq2 R package [184], which serves to evaluate the similarities among samples within a dataset. PCA enables us to assess the clustering of similar samples and verify that the experimental conditions are the primary source of variation. This technique is instrumental in identifying outlier samples, which can be scrutinised to decide whether their exclusion is necessary before proceeding with differential expression analysis. Additionally, PCA may reveal the presence of batch effects that require correction.

To facilitate PCA, we apply the Variance-Stabilising Transformation (VST) function from the DESeq2 package, preparing the data for analysis. The PCA plot is generated twice to distinguish variations by conditions and by batch, providing

a clearer understanding of the data's structure. Following the transformation, an interquartile range (IQR) analysis is conducted on the data. The IQR is utilised to examine data variability and pinpoint potential outliers for possible exclusion. This analysis adheres to the standard 1.5 threshold for identifying outliers.

3.3.2 Normalisation of the NGS Data

The further processing of NGS data critically involves filtering out lowly expressed genes, a task efficiently handled by the 'filterByExpr' function within the edgeR package [185]. After filtering, we apply normalisation using Counts Per Million (CPM) to adjust for variances in sequencing depth across samples. To further investigate the impact of this normalisation, we proceed with a new PCA. Creating a new PCA plot affords us the opportunity for a visual assessment of sample distribution and clustering post-normalisation. This identical process is also rigorously applied to the Isoform Expression data, ensuring consistency across our data analysis approach.

3.3.3 Gene Set Enrichment Analysis for the NGS Data

Our analysis includes a Gene Set Enrichment Analysis (GSEA) when necessary, a powerful method aiming at uncovering groups of genes that show significant over-representation within a broad gene set and may be linked to distinct phenotypes. To facilitate this analysis, we utilise g:Profiler along with its g:GOST tool [186], a publicly accessible web server designed for gene mapping against recognised databases of functional information. This tool enables us to identify and assess statistically significant enriched terms, providing insights into the biological significance of our gene sets in relation to various phenotypes. Through this approach, we can explore the potential functional implications of our data, shedding light on how certain gene groups may influence or correlate with specific biological traits or conditions. To streamline this process, we use the R package gprofiler2 [187], which provides programmatic access to g:Profiler's resources through a REST API, ensuring efficient preparation and submission of gene lists for analysis. This process culminates in the generation of a direct link to the g:Profiler web application, enabling immediate execution of GSEA and facilitating user access to the analysis results.

3.3.4 Exploratory Analysis for TGS Data

Following a similar approach detailed in Section 3.3.1, the nanopore sequencing data also undergo an extensive exploratory analysis. This process begins with the manipulation of the unfiltered data, adhering to TALON’s guidelines, to generate a gene-level expression matrix. This involves aggregating transcript counts to their respective originating genes. Subsequently, genes with a total count of fewer than 10 reads are excluded from further analysis to ensure data quality.

The exploratory phase encompasses several analytical techniques to assess data quality and structure. A PCA plot, similar to the one described in Section 3.3.1, is created to visualise the variance within the data and identify potential clusters or outliers based on gene expression profiles. Additionally, we construct a Sample-to-Sample correlation heatmap. This heatmap, derived from Pearson correlation coefficients, visually represents the gene expression similarities between samples. Hierarchical clustering is applied to this heatmap, organising samples into groups within a dendrogram, where the branch lengths indicate the level of similarity between sample gene expressions.

Furthermore, to assess the distribution of gene expression across the samples, we perform a boxplot analysis on the log₂-transformed counts of the raw expression matrix. This visualisation helps identifying any disparities in expression levels, potential outliers, or systematic biases within the dataset, providing a comprehensive overview of the data’s characteristics before proceeding to more targeted analyses.

3.3.5 Differential Expression and Usage Analysis for TGS Data

Utilising R, we conducted analyses for Differential Gene Expression (DGE), Differential Transcript Expression (DTE), and Differential Transcript Usage (DTU). For DGE analysis, transcript counts were aggregated to their respective genes to form a gene-level expression matrix. This matrix was subsequently refined by removing genes with low expression using the ‘filterByExpr’ function from the edgeR package. Following this filtration, the Trimmed Mean of M-values (TMM) method, also from the edgeR package, was applied to calculate scaling factors among samples.

These factors were then incorporated into the statistical tests to identify genes that are differentially expressed across the samples. Following the identification of differentially expressed genes, we embarked on Functional Enrichment Analysis to delve deeper into the biological significance behind these expression patterns. For this purpose, we employed the clusterProfiler R package [188], a comprehensive tool designed for systematic analysis and visualisation of functional profiles for genes and gene clusters. By utilising clusterProfiler, we can uncover the biological processes, cellular components, and molecular functions most associated with our set of differentially expressed genes, providing valuable insights into the underlying mechanisms of the studied condition or phenotype.

For DTE analysis, we proceed with the filtered expression matrix, specifically omitting transcripts labeled as ‘Genomic’ by TALON, following their recommendations. The approach for DTE mirrors that of the DGE analysis, employing identical filtering techniques and statistical testing methods to discern variations in transcript expression across the dataset.

In our DTU analysis, we applied the IsoformSwitchAnalyzeR package [189] using the pre-filtered abundance matrix as our starting point. To enrich our analysis, we integrated external transcript data from multiple databases: Pfam [190] for the identification of protein families and domains, IUPred2A [191] for the prediction of disordered protein regions, SignalP [192] for the detection of signal peptides, and CPC2 [193] for evaluating the coding potential of transcripts.

3.3.6 Novel Transcript Characterisation for TGS Data

Newly identified isoforms were subject to an in-depth functional characterisation and annotation process, leveraging the capabilities of the Trinotate framework [194]. Trinotate, a comprehensive annotation suite, was employed to systematically analyse these novel isoforms for a wide range of functional attributes. This process included the identification of protein families, prediction of protein domains, assessment of gene ontology terms, and evaluation of potential metabolic pathways.

3.3.7 PolyA-Tail Length, DPA, and APA Analysis for TGS Data

Polyadenylation tail lengths were quantified utilising the Nanopolish tool [195] [196], a methodology supported by its robust algorithms for accurate measurement of polyA tail lengths within our dataset. To explore significant variations in polyadenylation across different samples, we implemented a custom script designed to conduct the Mann-Whitney U statistical test, a non-parametric method chosen for its effectiveness in detecting differences between two independent groups. This was complemented by subsequent analysis using the LAPA tool [197] for a focused examination of Alternative Polyadenylation Analysis (APA). LAPA facilitated the identification of sites exhibiting differential APA, enabling us to pinpoint specific regions where alternative polyadenylation patterns may play a crucial role in gene expression regulation.

3.3.8 Methylation Detection for TGS Data

The process of methylation analysis began by aligning the corrected fastq files to a reference transcriptome provided by TALON, utilising Minimap2 as the alignment tool. Following alignment, Nanopolish Eventalign was employed to segment the sequencing signals, a critical step for identifying specific regions of interest within the data. Finally, the identification of sites exhibiting differential methylation across the dataset was performed using the xPore tool [198]. xPore specialises in analysing and quantifying differential RNA modifications, leveraging the alignment and signal segmentation data to pinpoint variations in methylation that may signify regulatory modifications or changes associated with different biological conditions.

3.3.9 Initial Quality Control and Exploratory Analysis for NanoString nCounter Data

The NanoString nCounter technology, which relies on direct digital detection to quantify gene expression, produces not only raw data, but also a suite of quality metrics designed to assess the fidelity of the detection process. These metrics provide valuable insights into the quality of the experiment and include Imaging QC,

which evaluates the clarity and integrity of the images captured during detection; Binding Density QC, which measures the efficiency of probe-target binding; Positive Control Linearity QC, assessing the linear response of the system across a range of known concentrations; and Limit of Detection QC, determining the lowest concentration at which targets can be reliably detected.

To ensure the reliability of our data, we generate several boxplots focusing on these metrics to pinpoint any samples that deviate significantly from the norm, potentially indicating issues with sample preparation, handling, or instrument performance.

Following this initial QC phase, we delve into a comprehensive exploratory analysis of the data. This includes examining the rlog-transformed expression matrix to better normalise the data distribution and mitigate the impact of large expression differences. A detailed boxplot analysis highlights the overall distribution of expression levels across samples, aiding in the visualisation of data spread and central tendencies.

Further, an IQR analysis is employed to identify samples that significantly deviate from the collective behaviour of the dataset, flagging potential outliers for further scrutiny. PCA and Multi-Dimensional Scaling (MDS) plots are then generated, akin to PCA in their goal to reduce data dimensionality and visually represent sample similarities and dissimilarities in a lower-dimensional space. These plots are instrumental in revealing underlying patterns, batch effects, or groups within the data, providing a visual summary of how samples relate to each other based on their gene expression profiles.

3.3.10 Normalisation and Differential Expression for NanoString nCounter Data

In our analysis of NanoString nCounter data, we employed a comprehensive suite of eight distinct normalisation methods, each tailored to address various aspects of technical and biological variability inherent in gene expression studies.

1. **Standard nSolver Normalisation:** This foundational approach combines Positive Control Normalisation, leveraging synthetic positive control targets,

with CodeSet Content Normalisation, which utilises housekeeping genes. Together, they adjust for technical variabilities across samples by applying a sample-specific correction factor to target probes.

2. **Housekeeping Scaling:** This normalisation technique transforms gene expression counts using a scaling factor derived from the geometric mean of the housekeeping genes within each sample. Housekeeping genes, selected for their stable expression across various conditions, act as a reliable baseline for normalisation. The pivotal aspect of this method is the division of the geometric mean of housekeeping gene expressions for each sample by the aggregate arithmetic mean of these geometric means across all samples. This procedure adeptly balances out inter-sample variability, curtailing the impact of technical variances ensuring the resultant data accurately mirrors true biological expression variations.
3. **Housekeeping with geNorm:** This approach employs the geNorm algorithm [199] to meticulously select housekeeping genes known for their stable expression across various conditions. By focusing on these genes, the method establishes a robust normalisation factor, calculated as the geometric mean of the expression levels of the chosen housekeeping genes. This calculated factor serves to standardise gene expression data across samples, providing a reliable basis for comparison. The strength of this method lies in its ability to minimise variability introduced by experimental conditions thereby enhancing the accuracy and consistency of expression level assessments across the dataset.
4. **Endogenous and Housekeeping Scaling:** This normalisation strategy integrates a holistic approach by utilising scaling factors derived from the geometric mean of counts from both endogenous genes and selected housekeeping genes within each sample. These scaling factors are then calibrated against the overall arithmetic mean of these geometric means across all samples. This method provides a balanced and comprehensive normalisation technique, ensuring that the variability due to technical factors is minimised while maintaining the biological integrity of the data. By incorporating both endogenous and housekeeping genes, this approach allows for a more nuanced adjustment of expression levels, reflecting a true representation of gene expression across the dataset. This dual consideration effectively addresses the

complexities of gene expression normalisation, offering a robust framework for accurate and reliable comparative analyses.

5. **Quantile Normalisation:** This method standardises the distribution of gene expression data across all samples within a dataset, achieving uniformity and comparability. Quantile normalisation operates by aligning each sample to a common distribution, utilising the average quantiles across the dataset to recalibrate individual data points. This technique effectively harmonises gene expression levels, ensuring that differences observed are due to biological variance rather than technical discrepancies. By aligning the expression profiles to a shared distribution curve, quantile normalisation mitigates the impact of outliers and batch effects, facilitating more accurate cross-sample comparisons and analyses.
6. **Cyclic Loess:** This method represents an advanced non-linear local regression technique designed to rectify variances among samples by applying pairwise normalisation. Cyclic Loess specifically targets the log-ratio of expression levels (M) and the mean average (A) of expression across samples, meticulously correcting for any non-linear distortions present within the dataset [200]. By iteratively adjusting these values, the method ensures a more accurate alignment of gene expression profiles between pairs of samples, effectively normalising the data across the entire study. This approach is particularly valuable in scenarios where the relationship between expression intensity and measurement error is complex and cannot be adequately addressed by simpler linear models. Cyclic Loess's capacity to fine-tune and harmonise the dataset underpins its utility in preparing nCounter data for subsequent analyses, ensuring that observed differences in gene expression are reflective of biological reality rather than artifactual variations.
7. **Variance Stabilisation Normalisation (VSN):** The VSN technique stands out for its sophisticated approach to data normalisation through a parametric transformation process. This method specifically targets the inherent variance-mean dependence observed in gene expression data, aiming to achieve a consistent variance across all levels of mean expression values [184]. By methodically modelling this relationship, VSN effectively stabilises the variance, ensuring that the data across the spectrum of mean values are rendered more comparable and analytically reliable. This stabilisation is crucial for enhancing the accuracy of downstream analyses, as it mitigates the effects

of heteroscedasticity—where the variability in gene expression data changes with the level of expression. Consequently, VSN facilitates a more robust and meaningful comparison of gene expression levels across different samples or experimental conditions, providing a solid foundation for the identification of genuine biological differences in gene expression studies.

8. **RUVSeq:** The RUVSeq package introduces a sophisticated approach to data normalisation through its RUVg function, specifically designed to mitigate technical biases in gene expression data [201]. This method hinges on the strategic use of reference genes to align the dataset, ensuring that technical variations do not overshadow biological signals. A key feature of this technique is the ability to select a subset of target genes, with the selection process augmented by geNorm’s algorithm to pinpoint the most stable genes. These stable genes serve as a reliable baseline for normalisation, effectively calibrating the dataset and enhancing the comparability of gene expression across samples. By focusing on these stable reference genes, RUVSeq addresses and corrects for unwanted variation, thereby improving the integrity and interpretability of the results. This approach is particularly valuable in NanoString nCounter-based studies, where unwanted variations can impact downstream analysis.

After the completion of the normalisation process, the analysis progresses to identifying differentially expressed genes, a task performed using the limma package in R [200]. Limma utilises an empirical Bayesian methodology to effectively identify genes that exhibit significant variations in expression across different experimental conditions. This approach is particularly adept at increasing the statistical power and reliability of the analysis, especially in studies with limited sample sizes, by borrowing strength across genes to stabilise variance estimates.

3.4 Machine Learning Methods

Throughout this thesis, ML techniques have been pivotal, serving as a foundational methodology across a variety of projects. The deployment of ML has spanned a range of essential processes: from the initial data preprocessing, which includes filtering to enhance dataset quality, to the meticulous selection of features and classifiers aimed at optimising model performance. Additionally, the adoption of

ensemble learning strategies has been crucial in improving predictive accuracy, while the employment of comprehensive evaluation metrics has facilitated a thorough assessment of each model's effectiveness. This holistic ML approach has formed the backbone of the analytical strategies employed, enabling a detailed and sophisticated exploration of complex datasets. In the ensuing sections, we will explore each of these fundamental ML aspects in greater depth, shedding light on their practical application and significance in the research presented.

3.4.1 Filtering Process

The filtering stage in our ML workflows involves two crucial steps aimed at refining the quality of the input datasets for more effective analysis. Firstly, the issue of multicollinearity is addressed by identifying and eliminating features that exhibit a high degree of correlation with one another. Specifically, any features demonstrating a Pearson correlation coefficient greater than 0.80 are excluded from further analysis. This threshold is selected to ensure that the remaining features provide unique information, enhancing the predictive power of our models while avoiding redundancy that could skew the results.

Secondly, we eliminate quasi-constant features from the analysed datasets. Quasi-constant features are those where a single value overwhelmingly dominates the observations, to the extent that the same value accounts for more than 99% of the data points for that feature. These features are considered to offer minimal variability and, by extension, limited predictive value for our models. By removing such features, we streamline our dataset, focusing on variables that contribute meaningfully to the variation in our data and thereby improving the efficiency and interpretability of subsequent analyses.

3.4.2 Feature Selection Techniques

Feature selection stands as a pivotal component in the ML pipelines across the projects detailed in this thesis, playing a vital role in improving model performance. By removing redundant or irrelevant instances (features), it simplifies model complexity and improves interpretability. Additionally, the practice of cross-validation, an essential technique in feature selection, is employed to ensure the robustness and generalisability of the model. Cross-validation systematically

partitions the data to validate the model on different subsets, providing a comprehensive assessment of its predictive power. Below, we delve into the specific feature selection methods employed, each chosen for its unique advantages in optimising our ML models.

Genetic Algorithm (GA)

The GA is a search heuristic inspired by the process of natural selection. This method is particularly effective for feature selection due to its ability to explore a vast search space and identify optimal feature subsets for model training. GA operates by generating a population of candidate solutions (feature sets) and iteratively improving them through operations such as selection, crossover, and mutation. By evaluating the fitness of each feature set based on model performance, GA selectively evolves towards the most promising feature combinations. This approach has been invaluable in our projects for tackling high-dimensional data, allowing us to navigate through thousands of potential features to uncover those that significantly contribute to predictive accuracy.

Recursive Feature Elimination with Cross-Validation (RFECV)

RFECV is a robust feature selection method that integrates the recursive feature elimination (RFE) process with cross-validated selection to find the optimal number of features. When used in conjunction with the Random Forest classifier, RFECV systematically removes the least important features (as determined by the classifier), while assessing model performance across multiple cross-validation folds. This ensures that the selected features are not only important, but also contribute to a stable and generalisable model. The Random Forest classifier, known for its excellent performance and feature importance measures, provides a solid basis for RFECV, allowing for the identification of a compact yet powerful subset of features that drive accurate predictions.

Permutation Feature Importance (PFI)

PFI is a model inspection technique that measures the increase in prediction error after shuffling each feature independently, thus breaking the relationship between the feature and the true outcome. When applied using the Gradient Boosting classifier, a powerful ensemble method that builds models sequentially to correct errors of the predecessors, PFI can accurately assess the value of each feature.

This method is particularly effective for identifying features that have a significant impact on model predictions, distinguishing between features that genuinely improve model performance and those that may contribute to noise.

Differentially Expressed (DE) Genes

In the context of biological data analysis, identifying DE genes is a critical feature selection strategy. DE genes are those that show significant differences in expression levels across different conditions or classes and are likely to be biologically informative. By focusing on DE genes as features, we can direct our ML models to concentrate on the most relevant biological signals, enhancing the relevance and specificity of our analyses to the underlying biological questions.

3.4.3 Classification Algorithms

In the diverse range of projects detailed within this thesis, a variety of classification algorithms have been employed, each selected for its specific strengths and suitability to the data at hand. These classifiers, central to the ML models developed, include AdaBoost Classifier, Logistic Regression, RandomForest Classifier, ExtraTrees Classifier, GradientBoosting Classifier, and KNeighbors Classifier. The choice of classifier was influenced by the characteristics of the biofeature matrices. Here, we delve into the specifics and advantages of each classifier used.

AdaBoostClassifier (AdaBoost)

The AdaBoost (Adaptive Boosting) Classifier is an ensemble technique that combines multiple weak classifiers to form a strong classifier. AdaBoost focuses on instances that are hard to predict, assigning higher weights to them in subsequent training rounds. This process creates a series of models that, when combined, improve the overall model's accuracy. AdaBoost's adaptability and ease of use have made it a valuable tool for classification tasks where the goal is to enhance model performance iteratively.

LogisticRegression (LR)

LR is a statistical method for analysing datasets in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (where there are only two possible outcomes). LR calculates the odds ratio in favour of the occurrence of an event, providing a powerful

framework for modelling binary outcomes with one or more explanatory variables. Moreover, it is used extensively in scenarios where the goal is to predict the presence or absence of a characteristic or outcome based on values of a set of predictor variables.

RandomForestClassifier (RF)

The RF classifier is a powerful estimator that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes of the individual trees. It is particularly well-suited to biofeature matrices due to its ability to handle high-dimensional data and its robustness to overfitting, making it a popular choice for complex classification tasks.

ExtraTreesClassifier (ET)

ET classifier is fundamentally similar to the RF Classifier, but introduces randomisation in the way splits are chosen at each node, making the trees more random. This method is often faster than its RandomForest counterpart and can achieve similar or sometimes better performance, especially in the presence of noisy features.

GradientBoostingClassifier (GB)

The GB classifier builds an additive model in a forward stage-wise fashion; it allows for the optimisation of arbitrary differentiable loss functions, making it a flexible choice for classification. Each new model incrementally reduces the loss function, making the overall model increasingly predictive. Its strength lies in its ability to capture complex interactions between features, offering precise predictions.

KNeighborsClassifier (KNN)

KNN implements the K-Nearest Neighbours vote, a type of instance-based learning or non-generalising learning. It does not attempt to construct a general internal model, but stores instances of the training data. Classification is computed from a simple majority vote of the nearest neighbours of each point. This method is highly intuitive and flexible, making it suitable for datasets where the decision boundary is irregular.

3.4.4 Ensemble Learning Techniques

Ensemble learning techniques have played a critical role in synthesising insights from complex datasets. These techniques, by combining the predictions from multiple models, have enhanced the predictive performance and reliability of our ML endeavours. Notably, we have leveraged majority voting, soft voting, and stacking approaches, each chosen for its ability to integrate and amplify the strengths of individual classifiers. Below is an elaboration on how these techniques were applied and their impact on the projects.

Majority Voting

Majority voting (or hard voting) is one of the simplest yet most effective ensemble methods. In this approach, each model in the ensemble votes for a single class, and the class receiving the majority of votes is chosen as the final prediction. This technique is particularly powerful when the models are diverse, as it can significantly reduce the variance and likelihood of overfitting, leading to more stable and reliable predictions. By aggregating the decisions of multiple classifiers, majority voting can often outperform even the best individual model in the ensemble.

Soft Voting

Soft voting refines the principle of majority voting through the integration of probability distributions for each class prediction made by models within the ensemble. Rather than each model contributing a singular vote towards a class, soft voting assigns weights to these votes based on the confidence or probability linked to each prediction. This technique enables a more sophisticated combination of predictions, factoring in the precision of each model's forecasts. In our application of soft voting for determining final label predictions, we discovered it offered greater adaptability and frequently improved accuracy over hard voting. This advantage was particularly pronounced in scenarios where predictions were tightly competitive, illustrating soft voting's capacity to harness the predictive strengths of individual models effectively.

Stacking

Stacking (stacked generalisation) involves layering models to use the predictions made by one layer of models as input for the next layer. Typically, the first layer consists of a diverse set of base models, and the final layer is a meta-model that learns how to best combine the predictions of the base models. In our projects,

we used a Random Forest classifier as the meta-classifier, due to its robust performance and ability to handle the feature importance variability of the base estimators' predictions effectively. The stacking classifier was built with estimators tailored to the specific requirements of the project, with the Random Forest meta-classifier trained to optimally integrate these base predictions. This stacking approach not only captured the strengths of individual models, but also learned the most effective way to combine these predictions for superior performance.

3.4.5 Evaluation Metrics

The assessment of ML models across various projects in this thesis was conducted using a comprehensive set of evaluation metrics. These metrics provide insight into different aspects of model performance, from overall accuracy to the balance between sensitivity and specificity. Below is a brief explanation of each metric and its significance:

Confusion Matrix (TN, FP, FN, TP)

The confusion matrix is a foundational tool in evaluating classification models, displaying the counts of true negatives (TN), false positives (FP), false negatives (FN), and true positives (TP). This matrix is crucial for calculating many other metrics and offers a detailed view of model performance.

Recall

Recall (or Sensitivity or True Positive Rate) measures the model's ability to identify all relevant instances, calculated as the number of true positives divided by the actual positives (true positives plus false negatives). It is critical in contexts where missing a positive instance has serious implications.

False Positive Rate

The False Positive Rate calculates the proportion of negative instances incorrectly classified as positive, emphasising the model's propensity to incorrectly signal a condition.

ROC AUC Score

The Receiver Operating Characteristic (ROC) curve visualises a model's capability to discriminate between positive and negative classes at various thresholds (different points at which the criteria for classifying observations into positive or

negative outcomes are set). The Area Under the ROC Curve (AUC) quantifies this discriminative power, with a score of 1 indicating perfect classification and 0.5 suggesting performance no better than random chance.

Accuracy

Accuracy measures the proportion of correct predictions (both true positives and true negatives) out of the total number of cases. It provides a straightforward metric of overall model performance, but may not fully reflect effectiveness in imbalanced datasets.

Balanced Accuracy

Balanced Accuracy amends the traditional accuracy metric by equally weighing the performance on each class. It calculates the average of the recall (or sensitivity) for each class, ensuring fair evaluation in datasets where class distribution is uneven.

Precision

Precision (or Positive Predictive Value) indicates the accuracy of positive predictions, calculated as the number of true positives divided by the total number of positive predictions (true positives plus false positives). It's especially relevant in situations where the cost of a false positive is high.

Average Precision Score

The average precision score summarises the precision-recall curve as the weighted mean of precision achieved at each threshold, reflecting the model's ability to identify positive instances with confidence across all levels of recall.

F1 Score

The F1 Score combines precision and recall into a single metric by taking their harmonic mean, offering a balance between the two. It's particularly useful when the costs of false positives and false negatives are similar.

F2 Score

The F2 Score places more emphasis on recall than precision, weighting false negatives more heavily than false positives. It is useful in scenarios where failing to detect a positive is more problematic than incorrectly identifying a negative as a positive.

False Negative Rate

The False Negative Rate measures the proportion of positives that were incorrectly classified as negatives, highlighting cases where the model misses a condition.

True Negative Rate

True Negative Rate (or Specificity) measures the proportion of actual negatives that are correctly identified, indicating the model's ability to recognise negative instances accurately.

Negative Predictive Value

This metric reflects the proportion of negative identifications that were actually correct, indicating the model's accuracy in predicting negative outcomes.

False Discovery Rate

The False Discovery Rate is the proportion of positive predictions that were false, providing insight into the model's error rate in falsely diagnosing conditions.

Cohen Kappa Metric

The Cohen Kappa Metric quantifies the agreement between two raters (or models) by adjusting for agreement that could occur by chance, offering a normalised measure of reliability beyond simple accuracy.

Matthews Correlation Coefficient

The Matthews Correlation Coefficient is a robust measure for binary classification, offering insight into the quality of predictions by considering all four quadrants of the confusion matrix. It's particularly effective for imbalanced datasets.

Log Loss

Log Loss, or logistic loss, measures the uncertainty of probability estimates by penalising false classifications, with a focus on the predicted probability's deviation from the actual label.

Brier Score

The Brier Score evaluates the accuracy of probabilistic predictions, penalising the squared difference between the predicted probability and the actual outcome, with lower scores indicating better predictions.

Likelihood Ratios

Likelihood Ratios (LRs) are pivotal in the medical field for interpreting diagnostic test results, offering insights into the test's ability to correctly identify those with and without a specific condition. They are particularly useful in assessing how likely a patient has a condition given a positive or negative test result. LRs come in two forms: Negative Likelihood Ratio (LR-) and Positive Likelihood Ratio (LR+), each providing specific insights into the test's diagnostic performance.

The Positive Likelihood Ratio (LR+) measures the extent to which the odds of having the condition increase when a test result is positive. It is calculated as the ratio of the test's sensitivity (true positive rate) to 1 minus its specificity (false positive rate). A LR+ greater than 1 indicates that a test result is more likely to be seen in a person with the condition than in one without, with higher values suggesting greater diagnostic accuracy.

Conversely, the Negative Likelihood Ratio (LR-) quantifies how much the odds of the condition decrease when a test result is negative. It is calculated as the ratio of 1 minus the test's sensitivity to its specificity. A LR- less than 1 suggests that a negative test result is less likely in individuals with the condition than those without, indicating the test's utility in ruling out the disease. The closer the LR- is to 0, the more effective the test is at correctly identifying those without the condition.

Diagnostic Odds Ratio

The Diagnostic Odds Ratio (DOR) is a comprehensive metric used to evaluate the effectiveness of a diagnostic test. It provides a single indicator that quantifies the test's discriminative power, essentially measuring how well the test can distinguish between individuals with and without the condition in question. The DOR is calculated by taking the ratio of the odds of a positive test result in individuals with the condition to the odds of a positive test result in individuals without the condition. This metric is particularly useful because it combines the test's sensitivity (True Positive Rate) and specificity (True Negative Rate) into a single number. A DOR of 1 indicates that a test does not discriminate between patients with and without the condition any better than random chance. A DOR greater than 1 suggests the test has discriminative power, with higher values indicating better discriminatory performance. Conversely, a DOR less than 1 would indicate poor test performance, where the test might incorrectly identify more individuals without the condition as having it, and vice versa.

Chapter 4

Assessing the complementary information of biologically relevant features in lbRNA-Seq data

“A thinker sees his own actions as experiments and questions—as attempts to find out something. Success and failure are for him answers above all.”

— FRIEDRICH NIETZSCHE

This Chapter introduces the Ensemble Learning for Liquid Biopsy Analysis (ELLBA) methodology, a novel and comprehensive approach for analysing liquid biopsy RNA sequencing (lbRNA-Seq) data in cancer diagnostics. ELLBA leverages six distinct biofeature types to capture diverse molecular characteristics of cancer. Employing robust intra-sample normalisation and ensemble classification methods, ELLBA surpasses traditional gene expression analysis in predictive accuracy. The methodology was evaluated across six datasets and four independent validation sets, covering a variety of cancer types and biosources. This approach marks a significant advancement in lbRNA-Seq analysis, offering a more holistic understanding of cancer biomarkers and paving the way for personalised cancer care.

4.1 Introduction

Human body biofluids, such as blood, urine, and saliva, have proven to be a rich and valuable source of information about an individual's health status [202]. The burgeoning field of liquid biopsy (LB) research has been actively exploring these resources with the aim of unlocking their full potential for diagnosis, monitoring, prognosis, and treatment response assessment in various diseases, including cancer [203], [204]. LB's most effective assets lie in its repeatability, cost-effectiveness, and minimal invasiveness.

Advancements in technology and computer science have propelled LB research. The advent of Next Generation Sequencing (NGS) technology, combined with continuous improvements in bioinformatics have deepened our understanding of the molecular landscapes in LB samples, revealing insights into disease mechanisms and the discovery of potential biomarkers [205]–[208]. Several studies investigating blood-based biosources like Tumour-Educated Platelets (TEPs), Extracellular Vesicles (EVs), Circulating Epithelial Cells (CECs), and Circulating Tumour Cells (CTCs) have notably unveiled a range of diagnostic signatures that hold promise for the early detection of prominent cancers, harnessing the power of mRNA sequencing (mRNA-Seq) [89], [209], [210]. For instance, Antunes-Ferreira, D'Ambrosi, Arkani, *et al.* [211], report an Area Under the ROC Curve (AUC) of 0.88 through an 881 RNA biomarker panel to predict outcomes in Non-small Cell Lung Carcinoma patients.

Despite the great promise and certain advancements in the field, the current focus on LB-based transcriptomics has predominantly been centred around gene expression profiling, partly due to the lack of comprehensive pipelines tailored for laboratories with limited bioinformatics resources. Consequently, biofeatures such as isoform expression, fraction of canonical transcript (FoCT), gene fusion, RNA editing, and single nucleotide variants (SNVs) have remained largely uncharted, representing untapped sources of valuable insights. Specifically, the simultaneous usage of these biofeatures has not been explored in previous research, promising substantial potential for enhancing our understanding of LB data.

Moreover, the predominant application of cross-sample normalisation methods, while beneficial for prediction accuracy in enclosed, frequently single lab study designs, fosters challenges for clinical applications where intra-sample normalisation is mandatory to classify individual clinical samples applying a fixed prediction

model. Additionally, the lack of independent test sets in many studies raises concerns about the reproducibility and generalisability of reported results, rendering metrics like AUCs potentially misleading.

The untapped wealth of information within LB-derived RNA-Seq (lbRNA-Seq) data presents an opportunity for more comprehensive and clinically relevant insights. To fully unlock the potential of LB data, a comprehensive exploration of complementary biological information available in a lbRNA-seq sample is imperative in order to gain an encircled understanding of such biosources. However, integrating this high amount of heterogeneous data into a single prediction model is challenging. Moreover, due to the high-dimensional nature of lbRNA-seq data, machine learning (ML) approaches have become indispensable for detecting patterns and gaining a deeper understanding of the underlying biological conditions [212].

In this study, we introduce ELLBA (Ensemble Learning for Liquid Biopsy Analysis), a methodology designed to tackle the complexities of LB data and enhance the predictive modelling of patient, with applicability to clinical settings. ELLBA encompasses six biologically motivated feature types: gene expression, isoform expression captures alternative splicing, FoCT quantifies predominant transcript shifts, gene fusion detects structural changes, RNA editing indicates post-transcriptional modifications, and SNVs unveil potential mutations. Each biofeature type addresses different molecular properties that can be altered in pathologies like cancer, offering therefore diagnostic and prognostic value. In the context of existing literature, it is noteworthy that virtually all published LB-based studies are based on Gene Expression or, at most, SNVs. No comparable study or methodology exists harnessing the complementary information contained in 6 different biofeatures providing a unified decision output and applicability to clinical data. This innovative strategy distinguishes our approach, marking a significant advancement in the field of lbRNA-Seq analysis. Given the absence of a comparable workflow, our methodology focuses on comparing the final ensemble output to the standard Gene Expression results. Finally, the modelling part of the methodology utilises Ensemble Classification Methods to combine complementary information from these features.

ELLBA was rigorously evaluated across six datasets and four independent validation sets, encompassing around 2,500 samples, covering various cancer types and biosources. Our work highlights the utility of the rather simple intra-sample Count

Per Million (CPM) normalisation in clinical settings. We show that while the best normalisation method depends both on the data type and employed ML model, in general, CPM performs equally well compared to more sophisticated cross-sample methods. Moreover, our study demonstrates that Ensemble Learning is effectively leveraging the complementary information contained in the different biofeature types, always improving the prediction power over the best individual biofeature type. Interestingly, the improvement seems to be especially pronounced when evaluating independent test sets, which might indicate the robustness and reproducibility of the discriminative biofeatures detected by ELLBA. In summary, our workflow improves prediction accuracy and streamlines clinical decision-making, contributing to personalised cancer care (Figure 1).

4.2 Materials and Methods

4.2.1 Workflow and Implementation

The ELLBA workflow can be easily installed using a Docker image. The source code and exact installation instructions are available on GitHub. The workflow integrates established bioinformatics tools with novel algorithms for data processing, biofeature generation, and ML analysis. ELLBA was primarily developed using Python (v3.8) with supplemental R (v3.6.3) scripts. ML analysis relies on the scikit-learn (v1.2.0) Python package [213]. Table 1 presents a comprehensive summary of all the software and packages utilised in the workflow, along with their corresponding versions.

We employed data from 10 different LB studies, encompassing a total of 2,479 publicly available samples from the SRA repository [214]. The studies span different types of LB data, including TEPs, EVs, and CECs. To initiate the ELLBA workflow, in addition to the raw fastq files, a sample sheet specifying at least the sample name and group label (e.g., control, cancer) is required. Although we conducted rigorous benchmarking in terms of normalisation and ML, this section outlines the final workflow configuration.

4.2.2 Data Preprocessing and Biofeature Extraction

Artificial adapter sequences and low-quality reads (average Q below 20 or shorter than 40nt) are automatically detected and removed by the BBDuk (v38.18) tool [162]. Furthermore, the workflow provides multi-sample quality reports through multiQC (v1.13) [163].

Genome mapping was performed by the STAR (v2.7.6a) aligner [164] using the human reference genome GRCh38.p13 primary assembly as well as the GENCODE v35 reference gene annotation [165]. STAR was employed with ‘GeneCounts’ and ‘TranscriptomeSAM’ parameters to obtain count matrices at both gene and transcript levels. Alignment quality metrics were extracted using RSeQC (v3.0.0) and Picard tools (v2.23.3) [167], [168] to evaluate the alignment process and a final summarised report is being generated.

Based on the STAR-generated BAM files, we generate a total of six biologically motivated feature types: (1) Gene expression, (2) Isoform expression, (3) FoCT, (4) Gene fusion, (5) RNA editing and, (6) SNV. A detailed overview of each biofeature type can be found in Table 4.1.

4.2.3 Gene Expression

To quantify gene expression, we collected read abundances from the GeneCounts-based generated files and created a single expression matrix that encompassed the quantification results for all samples. Subsequently, we performed a principal component analysis (PCA) on the gene expression matrix to identify potential batch effects in the data. Furthermore, an Interquartile Range (IQR) analysis was employed to identify any potential outlier samples.

Following the exploratory analysis, we utilised the ‘filterByExpr’ function from the edgeR (v3.28.1) package [185] to remove genes with low expression levels across samples. Subsequently, we applied the standard CPM normalisation followed by the MinMaxScaler function, from the sklearn package, to further transform the gene expression data ensuring that all feature values are on a comparable scale.

TABLE 4.1: Implementation details on the six biofeature types utilised in this study.

Biofeature type	Biological rationale
Gene expression	Gene-level expression profiles provide information on the transcriptional activity of each gene in the sample. It is a measure of how active a gene is and determines the abundance of RNA molecules produced from that gene. Gene expression plays a crucial role in determining an organism's traits and functions. Consequently, perturbations in gene expression, driven by diseases, can lead to substantial alterations.
Isoform expression	Isoform expression is the measurement of different splice variants or isoforms of a gene's RNA transcripts. Alternative splicing allows genes to produce multiple isoforms with sometimes different functional characteristics. Isoform expression profiling reveals gene product diversity and potential disease associations.
FoCT	FoCT is designed to assess the predominant canonical transcript shift in each gene using a default group. Changes in the frequency of alternative splicing events are quantified by means of the fraction of the canonical transcript. The rationale behind this metric is that in cancer, frequently the splicing pathway is affected, increasing the transcriptional variation for at least certain genes. Under this scenario, it might be less important to correctly identify the different isoforms, but to robustly quantify the existence of a differential amount of alternative transcripts.
Gene fusion	Gene fusion detection involves identifying abnormal fusion events between two genes, which can arise from chromosomal rearrangements or translocations. Fusion events can create chimeric RNA transcripts or fusion proteins that are often associated with disease.
RNA editing	RNA editing is a post-transcriptional modification process that alters the nucleotide sequence of RNA molecules, leading to changes in the encoded protein or functional non-coding RNA. This process is crucial for expanding the functional diversity of the transcriptome and can impact gene regulation, protein structure, and function.
SNV	Single Nucleotide Variants (SNVs) can alter protein structure, function, or gene regulation based on their location in the coding or regulatory sequences. SNV analysis aids in discovering disease-associated (driver) mutations.

4.2.4 Isoform Expression

For quantification at a transcript level, we employed Salmon (v1.9.0) software [169] in alignment-based mode, using STAR TranscriptomeSAM output. Default parameters were employed, along with additional settings including `seqBias` and `gcBias` enabled for sequence and GC content biases correction, `libType U` for unstranded data, and 100 bootstrap iterations for robust quantification. The resulting transcript-level expression matrix was obtained by merging quantification outputs using Salmon’s `quantmerge` script, selecting the `numreads` column for the final matrix.

Similar to gene expression analysis, we performed exploratory analysis, normalization, and transformation on the transcript-level expression matrix. These steps followed the same approach as described in Section 4.2.3 for gene expression analysis.

4.2.5 Fraction of Canonical Transcript

We randomly selected 20 control individuals and extracted the most abundant transcript of each gene (canonical transcript) based on the isoform expression levels in these samples. This list of transcripts was used to convert the transcript expression matrix into a canonical transcript matrix. The transformed matrix considered only the most abundant transcripts, dividing their expression levels by the total counts of all transcripts originating from the same gene. This yielded the FoCT for each sample, which was then combined into a single matrix. Any features with missing values (NA) were entirely removed. Finally, the `StandardScaler` function was applied to transform the frequency feature matrix.

4.2.6 Gene Fusion

Gene fusions were identified using the Arriba (v2.1.0) software [166], [215], with specific parameters adjusted within the STAR aligner for fusion gene detection. The Arriba software was utilised with default parameters. Following detection, gene fusions were sorted alphabetically, and gene IDs were appropriately adjusted. Specifically, the first gene ID within the fusion nomenclature was retained, while the second part was omitted. Furthermore, instances of the same first gene ID

were merged into a unified entry, resulting in a count matrix. This matrix was subsequently transformed into a binary format, featuring exclusively 0 or 1 values. In this binary configuration, a value of 0 denotes the absence of fusion detection within a sample, while a value of 1 signifies its presence.

4.2.7 RNA Editing Events

The genome-aligned BAM files were subjected to de-duplication using the `rmDup` function of `samtools` (v1.7) [170]. Additionally, GATK BaseRecalibrator (v4.1.9.0) [171] was employed to recalibrate the base quality scores. Subsequently, BCFtools `mpileup` (v1.7) [172] was utilised on the pre-processed BAM files with the following parameters: `min-MQ 15`, `min-BQ 15`, `redo-BAQ`, `per-sample-mF`, and `min-ireads 2`. Variant calling was performed using BCFtools `call` with the parameters `ploidy GRCh38`, `variants-only`, and `multiallelic-caller`. Known common variants from the dbSNP [173] database (`common_all_20180418`) were excluded from the analysis.

RNA editing events were detected using REDIttools (v2.0) software [174]. The software was configured with the following parameters: `min-edits 2`, `min-read-quality 18`, and `min-base-quality 15`. RNA editing events were filtered at the sample level based on a minimum depth per site of 10, a mean quality score per site of at least 20, and a minimum substitution frequency of 0.3. All RNA editing events that passed the quality filters were merged into an overall feature matrix and further filtered to include only events that were present in at least 20% of the samples. The resulting RNA editing event matrix was transformed into a binary format, where 0 indicated the absence of an event and 1 indicated its presence.

4.2.8 Single Nucleotide Variants

Up to the common variant (SNP) filtering step, an identical protocol to RNA editing analysis was followed. After this stage, positions with quality scores below 20 and a minimum depth of 2, along with previously identified RNA editing sites, were filtered out. BCFtools `merge` was then employed to consolidate the filtered VCF files, producing a unified matrix across all samples. Subsequently, GATK `VariantsToTable` (v4.1.9.0) [216] extracted the genotype field for each variant from

the filtered VCF file, converting the data into a tab-delimited table format. SNV values were discretised as follows: 0 for no alternative allele, 0.5 for heterozygous calls, and 1 for homozygous positions. Lastly, variants occurring in less than 20% of samples were filtered out.

4.2.9 Machine Learning and Ensemble Learning Implementation

To be able to generate a robust model that may classify each input sample within the datasets, ML techniques were employed on each of the six distinct biofeature spaces extracted from the data. Each dataset was initially split into training and test sets. For datasets with an independent external validation dataset, this was designated as the test set, while the main dataset served as the training set. In the absence of an external validation set, a random 70-30 split was performed, with 70% of the data used for training and the remaining 30% as an approximation to an independent test set.

Feature selection and model training were conducted on the training set, followed by final validation on the test set. Initially, highly correlated features (Pearson's correlation above 0.8) and quasi-constant features (with 99% similarity) were removed from further analysis.

The filtered training biofeature matrices underwent feature selection using the GeneticSelectionCV function, a Python implementation of the genetic algorithm (GA) [217]. The GA operates in a wrapper-like mode, systematically searching for the optimal set of features for the classification task. The choice to utilise the GA for feature selection stems from its efficiency in managing well vast high-dimensional biofeature spaces within a relatively short timeframe. As a population-based metaheuristic algorithm, the GA employs multiple candidate solutions during the search process. It excels in exploring diverse biofeature combinations comprehensively, proving notably faster and less computationally expensive—ideal for large-scale datasets. In contrast to standard methods like Recursive Feature Elimination (RFE) or Univariate methods, the GA offers unique advantages. RFE, while effective, can be computationally demanding and time-intensive, especially with high-dimensional data, such as Isoform Expression with

nearly 34,000 features. Univariate methods, on the other hand, may overlook intricate feature interactions, limiting their ability to capture nuanced patterns in the data. The GA, with its capacity to consider feature interactions and navigate vast feature spaces efficiently, provides a more robust and holistic approach to feature selection in the context of LB-based datasets.

The GA was utilised with stratified 5-fold cross-validation (CV), employing empirical parameters including `n_population = 130`, `n_generations = 130`, `scoring = "accuracy"`, and `max_features = 50`, a choice justified by balancing model complexity and prediction performance. For each training biofeature type matrix used in GA-based feature selection, an appropriate base estimator was selected. It's worth noting that the `GeneticSelectionCV` function employs the selected base estimator to evaluate the fitness of different feature subsets during the GA process. Specifically, the Random Forest classifier was selected as the base estimator for Gene Expression, the SVM classifier for Isoform Expression, and the Logistic Regression classifier for the remaining biofeatures (Table 2).

Kindly take note that, going forward, when we refer to selecting or utilising an appropriate base estimator, we specifically mean choosing or using the underlying method that is employed by algorithms like `GeneticSelectionCV` or `AdaBoost`. This subtle yet crucial distinction is pivotal for comprehending how ensemble learning harnesses the unique strengths of distinct base estimators to enhance overall predictive accuracy.

In particular, for both feature selection and modelling, we explored a range of standard and diverse classifiers to identify the most suitable ones. This set included classifiers such as `AdaBoost`, K-Nearest Neighbours (KNN), support vector machine with a linear kernel (`LinearSV`), Logistic Regression, Naive Bayes, and Random Forest. Model training was conducted using stratified 5-fold CV. Table 2 provides a detailed overview of the final feature selection and classifier combinations.

Ensemble learning was employed to combine the individual information from all six distinct biofeature types and enhance the predictive performance of each sample. To ensure the inclusion of reliable features, a minimum mean accuracy score of 0.65 during CV in the training process was set as an empirical eligibility criterion. The soft voting strategy was then applied to make the final label prediction, averaging the aggregating predictions based on the probability distribution of class labels.

4.2.10 Functional Enrichment Analysis

To perform functional enrichment analysis, we utilised the online GOst tool provided by the gProfiler [186] web service. This tool facilitated the Gene Ontology (GO) and pathway enrichment analyses on the genes selected through the GA for each biofeature type. Additionally, we conducted enrichment analysis on the combined biofeature set, incorporating all selected features.

4.3 Results

4.3.1 Data Collection and Description

Our study design comprises data from ten different LB-based studies encompassing a total of 2,479 samples [74], [76], [89], [127], [210], [218]–[222]. The data were obtained from publicly available sources and consisted of short-read RNA sequencing (RNA-Seq) data.

Various sequencing protocols, including mRNA-Seq, Extracellular Vesicles Long RNA Sequencing (exLR-Seq), single-cell RNA sequencing (scRNA-Seq), read lengths (100, 150, and 250), read types (single and paired-end), and sequencing methods (bulk and single-cell) were included into the analysis.

The collected data were derived from three distinct blood-extracted biosources: TEPs, EVs, and CECs. TEP-derived data comprise the majority accounting for over 1900 samples, while approximately 450 samples were derived from EVs, and the remaining samples originated from CECs (Figure 4.1).

Furthermore, our study was particularly focused on six different cancer types, each represented by a unique acronym: Non-small Cell Lung Carcinoma (NSCLC) for lung cancer, Glioblastoma Multiforme (GBM) for brain cancer, Colorectal Cancer (CRC) for colon cancer, Esophageal Squamous-Cell Carcinoma (ESCC) for esophageal cancer, Pancreatic Ductal Adenocarcinoma (PDAC) for pancreatic cancer, and Hepatocellular carcinoma (HCC) for liver cancer. These well-defined cancer types serve as the foundation for our analyses, and we refer to each of the six datasets by their respective cancer type acronyms: NSCLC, GBM, CRC, ESCC, PDAC, and HCC.

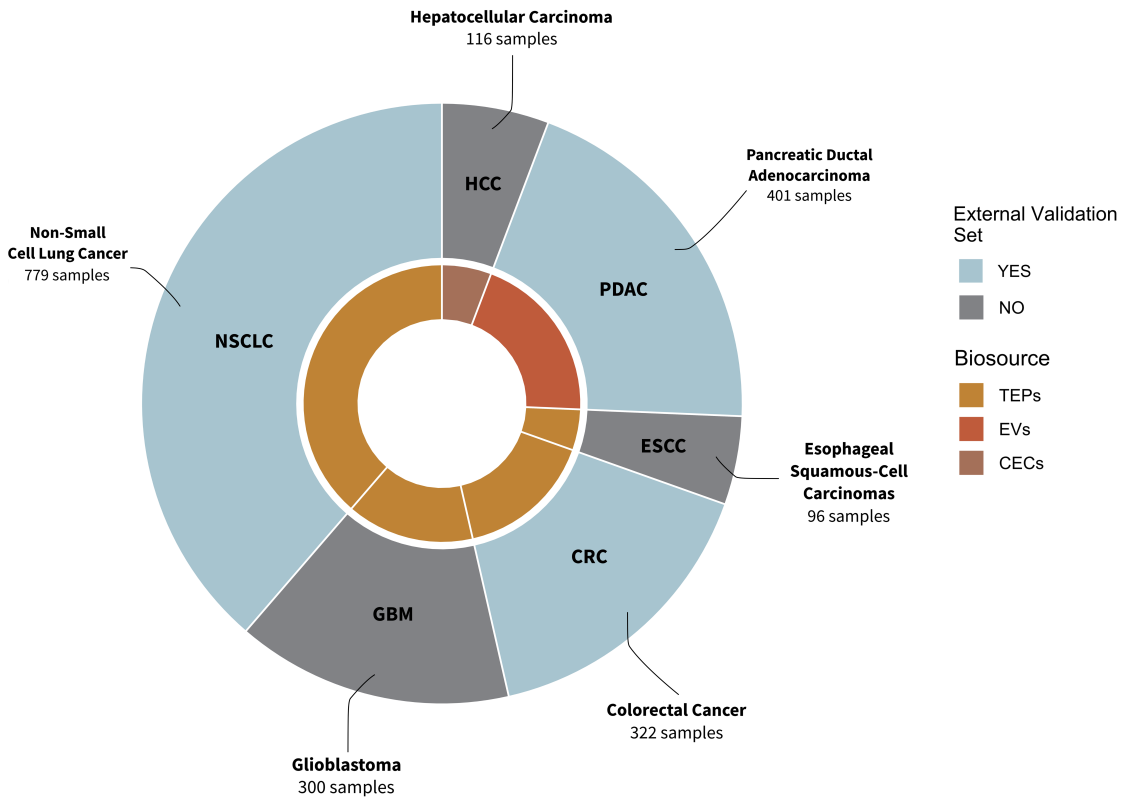


FIGURE 4.1: **Comprehensive Overview of Datasets in the Study.** A detailed overview of the datasets employed in this study, showcasing six distinct datasets: NSCLC, GBM, CRC, ESCC, PDAC, and HCC, as depicted in the outer donut plot. Light blue colouring (NSCLC, CRC, and PDAC) signifies datasets with independent external validation sets, while grey shading (GBM, ESCC, HCC) represents datasets without external validation. The inner circle categorises the biosource origin of each dataset: dark yellow for TEPs in NSCLC, GBM, CRC, and ESCC; cinnamon red for EVs in PDAC; and brown for CECs in HCC.

To ensure the robustness of our analysis, we divided the collected studies into two main subsets: a training set and an independent external validation testing set, when available. The training set comprised samples from six of the ten studies, totalling approximately 2000 samples. The remaining four studies were exclusively used for the external validation testing set. To be more precise, within the set of six datasets, three (NSCLC, CRC, and PDAC) are accompanied by independent external validation datasets. For instance, in the NSCLC dataset, which encompasses 779 samples comparing NSCLC to non-cancer TEP samples, sequenced using an SE100 mRNA-Seq protocol, we identified an external validation dataset that perfectly aligns with the same sequencing protocol and cancer type. Similarly,

the CRC dataset, comprising 322 samples, employed PE100 mRNA-Seq sequencing to distinguish between Colorectal Cancer and Non-Cancer samples. Its corresponding external validation set maintained the cancer type and utilised an SE100 mRNA-Seq protocol. In the context of PDAC, featuring 401 samples focused on Pancreatic Ductal Adenocarcinoma versus Healthy Controls, the sequencing followed a PE150 exLR-Seq protocol. The matching external validation set, derived from two publications, also maintained consistency in terms of cancer type and sequencing protocol. Table 3 provides a detailed overview of the datasets, including their configuration, utilisation, and accession information. Kindly consult the Supplementary Materials Section A.1.1 for a detailed account of the specific procedures and analyses applied to individual biofeatures in the utilised datasets.

4.3.2 Overview of the ELLBA Methodology

The ELLBA methodology is organised into two core components: bioinformatics and ML. It consists a total of seven distinct modules. The initial four modules, namely Input, Mapping, Biofeature Extraction, and Biofeature Processing, constitute the bioinformatics phase of the analysis. The subsequent three modules, Feature Selection, Classification, and Decision Output, are focused on ML. The entire workflow is depicted in Figure 4.2. Each module within this framework performs a specific set of tasks, which can be summarised as follows:

Input module: Adapter trimming and quality control (Figure 4.2A).

Mapping module: Genome alignment and gene profiling, including alignment quality controls (Figure 4.2B).

Biofeature extraction module: Six distinct biological features are extracted (Table 4.1): (i) gene-level expression profiles, (ii) isoform-level expression profiles, (iii) FoCT, (iv) gene fusion quantification, (v) RNA editing, and (vi) putative somatic SNV (Figure 4.2C). Table 4 provides a numerical overview of the extracted biofeatures before any filtering.

Biofeature processing module: Normalisations and discretisation techniques are applied prior to filtering low-quality and non-discriminative biological features like lowly expressed genes or common germline variants (SNPs) (Figure 4.2D).

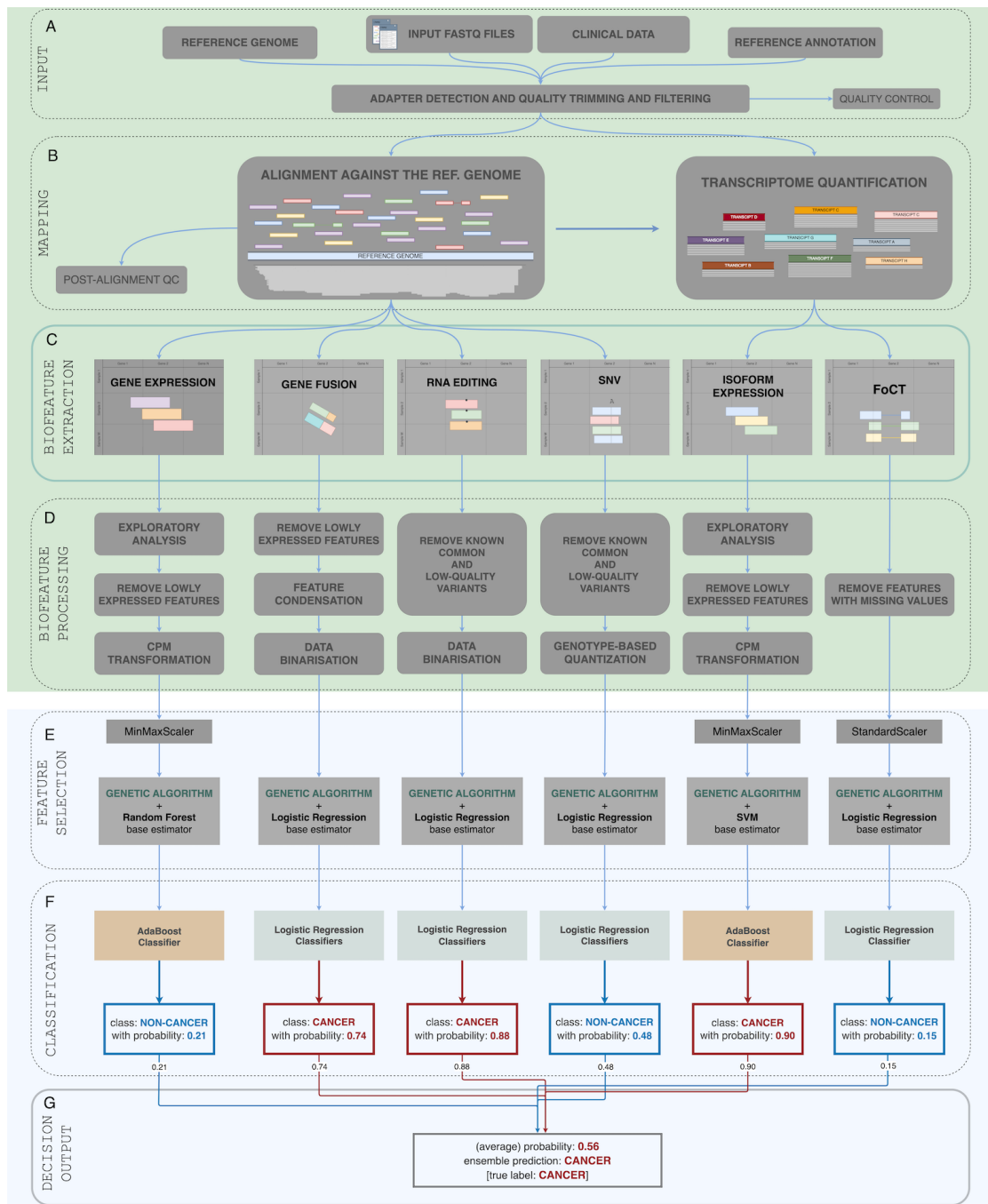


FIGURE 4.2: **Overview of the ELLBA Workflow.** ELLBA methodology features two main components: bioinformatics analysis (light green) and ML (light blue). The workflow includes seven modules. Bioinformatics analysis involves Input, Mapping, Biofeature Extraction, and Biofeature Processing. ML includes Feature Selection, Classification, and Decision Output. The process starts with data Input and progresses through Mapping, Biofeature Extraction, and Biofeature Processing for bioinformatics analysis. Then, Feature Selection, Classification, and Decision Output handle ML analysis. Biofeatures are individually processed in Biofeature Processing, involving data cleaning and normalisation or discretisation. In ML, Feature Selection and Classification are applied to each biofeature. The Decision Output combines individual classification outputs using ensemble learning (soft voting) for the final decision.

Feature selection: For each specific biological feature type, feature selection is performed using the GA in conjunction with a designated base estimator (Table 2 and Table 5) tailored to that particular biofeature type (Figure 4.2E).

Classification module: Following feature selection, each biofeature classification using standard ML models. The class confidences generated are then retained for subsequent use in the final Decision Output module (Figure 4.2F).

Decision Output module: To leverage the complementary information offered from each biofeature type, by default ensemble soft voting classification is applied. This method combines the predicted probabilities from all biofeature matrices, aggregating them into a single, consolidated average prediction. In this context (Figure 4.2G), each predictive model generates a label (either "Non-cancer", highlighted in blue, or "Cancer", highlighted in red) along with an associated probability displayed beneath the respective label. During the soft voting process, all these output predictive probabilities are consolidated through averaging, culminating in the ultimate decision (as demonstrated by "Cancer" in the figure). Additional details about soft voting can be accessed in the Supplementary Materials Section A.1.2.

4.3.3 Comparative Analysis of Normalisation Methods for Gene and Isoform Expression Data

Normalisation of gene and transcript expression data is a crucial step in analysing raw-count matrices. While the remaining biofeature types (FoCT, Gene fusion, RNA editing, and SNV) are inherently normalised as ratios or discrete values, count matrices require normalisation to account for variations in read yield and technical artefacts. Several methods with different assumptions have been implemented for this purpose [223], and their performance varies depending on whether these assumptions are met [224]. Most commonly, cross-sample normalisation methods are applied. Examples of such approaches include TMM and TMMwsp from the edgeR package, RLE from DESeq2, RUV from the RUVseq package (which can also address and rectify batch effects), as well as quantile-based methods like Full Quantile (FQ) and Upper Quartile (UQ) normalisation. While these cross-sample methods often exhibit superior performance in benchmark studies, they have a notable drawback: the normalisation outcome for a specific gene and

sample is influenced by the values of other samples. This characteristic hampers their utility in clinical settings, where the objective is the normalisation of individual samples and its application to fixed prediction models.

To address this limitation, we evaluated the performance of two intra-sample normalisation methods, CPM and Reads Per Kilobase per Million mapped reads (RPKM), and the aforementioned six cross-sample normalisation methods for all six datasets. This evaluation was carried out by incorporating these normalisation methods into the "Biofeature processing module" step of the pipeline and assessing their impact on the results. To mitigate the influence of specific ML models, our analysis encompassed six diverse algorithms: AdaBoost, KNN, LinearSV, Logistic Regression, Naive Bayes, and Random Forest.

Performance assessment was conducted on all six training datasets applying a 5-fold CV approach with exactly the same folds for each biofeature type. For an in-depth examination, refer to Section A.1.3 in the Supplementary Materials. The average AUC is then used as the principal quality measure. The results of the gene expression normalisation, presented in Figure 4.3, consistently demonstrated that CPM normalisation, despite variations between datasets and ML models, performed comparably to the more sophisticated cross-sample methods. Specifically, CPM normalisation yielded the highest collective mean AUC of 0.81 across all datasets, while RPKM exhibited the lowest performance with 0.58. RLE normalisation ranked second highest with a collective mean AUC of 0.79. On an individual dataset basis, CPM normalisation consistently exceeded other methods, except in the CRC dataset, where FQ normalisation achieved a slightly higher mean AUC (0.73) compared to CPM (0.70). It is worth noting that, despite employing RUVSeq normalisation to account for batch effect, our results showed that CPM normalisation outperformed RUVSeq. This observation, indicating that batch effect correction did not significantly impact the ML analysis, highlights the robustness of CPM normalisation in combination with ML downstream analysis.

Similar trends were observed in the analysis of isoform expression normalisation (Figure 2), following the same evaluation protocol as the gene expression analysis. More specifically, CPM normalisation demonstrated favourable performance, achieving the highest score with a collective mean AUC of 0.80 across all datasets, while RPKM displayed the lowest performance with 0.60. FQ normalisation was rated second, with a collective mean AUC of 0.78. Notably, when evaluating each dataset individually, CPM normalisation consistently outperformed the other

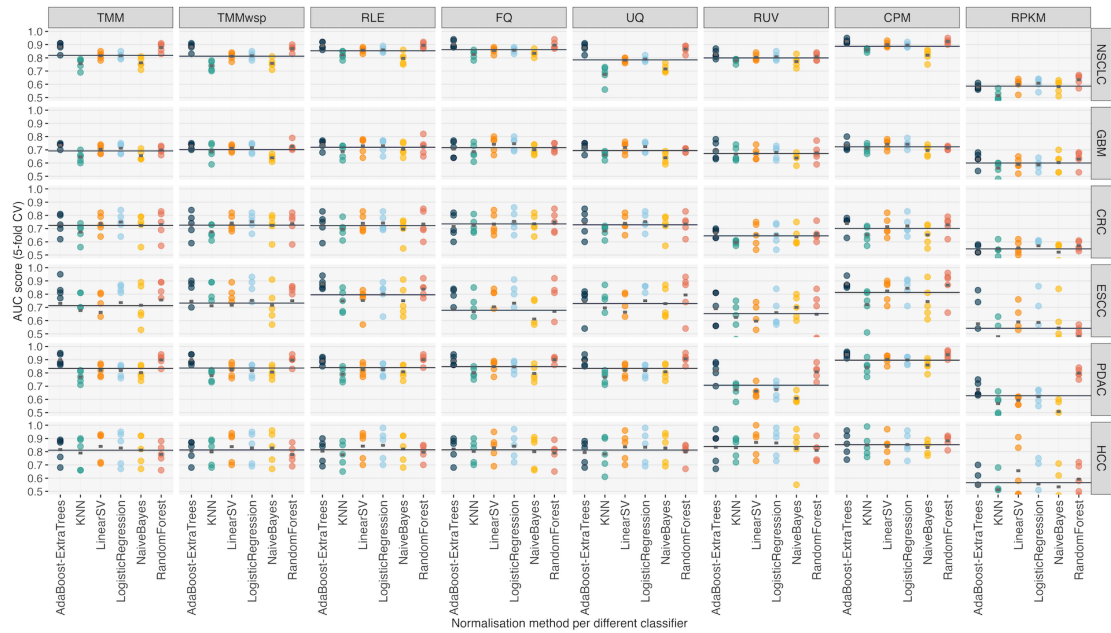


FIGURE 4.3: **Summary of Various Normalisation Methods.** A total of eight normalisation techniques were assessed across all six datasets. Each column corresponds to a distinct normalisation method, while each row represents a different dataset employed. The x-axis illustrates the various models utilised within each normalisation. and the v-axis depicts the mean AUC score achieved through 5-fold CV. Each model is represented by dots indicating the AUC for each CV fold. Additionally, a dashed line indicates the mean AUC across the 5 folds. while a solid line represents the mean AUC across all models.

methods, with the exception of the GBM and CRC datasets. In the GBM dataset, RLE normalisation achieved a slightly higher mean AUC of 0.72, compared to CPM’s 0.70. Similarly, in the CRC dataset, TMMwsp exhibited a mean AUC of 0.66, while CPM achieved 0.65. Notably, CPM normalisation not only provided very robust results, but also exhibited significant clinical value. Unlike other methods, CPM normalisation can normalise a single sample independently, making it more time-efficient for clinical applications.

4.3.4 Optimal Classification Models for the Different Biofeature Types

Having established CPM as the default normalisation method for expression-based features, we extensively explored the performance of six diverse classifiers: AdaBoost, KNN, LinearSV, Logistic Regression, Naïve Bayes, and Random Forest, using all six biofeature types and datasets. These six classifiers were meticulously

selected based on their algorithmic diversity, computational efficiency, and their capacity to provide a balanced approach to classification, incorporating both linear and non-linear techniques. Their selection also took into account the specific characteristics of our dataset and the nature of the features extracted. Emphasising diversity was crucial, as these classifiers comprise different modelling paradigms, extracting the most appropriate patterns and discrimination functions regarding the distinct feature spaces of the problems involved. Moreover, we have opted for the use of standard libraries and well-validated classification methods, ensuring the robustness and reliability of our software implementation. Furthermore, we acknowledge the need for diversity not only in individual classifier performance, but also in the success and good behaviour of the multi-classification approach, i.e., ensemble combination. Recognising that the joint use of different methods ensures a good trade-off among their potentially different predictions, we have placed a strong emphasis on diversity to enhance the effectiveness of the ensemble method. This approach aligns with the need for a comprehensive and well-balanced strategy, considering the varying strengths of each classifier within the ensemble framework. Additionally, their widespread use in similar bioinformatics contexts enhances comparability with existing literature. The performance of each classifier was evaluated using the same stratified 5-fold CV approach on each dataset, and the average AUC was reported (Figure 4.4). To determine the most suitable classifier for each biofeature type, we calculated the average mean AUC across all datasets for that biofeature type. Our findings revealed that expression-based feature matrices yielded the highest performance when paired with the AdaBoost classifier using the ExtraTrees as base estimator. Specifically, AdaBoost achieved the highest mean AUC of 0.86 for gene expression, while KNN exhibited the lowest mean AUC of 0.78. LinearSV, Random Forest, and AdaBoost demonstrated similar performance for isoform expression with a collective mean AUC of 0.83, while KNN yielded the lowest mean AUC of 0.77. Consequently, we selected the AdaBoost classifier with the ExtraTrees base estimator as the optimal choice for both gene and isoform expression analyses.

Conversely, Logistic Regression emerged as the most suitable classifier for the remaining biofeature matrices, including FoCT, fusion genes, RNA editing, and SNVs. More specifically, Logistic Regression achieved the highest collective mean AUC of 0.83 for FoCT, while Naïve Bayes exhibited the lowest mean AUC of 0.72. In the case of fusion genes, both Logistic Regression and Naïve Bayes performed equally, yielding a collective mean AUC of 0.54, while LinearSV demonstrated

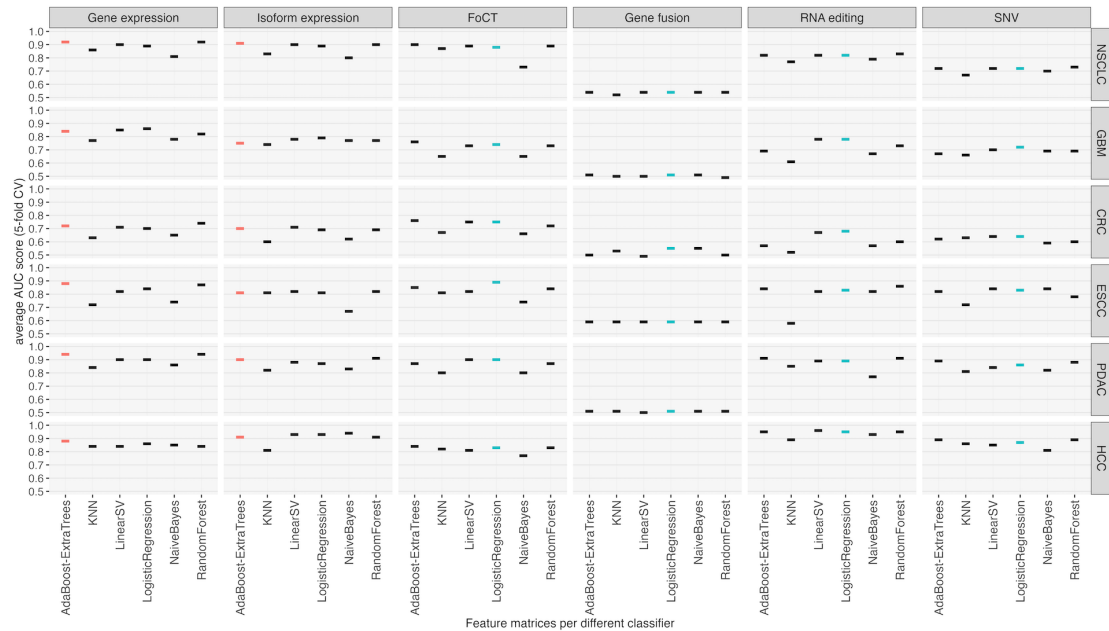


FIGURE 4.4: **Classifier Selection Overview.** Each column in the plot represents a different feature type, while each row corresponds to a different dataset. On the x-axis, six classifiers are evaluated, and the y-axis displays the mean AUC score achieved through 5-fold CV. The dashed lines represent the average AUC score from the 5-fold CV evaluation. For the first two feature types (gene and isoform expression), AdaBoost with ExtraTrees, is highlighted in red, as it is better suited for these feature types. For the remaining feature types, Logistic Regression, is highlighted in blue, as it is better suited for these types of features.

the lowest performance with a collective AUC of 0.52. Regarding RNA editing and SNVs, Logistic Regression outperformed other classifiers, obtaining a collective mean AUC of 0.82 and 0.77, respectively, whereas KNN exhibited the lowest performance with a collective AUC of 0.70 and 0.72, respectively.

4.3.5 Enhancing Predictive Output through Ensemble Learning

Having thoroughly assessed the performance of each of the six biofeature types in isolation, we delved into the potential benefits of combining these diverse, high-dimensional biofeature spaces, each characterised by different scales and distributions. To do so, we adopted three ensemble combination techniques: soft voting, majority voting, and stacking, each with its unique approach to integrating predictions from multiple models. Soft voting relies on averaging probabilities, majority

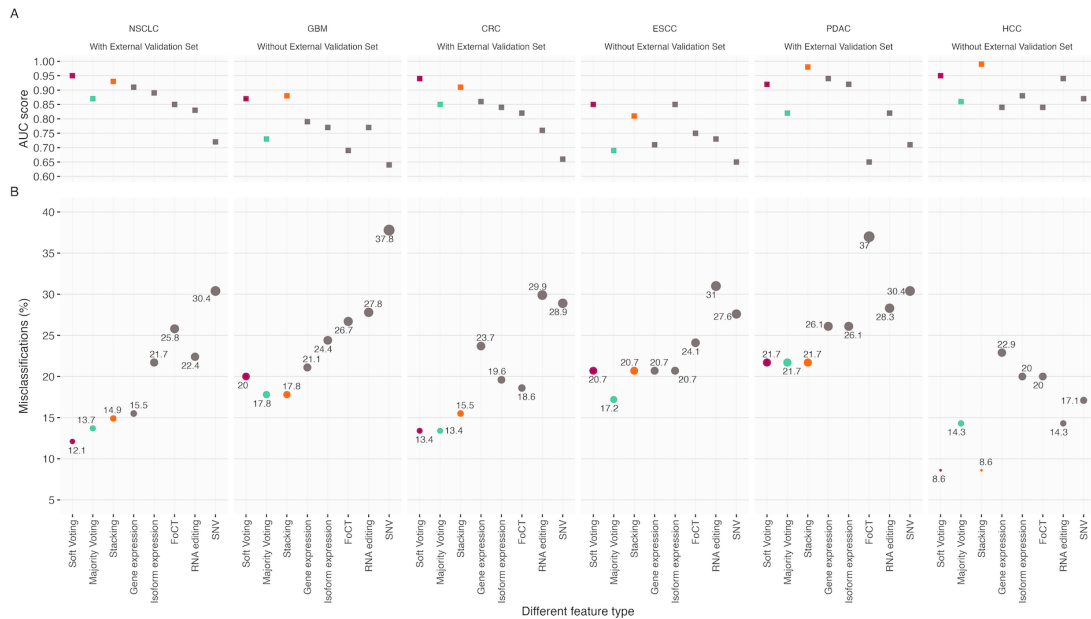


FIGURE 4.5: Performance Overview of Feature Types and Ensemble Learning Methods Across Datasets. Each column in this figure corresponds to a specific dataset, with dataset information and the presence of an independent validation set indicated above. (A) The plot displays AUC scores for the test sets across all feature types. (B) Dot plots illustrate the percentage of misclassifications for each feature type across different datasets. The colourful squares and dots in both figures A and B represent the ensemble learning techniques, while the grey squares and dots represent the remaining feature types.

voting on majority decisions, and stacking employs a meta-model to optimise the fusion of predictions (see Supplementary Materials Section A.1.4).

Our comprehensive evaluation spanned six datasets, assessing performance using key metrics: AUC and the percentage of misclassified samples (Figure 4.5A-B and Supplementary Materials Section A.1.5). While AUC offers a broad measure of model performance, its limitation in discerning specific error types prompted us to incorporate misclassification rates for a more nuanced evaluation. Strikingly, ensemble learning consistently outshone individual models across all datasets. The selection of the most suitable ensemble technique varied depending on the dataset and the metric considered. For AUC, soft voting excelled in NSCLC, CRC, and ESCC, whereas stacking proved superior in GBM, PDAC, and HCC. Conversely, when evaluating the percentage of misclassified samples, NSCLC, CRC, PDAC, and HCC demonstrated the best outcomes. In GBM and ESCC, majority voting slightly outperformed soft voting. In some instances, multiple ensemble techniques performed equally well, such as majority voting and stacking in GBM and CRC, and soft voting and stacking in PDAC and HCC.

A deeper dive into specific datasets revealed intriguing findings. In NSCLC, soft voting achieved an impressive low misclassification rate of 12.1%, outperforming gene expression at 15.5%. In GBM, majority and stacking techniques led with a 17.8% misclassification rate, surpassing soft voting at 20% and gene expression at 21.1%. For CRC, both soft and majority voting achieved a low misclassification rate of 13.4%, while gene expression yielded a higher rate of 23.7%. In ESCC, majority voting reached 17.2%, while soft voting and other ensemble techniques, as well as gene and isoform expression, hovered around 20.7%. In PDAC, all ensemble techniques achieved a comparable rate of 21.7%, while gene expression exhibited a higher rate of 26.1%. Lastly, in HCC, soft voting and stacking were equally successful, both at 8.6%, while gene expression performed the worst at 22.6%. For a detailed overview of the computed misclassification rates and comprehensive data, please refer to the Supplementary Materials and Figure 3. Additionally, various evaluation metrics are provided in Table 6.

To reach a broader consensus on which ensemble learning technique consistently outperforms the others, we aggregated the misclassification rates across all datasets and calculated their average. The same analysis was extended to the individual models for a comprehensive comparison. Figure 4.6 presents our final findings, illustrating the overall performance of each model. As previously noted, all ensemble learning techniques demonstrated superiority over individual models. Soft voting, with an average rate of 16.08, emerged as the frontrunner, followed by majority voting at 16.35, and stacking at 16.53.

o further reinforce our findings, we conducted a similar analysis, but limited it to datasets that possessed an independent external validation set. Once again, soft voting demonstrated superior performance with an average rate of 15.73, while majority voting and stacking followed with rates of 16.27 and 17.37, respectively.

It's important to emphasise that even the top-performing results achieved by individual biofeatures fall short, or at the very least, are on par with the ensemble learning's average performance. In light of these discoveries, it becomes evident that ensemble techniques significantly enhance model performance by capitalising on the diverse nature of biofeature spaces and adeptly managing their inherent heterogeneity.

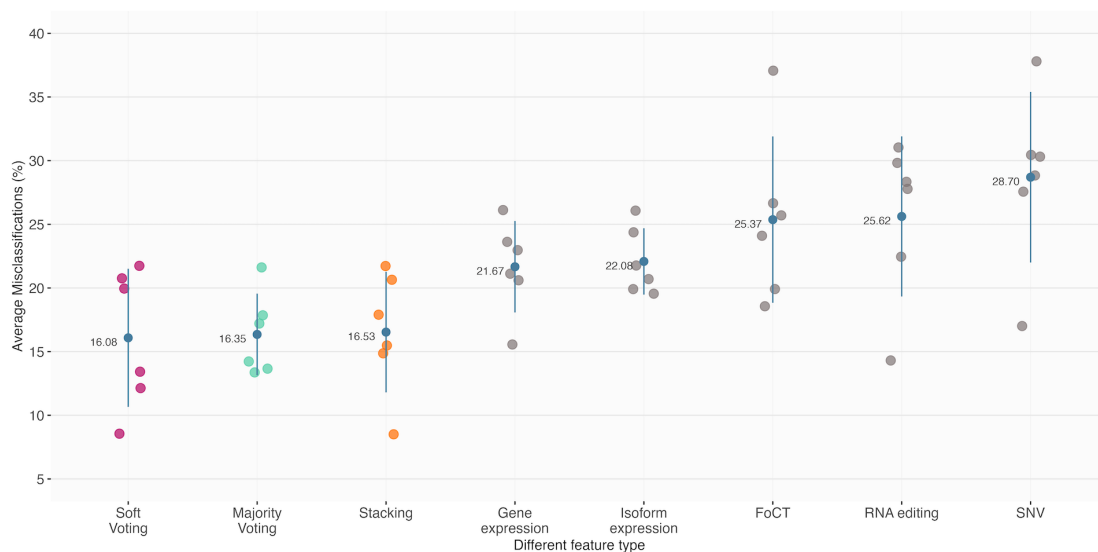


FIGURE 4.6: **Average Misclassification Percentage Across All Datasets.** This graph displays the percentage of average misclassifications for each feature type on the x-axis, with feature types arranged from the least to the most misclassifications from left to right. Each dot represents the percentage of average misclassifications, providing an overview of the classification performance across all datasets.

4.3.6 Biological Feature Selected Interpretation

Genes detected as discriminative biofeatures should ideally have a causal relation to the analysed phenotype. For gene-expression based features putative relations can usually be shown by means of functional enrichment analyses, however for the other biofeature types explored in this study, this kind of analysis was not performed before. For example, genes showing RNA editing events with significant differences between control and cancer samples should act in relevant cancer pathways. To test this hypothesis, we analysed the 216 genes selected as features from the NSCLC dataset by means of the gProfiler. Figure 4.7 and Supplementary Materials Section A.1.6 shows the overrepresented functional annotations among these genes. Selected genes are mainly related to immune system functions, signal transduction and regulation of vesicle-mediated transport, which might indeed indicate that biologically meaningful features were extracted by means of our workflow.

The analysis of overrepresented functional annotations was specifically conducted for the NSCLC dataset alone, but we examined which genes are selected in more than one dataset (see Table 7 for complete list). Interestingly, many genes, mainly related to immune system functions, are selected from several studies and could

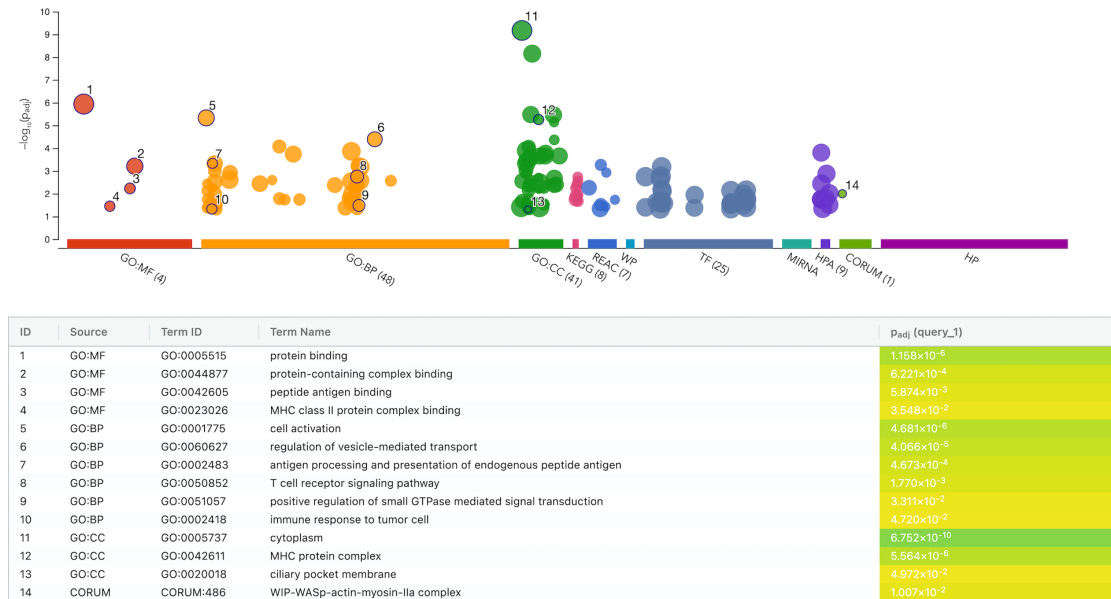


FIGURE 4.7: **Enrichment Analysis results.** Presented is a Manhattan plot illustrating the results of an enrichment analysis performed on the genes selected for ensemble learning within the NSCLC dataset. The x-axis is dedicated to functional terms, which have been meticulously organised and colour-coded based on their respective data sources. Simultaneously, the y-axis represents the adjusted enrichment p-values, thoughtfully presented in a logarithmic negative scale. Terms of lesser significance are discreetly depicted as faint circles, while encircled numbers within the figure denote statistically significant enriched GO terms.

TABLE 4.2: Concise summary of genes recurring in more than three, up to six out of six datasets. The table includes the following attributes: GeneID, Gene Name, Gene Type, Occurrence (indicating in how many datasets the gene was selected as a feature), and a brief description of each gene’s function.

GeneID	Gene Name	Gene Type	Occurrence	Description
ENSG00000234745.11	HLA-B	protein coding	6/6	major histocompatibility complex, class I, B
ENSG00000206503.13	HLA-A	protein coding	5/6	major histocompatibility complex, class I, A
ENSG00000213492.2	NT5C3AP1	transcribed processed pseudogene	3/6	NT5C3A pseudogene 1
ENSG00000166710.20	B2M	protein coding	3/6	beta-2-microglobulin
ENSG00000160014.17	CALM3	protein coding	3/6	calmodulin 3
ENSG00000162852.14	CNST	protein coding	3/6	consortin, connexin sorting protein
ENSG00000115956.10	PLEK	protein coding	3/6	pleckstrin
ENSG00000196126.11	HLA-DRB1	protein coding	3/6	major histocompatibility complex, class II, DR beta 1
ENSG00000117640.18	MTFR1L	protein coding	3/6	mitochondrial fission regulator 1 like
ENSG00000149781.12	FERMT3	protein coding	3/6	FERM domain containing kindlin 3
ENSG00000115310.18	RTN4	protein coding	3/6	reticulon 4
ENSG00000150867.14	PIP4K2A	protein coding	3/6	phosphatidylinositol-5-phosphate 4-kinase type 2 alpha

therefore serve as pan-cancer markers. Table 4.2 offers a concise summary of genes recurring in more than three, up to six out of six datasets.

4.4 Discussion

In this study, we introduced the ELLBA methodology, a robust and comprehensive multi-level bioinformatics approach tailored for analysing lbRNA-Seq data to predict patient outcomes. ELLBA's core framework centres on the extraction of six distinct biofeature types: gene expression, isoform expression, FoCT, gene fusion, RNA editing, and somatic SNVs. These diverse biofeatures aim to capture distinct molecular and functional characteristics. Our study design encompassed different cancer types and blood-based biosources, employing multi-condition samples from six datasets, followed by comprehensive evaluation across four independent validation sets. A crucial aspect of our study was to explore the feasibility of intra-sample normalisation methods for the count-based biofeature types to improve the clinical applicability of the pipeline. Through the assessment of eight diverse normalisation methods, CPM normalisation emerged as a robust method for both gene and isoform expression analyses, performing comparably to the more sophisticated cross-sample normalisation techniques like TMM or RUVSeq, which additionally corrects for batch effects. Notably, the clinical utility of intra-sample normalisation, such as CPM, deserves highlighting, as it enables rapid individual sample normalisation and seamless integration with instant predictions using pre-trained ML models. This property renders it highly suitable and time-efficient for clinical applications.

Further, we extensively benchmarked the choice of optimal classifiers for each biofeature type. Gene and isoform expression with continuous values demonstrated a strong compatibility with the AdaBoost classifier using the ExtraTrees as base estimator. Conversely, Logistic Regression exhibited superior performance for biofeature types with ratios (FoCT) or discrete values (remaining biofeatures). While most evaluated models performed comparably across each biofeature type, Naive Bayes and KNN consistently ranked lower. However, it is important to note that fluctuations in individual biofeature type performance might occur depending on the dataset, with no one biofeature type universally excelling. We believe that the variations in classifier performance across the different biofeature types in our benchmarking experiment can be attributed to several factors. Gene and isoform expression data might perform well with Adaboost due to their complexity and non-linear separability. On the other hand, Logistic Regression might excel with biofeatures like FoCT, gene fusion, RNA editing events, and SNVs because of their

simpler, more linear relationships. Data size, feature importance, noise levels, and the presence of interaction effects could also contribute to these variations.

Our core finding underscores the superiority of combining information from several biofeatures over relying solely on standard Gene Expression. To mitigate fluctuations originating from individual biofeature space limitations and harness their complementary information, we introduced three distinct ensemble classification approaches within the ELLBA methodology. These ensemble methods, namely soft voting, majority voting, and stacking, integrate predictions derived from all six biofeature types, employing diverse strategies. Our findings showed, initially, that ensemble classification always improved predictive accuracy compared to gene expression alone. Furthermore, these ensemble learning techniques effectively reduce misclassification rates and enhance overall prediction accuracy, underscoring the value of multi-view analysis for LB data. It is important to note that the choice of ensemble classification technique may vary depending on the dataset. In general, our findings suggest that soft voting is a robust and versatile ensemble learning method, a conclusion substantiated by datasets featuring an external validation set. This observation underscores the pivotal role of ensemble learning in enhancing the reproducibility and reliability of our results, further solidifying its importance in the context of LB data analysis. Moreover, we believe that this superiority arises from several key advantages of ensemble methods. Firstly, they leverage the complementary information inherent in each biofeature type, capturing distinct aspects of the data and thus improving prediction accuracy. Secondly, ensemble methods reduce the impact of noise in individual models by combining multiple predictions, enhancing robustness. Additionally, ensemble techniques improve generalisation, reduce variance, and handle imbalanced data more effectively. Overall, ensemble learning emerged as a powerful strategy for harnessing the full potential of multi-feature data in cancer diagnostics.

Upon consideration of additional complementary metrics, including balanced accuracy, F1 score, and average Precision score, and comparing them with the corresponding metrics from the standard Gene Expression, as well as other biofeatures, we consistently observed unaltered results. It is crucial to emphasise that the analysis remains invariant due to the globally robust behaviour of the methodology. Our approach is dedicated to enhancing performance across diverse scenarios, irrespective of class distribution, with particular attention to cancer precision. This steadfast commitment to robustness underscores the methodology's reliability.

To gain deeper insights into the biological significance of our selected features, we conducted a functional enrichment analysis, focusing on features that consistently recurred across datasets. The results highlighted that the genes identified as discriminative features, particularly those associated with immune system functions may hold significant relevance in cancer pathways. Furthermore, the recognition of genes recurring across multiple studies underscores their potential as valuable pan-cancer markers, underscoring the strength and applicability of our workflow in uncovering biologically meaningful features.

Several important considerations emerge from our study. Firstly, the restricted availability of gene fusion data may have affected the results, possibly due to the limited depth of sequencing in certain datasets or the use of single-end sequencing in cases such as NSCLC, GBM, and CRC datasets. While our fusion detection criteria were not overly stringent, the challenge of accurately capturing fusion events with high confidence warrants future exploration, particularly with the potential benefit of paired-end deep sequencing for enhancing the identification of these putatively important biomarkers. It is worth noting, that despite the depth of the CRC and the switching from PE to SE, we noticed that the Accuracy in the training is 0.60 by selecting 3 features out of the initial filtered ones which were 12. Secondly, the robust performance of our workflow on independent validation sets underscores the significance of maintaining uniform handling and sequencing protocols for validation data. This is exemplified by the closely aligned accuracy observed in the NSCLC dataset and its validation set, both managed by the same group, reinforcing the imperative for standardised procedures even across diverse laboratory settings. Finally, despite our comprehensive investigation of normalisation techniques, ML algorithms, and ensemble learning, there remains potential for enhancing our ensemble learning classifier's performance. This might involve exploring newer algorithms, adapting existing ones, exhaustive hyperparameter exploration, evolutionary-based feature selection, or integration of additional genomic data types. Through ongoing refinement of our ML pipeline, we anticipate continued progress in the precision and reliability of cancer predictions across diverse datasets.

4.5 Conclusion

The ELLBA workflow offers a significant contribution to LB data bioinformatics analysis. Through multi-feature integration and ensemble learning, ELLBA presents a comprehensive avenue for patient outcome prediction in LB-based cancer research. Its flexibility aligns well with upcoming LB advancements, facilitating analysis across diverse cancer types and biosources. As LB gains traction in cancer diagnostics, ELLBA holds promise for advancing precision oncology. Rigorous validation in even larger, diverse cohorts, complemented by experimental confirmation, will be pivotal to establishing ELLBA's clinical utility and reliability in LB data analysis. With ongoing strides in LB technologies and ML, ELLBA's continued evolution holds potential as an indispensable tool in LB-based cancer research and clinical applications.

Chapter 5

Expanding the landscape of cancer transcriptome by native RNA sequencing of NSCLC tissue samples

“Science knows no country, because knowledge belongs to humanity, and is the torch which illuminates the world.”

— LOUIS PASTEUR

This Chapter introduces a pilot project that utilises the Nanopore Direct RNA Sequencing protocol for the transcriptomic profiling of NSCLC adenocarcinoma. This relatively new approach has not been widely applied in cancer research studies. Recognising the need for specialised bioinformatics tools due to the innovative nature of this method, we developed "DRseeker", a pipeline specifically designed to harness the theoretical advantages of Nanopore technology in capturing complex transcriptomic data. The study compares transcriptome profiles between lung cancer tissues and adjacent non-transformed tissues, revealing significant findings such as novel transcripts, alterations in the AGER and MEST genes, and variations in polyadenylation patterns. Insights into deviations in epitranscriptomic modification patterns, identified by DRseeker, underscore the pipeline's effectiveness and point to promising directions for future lung cancer research to enhance our understanding and treatment of the disease.

5.1 Introduction

Lung cancer stands as one of the most prevalent and deadly cancers worldwide, representing a critical challenge in both clinical and research settings. Despite significant advancements in our understanding and treatment of the disease, lung cancer continues to have a high mortality rate, primarily due to late diagnosis and the complex nature of its pathogenesis [1]. This complexity is underscored by the diverse genetic and environmental factors contributing to its development and progression [225], [226]. The imperative to delve deeper into the molecular mechanisms of lung cancer is clear, as it holds the key to unlocking more effective diagnostic and therapeutic strategies.

Transcriptome analysis occupies a prominent position in advancing our comprehension of lung cancer at the molecular level. By providing a holistic perspective of gene expression within cancerous tissues, it illuminates the specific pathways and mechanisms that underlie cancer's development and progression. This analytical approach is indispensable for the discovery of novel biomarkers and potential therapeutic targets, significantly enriching our insights into the intricate mechanisms at play in cancer [227]–[229].

The transcriptome's intricate and multi-layered nature extends beyond mere gene expression, encompassing a spectrum of factors, including alternative splicing, polyadenylation, fusion transcripts, and the burgeoning field of epitranscriptomics. This complexity unveils the remarkable variability inherent in lung cancer subtypes. Scrutinising this variability not only furthers our grasp of cancer's molecular basis, but also is pivotal for the implementation of personalised medicine, underscoring the importance of adapting treatment strategies to align with each patient's distinctive genetic landscape.

Emerging technologies in transcriptome analysis, particularly through advanced sequencing methods, have opened new avenues for understanding the disease at a molecular level. The field of sequencing technologies has witnessed remarkable advancements in recent years, revolutionising the way researchers understand genetic and molecular structures in various diseases, including lung cancer. These advancements are characterised by significant improvements in accuracy, speed, and cost-effectiveness. High-throughput sequencing technologies, such as Next-Generation Sequencing (NGS), have enabled the analysis of genetic material at an unprecedented scale and resolution. However, the recent advent of Nanopore

Direct RNA Sequencing represents a significant leap forward. This technology offers real-time sequencing, providing a rapid and detailed view of the transcriptome. Its ability to read long sequences of RNA in their native form, without prior amplification or conversion to DNA, is a major advancement, allowing for a more accurate and comprehensive understanding of RNA molecules, including their modifications [230].

This study aims to contribute to this evolving landscape by exploring the transcriptome profile of lung cancer tissues. Our work represents a pioneering effort in harnessing the advanced capabilities of Nanopore Direct RNA Sequencing to conduct an unprecedentedly detailed comparative analysis of the transcriptome profiles of lung cancer tissues and adjacent non-transformed lung tissues. It stands as the first of its kind to delve into the transcriptome of lung cancer (NSCLC adenocarcinoma) with such resolution, comprehensively surveying all the layers of information that this technology offers. The primary goal is to uncover unique transcriptomic changes in NSCLC adenocarcinoma, enhancing our understanding of the disease and potentially uncovering new biomarkers and therapeutic targets. Initially, an advanced bioinformatics pipeline was developed, called DRseeker, to simultaneously analyse multiple layers of the transcriptomic data, ensuring a comprehensive and holistic view of the transcriptome. Alongside, we validated our gene-level analysis findings by comparing it with similar short-read sequencing studies, confirming the reliability of our data. Our approach included examining differential transcript expression and usage, particularly focusing on cancer-related genes like *AGER* and *MEST*. We also explored variations in the polyA tail of transcripts, a key area in genetic research. This method enabled us to analyse the behaviour of the *TXNRD1* gene in different forms. Finally, we delved into epitranscriptomics, assessing the modifications in genetic transcripts to better understand their role in the progression of lung cancer. Through this comprehensive study, we aim to provide a deeper insight into the molecular mechanisms of lung cancer and open new avenues for early detection and treatment.

5.2 Materials and Methods

5.2.1 Methodology for Lung Tissue Collection and RNA Isolation in NSCLC Patients

In this study, we focused on three patients (average age of 57.3 years), all diagnosed with NSCLC, specifically categorised as KRAS-mutant adenocarcinomas. The patients underwent tumour resection surgery, during which we meticulously collected both tumour tissues and corresponding adjacent non-transformed (control) lung tissues from each participant. To preserve the integrity of the samples, we immediately snap-froze the tissues in liquid nitrogen, ensuring their long-term preservation at -80°C .

For histological examination and to ascertain the viability of the samples for subsequent RNA extraction, we prepared cryosections from both the peripheral and central regions of each tissue sample. These sections underwent hematoxylin and eosin staining, followed by detailed microscopic analysis. This step was crucial in confirming the minimal presence of necrosis in the collected samples, thereby validating them for RNA isolation.

Throughout the tissue processing stages, stringent measures were taken to prevent RNA degradation. This involved the use of RNase-free water and thoroughly cleansing all equipment with RNA zap. For the actual collection of tissue sections for RNA isolation, we utilised a lysis buffer from the MirVana kit [231]. The sections were immediately placed at -20°C , kept on ice during the entire processing phase, and subsequently stored at -80°C overnight or until the RNA isolation procedure was initiated.

5.2.2 RNA Isolation from Lung Tissue Samples

RNA isolation from the tissue samples was executed using the MirVana Total RNA Isolation Kit, strictly following the manufacturer's protocol. To evaluate the quality of the extracted RNA, we utilised the Bioanalyzer RNA 6000 Picochip from Agilent Technologies. For quantifying the RNA, the NanoDrop 2000 spectrometer by Thermo Scientific was employed. These techniques ensured precise assessment of the RNA's quality and quantity.

5.2.3 Direct RNA Sequencing of NSCLC Samples Using Oxford Nanopore MinION Technology

Direct RNA sequencing was performed using 500 ng of Poly-A RNA, which was enriched with magnetic beads to construct our RNA libraries. The process complied with the Direct RNA Sequencing Kit (RNA Kit SQK-RNA002) instructions from Oxford Nanopore Technologies Ltd. As part of this protocol, we incorporated 0.25 µl of the RNA Calibration Strand (RCS) for Enolase II (ENO2).

The sequencing process involved priming the MinION flow cell from Oxford Nanopore Technologies Ltd. using the Flow Cell Priming Kit (EXP-FLP002) from the same company. Following the priming, 75 µl of the RNA library was loaded into the SpotON sample port of the MinION device.

Subsequently, we initiated the sequencing run with the MinKNOW software (v3.2.6), which controlled and monitored the sequencing process. This method enabled us to perform direct RNA sequencing effectively, utilising the capabilities of the Oxford Nanopore MinION technology. Among the three samples analysed, two produced an adequate number of reads. However, the matching non-transformed sample from the third set generated a low quantity of reads, prompting us to replicate the process using a non-transformed sample from a fourth subject with comparable characteristics.

5.2.4 Re-Basecalling and Initial Filtering Process

We processed the MinION fast5 files for basecalling on a local setup using the ONT Guppy v6.3.7 software [175]. Our configuration of Guppy ensured that only reads meeting the ONT-recommended minimum Phred quality score of 7 were included, thus enhancing the quality of the data. The basecalling parameters were tailored to replace U bases with T in the sequence output and to deliver reads in the reverse orientation. Specifically, our Guppy parameters included `-flowcell FLO-MIN106, -kit SQK-RNA002, -calib_detect, -trim_strategy rna, -reverse_sequence true, -qscore_filtering, and -u_substitution false`

5.2.5 Initial Quality Control and Preprocessing

To ascertain the integrity of our sequencing data and the efficacy of the sequencing process, we utilised the capabilities of NanoPlot (v1.40.0) [176]. This software provided a comprehensive statistical summary alongside a suite of quality control visuals. Included in the output were histograms charting read lengths, graphs showing the cumulative yield, and violin plots that illustrated variations in read length and quality over the duration of sequencing. Additionally, bivariate plots were used to examine the relationships between read lengths and quality scores in the context of reference identity and mapping quality. NanoPlot also furnished several supplementary quality metrics that were instrumental in further evaluating the sequence data's quality.

Subsequently, we filtered out reads under 50 nucleotides in length and applied a conservative error correction using IsONcorrect (v0.0.6.1) [177]. This software uniquely leverages the full spectrum of gene isoforms for error correction, which is particularly effective for correcting reads even at low sequencing coverage.

5.2.6 Genome Alignment and Post-Alignment Quality Control

We aligned the 'pass' marked compressed fastq files to the reference human genome (GRCh38, primary assembly) using Minimap2 (v2.17) aligner [178]. The alignment was executed in splice-aware mode with k-mers set to a length of 14, and the software was configured to only consider the forward transcript strand without reporting any secondary alignments. The specific parameters used were `-ax splice`, `-k 14`, `-uf`, and `-secondary = no`. For sorting and compressing the output sam files, along with exporting various mapping quality metrics for further evaluation of the sequencing experiment, we utilised Samtools (v1.12) [170], PycoQC (v2.5.2) [179], and RSeQC (v4.0.0) [180].

5.2.7 Gene and Transcript Identification and Abundance Estimation

TALON (v5.0) [181], a Python package tailored for long-read transcriptomes, was utilised for identifying known and novel genes/transcripts. Initially, following the developers' recommendation, we employed TranscriptClean (v2.0.2) [182] with default settings to further correct the aligned reads, aiming to rectify non-canonical splice junctions. The corrected aligned bam files from all six samples, along with the GENCODE v35 human reference annotation [165], were input into TALON.

Within TALON's analytical process, transcript abundance was quantified for each sample, resulting in the generation of two distinct matrices. The initial matrix presented a comprehensive record of transcript expression levels, both for established and novel transcripts, all aligned to recognised loci and presented without the application of any count-based or additional filters. The subsequent matrix refined the original by incorporating more stringent criteria, particularly for novel transcripts. For inclusion in this refined matrix, a novel transcript was required to be present in at least one group with a minimum presence of 5 counts.

A tailored script was also employed to further refine the selection of novel transcripts. Initially, we required that all transcripts, regardless of their novel classification, exhibit a measurable polyA tail length in at least 10 transcripts. In addition, a minimum threshold of 20 counts per isoform was established for novel transcripts within the ISM Prefix, Suffix, None, and Both categories to qualify for inclusion. For the novel ISM Suffix category, to ensure transcript completeness, we conducted a comprehensive analysis by comparing the 5' of these transcript models against CAGE assay data, as detailed in Section 5.2.12. Ultimately, we consolidated the ISM None and Both categories into a single category, which we designated as ISM Both.

5.2.8 Exploratory Analysis

We commenced our study with an exploratory analysis using R (v4.0.5). This initial phase of investigation was grounded in gene expression data, which we derived by summing up the transcript counts corresponding to their originating genes. Various analytical techniques were applied to this gene expression data, including Principal Component Analysis and heatmap analysis for visualising pairwise

correlation values of rlog-transformed counts. Additionally, we explored the distribution of raw counts by creating boxplots of log₂-transformed Counts Per Million (log₂CPM). Further to these analyses, we generated a Multi-Dimensional Scaling (MDS) plot and also constructed a heatmap to display the top 100 genes with the highest variability.

5.2.9 DGE, DTE, and DTU

For our analyses involving Differential Gene Expression (DGE), Differential Transcript Expression (DTE), and Differential Transcript Usage (DTU), we once again utilised R. The base data for gene expression calculations originated from the unfiltered abundance matrix generated by TALON. In these calculations, isoform abundances were cumulatively assigned to their respective genes, with any transcripts classified as "Genomic" being omitted. Initially, we applied the 'filterByExpr' function from the edgeR package (v3.32.1) [185] to filter out genes with low expression. Subsequently, edgeR was used in the Trimmed Mean of M-values (TMM) mode for normalisation and differential expression analysis within a quasi-likelihood framework. Various plots, including MA and Volcano plots, along with enriched MA plots highlighting DE genes with at least one novel transcript, were produced from this analysis.

The methodology for DTE mirrored that of DGE, but relied on the filtered TALON abundance matrix, and similar plots were generated for this analysis as well. For the purpose of visualising the array of isoforms identified through DRS, the Ensembl Genome Browser was employed [232]. The Ensembl Genome Browser is a comprehensive resource offering access to a wealth of sequence data, including gene annotations and predictions.

In conducting DTU analysis, we utilised the IsoformSwitchAnalyzeR package (v1.12.0) [189], again based on the filtered abundance matrix. To enhance our analysis, we incorporated external transcript data from several sources: Pfam [190] for protein family and domain identification, IUPred2A [191] for predicting disordered protein regions, SignalP [192] for detecting signal peptides, and CPC2 [193] for assessing coding potential of transcripts. We ran IsoformSwitchAnalyzeR in its default mode for this analysis.

5.2.10 Gene Functional Enrichment Analysis

For the gene-level functional enrichment analysis, we utilised R and implemented various packages, specifically clusterProfiler (v3.18.1) [233] and enrichplot (v1.11.2) [234]. These packages were used to identify enriched GO categories following the DGE analysis.

5.2.11 Transcript Functional Annotation and Functional Enrichment Analysis

The functional characterisation and annotation of novel isoforms were carried out using the Trinotate suite (v3.2.1) [194]. This suite employs multiple features for annotation, including prediction of putative coding regions, homology searches against known sequences, identification of protein domains, and prediction of protein signal peptides and transmembrane domains, to functionally annotate the query transcripts.

In conducting the functional enrichment analysis of the novel DE transcripts, we took advantage of the GOst tool available through the gProfiler web service [186]. This utility enabled us to perform Gene Ontology (GO) and pathway enrichment analyses on the cohort of genes from which the chosen novel transcripts are derived. The analysis was carried out separately for each group: once for the novel transcripts that were up-regulated and once for those that were down-regulated.

5.2.12 CAGE Analysis

We acquired well-defined CAGE (Cap Analysis Gene Expression) peak data for humans from the FANTOM5 database [183], formatted as BED files. These coordinates, initially based on the hg19 human genome reference, were updated to the hg38 version using the LiftOver tool from the UCSC Genome Browser [235]. For each long-read transcript model in our GTF-based transcriptomes, we identified the transcription start site (TSS). Utilising Bedtools (v2.26.0) [236], we then investigated the presence of any CAGE peaks within a 100 bp range both upstream and downstream of these TSS locations.

5.2.13 PolyA-Tail Length Estimation, DPA and APA Analysis

PolyA-tail length estimation was conducted for each sequenced read. We utilised the raw fast5 files and the aligned lam files as inputs for the Nanopolish polyA (v0.14.0) software [195], [196]. From the output of this software, we only retained the polyA length estimations for isoforms that matched those in the TALON filtered abundance matrix for further analysis.

Additionally, we employed a customised version of the ONT `polyA_diff.py` script [237] for Differential Polyadenylation Analysis (DPA). To begin with, transcripts exhibiting fewer than 10 counts were filtered out from the DPA process. This analysis involved using the Mann-Whitney U nonparametric test to statistically examine the variations in polyA tail lengths among different samples. The evaluation was conducted in two distinct stages. Initially, a global analysis was performed to compare the two groups as a whole, contrasting the cancer samples against the non-transformed samples. Subsequently, the analysis culminated in a more detailed, transcript-by-transcript comparison between the two groups.

Alternative Polyadenylation Analysis (APA) was conducted using the LAPA software [197], which identifies and clusters reads with poly(A) tails and performs peak-calling to pinpoint poly(A) sites. LAPA annotates these peaks with genomic features and regulatory elements, including the poly(A) signal. It also conducts statistical tests with multiple testing correction to detect differential APA. We ran LAPA using its default settings and employed a custom script to represent the differential APA results in a Volcano plot.

5.2.14 Methylation Detection

For the detection and analysis of methylation, we initially processed the corrected fastq files by realigning them against a reference transcriptome. This reference was obtained using the TALON software, accompanied by its exported TALON annotation gtf file. The alignment was performed utilising Minimap2 software, employing specific parameters including `-ax map-ont`, `-k 14`, `-secondary = no`, and `-MD`.

Following the alignment, the output BAM files were sorted with the aid of Samtools (v1.12). The next step involved the use of Nanopolish Eventalign (v0.14.0), which was employed for effective signal segmentation.

Both aligned reads and event data, in conjunction with the reference files, were processed using xPore [198] to determine differential modification sites.

5.3 Results

5.3.1 Comprehensive Analytical Pipeline for Direct Long-Read RNA-Seq Data Investigation

To fully harness the capabilities of direct long-read RNA sequencing (DRS), we formulated a detailed computational workflow. This workflow was designed to methodically identify, quantify, and analyse several transcriptome features, as illustrated in Figure 5.1. Our pipeline is structured into four primary modules: 1) data preprocessing; 2) alignment; 3) core analyses; 4) complementary analyses.

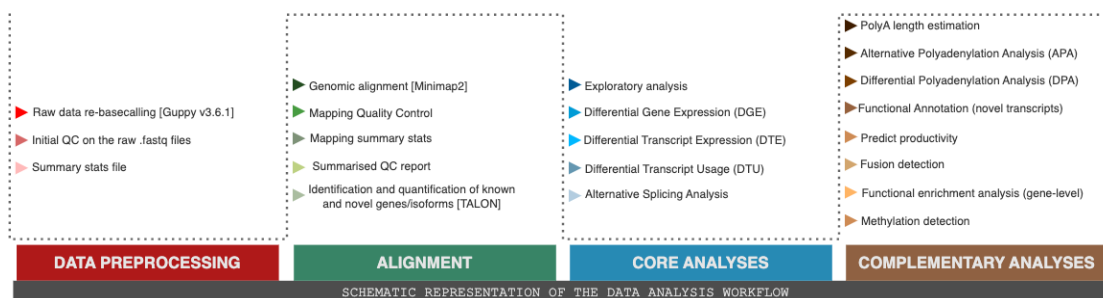


FIGURE 5.1: **Schematic representation of the data analysis workflow.** This figure presents a structured overview of the data analysis workflow divided into four main stages: Data Preprocessing, Alignment, Core Analyses, and Complementary Analyses. Each stage outlines the key steps taken, from initial raw data re-basecalling and QC, through genomic alignment and genes/isoforms quantification, to the core examination of transcriptomic features including differential expression and alternative splicing, and the complementary analyses such as polyadenylation assessment, functional annotation, and methylation detection.

The initial *data preprocessing* stage of our analysis pipeline is meticulously designed to refine the sequencing data, initiating with the re-basecalling of raw sequencing signals. This stage incorporates the elimination of shorter reads and

culminates with a thorough quality control (QC) evaluation to verify the data's integrity and the sequencing process's accuracy.

Proceeding to the *alignment* module, our efforts are centred on the precise mapping of transcriptomic data against the reference genome. This essential step forms the foundation for subsequent analytical procedures and involves a comprehensive QC process to ensure the accuracy and quality of the alignment. We generate a detailed report encompassing crucial quality metrics for each sample, such as the rates of aligned and duplicate reads, the genomic distribution of reads, and gene body coverage.

The workflow then transitions to leveraging the strengths of the TALON pipeline for robust identification and quantification of both known and novel isoforms. Within this segment, we have integrated strict filtering methods aimed at significantly minimising the potential for artefacts. This filtering approach is thoroughly documented in the Materials and Methods Section 5.2.7.

Additionally, we employed TALON's nomenclature to organise transcripts into categories based on their distinctive characteristics. This taxonomy is visually represented in Figure 4, with further details available in the Supplementary Materials section A.2.1.

The third component of our workflow, designated as the *core analyses*, begins with an extensive exploratory phase. This phase employs analytical methods such as PCA and MDS to uncover patterns and relationships in the data. The pivotal aspect of this module is the differential expression (DE) analysis, which is conducted on both gene (DGE) and transcript (DTE) levels under various experimental conditions. This is essential for identifying genes and transcripts that are crucial to specific phenotypic traits. The analysis extends to DTU, which reveals variations in the expression of isoforms within a gene that may not affect the gene's overall expression level [238]. The final piece of the module is an Alternative Splicing Analysis, which provides insights into the diversity of transcript isoforms generated from the same genetic sequences, highlighting the complexity of gene expression [239].

The concluding module, termed *complementary analyses*, integrates advanced tools for a detailed examination of various transcriptome features and characteristics. This involves estimating the polyadenylated tail length of each transcript, which is critical to mRNA stability and its subsequent translation. The analysis

of APA delves into variations in polyadenylation that affect post-transcriptional regulation, while DPA probes the regulatory significance of these variations [240].

This module also includes an evaluation of the capacity for novel transcripts to produce functional proteins. Functional enrichment analysis is another key component, complementing the DGE findings by identifying associated biological pathways and processes.

In addition, the module encompasses the detection of fusion transcripts, which result from the merging of two separate transcripts and can be significant in the development of diseases like cancer. Lastly, an analysis of methylation patterns across the transcriptome is conducted, providing an overview of epigenetic modifications and their impact on gene expression [241].

5.3.2 Transcriptome Profiling of Lung Cancer Tissue and Adjacent Non-Transformed Lung Tissue Using ONT Direct RNA Sequencing Protocol

The transcriptomic landscape of cancer is marked by a myriad of alterations mirroring the underlying cellular transformations within tumours. Deciphering this intricate RNA signature is pivotal for uncovering novel biomarkers and therapeutic targets. Consequently, we employed our bespoke workflow to thoroughly scrutinise the transcriptomes from lung cancer patients.

In our study, we used the DRS technique to construct sequencing libraries from polyadenylated RNA isolated from three lung cancer tissues and an equal number of non-cancerous lung tissues (referred to as non-transformed) (see Figure 5.2A). Tissue origins were verified via Immunohistochemistry (IHC) staining.

Sequencing produced between 844,000 and 1.67 million reads per sample, amounting to a total of roughly 7.72 million reads, as outlined in Table 5.1. From this total, cancerous tissues contributed 4.26 million reads, and non-transformed tissues provided 3.45 million reads. Furthermore, approximately 1.1 million CSRs were excluded from the analysis (refer to Materials and Methods Section 5.2.3). The reads that remained were then processed through the first stage of our analytical pipeline.

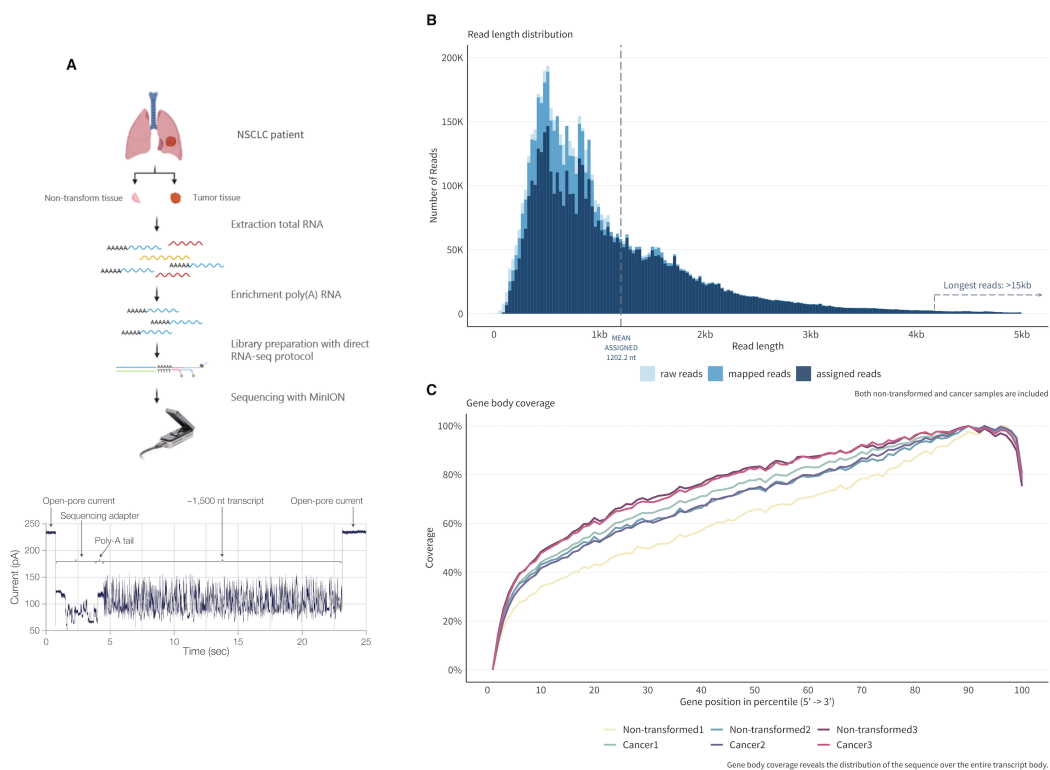


FIGURE 5.2: Overview of RNA sequencing workflow and data characteristics. (A) Illustrating the process flow from tissue collection in NSCLC patients, including tumour and non-transformed tissue, through RNA extraction, library preparation, and sequencing using the MinION platform. The inset graph shows a representative sample of sequencing current over time. (B) Displaying a histogram of read lengths for the combined set of raw, mapped, and assigned reads from both tumour and non-transformed samples. (C) Depicting the gene body coverage profiles of all six samples.

Among the 5.96 million high-quality reads obtained (with a Q-score above 7), a substantial 95% were classified as primary alignments, boasting an average read length of approximately 1100 bases. Out of the total mapped reads, 78.7% were accurately mapped to the human genome with an average read length of 1127 bases. We managed to identify 35,220 genes, representing about 60% of the genes identified in the GENCODE v35 database, covering a substantial portion of the human genome. The median read length was 1202 nucleotides (illustrated in Figure 5.2B), with the longest transcript being from the AHNAK gene (ENST00000378024.9), measuring 18,761 nucleotides. This gene encodes a protein implicated in various cellular functions, including structural integrity, cell migration, and tumour metastasis [242], [243].

Analysis of gene body coverage revealed a characteristic pattern of high read density at the 3'-end, diminishing towards the 5'-end (Figure 5.2C), consistent with previous findings [244]–[246]. This pattern can be partly attributed to the direct RNA library preparation method, which selects for transcripts with intact 3' poly-A tails, necessary for adapter ligation, potentially leading to a selection bias towards transcripts with better-preserved 3' ends. Given this bias, we incorporated rigorous filtering criteria to ensure the selection of full-length transcripts for subsequent analyses.

TABLE 5.1: Direct RNA sequencing general overview.

Attributes	Non-transformed1	Cancer1	Non-transformed2	Cancer2	Non-transformed3	Cancer3*
Number of raw reads	1,162,330	924,243	1,447,544	1,662,661	844,464	1,674,937
Mean read length (nt)	984.0	1,081.9	1,085.6	1,018	1,265.7	1,155.1
Mean read quality	10.0	9.7	10.0	10.0	9.5	9.8
Calibration reads	406,499	22,469	547,933	48,605	30,322	61,782
-						
Number of reads above Q7	1,057,135 (90.9%)	841,308 (91.0%)	1,341,554 (92.7%)	1,518,181 (91.3%)	774,141 (91.7%)	1,532,611 (91.5%)
Number of useful reads above Q7	655,531	818,924	799,004	1,469,763	744,026	1,471,477
Reads aligned to genome**	548,616 (83.6%)	809,030 (98.7%)	693,785 (86.8%)	1,454,216 (98.9%)	736,127 (98.9%)	1,456,822 (99.0%)
Number of assigned reads**	453,410 (82.6%)	685,854 (84.7%)	569,851 (82.1%)	1,217,954 (83.7%)	631,574 (85.7%)	1,244,762 (85.4%)
Longest assigned read length (nt)	10,129	16,235	12,385	18,329	18,578	17,099

* Sample Cancer3 and Non-transformed3 are mismatched, originating from different patients.

** The calculation was based on the percentage of "useful reads above Q7".

5.3.3 TALON: Detecting and Measuring Both Known and Novel Transcripts

Long-read sequencing technology presents the significant benefit of capturing entire transcript sequences, enabling the resolution of intricate RNA isoform structures and the discovery of previously unidentified transcripts unattainable with short-read sequencing [247].

For the task of detecting and quantifying known and novel RNAs, we utilised the TALON pipeline. This software differentiates unannotated transcripts by the unique aspects of their sequences, such as previously undetected splice donor and acceptor connections. The taxonomy adopted by TALON for these classifications is detailed in Figure 4.

Our analysis identified 391,234 transcripts originating from 35,220 distinct genes (Figure 5.3A). These transcripts encompassed a diverse array of RNA types, including protein-coding genes, pseudogenes, long non-coding RNA, non-coding RNA, among other categories, as shown in Figure 5. Given the relevance of all RNA types in cancer research, we nonetheless narrowed our focus to solely the

mRNA transcriptome for these samples. In the initial count table, a dominant 91% of the detected transcripts were deemed novel, many of which appeared only in a single sample and typically at low counts. We instituted stringent filtering criteria to eliminate low-abundance transcripts and potential sequencing artefacts, as elaborated in the Materials and Methods Section 5.2.7. Moreover, we subjected the ISM category to additional scrutiny, employing even stricter filters.

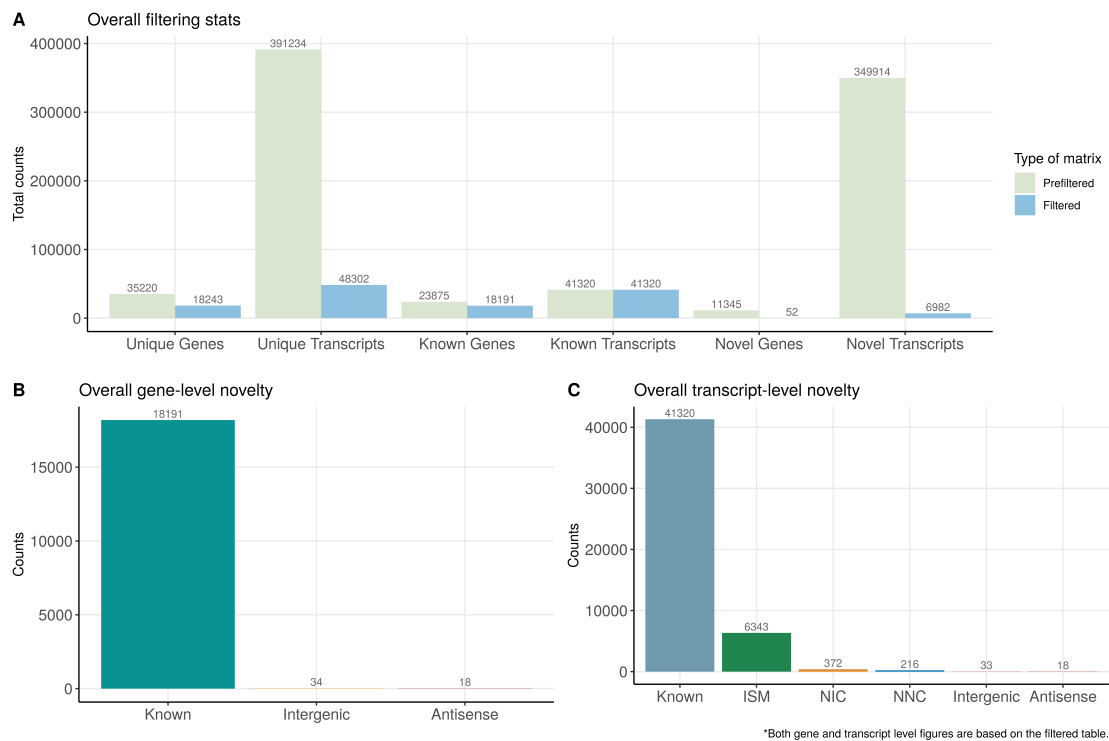


FIGURE 5.3: Overview of RNA sequencing workflow and data characteristics. (A) Depicting the overall filtering statistics, contrasting prefiltered (light green) and filtered (dark blue) counts for unique genes, unique transcripts, known genes, known transcripts, novel genes, and novel transcripts. (B) Illustrating the overall gene-level novelty, depicting counts for known genes, intergenic regions, and antisense transcripts. (C) Presenting the overall transcript-level novelty, with counts for known transcripts, ISMs, NICs, NNCs, Intergenic and Antisense transcripts. Please note that both gene and transcript level figures are based on the filtered data.

The filtered expression matrix featured a total of 48,303 distinct transcripts, with 41,320 known annotations and 6,983 novel ones. The lion’s share of these novel transcripts was ascribed to the ISM category (90.83%), succeeded by NIC (5.32%), NNC (3.09%), Intergenic (0.47%), and Antisense (0.25%). There were no marked differences in the distribution of these categories across the samples, as illustrated in Figure 6.

5.3.4 Exploratory and DGE Analysis in NSCLC and Adjacent Non-Transformed Tissues via Long-Read Sequencing

Short-read sequencing is renowned for its precise gene expression analysis capabilities. Recent literature has validated that DGE analysis using long-read sequencing data is reliable and consistent with results from short-read sequencing [244], [248]–[250]. In our study, edgeR was employed to pinpoint and scrutinise genes expressed differently between tumour and non-transformed tissue samples.

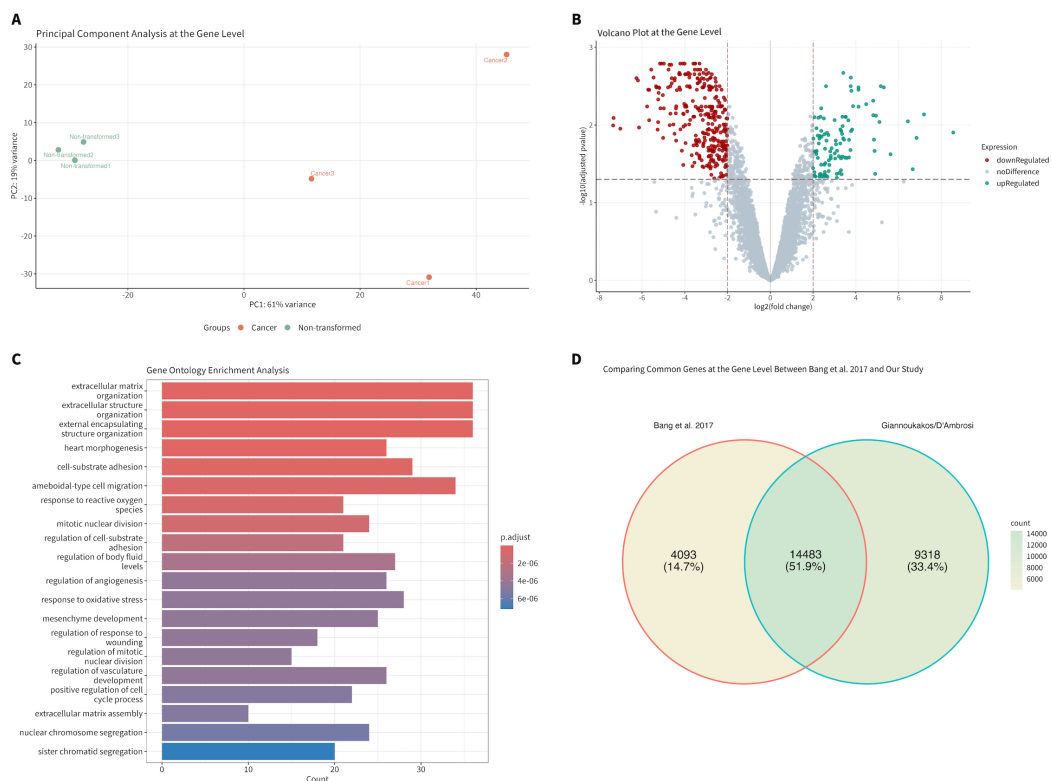


FIGURE 5.4: Composite Overview of Gene-Level Analysis in cancer versus non-transformed samples. (A) Shows a PCA at the gene level that distinguishes cancerous samples from non-transformed tissues, noting the percentage of variance explained by the first principal component. (B) Displays a Volcano Plot illustrating differences in gene expression between the two sample types, with genes categorised as up-regulated (in green), down-regulated (in red), or unchanged (in grey), based on a log₂ fold-change threshold of 2 and an adjusted p-value cutoff of 0.05. (C) Details a Gene Ontology Enrichment Analysis, with bars representing the frequency of genes involved in various biological processes, coloured by significance of enrichment. (D) Presents a comparative Venn diagram that quantifies and contrasts genes identified in this study with those found in prior research by Bang et al. in 2017, emphasising the unique and overlapping findings relevant to NSCLC genomics.

Multivariate analysis techniques like PCA and hierarchical clustering (referenced in Figure 5.4A and supplementary Figure 7) delineated tissue-specific clusters grounded on gene expression profiles. Notably, non-transformed tissues demonstrated greater sample-to-sample expression consistency in comparison to the tumour tissues, likely reflective of the latter's intrinsic heterogeneity.

Our criteria for identifying differentially expressed genes included an absolute log₂ fold-change threshold greater than two and an adjusted p-value of less than 0.05. Our analysis revealed 400 genes with differential expression; of these, 103 were up-regulated and 297 down-regulated in cancer tissues (as shown in Figure 5.4B). Gene ontology analysis indicated these differentially expressed genes participate in pathways related to extracellular matrix composition, structural organisation, and cellular adhesion and migration (presented in Figure 5.4C).

In assessing the efficacy of DRS with ONT for DGE analysis, we drew comparisons with short-read sequencing studies, such as the one by Bang, Kang, Lee, *et al.* (2017) [251]. They conducted similar DGE analysis on NSCLC samples and corroborated their findings with data from additional datasets, including 71 paired samples from Gene Expression Omnibus (GSE40419) and 58 paired samples from The Cancer Genome Atlas Program. In our dataset, a significant proportion (51.9%) of genes overlapped with those identified by Bang, Kang, Lee, *et al.* (referenced in Figure 5.4D). Furthermore, the GO analysis from both studies showed a high concordance in the significantly altered pathways (illustrated in Figure 5.4C and supplementary Figure 8). Bang, Kang, Lee, *et al.* identified 10 genes as potential NSCLC biomarkers, with specific genes such as MFAP4, AGER, GPX3, SPTPC, and A2M being associated with poor prognosis when down-regulated, and SPP1 when up-regulated. Our data mirrored these expression patterns, reinforcing the notion that DGE analysis via long-read sequencing is a robust alternative to traditional short-read sequencing approaches.

5.3.5 Assessment of Differential Transcript Expression in NSCLC Using Long-Read Sequencing

Contrary to gene expression analysis, estimating transcript expression is notably more complex with short-read sequencing due to the difficulty in uniquely assigning similar isoforms or those with repetitive sequences [252], [253]. ONT and other

long-read sequencing methods offer a solution by capturing the entire transcript length, thus facilitating precise isoform identification. The DRS approach also provides an accurate transcript quantification, free from biases typically introduced by retro-transcription or amplification processes [244]–[246], [249], [254].

Utilising TALON to detect and quantify both known and novel transcripts, our subsequent DTE analysis with edgeR highlighted 369 transcripts exhibiting significant differential expression between cancer and non-transformed samples, as depicted in Figure 5.5A). These transcripts, selected based on an absolute log₂ fold change exceeding 2 and an adjusted p-value under 0.05, represent a fraction of the total 6376 filtered transcripts that were examined for DE. Similar to our gene expression results, the majority of these transcripts (287 out of 369) were found to be down-regulated in the cancer samples, while a smaller portion (82 out of 369) were up-regulated. Notably, a significant portion (49.6%) of these significantly expressed transcripts constituted isoforms that had not been characterised before. In particular, our analysis identified 157 novel transcripts that were down-regulated and 26 that were up-regulated in the NSCLC samples, as presented in Figure 5.5B).

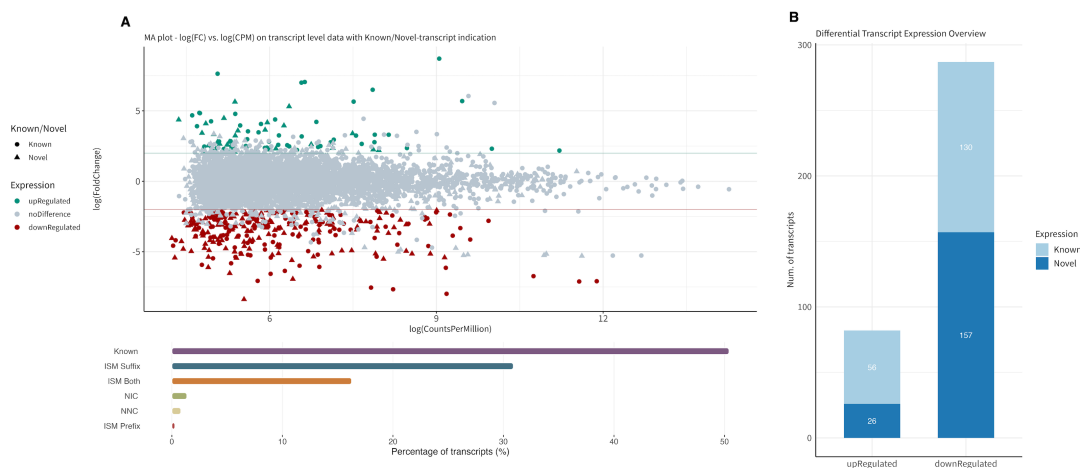


FIGURE 5.5: Transcriptomic Landscape of Differential Expression in Cancer. (A) MA plot illustrating the log₂ fold change against log Counts Per Million, distinguishing between known and novel transcripts. The significant differentially expressed transcripts are marked as up-regulated (green) and down-regulated (red), with novel transcripts denoted by triangles. (B) Bar graph summarising the number of known and novel transcripts that are up-regulated and down-regulated, providing a clear distribution of expression changes in the cancer samples.

The gene ontology analysis revealed that these novel isoforms with altered expression are implicated in essential biological functions. For the up-regulated novel transcripts, terms like extracellular matrix structural constituent and cell adhesion molecule binding were markedly enriched, indicating a significant role for these transcripts in the maintenance of cellular architecture and interactions (Figure 9A). In contrast, the down-regulated novel transcripts are associated with fundamental biological processes, including cell motility, maintenance of organismal homeostasis at the multicellular level, and protein metabolism regulation. These associations suggest that the reduction in expression of these transcripts may contribute to the disruption of standard cellular operations and homeostasis within the cancerous environment (Figure 9B).

A gene of particular interest due to its significant DTE is *AGER*, also known as the Advanced Glycosylation End Product-Specific Receptor, or RAGE. *AGER* is part of the immunoglobulin superfamily and functions as a multi-ligand cell surface molecule. It interacts with a variety of ligands, including advanced glycation end-products (AGEs), which accumulate in various tissues and organs over time [255]. The engagement of *AGER* with AGEs is implicated in the modulation of several chronic diseases, most notably diabetes [256], and Alzheimer's disease [257].

In the context of cancer, *AGER* is known for its involvement in cancer progression, exhibiting a dual expression pattern: it is naturally up-regulated in normal lung tissue, but down-regulated in lung cancer [258]–[260]. This observation aligns with our data. Chen, Chen, Chang, *et al.* (2020) in their study published in *Cell Death and Disease*, found that overexpression of RAGE, the product of the *AGER* gene, slows the proliferation of lung cancer cells and leads to cell cycle arrest. Furthermore, loss of RAGE during the initial stages of tumour development interrupts these regulatory mechanisms, leading to unregulated cell division and the emergence of lung tumours. Interestingly, increased RAGE expression was observed in cell lines with greater invasiveness. Moreover, the research noted that the interaction between the RAGE protein and certain immune cells could speed up tumour growth, thus indicating a complex role in both suppressing and promoting lung cancer development [261].

Figure 5.6 presents a comparison of the *AGER* gene's expression in non-transformed and cancerous tissues. The left-side bar chart reveals a pronounced decrease in *AGER*'s overall gene expression in cancer samples compared to non-transformed ones, as highlighted by the significance markers. Specifically, non-transformed

samples exhibit an average normalised expression of 2018.7, which significantly reduces to 28.2 in cancer samples, with a log2 fold change of -6.1 and an adjusted p-value of 0.0107. Such down-regulation of AGER in cancerous tissues aligns with findings from other studies, implying a potentially intricate regulation process. Notably, this down-regulation is even more pronounced at the isoform level, adding another layer to the gene's complex expression dynamics in cancer.

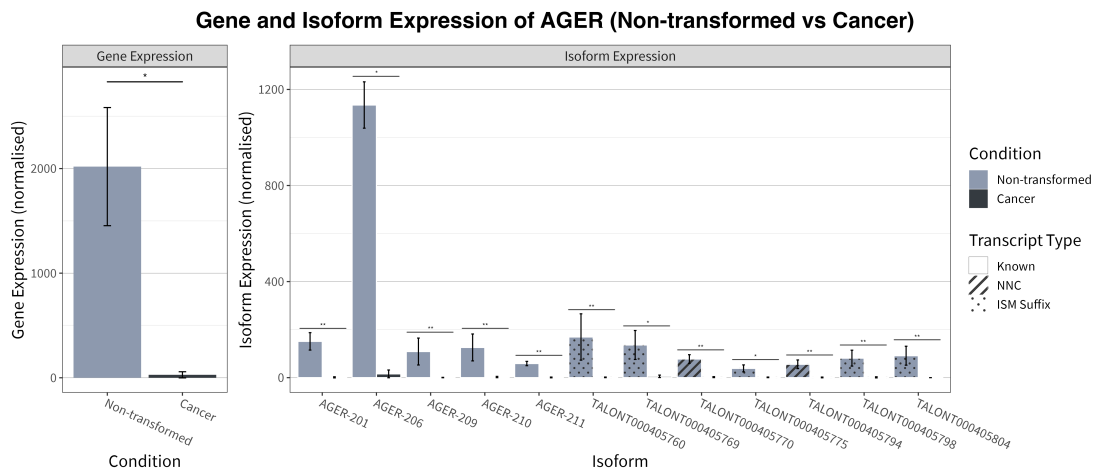


FIGURE 5.6: Differential Expression Analysis of AGER in non-transformed and cancer States. This composite bar chart illustrates the gene and isoform expression levels of AGER between non-transformed and cancer conditions. The bars indicate the mean expression levels, and the error bars represent the variability within the sample conditions, reflecting the standard deviation of the measurements. The condition of the samples is colour-coded for clarity. The left panel shows the aggregated gene expression, indicating a statistically significant decrease (* $p < 0.05$) in cancer compared to non-transformed samples. The right panel details the expression of individual isoforms, with the y-axis representing normalised expression levels. The various isoforms are distinguished by their transcript types (Known, NNC, ISM Suffix). Notable variations in expression are marked with significance levels, where a single asterisk (*) denotes $p < 0.05$ and double asterisks (**) denote $p < 0.01$.

Moving to the right side of the figure, we observe a more granular view with the expression levels of various AGER isoforms. It is worth noting, that DRS captured 12 different isoforms of AGER. Five are known and seven are novel. From these seven, two belong to the Novel Not in Category and the rest five belong to ISM Suffix category. The expression patterns of various isoforms of the AGER gene differ markedly between non-transformed and cancer states, with a notable and significant reduction observed in cancer samples, as indicated by the asterisks signifying statistical significance. This pattern suggests that both the AGER gene

and all its isoforms are consistently suppressed in the cancerous environment, implying a systematic down-regulation by the cancerous cells.

The presence of different transcript types (Known, NNC, and ISM Suffix) reflects the diversity in the mRNA splicing patterns leading to the different AGER isoforms (Figure 10). Known transcripts are those that have been well-characterised and annotated in genomic databases. In contrast, NNC and ISM indicate novel or less-characterised transcripts, which might represent alternative splicing events or variations that are not fully understood yet.

5.3.6 Transcriptomic Diversity in Cancer: Investigating Differential Transcript Usage

DTU is a critical aspect of gene expression dynamics, where various isoforms of a single gene are differentially expressed across distinct biological conditions. DTU refers to the phenomenon where different versions (transcripts/isoforms) of a single gene are expressed in varying patterns across different cell types, developmental stages, or in response to certain conditions. This variation is orchestrated through mechanisms such as alternative splicing, as well as the use of alternative transcription start and termination sites, giving rise to multiple mRNA variants from a single pre-mRNA template. The resulting isoforms diversify in function, localisation within the cell, and potential interactions with other biomolecules, thereby influencing cellular function and organismal homeostasis [262], [263].

Comprehending the nuances of DTU is indispensable for unraveling the complexities of gene expression regulation and its consequent impact on health and disease. In our study, we leveraged the detailed transcriptomic data provided by nanopore technology to compare isoform expression between the non-transformed and cancer samples. This technology affords an unprecedented opportunity to quantify the full length of transcripts, enriching our understanding of DTU which was previously challenging to capture.

Utilising isoformSwitchAnalyzer, detailed in Materials and Methods Section 5.2.9, we scrutinised an initial dataset of 48,146 isoforms spanning 18,087 genes. The software's stringent filtering process removed any genes expressing a solitary isoform or those without detectable expression, thereby refining our analysis to 28,156

isoforms from 7,378 genes. This curation was pivotal for ensuring an intricate examination of transcriptomic diversity.

The analysis unveiled striking DTU patterns within a subset of ten genes, notably including MBD3, MEST, and PRMT1, among others. For an extended overview of all genes involved, refer to Supplementary Table 10.

Among these, the mesoderm-specific transcript (MEST) gene emerged as an interesting case of DTU between non-transformed and cancer samples. The MEST gene, a subject of interest within cancer research, has recently been implicated as a regulator of invasiveness in lung cancer [264]. This gene is typically regulated through genomic imprinting, where only one allele—either maternal or paternal—is expressed. However, in various cancers, including lung adenocarcinomas, a frequent loss of imprinting occurs. This loss results in the activation of both alleles, leading to a phenomenon known as promoter switching, particularly from isoform 1 to isoform 2 of MEST [265], [266].

Our investigation into the MEST gene has uncovered a fascinating instance of DTU when contrasting non-transformed and cancer cell samples. As depicted in Figure 5.7, a closer inspection reveals no apparent difference in the overall gene expression levels of MEST between the two sample types. DGE analysis corroborates this, indicating no significant changes in MEST expression, with a log₂ fold change of 0.4205 and an adjusted p-value of 0.6633. These findings suggest that standard gene-level analyses might overlook crucial variations in MEST activity.

Nevertheless, a significant shift is observed at the isoform level, where isoform usage between the conditions dramatically differs. In non-transformed samples, isoform ENST00000223215.10 (MEST-201) is predominantly utilised. Conversely, cancer samples exhibit a marked preference for isoform ENST00000341441.9 (MEST-202), which is notably longer and includes a Signal Peptide domain. This domain is typical of proteins destined for the secretory pathway [267], hinting at the isoform's potential to mediate interactions with a distinct array of molecular partners in cancerous cells.

Given the complex role of MEST in cancer, where it is expressed in tumours via promoter switching and is associated with invasive cancer properties, the switch to the longer MEST-202 isoform could be of particular importance. The presence of the Signal Peptide domain in MEST-202 raises the hypothesis that this isoform may facilitate unique interactions within the cancer microenvironment, potentially

contributing to the mechanisms of tumourigenesis and metastasis. Understanding these interactions and the signalling pathways they involve is crucial for developing targeted cancer therapies. Thus, while MEST’s expression level remains consistent across cell types, the shift in isoform usage signifies a nuanced layer of regulation, possibly linked to the gene’s regulatory functions in tumourigenic processes.

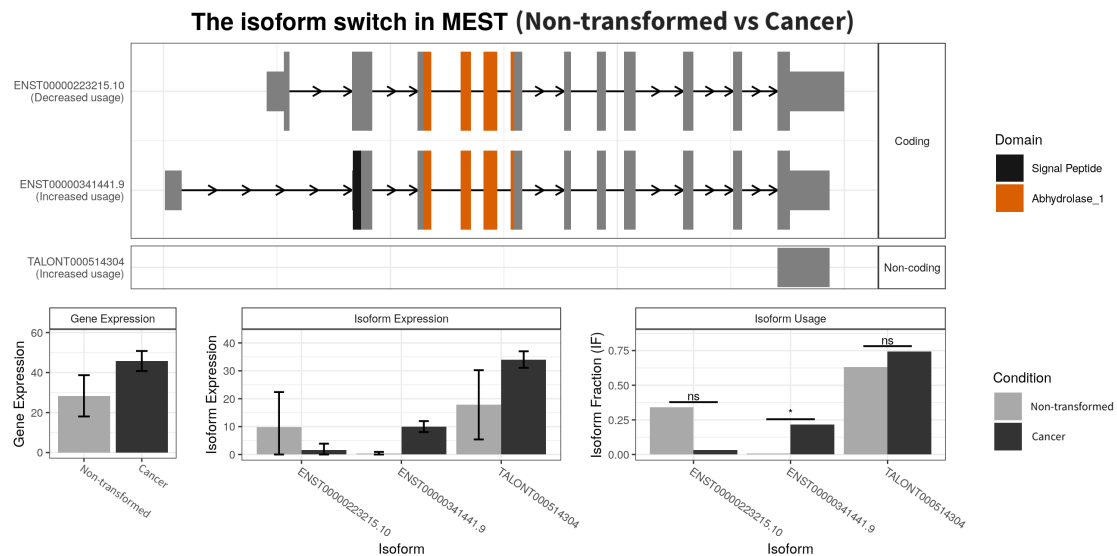


FIGURE 5.7: Isoform Usage Shift in MEST Gene between Non-transformed and Cancer Conditions. The figure depicts the isoform switch in the MEST gene between conditions. The top panel illustrates exon-intron structures of three distinct MEST isoforms, with coding regions in grey and the Abhydrolase_1 domain in orange. The lower panels show quantified gene and isoform expressions, and isoform usage fractions (IF), contrasting non-transformed with cancer conditions. Bars represent mean expression levels, and error bars indicate variability. Significant changes in isoform usage are marked by asterisks, underscoring the potential regulatory implications of isoform switching in oncogenesis.

5.3.7 Unraveling the Complexity of NSCLC: Insights from PolyA Tail Length Variations

In the pursuit to deepen our understanding of NSCLC, the DRS technology allowed us to delve into the intricacies of polyadenylation among the transcriptomes of both cancerous and non-transformed samples.

Polyadenylation, the process of adding a polyA tail to RNA transcripts, is a critical post-transcriptional modification that can influence RNA stability, nuclear export, and translation efficiency. Variations in polyA tail length have been associated

with diverse cellular states and can have profound implications for gene regulation and expression [268].

In our analysis, we meticulously estimated the polyA tail length of the transcriptome, uncovering subtle yet significant differences between cancerous and non-transformed tissues. The DPA analysis revealed a statistically significant disparity in the polyA tail lengths across the two conditions, pointing to a potential mechanism of post-transcriptional regulation that may contribute to the pathophysiology of NSCLC.

In a more detailed look, the overall evaluation of the two groups using the DPA revealed striking outcomes. The application of the Mann-Whitney U test yielded a statistic of 2,727,598 and a p-value of $2.2e-16$ (Figure 5.8A). This result strongly suggests a substantial and statistically significant difference in polyA tail lengths at a broader level when comparing cancerous to non-transformed samples. These striking results strongly suggest that variations in polyA tail length are not merely incidental, but are systematically associated with the cancerous state.

Further delving into the data, our findings from the DPA illuminated the scale of polyadenylation heterogeneity within these groups. An astonishing 2254 out of 8821 transcripts, approximately a quarter of the analysed transcriptome, showed significant variations in polyA tail length (Figure 5.8B). This substantial fraction of the transcriptome exhibiting alternative polyadenylation events in NSCLC is indicative of the complex molecular alterations that occur in cancerous cells compared to their non-transformed counterparts.

Additionally, we conducted an APA analysis, which investigates the usage of different polyadenylation sites within the same gene, a phenomenon known to generate transcript diversity. APA is a crucial post-transcriptional regulatory mechanism that produces transcripts with distinct 3' ends. The RNA molecules undergo cleavage at specific poly(A) sites, followed by the addition of a poly(A) tail at their 3' ends [269]. By applying the LAPA method (refer to Materials and Methods, Section 5.2.13 for details), we conducted a Differential Alternative Polyadenylation Analysis (DAPA). Out of 1178 filtered transcripts assessed using Fisher's exact test, 247 showed evidence of DAPA, meeting criteria of an adjusted p-value of 0.05 and an absolute delta Usage threshold above 0.3. More specifically, 130 transcripts were significantly down-regulated, while a nearly equal number, 117, were significantly up-regulated, as elaborated in Figure 5.8C.

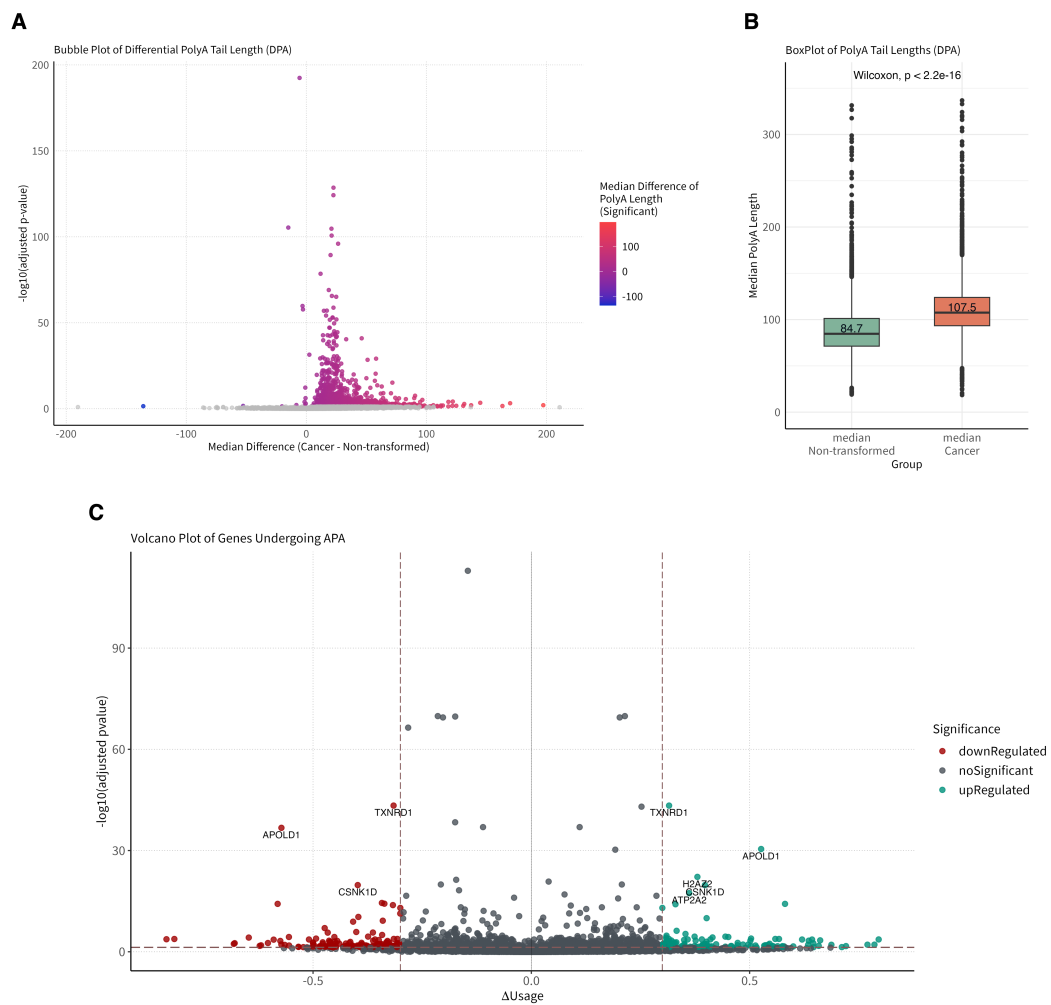


FIGURE 5.8: Comprehensive Analysis of Polyadenylation Variability in Cancer. (A) Bubble plot displaying the differential polyA tail length (DPA) across all transcripts, with bubble size corresponding to the median cancer PolyA length and colour intensity indicating significance level. (B) Boxplot of global comparison of polyA tail lengths (DPA) between non-transformed and cancer groups, with median values annotated, showcasing a statistically significant difference as determined by the Wilcoxon test. (C) Volcano plot highlighting genes with alternative polyadenylation (APA) patterns, where the vertical axis represents the negative log₁₀ adjusted p-value and the horizontal axis shows the delta usage. Points are coloured based on regulatory direction, with labels identifying notable genes with significant APA events.

The Volcano plot illustrated in Figure 5.8C enables the identification of two distinct isoforms of the gene TXNRD1, both demonstrating considerable APA. One isoform, situated to the left (TXNRD1-210), is markedly down-regulated, while the other, on the right, is significantly up-regulated (likely TXNRD1-204). Thioredoxin reductase-1 (TXNRD1) is known for its overexpression in several cancer

types, including lung cancer, and is believed to be a critical factor in the proliferation and survival of cancer cells [270]. It is essential to delve further into understanding the role of APA in these TXNRD1 isoforms. Investigating how the noticeable changes in the length of the polyA tail can substantially influence these isoforms, particularly in terms of mRNA stability and translation efficiency, is crucial. These changes may also be linked to the notable up-regulation of the gene observed in the cancer state, which is characterised by a log₂ fold change of 1.969 and an adjusted p-value of 0.0432.

5.3.8 Deciphering RNA Methylation Patterns in the NSCLC Transcriptome: A New Dimension of Analysis

The study of RNA modifications, particularly methylation, has become a frontier in understanding the regulatory complexities of the transcriptome. Our investigation primarily utilised the DRS technology to probe the methylome of the NSCLC transcriptome. This innovative approach allowed us to directly observe and quantify RNA modifications without the need for reverse transcription, which can often obscure the detection of such modifications.

Methylation, the addition of a methyl group to the nucleotide bases of RNA, is a critical post-transcriptional modification that can influence RNA stability, splicing, localisation, and translation. In healthy cells, methylation plays a crucial role in the normal functioning of RNA molecules, aiding in the fine-tuning of gene expression and cellular homeostasis. However, in the context of cancer, aberrant methylation patterns have been observed. Such alterations can disrupt the normal regulatory processes and contribute to the onset and progression of malignancy. In NSCLC, changes in RNA methylation could potentially serve as biomarkers for diagnosis, prognosis, and therapeutic targets.

In our quest to unravel the complexity of RNA methylation in NSCLC, we performed comprehensive methylation detection and analysis using the software tool xPore, as detailed in Materials and Methods Section 5.2.14. xPore's sophisticated analysis algorithm enabled us to detect differential methylation patterns between non-transformed and cancerous tissues with high precision. By comparing these patterns, we aimed to identify specific methylation signatures that are characteristic of the cancerous state.

Our analysis with xPore encompassed 1263 distinct transcripts derived from 1192 genes. Notably, xPore identified over 1000 different kmers. To our surprise, of the total 1024 kmers analysed, only 18 were classified as DRACH kmers, indicative of m6A methylation, while the remainder were associated with various other modifications.

Focusing on the most significantly modified transcripts, we identified ten key transcripts: GNLY-201, ANGPTL2-202, SMIM26-201, ATXN7L3B-201, TRIM56-201, C11orf98-201, TALONT000418557, TCEAL4-208, TXNDC17-201, and UBA52-202. These transcripts represent a diverse array of genes, each potentially playing a unique role in the biological processes under investigation. We proceeded to further explore the genes from which these top ten modified transcripts originate, aiming to better understand their functional implications and potential roles in disease processes, particularly focusing on their involvement in the complex molecular landscape of NSCLC. Those top ten most extensively modified transcripts, which are associated with a diverse set of genes: GNLY, ANGPTL2, SMIM26, ATXN7L3B, TRIM56, C11orf98, FOXP1, TCEAL4, TXNDC17, and UBA52. These genes represent a broad spectrum of biological functions and implications, highlighting the pervasive influence of RNA modifications in cellular processes:

1. **GNLY (Granulysin):** This gene plays a critical role in the immune response, known for its involvement in cytotoxic activity against various tumour cells and microbes. RNA modifications in its transcript could impact how the immune system responds to pathogens and malignant cells [271]–[273].
2. **ANGPTL2 (Angiopoietin-Like 2):** This gene plays a crucial role in angiogenesis and inflammation, with observational clinical studies highlighting its significant increase in various chronic inflammatory diseases [274]. Elevated ANGPTL2 levels have been linked to the diagnosis and prognosis of cardiovascular diseases, diabetes, chronic kidney disease, and various cancers, suggesting that alterations in its transcript are vital for understanding its impact on these diseases. Additionally, ANGPTL2, secreted from cancer cells, leads to increased tumour cell invasion, motility, and angiogenesis, contributing to tumour metastasis. Notably, high ANGPTL2 expression within primary tumour sites in lung cancer patients correlates with poorer disease-free survival outcomes [275].

3. **SMIM26 (Small Integral Membrane Protein 26):** A relatively obscure gene, may play a role in membrane biology or cellular signalling. Recent research reveals that SMIM26 is a microprotein encoded by LINC00493, situated in mitochondria. It is notably down-regulated in clear cell renal cell carcinoma (ccRCC), with this decreased expression linked to poorer overall survival rates. Functionally distinct from its encoding lncRNA, SMIM26 curbs the growth and metastasis of ccRCC [276].
4. **ATXN7L3B (Ataxin 7-Like 3B):** A gene involved in transcriptional regulation, has functions that are not yet fully understood. However, it has been implicated in various disorders, including cancer. For instance, research by Chen, Cha, Yan, *et al.* (2021) [277] revealed that ATXN7L3B expression is significantly inversely correlated with survival in liver cancer patients. Moreover, Leberfarb, Degtyareva, Brusentsov, *et al.* (2020) also showed its association with an increased risk of colorectal cancer [278].
5. **TRIM56 (Tripartite Motif Containing 56):** Belonging to the TRIM protein family and known for its involvement in antiviral defence mechanisms, TRIM56 has been associated with various cancers such as lung adenocarcinoma, , hepatocellular carcinoma, and multiple myeloma. Research indicates that TRIM56 is crucial in hindering tumour progression, achieving this through the modulation of the Wnt and TLR3/TRIF signalling pathways [279]–[281].
6. **C11orf98 (Chromosome 11 Open Reading Frame 98):** While the specific function of this gene remains largely unclear, it is known to play a role in regulating the activity of DNA-binding transcription factors, a function influenced by its range of protein partners including ESR1, ESR2, FOXA1, JUN, and WWP2 [282].
7. **FOXP1 (Forkhead Box Protein P1):** Is a highly conserved transcription factor within the Forkhead Box P (FOXP) family, playing key roles in regulating gene transcription in various tissues and cell types throughout development and adulthood. It is suspected to function as a tumour suppressor due to its loss in several cancer types and its location in a chromosomal region known for housing tumour suppressor genes [283]. In the lung, FOXP1 is known to be a key regulator of epithelial gene transcription [284]. In lung cancer, reduced expression of FOXP1 has been linked to poorer survival outcomes [285].

8. **TCEAL4 (Transcription Elongation Factor A Like 4):** Is a gene involved in transcriptional regulation, but has an undefined role in cancer. Research suggests that it often shows decreased expression in various cancer types, indicating a potential link to oncogenesis [286]–[288].
9. **TXNDC17 (Thioredoxin Domain Containing 17):** A highly conserved oxidoreductase protein present in mammalian tissues, plays a key role in cellular processes. While its specific functions are not yet fully understood, recent studies have linked TXNDC17 to the TNF signaling pathway, which is known to stimulate cellular autophagy [289].
10. **UBA52 (Ubiquitin A-52 Residue Ribosomal Protein Fusion Product 1):** Is a key in targeting cellular proteins for degradation and maintaining cellular homeostasis. A recent study highlighted its indirect yet crucial involvement in the cell cycle progression and proliferation within NSCLC cell lines, pointing to its potential impact in cancer biology [290].

The extensive modification of transcripts from these genes underlines the complexity and importance of RNA modifications in regulating gene expression and function across a wide range of biological contexts. This area is ripe for further exploration to unravel the intricate mechanisms by which RNA modifications influence cellular and molecular processes, potentially offering new avenues for therapeutic interventions.

Furthermore, we conducted an exclusive investigation into known tumour - suppressive and oncogenic genes related to lung cancer. We retrieved a list of 916 genes from the Lung Cancer Gene Literature Database [291] and only 79 were detected by xPore. Notably, four genes – BTG2, IGFBP7, RHOC, and SFN – showed significant differential modifications between cancerous and non-transformed tissues with p-values: 0.0375, 0.0471, 0.0174, 0.0176 respectively (Figure 5.9). These genes have crucial roles in lung cancer:

- **BTG2 (B-cell Translocation Gene 2):** BTG2 is known to play a significant role in cell cycle regulation and apoptosis. In the context of lung cancer, BTG2 acts as a tumour suppressor [292]. Its expression is often found to be down-regulated in cancerous tissues compared to normal tissues, suggesting that loss of BTG2 function contributes to tumourigenesis. BTG2's role in inhibiting cell proliferation and promoting apoptosis makes it a potential

target for therapeutic strategies aiming to restore its function in lung cancer cells [293], [294].

- **IGFBP7 (Insulin-like Growth Factor Binding Protein 7):** A member of the IGFBP superfamily, also known as GFBP-related proteins, is a widely secreted protein with a complex role in multiple biological functions. This protein is crucial to the insulin-like growth factor (IGF) system [295] and is actively involved in regulating key cellular processes such as proliferation, differentiation, angiogenesis, cell adhesion, and senescence across various cell types [296]. Its involvement in these diverse mechanisms underlines its significance in both normal physiological and pathological conditions. In healthy lung tissues, particularly within small airway bronchial epithelial cells, IGFBP7 is typically expressed at higher levels, but this expression significantly decreases in primary lung cancer tissues [297]. Research indicates that IGFBP7 plays a vital role in lung cancer, influencing its progression, response to treatment, and metastasis [298]–[300].
- **RHOC (Ras Homolog Family Member C):** This protein is part of the Rho GTPase family, plays a key role in cytoskeletal regulation, impacting cell morphology and movement [301]. The abnormal activation of Rho GTPases, including RHOC, is associated with increased movement, invasion, and metastasis of cancer cells [302]. RHOC undergoes post-translational modifications such as methylation, essential for its localisation and function [303]. In lung cancer, overexpression of RHOC has been also associated with increased tumour metastasis. RHOC facilitates the aggressive behaviour of cancer cells, including enhanced migration and invasion capabilities, which are key factors in the spread of cancer to other parts of the body. Therefore, targeting RHOC's signalling pathways presents a promising therapeutic approach to manage lung cancer metastasis [304].
- **SFN (Stratifin, also known as 14-3-3 sigma):** Belongs to the 14-3-3 protein family, is involved in regulating the cell cycle and apoptosis [305]. Often considered a potent oncogene, its elevated expression in lung cancer is believed to play a significant role in the development and progression of the disease, highlighting its potential impact as a target in lung cancer therapy and research [306]–[308].

Deciphering the RNA modification landscape in NSCLC is pivotal for understanding cancer pathogenesis and developing new therapeutic approaches. Targeting specific methylation sites or related enzymes could help normalise aberrant gene expression and slow cancer progression. Integrating methylation data with gene expression profiles and clinical outcomes promises a more customised cancer treatment approach, tailoring therapies to each patient’s unique molecular tumour profile.

While this research paves the way for innovative diagnostic and therapeutic methods anchored in RNA modification patterns, potentially enhancing patient outcomes in NSCLC, it is clear that we have only just begun to scratch the surface of the NSCLC methylome. There is a pressing need for more extensive and in-depth research to comprehensively understand the breadth of information uncovered and to effectively translate these findings into clinical practice.

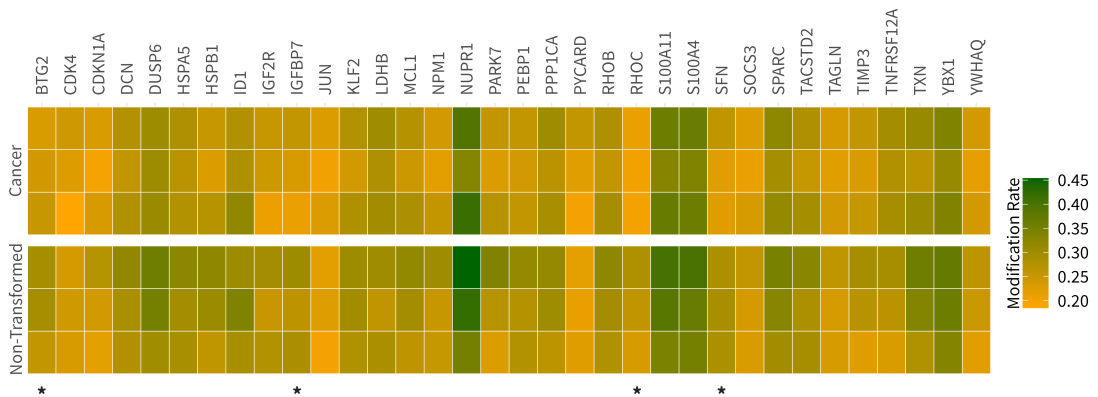


FIGURE 5.9: Comparative Heatmap of Methylation Rates in Cancer vs Non-Transformed samples. This heatmap displays varying methylation rates across a spectrum of tumour-suppressive and oncogenic genes related to lung cancer, comparing cancerous to non-transformed samples; each horizontal band represents a sample, with upper bands for cancer and lower for non-transformed samples; vertical bands correspond to different genes. Colour gradients from dark green to bright orange reflect methylation rates, with darker tones indicating higher methylation. Genes marked with asterisks below the heatmap show significant methylation differences.

5.4 Discussion

This study marks an advancement in lung cancer research, particularly focusing on the transcriptomic profiling of NSCLC adenocarcinoma tissues. Employing the

advanced Nanopore Direct RNA Sequencing, this research stands as an initiative in comparing the transcriptome profiles of lung cancer tissues and adjacent non-transformed lung tissues. This approach enhances our understanding of NSCLC adenocarcinoma, potentially aiding in identifying subtle changes in the disease's pathogenesis.

The creation of a specialised bioinformatics pipeline was a crucial aspect of this project. This pipeline is unique in its integration of various modules and pipelines for a simultaneous extraction of multi-layered transcriptomic information. At its core, the pipeline utilises TALON software for the characterisation and quantification of transcripts. The inclusion of multiple steps, such as CAGE analysis for 5' end verification of novel RNAs and polyA length estimation for 3' end verification, ensures robustness and accuracy of data. This comprehensive pipeline enables an extensive range of analyses including DGE, DTE, and DTU, alternative splicing, and post-transcriptional modification detection, among others.

Our analysis, focused on KRAS-mutant adenocarcinomas and adjacent non-transformed lung tissues, yielded rich data. From approximately 7.72 million sequenced reads, we initially quantified over 35,000 genes and 390,000 transcripts. Post stringent filtering, these numbers refined to about 18,200 genes and 48,300 isoforms, with 15% of the isoforms classified as novel. This significant discovery of novel transcripts, categorised into distinct groups by TALON, underscores the depth of our investigation. Furthermore, the validation of our results through comparison with the Illumina short-read RNA-Seq data from Bang, Kang, Lee, *et al.* 2017 reinforced the reliability of our findings.

One of the most striking discoveries in our study was related to the AGER gene. Our findings not only confirmed its down-regulation at the gene level in cancer tissues, but also revealed the down-regulation of all its 14 captured isoforms. This observation is particularly notable considering the role of AGER in cancer progression. Additionally, we identified an isoform switch in the MEST gene, which is linked to tumourigenesis, providing further insights into the molecular mechanisms underpinning lung cancer.

In our investigation, we encountered an intriguing instance of Differential Transcript Usage involving the mesoderm-specific transcript (MEST) gene when comparing non-transformed and cancer cell samples. Despite observing no apparent differences in the overall gene expression levels of MEST between the two sample

types, a significant shift emerged at the isoform level, revealing marked differences in isoform usage between the conditions. Non-transformed samples predominantly expressed isoform MEST-201, while cancer samples displayed a notable preference for isoform MEST-202, which notably features a longer sequence and includes a Signal Peptide domain associated with proteins destined for the secretory pathway. This observation suggests the potential of the MEST-202 isoform to facilitate distinct interactions within the cancer microenvironment, potentially influencing tumourigenesis and metastasis mechanisms.

A critical aspect of our research was the analysis of polyadenylation in the transcriptome. We observed significant differences in polyA tail length between cancerous and non-transformed tissues, with cancer tissues showing a median polyA length notably longer than that of non-transformed tissues. This difference in polyadenylation patterns offers valuable insights into the regulatory mechanisms at play in cancer cells. Furthermore, our analysis extended to DPA and APA patterns. The examination of these patterns, particularly in the context of genes like TXNRD1, known for its role in lung cancer, provided a deeper understanding of the gene expression regulation in NSCLC. Identifying differentially expressed isoforms of this gene, with variations in polyadenylation, underscores the complexity of gene regulation in cancer.

Finally, our study also delved into the realm of epitranscriptomics, revealing that the most common post-transcriptional modifications in lung cancer did not follow the typical DRACH motif, suggesting a divergence from established m6A modification patterns. Moreover, the analysis of the top ten significantly modified transcripts, many of which are involved in tumourigenesis, and the assessment of post-transcriptional modification changes in 916 known tumour-suppressive and oncogenic genes, provided novel insights into the epitranscriptomic landscape of lung cancer.

As we reflect on the findings of this study, it's evident that our exploration of the transcriptome in NSCLC adenocarcinoma has yielded valuable insights. However, the path to fully understanding this complex disease and translating these insights into clinical applications is a continuous journey. To enhance the outcomes of our research and further solidify our findings, several steps can be undertaken in the future. Firstly, expanding the sample size would be a critical step. A larger cohort of patients, encompassing diverse demographics and genetic backgrounds, would provide a more robust data set. This expansion would not only reinforce the

validity of our findings, but also help in identifying more nuanced transcriptomic variations that might be present in smaller subgroups of the population. Secondly, increasing the depth of sequencing could uncover additional layers of complexity in the NSCLC transcriptome. Deeper sequencing might reveal rare transcripts or subtle alterations that are not detectable with the current depth. This could be particularly beneficial for understanding the role of less abundant RNA species and their contributions to NSCLC pathogenesis and progression. Thirdly, enhancing the sophistication and efficiency of the existing bioinformatics pipelines represents a critical area for future improvement. By refining these pipelines, we can achieve greater accuracy and precision in our analyses. This enhancement will not only yield better results, but also deepen our understanding of the capabilities and full potential of this technology. Such advancements in our analytical tools are essential for extracting richer insights from the data, ultimately leading to a more comprehensive understanding of NSCLC at the molecular level.

Ultimately, this study lays a solid groundwork for future investigations and stands as a testament to the continuous evolution of lung cancer research. The insights gleaned here are expected to contribute to the ongoing efforts in developing more sophisticated diagnostic methods and treatment strategies for lung cancer, emphasising our commitment to tackling this challenging disease with an ever-evolving scientific approach.

5.5 Conclusion

This study represents a significant advancement in lung cancer research, particularly in understanding NSCLC adenocarcinoma at a molecular level. By employing Nanopore Direct RNA Sequencing and developing a specialised bioinformatics pipeline, we have provided valuable insights into the transcriptomic intricacies of this disease. Our comprehensive analysis, which includes characterising a wide array of genes and isoforms as well as delving into epitranscriptomics, highlights the potential of this technology to enhance current research and suggests new directions for developing improved diagnostic and therapeutic approaches. This research not only contributes substantially to the existing body of knowledge in lung cancer, but also stands as a noteworthy step towards more precise and effective treatment strategies, underlining its importance in guiding future efforts for better management and treatment of this complex disease.

Chapter 6

NanoInsights: A Web Platform for Advanced NanoString nCounter Data Analysis

“In science, we must be interested in things, not
in persons.”

— MARIE CURIE

This Chapter introduces ‘NanoInsights’, a sophisticated web service designed to simplify NanoString nCounter data analysis, merging advanced bioinformatics and machine learning into an accessible platform for both experienced researchers and novices. It offers a comprehensive QC process, a variety of normalisation methods, and a unique auto-detection feature to identify the most suitable normalisation technique for each dataset. The platform also incorporates diverse machine learning algorithms to uncover complex patterns in the data, aiding in predictive modelling and biomarker discovery. Interactive visualisations enhance user experience, facilitating the interpretation of results.

6.1 Introduction

In the intricate landscape of molecular biology, the detailed analysis of RNA expression is critical for distinguishing the molecular basis of health and disease.

Variations in expression, primarily assessed through RNA abundance, illuminate the complex regulatory processes that govern cellular functions. Far from being a mere procedural task, this analytical focus is vital for grasping the molecular shifts that drive pathological changes [309]. As such, technologies facilitating accurate RNA expression profiling are indispensable, playing a pivotal role in revealing cellular transcriptional dynamics [310]. Their use extends beyond basic observation, enabling researchers to intricately dissect the genetic orchestration that defines cellular behaviour in various physiological and pathological contexts [311].

RNA expression studies have traditionally leaned on three primary technologies: PCR, microarrays, and the more recent high-throughput sequencing. PCR and microarrays, while cost-effective and suitable for broad-scale transcriptomic analysis, suffer from limited dynamic range, potentially obscuring subtle gene expression nuances [312], [313]. In contrast, high-throughput sequencing offers intricate digital quantification of transcripts, but it comes with the requisites of substantial laboratory resources, higher associated costs, and computational complexities associated with data analysis [314], [315].

Amidst this technological landscape, NanoString has emerged as a compelling medium-throughput alternative for comprehensive RNA analysis. The NanoString nCounter platform, in particular, offers a targeted approach to RNA expression quantification, capable of analysing up to 800 targets in a single assay (refer to Section A.3.1 of the Supplementary for additional information). Distinguished by its ability to directly quantify RNA or DNA molecules in a single reaction, the nCounter system offers notable advantages over PCR, microarray, and RNA sequencing methods.

Primarily, its parallelised design and minimal manual intervention streamline data acquisition compared to many PCR-based techniques. Additionally, the hybridisation method directly interrogates target sequences, eliminating the need for potentially biased amplification steps, even for low-abundance transcripts. This feature makes it particularly effective for applications like liquid biopsy, which typically involves analysing samples with low quantity. Furthermore, it utilises digital detection of uniquely bar-coded probes, ensuring absolute quantification. In contrast to RNA sequencing (RNA-Seq) methods, nCounter RNA analysis bypasses cDNA retrotranscription and library construction, reducing potential bias and enhancing target quantification precision. With these advantages, the nCounter system has emerged as a powerful tool, ushering in a new era in RNA expression analysis and

molecular profiling [316], [317]. It has firmly established itself as a robust and reliable platform suitable for both research and clinical applications. Importantly, it has found acceptance in translational research due to its capability to effectively handle highly degraded, low-quantity samples. Recent studies have expanded its utility to investigate diverse bioanalytes, including mRNA [157], [159], miRNA [318], and circRNA [156], [158], [160], among others.

Despite its substantial benefits, the NanoString nCounter platform poses challenges for researchers, especially in the post-experiment phase. The data it generates is not readily interpretable and requires compilation, normalisation, and analysis. While the company offers the nSolver software as a data analysis solution, user experience and accessibility are critical in fully harnessing the potential of this groundbreaking technology. At present, NanoTube [319] stands as the sole platform that has made strides in this direction, offering a limited set of analytical features. However, a considerable gap persists in delivering a comprehensive and automated solution tailored to the diverse and evolving needs of researchers.

Acknowledging the necessity for a more user-friendly and intuitive approach to NanoString nCounter data analysis, this study introduces "NanoInsights", a sophisticated web service that seamlessly integrates Bioinformatics and Machine Learning for enhanced NanoString Data Analysis. This platform is expertly crafted to simplify the intricacies of data pre-processing and analysis, thereby broadening the accessibility of RNA expression analysis through NanoString technology to a diverse range of researchers. NanoInsights distinguishes itself with a user-centric interface, tailored for ease of use regardless of the user's level of expertise. This makes it an ideal tool for both experienced researchers and those new to the field, democratising the process of complex genomic data analysis. The platform encompasses a broad array of features, including comprehensive quality control and exploratory analysis, along with a selection of eight distinct normalisation methods. Its standout feature is the auto-detection capability, which intelligently identifies the most suitable normalisation method for each dataset, thus optimising data handling and minimising the need for specialised analytical skills. A key strength of NanoInsights is its integration of Machine Learning, capable of applying various classification techniques. This functionality allows researchers to uncover subtle, hidden patterns within their data. Particularly transformative, this capability opens new pathways for predictive modelling and biomarker discovery across multiple disciplines, from oncology to developmental biology, enhancing

both the interpretability and accuracy of analysis. Additionally, NanoInsights concludes the analytical process with gene enrichment analysis, providing a holistic suite of tools for researchers to investigate the functional implications of gene expression data. The platform offers interactive plots for a dynamic and engaging user experience and enables the downloading of figures in a high-resolution format that is visually appealing and suitable for publication.

6.2 Materials and Methods

6.2.1 Web-Service Development and Hosting Details

NanoInsights has been developed using a diverse stack of programming languages, including Django (v4.2.6) for the frontend framework, alongside HTML5 for markup, Cascading Style Sheets (CSS) for styling, and JavaScript for bespoke interactive visualisations. At its core, NanoInsights is powered by a backend framework combining Python (v3.11.4) and R (v4.3.1), leveraging the strengths of both languages for computational tasks. Key Python libraries utilised in the development include plotly (v5.18.0) for interactive visualisations [320], and sklearn (v1.3.2) for machine learning [213]. A detailed listing of all the packages employed is systematically presented in Table 12.

6.2.2 Quality Control and Preliminary Data Exploration

NanoInsights executes a standard quality control (QC) procedure, which encompasses NanoString's established general assay performance checks. This thorough QC stage includes evaluations of Imaging Quality, Binding Density, Positive Control Linearity, and the Limit of Detection. For each quality metric, a boxplot is generated in both interactive and static forms, showcasing the entirety of the data along with NanoString's recommended ranges. Additionally, a detailed data table is provided, offering an extensive review of each sample's performance within the dataset. Following this, an in-depth exploratory examination is conducted as an initial assessment of the data. This examination incorporates critical elements including boxplots of the unprocessed data, Principal Component Analysis (PCA) plots, Multidimensional Scaling (MDS) plots, and Interquartile Range (IQR) analyses.

6.2.3 Filtering Process for Genes and Samples

NanoInsights offers a range of both standard and mild gene and sample filters to refine data analysis. The platform features four primary filters, each of which can be fine-tuned with ease to improve the integrity of the input dataset:

Filter Lowly Expressed Genes: Utilising the edgeR package’s ‘filterByExpr’ function, this filter identifies and removes genes with low expression levels.

Filter Genes Based on Negative Controls: This filter leverages the Negative Control (NC) probes within the assay to detect and exclude genes that show insufficient expression or excessive background noise. It calculates the background noise threshold by averaging the NC values for each sample, adding twice the standard deviation to this mean, and then deducting this value from each gene’s expression level in the sample. Any gene with an adjusted expression level of zero or less in more than 85% of samples is removed.

Filter Samples Based on the Negative Controls: Similar to the gene filter, this filter excludes any samples where more than 85% of genes do not surpass the expression threshold established by the Negative Control-Based Gene Filtering process.

Remove Outlier Samples: This filter utilises IQR analysis to detect and eliminate outlier samples. NanoInsights enables the customisation of the IQR threshold, allowing for the adjustment of the outlier detection stringency to ensure the dataset’s integrity is tailored to the specific requirements of the analysis.

After the application of any of these filters, NanoInsights automatically compiles additional documentation that records which genes and/or samples were excluded in the process. The filtered dataset is then reassessed with various visualisations, such as PCA plots, MDS plots, and sample correlation heatmaps. These visual aids are designed to clarify the effects of the applied filters on the dataset, providing a transparent view of the data’s configuration after the filtering process.

6.2.4 Normalisation and Differential Expression Analysis

NanoInsights provides a selection of eight unique normalisation methods. These methods are comprehensively detailed in Table6.1, which includes an array of

TABLE 6.1: Comprehensive Overview of Normalisation Methods and Their Explanations in Data Analysis

Normalisation Method	Explanation
nSolver	Employs the standard nSolver normalisation, a conventional method for normalising data.
Housekeeping Scaling	Normalises counts by utilising a scaling factor, calculated by dividing the geometric mean of built-in housekeeping genes per sample by the arithmetic mean of their geometric means.
Housekeeping geNorm	Normalises counts based on housekeeping genes selected for their stable expression by the geNorm algorithm [199]. Scaling factors depend on the geometric mean of these steadfast housekeeping genes, aligning with the approach used in Housekeeping Scaling.
Endogenous and Housekeeping Scaling	Normalises counts using scaling factors generated by the ratio of the geometric mean of all counts (endogenous and housekeeping) per sample to the arithmetic mean of their geometric means.
Quantile	Data is normalised by giving each sample in the dataset the same distribution. The mean quantiles are used to substitute the value of the data point in the original sample.
Cyclic Loess	Applies cyclic Loess, a non-linear, local regression normalisation technique. Pairwise normalisation is performed based on differences between samples, transforming data onto log expression (M) and mean average of expression (A) scales [200].
Variance Stabilising Normalisation (VSN)	Data is normalised by parametric transformation based on a model of variance-versus-mean dependence [184].
RUVSeq	Utilising the RUVg function from the RUVSeq package to account for technical bias [201]. The RUVg method normalises data based on reference genes. The user can choose the target genes and geNorm will detect most stable ones that will be used in the RUVg function.

normalisation techniques ranging from conventional to more advanced approaches.

Additionally, NanoInsights introduces the "Auto-detection" feature. This option automatically evaluates all available normalisation methods, selecting the most suitable one based on the lowest mean Relative Log Expression (MRLE) score [321].

The RUVSeq method is specifically crafted to detect and rectify unintended variations, employing the RUVg function to estimate these variations accurately. NanoInsights offers a broad selection of choices for selecting the essential reference genes required for the function's operation.

Following the normalisation process, differential expression (DE) analysis is conducted using the limma (v3.58.1) R package [200]. Limma employs an empirical Bayesian method to identify differentially expressed genes. The output of this analysis includes a data table with normalised counts and relevant statistical measures. Also, several interactive and static visualisations are provided for an in-depth review of the normalisation process and an overview of the DE analysis. These visualisations, based on the normalised data, encompass a PCA plot, a RLE plot, a Density plot, Hierarchical Clustering Analysis of the samples, and a Volcano plot.

6.2.5 Optional Preprocessing and Feature Selection

At the outset of the classification analysis, there is an optional preprocessing step which includes the elimination of genes that exhibit high correlation or are nearly constant. This is followed by feature selection that can be optionally conducted. NanoInsights provides a choice of four feature selection methods: Recursive Feature Elimination with Cross-Validation (RFECV), Permutation Feature Importance (PFI), Differentially Expressed genes (DE), or the option to proceed without any feature selection.

For RFECV, various cross-validation (CV) strategies are available, such as Leave-One-Out Cross-Validation (LOOCV), 5-fold Cross-Validation (5-fold CV), and 10-fold Cross-Validation (10-fold CV). This method also allows the specification of the minimum number of features to be selected in RFECV. In the case of PFI, it is necessary to determine the minimum number of features to be included. The ‘DE’ method selects features based on differentially expressed genes. Opting for ‘no feature selection’ means all features will be used in the subsequent analysis. However, this approach is generally not recommended as it might lead to overfitting.

6.2.6 Selection of Classification Algorithm

Subsequently, NanoInsights provides a selection of five algorithms for the classification task, which includes Random Forest (RF), K-Nearest Neighbours (KNN), Gradient Boosting (GB), Extra Trees Classifier (ETC), and Logistic Regression (LR). To thoroughly evaluate the training efficacy, NanoInsights employs CV as

a test method. The training process is assessed based on the average metrics derived from these CV sets. There are three options for CV: 3-fold Cross-Validation (3-fold CV), 5-fold Cross-Validation (5-fold CV), and 10-fold Cross-Validation (10-fold CV). Crucially, the chosen model is also put through a rigorous evaluation using an independent test set, which acts as the conclusive assessment stage for the classification procedure.

6.2.7 Selection of Test Set for final classifier evaluation

Additionally, the selected trained model undergoes further evaluation using an independent test set, which is established at the beginning of the data analysis process. This step is crucial for providing a definitive evaluation of the model's classification performance. NanoInsights offers several options for configuring the test set in the classification modelling process.

The "Split" option divides the input data into two portions, with 80% used for training and 20% for testing. The "Runs" option allows for the selection of samples from a specific run or multiple runs (loaded in a single cartridge) to be used as the test set, while the remaining data forms the training set. Another choice is the "External Set", which enables the use of additional raw and clinical data as the test set.

There's also the 'Only Normalisation' option, which is designed for scenarios where only normalisation analysis is required, bypassing the classification analysis entirely.

It's important to highlight that NanoInsights treats the training and test sets independently. This means that normalisation is conducted separately for each set, and the test set is strictly used for testing purposes without any involvement in the training phase. This approach ensures the integrity and independence of the test set in evaluating the model's performance.

6.2.8 Classification Output

The classification procedure yields various insightful outputs, including a plot demonstrating the optimal number of features based on RFECV or PI, a table

of the selected features, and a comprehensive file detailing the filtering, feature selection, and training processes. Detailed data tables present the classification outcomes for both the training and test sets. A final table consolidates key metrics like Balanced Accuracy, F1-score, Precision, Recall, among others. The results are further elucidated through interactive and static visualisations, including a Class Probabilities plot, a Confusion Matrix (CM) Heatmap, and Receiver Operating Characteristic (ROC) plots for both training and testing phases, enhancing the interpretation and understanding of the classification outcomes.

6.2.9 Analysis of Gene Set Enrichment

Gene set enrichment analysis (GSEA) is carried out in two distinct stages. Initially, it focuses on genes that show differential expression. This step is followed by a repeated analysis on features identified through RFECV or PI feature selection methods. To conduct these analyses, the capabilities of gProfiler and g:GOST are employed [186]. These tools are crucial for functional enrichment analysis, a process key to revealing biological insights from gene lists.

6.3 Results

6.3.1 Overview of the NanoInsights Platform Workflow

The NanoInsights workflow is an intricate and comprehensive process, methodically designed to facilitate in-depth RNA analysis through a series of eight distinct stages 6.1:

- **Stage 1: Data Collection**

The workflow begins with the collection of the necessary .RCC files and the associated clinical table file. This stage is essential as it lays the foundation for the entire analysis, ensuring a robust and relevant dataset for processing.

- **Stage 2: Data Upload and Pipeline Initiation**

In this stage, the collected data is uploaded to the NanoInsights platform. Here, the analytical pipeline is initiated. Researchers are provided the option to either customise various parameters according to their specific research

needs or to proceed with default settings, striking a balance between customisation and user-friendliness.

- **Stage 3: Preprocessing, Quality Control, and Exploratory Analysis**

This stage involves preprocessing the data, conducting thorough quality control, and undertaking an in-depth exploratory analysis. These steps are key to assessing data quality, exploring relationships among samples, identifying potential outliers, and detecting any batch effects. This stage is vital for setting the groundwork for the data analysis, as it ensures the data integrity and lays out preliminary insights that guide the direction of the further analysis.

- **Stage 4: Selective Filtering**

Moving forward, the workflow incorporates selective filtering of genes or samples, based on the parameters set in the second stage. This process allows for a more concentrated analysis pertinent to the specific research and data quality objectives.

- **Stage 5: Normalisation and Differential Expression**

In this stage, the emphasis is on data normalisation. Users have the option to choose from a diverse range of normalisation algorithms, or they can rely on the platform's automated algorithm selection feature. Ensuring the data is accurately normalised during this phase is vital for the precision of the subsequent analysis. After the normalisation process, the data undergoes DE analysis to pinpoint genes that exhibit differential expression.

- **Stage 6: Machine Learning**

The sixth stage of the workflow focuses on implementing Machine Learning, with an emphasis on supervised learning and classification. In this stage, a chosen model undergoes training using a predefined dataset. Once trained, the model's accuracy and effectiveness are then assessed using a separate test dataset. The final step involves the model making predictions on new, unseen data, referred to as the validation set. In this phase, normalised data is used for both training and testing the selected classifier. The ultimate aim is to utilise the trained classifier for making definitive predictions on the validation set (unseen data), which is predefined in Stage 2. Machine learning introduces an advanced level of analytical sophistication to our process, essential for developing models that not only analyse the dataset effectively, but also deliver accurate predictions.

- **Stage 7: Integration for Gene Set Enrichment Analysis**

In this stage, the workflow utilises the results obtained from DE analysis and the ML processes to perform gene set enrichment analysis. This stage allows for a deeper understanding of the gene sets and their biological significance, providing valuable insights into the data.

- **Stage 8: Data Visualisation and Insight Extraction**

The final stage is dedicated to data visualisation and the extraction of insights. This phase equips users with interactive tools to delve into the findings of the analysis in a user-friendly and comprehensive manner. It also allows for the download of the fully processed results for further utilisation.

This detailed overview of the NanoInsights workflow showcases its methodical and thorough process, encompassing every aspect from the initial data intake to the final extraction of insightful conclusions, ensuring a comprehensive approach to RNA analysis.

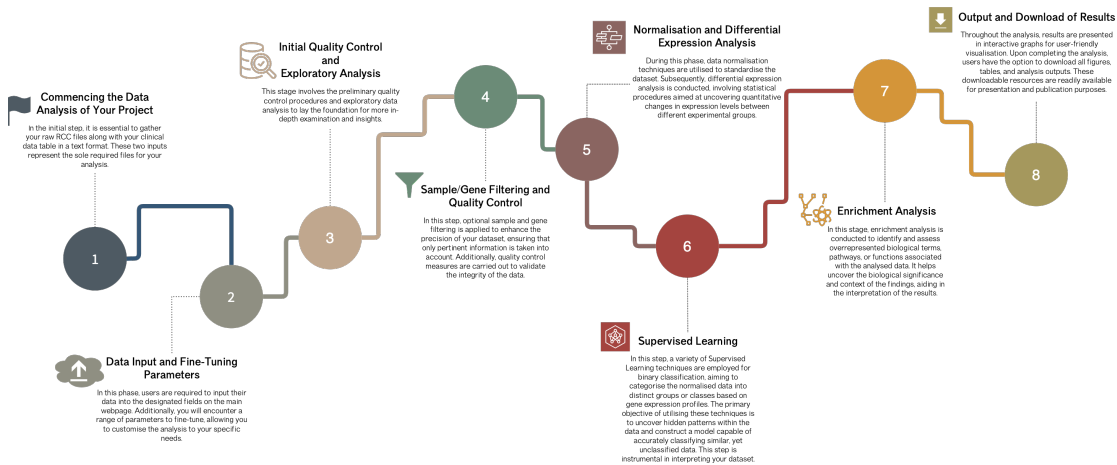
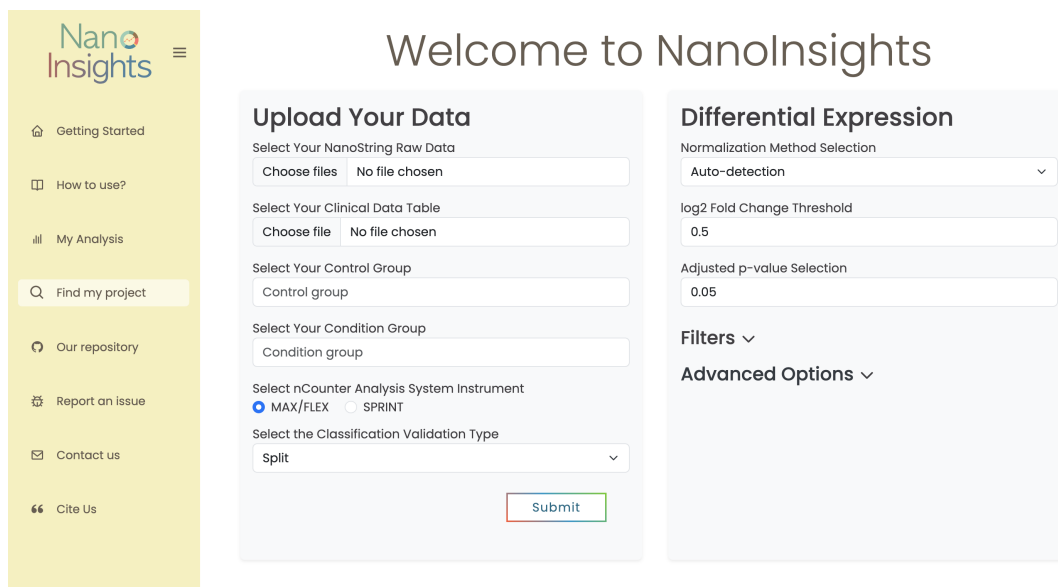


FIGURE 6.1: Detailed Flowchart of the NanoInsights Analytical Process. The figure presents an eight-stage progression through NanoInsights' analytical platform. The initial phase (Stage 1) involves the collection of .RCC files and the associated clinical data. In Stage 2, this data is uploaded for processing, and the analysis pipeline is initiated, with an option to customise or utilise default parameters. Preprocessing, quality control, and exploratory analysis define Stage 3, setting the groundwork for data analysis. Stage 4 applies selective gene or sample filters based on earlier parameter settings. Normalisation is the focus of Stage 5, where users can select from a suite of algorithms or rely on the system's automatic algorithm selection. Machine Learning takes centre stage in Stage 6, employing data to train and evaluate a selected classifier. Stage 7 integrates the outcomes of differential expression and Machine Learning for gene set enrichment analysis. The final phase, Stage 8, offers data visualisation and insight extraction, providing users with interactive tools to explore the analysis findings and facilitating the download of the fully processed results. This overview encapsulates NanoInsights' methodical process for thorough RNA analysis, spanning from data intake to the derivation of insightful conclusions.

6.3.2 User Interface Overview

The NanoInsights web platform is crafted to be exceptionally user-centric and straightforward, ensuring that even those new to bioinformatics analysis can navigate it with ease (Figure 6.2). Upon arriving at the "Getting Started" interface, users are warmly invited to commence their analytical journey by uploading their data. The interface is deliberately designed to offer users the autonomy to either fine-tune a multitude of analysis parameters to suit their individual project needs or to employ the preset default parameters for a more streamlined experience.



The screenshot displays the NanoInsights web service main page. On the left is a yellow sidebar with the NanoInsights logo and a menu containing: Getting Started, How to use?, My Analysis, Find my project (with a search icon), Our repository, Report an issue, Contact us, and Cite Us. The main content area is titled 'Welcome to NanoInsights' and is divided into two primary panels. The 'Upload Your Data' panel on the left includes sections for: 'Select Your NanoString Raw Data' (with a 'Choose files' button and 'No file chosen' text), 'Select Your Clinical Data Table' (with a 'Choose file' button and 'No file chosen' text), 'Select Your Control Group' (with a 'Control group' input field), 'Select Your Condition Group' (with a 'Condition group' input field), 'Select nCounter Analysis System Instrument' (with radio buttons for 'MAX/FLEX' (selected) and 'SPRINT'), and 'Select the Classification Validation Type' (with a dropdown menu set to 'Split'). A 'Submit' button is located at the bottom of this panel. The 'Differential Expression' panel on the right includes: 'Normalization Method Selection' (with a dropdown menu set to 'Auto-detection'), 'log2 Fold Change Threshold' (with an input field set to '0.5'), and 'Adjusted p-value Selection' (with an input field set to '0.05'). Below these are expandable sections for 'Filters' and 'Advanced Options'.

FIGURE 6.2: **Screenshot of the NanoInsights Platform Interface.** Screenshot of the NanoInsights web service main page interface, providing an overview of the user experience upon entering the website.

The user interface is equipped with interactive features that provide on-demand explanations for each parameter. For instance, a simple hover over a parameter title triggers a pop-up with detailed information, enhancing user understanding effortlessly. This helpful feature is exemplified in the interface concerning the "Select nCounter Analysis System instrument" option in Figure 6.2.

At the heart of the initial interface is the "Upload your data" panel. Here, users are prompted to provide the RCC files and associated clinical data table pertinent to their study. The platform supports a range of upload techniques, from traditional file selection dialogs to drag-and-drop interfaces, and accepts both individual and compressed file uploads in .zip or .tar.gz formats.

Following the data upload, users are asked to define the Control and Condition Group labels in sync with the clinical data file, select the nCounter system utilised for their study, and determine the Classification Validation Type. This latter step is critical as it delineates the methodology for the final evaluation of the classification process, and thus, how the test set will be segmented. This parameter is a compulsory field, and users are guided to refer to the Materials and Methods Section 6.2.7 for comprehensive information.

The platform's left-hand panel is home to further analytical parameters that cover DE, Filters, and Advanced Options pertaining to the Machine Learning component of the analysis. It is noteworthy that the user retains full control over the parameters, dictating the course of normalisation, filtering, and classification processes that the data will undergo. The flexibility extends to the point where a user can elect to bypass the machine learning and classification stages entirely if their analysis calls for it.

For detailed instructions on navigating the platform, the "How to use?" tab (Figure 11) presents an exhaustive analysis of each processing step, encompassing each parameter and its function, the requirements for input files, and a granular breakdown of the output visualisations. This section is replete with hyperlinks to external documentation for an expanded explanation of the processes in use. To further aid users, NanoInsights has curated a video tutorial that visually demonstrates navigating the website.

Additionally, for users who do not yet have their own data to analyse or who simply wish to practice using the platform, NanoInsights offers a test dataset consisting of 144 Colorectal Cancer samples, originally studied by Low, Blöcker, McPherson, *et al.* in 2017 [322]. This dataset serves as a practical tool for users to familiarise themselves with the platform's analysis process and to explore its comprehensive functionalities while they prepare for their personal data analyses.

After the data analysis is finalised, users are directed to the "My Analysis" tab, where they can interact with the visualisations and download the comprehensive analysis of their data.

Each user's analysis is uniquely encrypted with a fourteen-character project identifier comprising both numbers and letters. Projects executed on NanoInsights are securely archived on our servers, and users can access their results anytime using their unique alphanumeric code through the "Find my project" tab.

The platform extends its resources through the "Our repository" tab, which leads to the GitHub repository containing all the source code of the website, including the test set. Should users encounter any technical issues or wish to offer feedback, they can utilise the "Report an issue" function or engage with the NanoInsights team via the "Contact us" tab. For those who leverage our web service in their research and wish to acknowledge it, the "Cite us" tab provides all the necessary citation information to facilitate proper referencing in scientific discourse.

6.3.3 Case Study

In order to illustrate the practicality and effectiveness of NanoInsights, we conducted a re-analysis of a dataset originally published by Park, Yu, Mustafa, *et al.* in the year 2020, featured in the journal ‘Cancers’. This particular study included a cohort of 60 patients diagnosed with Locally Advanced Rectal Cancer (LARC). These patients were classified into two distinct response categories based on their reaction to Preoperative Chemoradiotherapy: those who were good responders and those who were non-responders. A distinctive feature of this dataset is the inclusion of an independent test set, which contains an additional 96 samples. Detailed information regarding the clinical data from this study is meticulously outlined in Table 6.2.

TABLE 6.2: Detailed Clinical Information of Patients in Training and Test Cohorts from the Study.

Variable	Training Cohort	Test Cohort
General Info		
Males	27 (45.0%)	53 (55.2%)
Females	33 (55.0%)	43 (44.8%)
Total Samples	60	96
Clinical T-stage		
TI	0 (0.0%)	0 (0.0%)
TII	4 (6.7%)	9 (9.4%)
TIII	53 (88.3%)	78 (81.3%)
TIV	3 (5.0%)	4 (9.0%)
Clinical N-stage		
N0	2 (3.3%)	10 (10.4%)
N1	24 (40.0%)	40 (41.7%)
N2	34 (56.7%)	46 (47.9%)
Clinical M-stage		
M0	58 (96.7%)	94 (97.9%)
M1	2 (3.3%)	2 (2.1%)
Assigned Class		
Good Responders	27 (45.0%)	62 (64.6%)
Non Responders	33 (55.0%)	34 (35.4%)

In our re-analysis using NanoInsights, we imported the raw RCC files, provided in a zipped file format, as well as the accompanying clinical data, which was uploaded in a text file format. The clinical file included essential details such as the Filename and the assigned Condition category for each sample, distinguishing between Good-Response and Non-Response groups. For the purpose of classification validation, we selected the ‘Run’ option. This involved choosing eight specific

runs ('run2845', 'hsm-20160622', 'run2854', 'cancerpathwayHSM1', 'cancerpathwayHSM2', '20161018AsanKMJ1', '20161018AsanKMJ2', '20161019AsanKMJ3'), which pertained to the 96 samples in the test dataset. The remaining samples, sourced from the remaining runs, were utilised to form the training set comprising 60 samples.

Regarding the analytical parameters within NanoInsights, we opted to use the default settings for several aspects of the analysis. This included selecting 'Auto-Detection' for the normalisation process of the data and choosing RFECV for the feature selection phase. To enhance the depth and breadth of our classification analysis, we expanded our choice of classifiers. Alongside the default Random Forest classifier, we also incorporated two additional classifiers: Extra Trees and Logistic Regression.

The re-evaluation of our dataset commenced with a primary QC check utilising NanoString's established metrics, as visualised in Figure 12A-D. This initial QC revealed that all samples conformed to the expected standards, with no samples flagged as potential outliers. However, during the exploratory data analysis, certain samples, specifically those from CartridgeID 20171020CBStest05 (depicted in brown), were noted to have expression levels lower than the average (the average denoted by the red horizontal line in Figure 6.3A). Although the IQR analysis, set with a threshold of 2, did not classify any samples as outliers (Figure 6.3B), a subsequent PCA revealed a distinct grouping of samples in the upper left quadrant of Figure 13, suggesting they could be outliers, which interestingly all originated from the same batch.

These four samples, identified as T136_01, T155_05, T164_07, and T177_11, were subsequently excluded in the sample-based filtering step that considers NCs. For an in-depth explanation of this process, please refer to Section 6.2.3 in the Materials and Methods. Furthermore, the gene-based filter, also based on NC, eliminated 105 genes that exhibited very low expression levels. The process of normalisation was addressed next, with the auto-detection feature suggesting Loess Normalisation as the optimal method for this dataset, indicated by a mRLE value of 0.01 (Figure 6.3C). This was corroborated by the RLE plot, which showed all medians closely aligned around the zero mark.

The differential expression analysis then identified five genes as differentially expressed, with CD40 and STAT1 being down-regulated, while MAPK8IP1, CACNA1,

and ID1 were found to be up-regulated, as detailed in Figure 6.3D. These findings were determined using an adjusted p-value threshold of 0.05 and an absolute log₂ Fold Change threshold of 0.5.

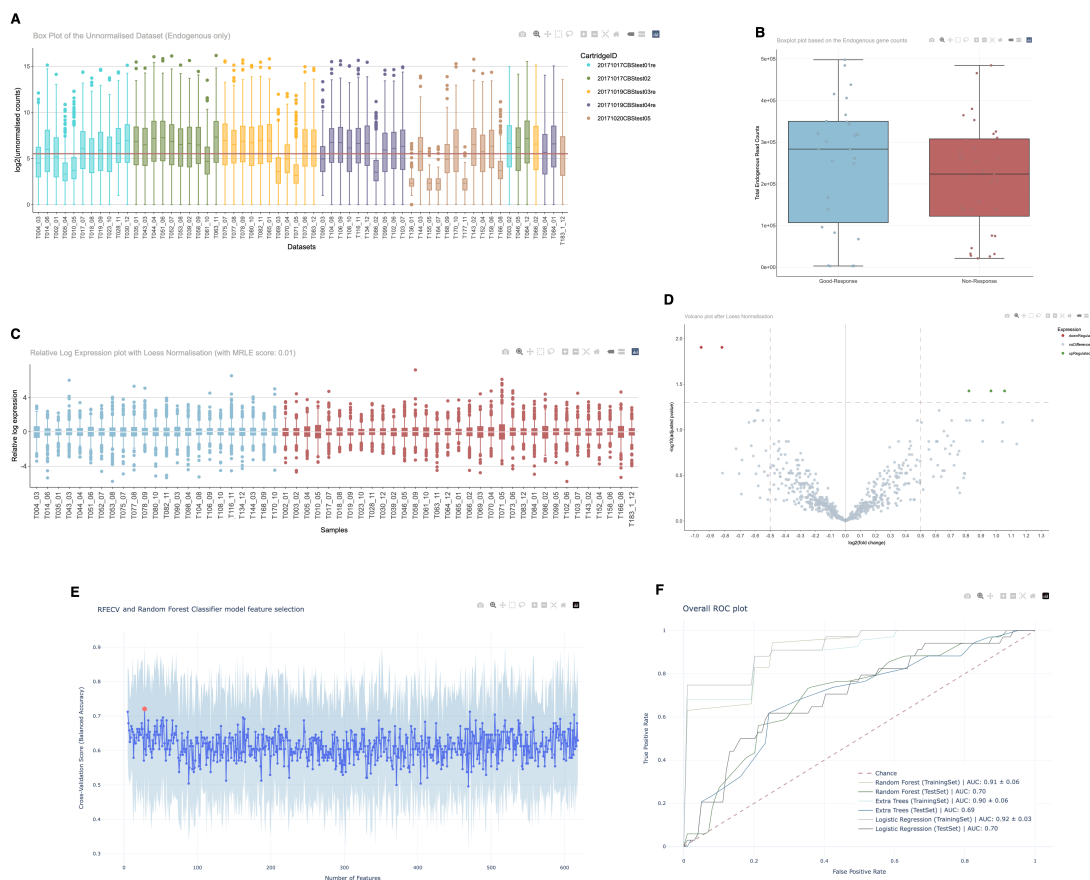


FIGURE 6.3: Visual Analytics of RNA Expression Data in Response to Chemoradiotherapy. This compilation of visualisations showcases the data analysis conducted with NanoInsights for patients with LARC, segmented by their response to Preoperative Chemoradiotherapy. **(A)** A box plot visualises the unnormalised data, delineating expression levels across the samples, with colours corresponding to the CartridgeIDs denoting different Runs. **(B)** A box plot derived from IQR analysis contrasts the expression profiles between good responders and non-responders. **(C)** Post-Loess normalisation, this box plot depicts RLE levels, with good responders in blue and non-responders in red. **(D)** A volcano plot captures DE genes post-Loess normalisation; red points indicate significantly down-regulated genes, and green points represent up-regulated ones, based on an adjusted p-value of 0.05 and an absolute log₂ fold change of 0.5. **(E)** The feature selection process is illustrated using RFECV with a Random Forest Classifier; it plots cross-validated scores against the number of features, using ‘balanced accuracy’ as the scoring metric. The red point marks the optimal number of selected features. **(F)** A ROC plot provides a comparative overview of different classifiers’ abilities to differentiate between good and non-responders, with varying colours representing the different classifiers and datasets, such as the training or test sets.

As we advanced to machine learning analysis, RFECV was employed, utilising LOOCV to identify a subset of 28 genes that offered a balanced accuracy of 0.72, as shown in Figure 6.3E and detailed in Table 13.

The ROC curve analysis, depicted in Figure 6.3F, compared the performance between Random Forest and Logistic Regression classifiers. It is noteworthy that although all classifiers demonstrated exemplary performance in the training set, with Area Under the Curve (AUC) values exceeding 0.90, they exhibited reduced effectiveness in the test set, with AUC values at or below 0.70. A closer examination of the classifiers revealed Random Forest as the more capable model, evidencing a higher True Positive Rate (TPR) at 0.74 compared to 0.54 for Logistic Regression. Furthermore, Random Forest achieved higher F1 and F2 scores of 0.62 and 0.68, respectively, outperforming Logistic Regression, which scored 0.58 on both accounts. Table 14 furnishes a comprehensive list of the metrics evaluated.

6.4 Discussion

NanoInsights emerges as a sophisticated web service designed to streamline the intricate process of NanoString nCounter data. Its interface and workflow are meticulously crafted to make advanced bioinformatics tools accessible to both seasoned researchers and newcomers to the field. The service integrates a variety of analytical methods and machine learning techniques into a single, cohesive platform, ensuring that users can navigate from data upload to in-depth analysis with relative ease.

The analytical journey begins with a thorough QC assessment, employing standard NanoString metrics to establish a solid foundation for the accuracy and reliability of further analysis. The platform enhances user engagement through its exploratory data analysis tools, featuring various methods like PCA, IQR, and MDS. These tools are instrumental in identifying potential batch effects or outliers, common issues in high-throughput data analyses, which could otherwise compromise the validity of the results. Not only do these features underpin quality assurance, but they also shed light on the underlying structure of the data, revealing technical variations that might influence the final outcomes.

As the analysis progresses, NanoInsights guides users through the critical step of data normalisation, offering an array of eight distinct normalisation methods,

ranging from standard NanoString normalisation to more sophisticated options like RUVSeq. The platform's auto-detection functionality stands out, identifying the most suitable normalisation method for each dataset. This feature ensures the highest standard of data processing, simultaneously alleviating the need for in-depth analytical expertise.

After normalisation and differential expression analysis, NanoInsights introduces its machine learning module. This module employs standard algorithms to delve deeper into the NanoString data, uncovering intricate patterns and relationships that might elude conventional statistical approaches. This step could provide more profound biological understanding, thereby refining the accuracy of gene expression profiling. The platform offers various preprocessing options, including feature selection methods such as RFECV, Permutation Feature Importance, and utilising the differentially expressed genes, enabling users to refine their models and potentially boost their predictive accuracy.

The diversity of machine learning algorithms available on NanoInsights, including Random Forest, Extra Trees, Gradient Boosting, K-Nearest Neighbours, and Logistic Regression, allows for the development and rigorous evaluation of robust models. The platform's detailed approach to cross-validation and model assessment is indicative of its commitment to providing reliable results. Importantly, models are evaluated using unseen data, ensuring their robustness and reproducibility in real-world scenarios.

In addition to its robust analytical capabilities, NanoInsights greatly enhances the user experience with its array of interactive visualisations. These visual tools play a pivotal role in making the journey through data analysis both intuitive and user-friendly. This interactive element is particularly beneficial in simplifying the understanding of intricate patterns and trends within the data, making it easier for researchers to draw meaningful conclusions.

Furthermore, NanoInsights recognises the importance of sharing and publishing research findings. To facilitate this, the platform enables users to download all analyses and results in high-resolution formats. This feature is particularly advantageous for researchers looking to include their findings in publications or presentations. The ability to export ready-to-use, publication-quality figures ensures that the output from NanoInsights not only meets, but exceeds the standards expected in scientific communication. This focus on both the analytical depth

and the practical utility of the output underscores NanoInsights' commitment to supporting researchers throughout their investigative journey, from initial data exploration to the final stages of dissemination.

While NanoInsights offers a robust platform for NanoString nCounter data analysis, it does have its limitations. One notable constraint is its processing capacity, particularly with larger datasets containing several hundreds of samples, which may result in longer times to yield results. Steps like feature selection and model training become more computationally demanding with increased sample sizes. Additionally, while the platform currently offers three feature selection methods, there might be a need for more advanced techniques tailored to the specific nature of this data. Fine-tuning of hyperparameters could also be an area for future development. Another challenge arises when applying machine learning to smaller datasets. In such cases, there is a risk of overfitting, where the model becomes overly adapted to the training data, failing to generalise well to new data. This issue, coupled with the potential unrepresentativeness of small datasets, can lead to models that are less effective or biased.

Looking ahead, the future of NanoInsights is bright with opportunities for expansion and enhancement. Future updates could focus on increasing computational efficiency for large datasets, integrating more sophisticated feature selection methods, and refining machine learning algorithms to better handle datasets of varying sizes. These enhancements will further cement NanoInsights as a versatile and powerful tool in the realm of RNA analysis, adapting to the evolving needs of researchers in this dynamic field.

6.5 Conclusion

NanoInsights stands as a groundbreaking web service, skilfully bridging the gap between complex NanoString nCounter data analysis and user accessibility. Its integration of comprehensive quality control, diverse normalisation methods, and advanced machine learning techniques within an intuitive platform revolutionises the way researchers approach RNA data analysis. While acknowledging its current limitations, particularly in handling large datasets and potential overfitting with smaller ones, NanoInsights is poised for future advancements. Anticipated enhancements aimed at improving computational efficiency and incorporating more

sophisticated analysis tools are expected to solidify its position as an indispensable asset in RNA research. As NanoInsights continues to evolve, it promises to not only meet, but exceed the ever-growing demands and intricacies of bioinformatics, charting a path towards more insightful and impactful scientific discoveries.

Chapter 7

Brief Discussion and Future Perspectives

Our progress in cancer diagnostics has been significant, driven by our development and application of various methodologies in liquid and tissue biopsy-based transcriptomics and sequencing. A key advancement is the Ensemble Learning for Liquid Biopsy Analysis (ELLBA), which has significantly improved the precision and reliability of cancer diagnostics. ELLBA uniquely combines a diverse array of biofeatures using ensemble learning techniques, providing a detailed view of cancer's molecular complexity. This methodology has been extensively tested and validated across a wide range of datasets, demonstrating its effectiveness and robustness.

Looking to the horizon, there are many ways to further amplify the capabilities of ELLBA. Laboratory advancements beckon, with potential optimisations in the protocols for RNA library preparations derived from liquid biopsies and the adoption of paired-end deep sequencing technologies. These refinements are poised to enhance our detection capabilities, particularly for identifying intricate molecular features such as gene fusions, RNA editing, and SNVs, thereby unveiling even the rarest mutations. This evolution in our laboratory practices, especially the strategic shift from polyA enrichment to ribosomal RNA depletion, is expected to widen our biomarker discovery net, incorporating novel entities like circular RNAs into our analytical purview.

On the computational front, ELLBA's framework is ripe for enhancement. This includes delving into more sophisticated feature selection methods and classifiers,

coupled with a comprehensive approach to hyperparameter optimisation. A focused effort on addressing the challenges presented by unbalanced datasets will be critical in significantly lifting ELLBA's predictive performance. These computational advancements are designed to not only refine our diagnostic tools, but also to deepen our insights, thereby enabling more nuanced and accurate predictions in the ongoing battle against cancer. Additionally, incorporating elements of Trustworthy AI—focusing on transparency, fairness, and security—through methods like Federated Learning could further enhance ELLBA's applicability in clinical settings by ensuring robust, unbiased, and secure data analysis.

Our engagement with Third Generation Sequencing, especially through the lens of Oxford Nanopore Technologies' direct RNA sequencing, has been transformative, bridging critical gaps in our understanding of the cancer transcriptome. The development of the 'DRseeker' bioinformatics pipeline is a testament to our unwavering commitment to exploring and harnessing new frontiers in cancer research. By providing an intricate analysis of the lung cancer transcriptome, we have unearthed pivotal shifts in transcript expression and post-transcriptional modifications, significantly advancing our grasp of the molecular underpinnings of cancer.

The path forward for DRseeker involves the incorporation of a broader and more diversified dataset, which will not only refine the pipeline, but also deepen our insights into the NSCLC transcriptome. The advent of newer sequencing chemistries, like the promising R10, and the potential utilisation of higher-output sequencers such as PromethION, herald a new era of transcriptomic analysis characterised by unprecedented depth and precision. Moreover, the proposed upgrades to DRseeker, including the integration of new analytical features and the enhancement of data visualisation techniques, aim to streamline the exploration of the vast datasets we generate. By melding machine learning with TALON, the core analytical engine of DRseeker, we anticipate a significant leap in the accuracy of isoform classification.

Moving forward, our 'NanoInsights' platform underscores our commitment to democratising access to advanced bioinformatics tools. By simplifying the analysis of NanoString nCounter data, we aim to empower researchers to unravel complex gene expression patterns, propelling forward the frontiers of cancer research and diagnostics. Future iterations of NanoInsights will focus on incorporating more normalisation methods tailored for nCounter data, enhancing interactive visualisation features for on-the-fly analysis, and broadening the platform's capability

to encompass a wider array of NanoString data types, including spatial genomics and proteomics.

As we forge ahead, the seamless integration of these avant-garde technologies and methodologies is set to revolutionise the landscape of personalised medicine, prognostics, and targeted therapy. The insights derived from our research not only deepen our comprehension of cancer biology, but also pave the way for novel diagnostic and therapeutic strategies. Our continued endeavours will focus on harnessing these innovations to refine diagnostic accuracy, enhance patient outcomes, and inject new hope into the global fight against cancer, thereby cementing our role in shaping the future of cancer diagnostics.

Chapter 8

Conclusions

Throughout this doctoral thesis, several bioinformatics applications have been conceived and transformed into user-friendly tools. The primary objective was to facilitate the analysis of transcriptomics data, making it more accessible, dependable, and reproducible. The resulting tools encompass: 1. ELLBA: This methodology and pipeline were designed for the analysis of lbRNA-Seq data, providing a novel approach to liquid biopsy analysis; 2. DRseeker: A dedicated pipeline specialising in the analysis of Nanopore DRS data. It peels back multiple layers of information embedded in RNA, allowing for a comprehensive examination; 3. NanoInsights: An intuitive web-based application that combines bioinformatics and machine learning for in-depth analysis of NanoString nCounter data. A detailed summary of these contributions includes:

1. The ELLBA methodology was developed to bridge the gap in liquid biopsy-based transcriptomics by utilising key discriminative features. This approach leverages gene and isoform expression, FoCT, gene fusion, RNA editing, and SNVs to capture essential molecular characteristics for cancer diagnosis.
2. The intra-sample CPM normalisation method proved effective, competing well with more complex approaches. Its simplicity adds convenience for clinical settings, allowing individual samples to be tested and significantly reducing run time—an important factor in clinical environments.
3. The application of Ensemble Classification to integrate diverse biofeatures significantly improved predictive accuracy, especially in external validation sets.

4. The "DRseeker" software was developed to analyse data derived from the newly introduced DRS technology. This software offers comprehensive functionalities enabling in-depth insights into the genetic underpinnings of disease.
5. Using DRseeker to analyse NSCLC adenocarcinoma tissues and adjacent non-cancerous tissues enabled the identification of significant transcriptomic alterations. Our focused analysis on genes like *AGER* and *MEST* revealed differential expression and isoform switching, helping to identify potential new biomarkers and therapeutic targets.
6. Additionally, DRseeker facilitated the exploration of differential transcript usage, polyA-tail analysis, and the examination of post-transcriptional modifications. Significant findings included variations in polyadenylation, specifically polyA tail lengths, and distinct post-transcriptional modification patterns that were unique to lung cancer.
7. We developed 'NanoInsights', a user-friendly web application to improve the analysis of data from NanoString's nCounter system. This tool tackles the challenges of analysing nCounter data, making clinical transcriptomics more accessible.
8. NanoInsights combines bioinformatics analysis with standard machine learning classification methods. This integration allows users to leverage advanced analytical techniques seamlessly within the platform.
9. The application offers multiple features, including eight unique normalisation techniques with an auto-detection option that selects the best method based on the input data. It also supports dynamic visualisations and produces high-resolution figures suitable for publication.

Chapter 9

Conclusiones

A lo largo de esta tesis doctoral, se han concebido y transformado varias aplicaciones bioinformáticas en herramientas amigables para el usuario. El objetivo principal fue facilitar el análisis de datos transcriptómicos, haciéndolos más accesibles, fiables y reproducibles. Las herramientas resultantes incluyen: 1. ELLBA: Esta metodología y su pipeline se diseñaron para el análisis de datos de lbRNA-Seq, ofreciendo un enfoque novedoso para el análisis de biopsias líquidas; 2. DRseeker: Un pipeline dedicado a la especialización en el análisis de datos de DRS de Nanopore. Revela múltiples capas de información incrustadas en el RNA, permitiendo un examen exhaustivo; 3. NanoInsights: Una aplicación web intuitiva que combina bioinformática y aprendizaje automático para un análisis profundo de datos de NanoString nCounter. Un resumen detallado de estas contribuciones incluye:

1. La metodología ELLBA se desarrolló para cerrar la brecha en la transcriptómica basada en biopsias líquidas, utilizando características discriminativas clave. Este enfoque aprovecha la expresión de genes e isoformas, FoCT, fusión de genes, edición de RNA y SNVs para capturar características moleculares esenciales para el diagnóstico del cáncer.
2. El método de normalización intra-muestra CPM demostró ser efectivo, compitiendo bien con enfoques más complejos. Su simplicidad resulta conveniente en entornos clínicos, permitiendo que se analicen muestras individuales y reduciendo significativamente el tiempo de ejecución, un factor importante en los entornos clínicos.

3. La aplicación de la Clasificación Combinada (ensemble) para integrar diversas bio-características mejoró significativamente la precisión predictiva, especialmente en conjuntos de validación externos.
4. El software 'DRseeker' fue desarrollado para analizar datos derivados de la nueva tecnología DRS. Este software ofrece funcionalidades completas que permiten obtener percepciones profundas sobre las bases genéticas de las enfermedades.
5. Utilizando DRseeker para analizar tejidos de adenocarcinoma de NSCLC y tejidos no cancerosos adyacentes, se logró identificar alteraciones transcripcionales significativas. Nuestro análisis enfocado en genes como AGER y MEST reveló expresión diferencial y cambio de isoforma, ayudando a identificar posibles nuevos biomarcadores y objetivos terapéuticos.
6. Además, DRseeker facilitó la exploración del uso diferencial de transcritos, análisis de colas de poliA y la examinación de modificaciones postranscripcionales. Los hallazgos importantes incluyeron variaciones en la poliadenilación, específicamente longitudes de colas de poliA, y patrones de modificaciones postranscripcionales distintos que eran únicos para el cáncer de pulmón.
7. Desarrollamos 'NanoInsights', una aplicación web amigable para mejorar el análisis de datos del sistema nCounter de NanoString. Esta herramienta aborda los desafíos de analizar datos de nCounter, haciendo la transcripción clínica más accesible.
8. NanoInsights combina análisis bioinformáticos con métodos estándar de clasificación de aprendizaje automático. Esta integración permite a los usuarios aprovechar técnicas analíticas avanzadas de manera fluida dentro de la plataforma.
9. La aplicación ofrece múltiples características, incluyendo ocho técnicas de normalización únicas con una opción de autodetección que selecciona el mejor método basado en los datos de entrada. También soporta visualizaciones dinámicas y produce figuras de alta resolución adecuadas para publicación.

Bibliography

- [1] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, “Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries,” en, *CA Cancer J. Clin.*, vol. 71, no. 3, pp. 209–249, May 2021.
- [2] J. Ferlay, M. Colombet, I. Soerjomataram, D. M. Parkin, M. Piñeros, A. Znaor, and F. Bray, “Cancer statistics for the year 2020: An overview,” en, *Int. J. Cancer*, vol. 149, no. 4, pp. 778–789, Apr. 2021.
- [3] World Health Organisation, *Healthy Topics/Cancer*, Dec. 2015. [Online]. Available: https://www.who.int/health-topics/cancer#tab=tab_1.
- [4] S. Elmore, “Apoptosis: A review of programmed cell death,” en, *Toxicol. Pathol.*, vol. 35, no. 4, pp. 495–516, Jun. 2007.
- [5] G. Cooper, *The cell: A molecular approach*, 2nd ed. Sunderland, MA: Sinauer Associates, Aug. 2000.
- [6] G. N. Wogan, S. S. Hecht, J. S. Felton, A. H. Conney, and L. A. Loeb, “Environmental and chemical carcinogenesis,” en, *Semin. Cancer Biol.*, vol. 14, no. 6, pp. 473–486, Dec. 2004.
- [7] B. K. Prasanth, S. Alkhowaiter, G. Sawarkar, B. D. Dharshini, and A. R. Baskaran, “Unlocking early cancer detection: Exploring biomarkers, circulating DNA, and innovative technological approaches,” en, *Cureus*, Dec. 2023.
- [8] J. Hausser, P. Szekely, N. Bar, A. Zimmer, H. Sheftel, C. Caldas, and U. Alon, “Tumor diversity and the trade-off between universal cancer tasks,” en, *Nat. Commun.*, vol. 10, no. 1, p. 5423, Nov. 2019.
- [9] C. E. Meacham and S. J. Morrison, “Tumour heterogeneity and cancer cell plasticity,” en, *Nature*, vol. 501, no. 7467, pp. 328–337, Sep. 2013.

- [10] R. Fisher, L. Pusztai, and C. Swanton, “Cancer heterogeneity: Implications for targeted therapeutics,” en, *Br. J. Cancer*, vol. 108, no. 3, pp. 479–485, Feb. 2013.
- [11] J. Brierley, B. O’Sullivan, H. Asamura, D. Byrd, S. H. Huang, A. Lee, M. Piñeros, M. Mason, F. Y. Moraes, W. Rösler, B. Rous, J. Torode, J. H. van Krieken, and M. Gospodarowicz, “Global consultation on cancer staging: Promoting consistent understanding and use,” en, *Nat. Rev. Clin. Oncol.*, vol. 16, no. 12, pp. 763–771, Dec. 2019.
- [12] J. Almagro, H. A. Messal, A. Elosegui-Artola, J. van Rheenen, and A. Behrens, “Tissue architecture in tumor initiation and progression,” en, *Trends Cancer*, vol. 8, no. 6, pp. 494–505, Jun. 2022.
- [13] R. Wei, S. Liu, S. Zhang, L. Min, and S. Zhu, “Cellular and extracellular components in tumor microenvironment and their application in early diagnosis of cancers,” en, *Anal. Cell. Pathol. (Amst.)*, vol. 2020, p. 6 283 796, Jan. 2020.
- [14] U. Ghoshdastider, N. Rohatgi, M. Mojtabavi Naeini, P. Baruah, E. Revkov, Y. A. Guo, S. Rizzetto, A. M. L. Wong, S. Solai, T. T. Nguyen, J. P. S. Yeong, J. Iqbal, P. H. Tan, B. Chowbay, R. Dasgupta, and A. J. Skanderup, “Pan-cancer analysis of ligand-receptor cross-talk in the tumor microenvironment,” en, *Cancer Res.*, vol. 81, no. 7, pp. 1802–1812, Apr. 2021.
- [15] C. E. Weber and P. C. Kuo, “The tumor microenvironment,” en, *Surg. Oncol.*, vol. 21, no. 3, pp. 172–177, Sep. 2012.
- [16] F. Spill, D. S. Reynolds, R. D. Kamm, and M. H. Zaman, “Impact of the physical microenvironment on tumor progression and metastasis,” en, *Curr. Opin. Biotechnol.*, vol. 40, pp. 41–48, Aug. 2016.
- [17] C. Belli, G. Antonarelli, M. Repetto, L. Boscolo Bielo, E. Crimini, and G. Curigliano, “Targeting cellular components of the tumor microenvironment in solid malignancies,” en, *Cancers (Basel)*, vol. 14, no. 17, p. 4278, Sep. 2022.
- [18] M. W. Pickup, J. K. Mouw, and V. M. Weaver, “The extracellular matrix modulates the hallmarks of cancer,” en, *EMBO Rep.*, vol. 15, no. 12, pp. 1243–1253, Dec. 2014.

- [19] F. Galli, J. V. Aguilera, B. Palermo, S. N. Markovic, P. Nisticò, and A. Signore, “Relevance of immune cell and tumor microenvironment imaging in the new era of immunotherapy,” en, *J. Exp. Clin. Cancer Res.*, vol. 39, no. 1, p. 89, May 2020.
- [20] K. Sobierajska, W. M. Ciszewski, I. Sacewicz-Hofman, and J. Niewiarowska, “Endothelial cells in the tumor microenvironment,” in *Advances in Experimental Medicine and Biology*, ser. Advances in experimental medicine and biology, Cham: Springer International Publishing, 2020, pp. 71–86.
- [21] A. Pender and S. Popat, “Understanding lung cancer molecular subtypes,” en, *Clin. Pract. (Lond.)*, vol. 11, no. 4, pp. 441–453, Jul. 2014.
- [22] American Cancer Society, *What Is Lung Cancer?* Jan. 2023. [Online]. Available: <https://www.cancer.org/cancer/types/lung-cancer/about/what-is.html> (visited on 11/20/2023).
- [23] —, *What Are the Risk Factors for Lung Cancer?* Jul. 2023. [Online]. Available: https://www.cdc.gov/cancer/lung/basic_info/risk_factors.htm (visited on 12/18/2023).
- [24] National Health Service (NHS) UK, *Lung cancer - Diagnosis*, Dec. 2023. [Online]. Available: <https://www.nhs.uk/conditions/lung-cancer/diagnosis/> (visited on 12/18/2023).
- [25] S. M. Gadgeel, S. S. Ramalingam, and G. P. Kalemkerian, “Treatment of lung cancer,” en, *Radiol. Clin. North Am.*, vol. 50, no. 5, pp. 961–974, Sep. 2012.
- [26] S. Hammerschmidt and H. Wirtz, “Lung cancer: Current diagnosis and treatment,” en, *Dtsch. Arztebl. Int.*, vol. 106, no. 49, 809–18, quiz 819–20, Dec. 2009.
- [27] H. Lemjabbar-Alaoui, O. U. Hassan, Y.-W. Yang, and P. Buchanan, “Lung cancer: Biology and treatment options,” en, *Biochim. Biophys. Acta*, vol. 1856, no. 2, pp. 189–210, Dec. 2015.

- [28] Cleveland Clinic, *Lung Cancer: Types, Stages, Symptoms, Diagnosis & Treatment*, Oct. 2022. [Online]. Available: <https://my.clevelandclinic.org/health/diseases/4375-lung-cancer> (visited on 12/18/2023).
- [29] American Cancer Society, *Signs and Symptoms of Lung Cancer*, Jan. 2019. [Online]. Available: <https://www.cancer.org/cancer/types/lung-cancer/detection-diagnosis-staging/signs-symptoms.html> (visited on 12/18/2023).
- [30] F. C. Detterbeck and C. J. Gibson, "Turning gray: The natural history of lung cancer over time," en, *J. Thorac. Oncol.*, vol. 3, no. 7, pp. 781–792, Jul. 2008.
- [31] G. J. Criner, A. Agusti, H. Borghaei, J. Friedberg, F. J. Martinez, C. Miyamoto, C. F. Vogelmeier, and B. R. Celli, "Chronic obstructive pulmonary disease and lung cancer: A review for clinicians," en, *Chronic Obstr. Pulm. Dis.*, vol. 9, no. 3, pp. 454–476, Jul. 2022.
- [32] A. C. Marin, A. Prasad, V. Patel, C. Lwoodskey, S. Hechter, A. Imtiaz, P. Patel, V. Shah, J. Appiah, and P. Cheriyaath, "Pulmonary adenocarcinoma mimicking pneumonia in a young adult," en, *Cureus*, vol. 15, no. 2, e35267, Feb. 2023.
- [33] A. Molassiotis, M. Lowe, J. Ellis, R. Wagland, C. Bailey, M. Lloyd-Williams, C. Tishelman, and J. Smith, "The experience of cough in patients diagnosed with lung cancer," en, *Support. Care Cancer*, vol. 19, no. 12, pp. 1997–2004, Dec. 2011.
- [34] M. O'Driscoll, J. Corner, and C. Bailey, "The experience of breathlessness in lung cancer," en, *Eur. J. Cancer Care (Engl.)*, vol. 8, no. 1, pp. 37–43, Mar. 1999.
- [35] L. Hyde and C. I. Hyde, "Clinical manifestations of lung cancer," en, *Chest*, vol. 65, no. 3, pp. 299–306, Mar. 1974.
- [36] S. Sone, F. Li, Z. G. Yang, T. Honda, Y. Maruyama, S. Takashima, M. Hasegawa, S. Kawakami, K. Kubo, M. Haniuda, and T. Yamanda, "Results of three-year mass screening programme for lung cancer using mobile low-dose spiral computed tomography scanner," en, *Br. J. Cancer*, vol. 84, no. 1, pp. 25–32, Jan. 2001.

- [37] V. Artinian and P. A. Kvale, "Update in screening of lung cancer," en, *Respirology*, vol. 10, no. 5, pp. 558–566, Nov. 2005.
- [38] M. E. Mayerhoefer, H. Prosch, L. Beer, D. Tamandl, T. Beyer, C. Hoeller, D. Berzaczy, M. Raderer, M. Preusser, M. Hochmair, B. Kiesewetter, C. Scheuba, A. Ba-Ssalamah, G. Karanikas, J. Kesselbacher, G. Prager, K. Dieckmann, S. Polterauer, M. Weber, I. Rausch, B. Brauner, H. Eidherr, W. Wadsak, and A. R. Haug, "PET/MRI versus PET/CT in oncology: A prospective single-center study of 330 examinations focusing on implications for patient management and cost considerations," en, *Eur. J. Nucl. Med. Mol. Imaging*, vol. 47, no. 1, pp. 51–60, Jan. 2020.
- [39] S. Purohit, N. Dutt, and L. K. Saini, "Transbronchial lung biopsy in diffuse parenchymal lung disease - question still remains whether to go for surgical lung biopsy or not?" en, *Lung India*, vol. 33, no. 1, pp. 117–118, Jan. 2016.
- [40] R. S. Wiener, D. C. Wiener, and M. K. Gould, "Risks of transthoracic needle biopsy: How high?" en, *Clin. Pulm. Med.*, vol. 20, no. 1, pp. 29–35, Jan. 2013.
- [41] Z. Gandhi, P. Gurram, B. Amgai, S. P. Lekkala, A. Lokhandwala, S. Manne, A. Mohammed, H. Koshiya, N. Dewaswala, R. Desai, H. Bhopalwala, S. Ganti, and S. Surani, "Artificial intelligence and lung cancer: Impact on improving patient outcomes," en, *Cancers (Basel)*, vol. 15, no. 21, Oct. 2023.
- [42] S. Chandrika and L. Yarmus, "Recent developments in advanced diagnostic bronchoscopy," en, *Eur. Respir. Rev.*, vol. 29, no. 157, p. 190 184, Sep. 2020.
- [43] A. Buder, C. Tomuta, and M. Filipits, "The potential of liquid biopsies," en, *Curr. Opin. Oncol.*, vol. 28, no. 2, pp. 130–134, Mar. 2016.
- [44] E. R. Malone, M. Oliva, P. J. B. Sabatini, T. L. Stockley, and L. L. Siu, "Molecular profiling for precision cancer therapies," en, *Genome Med.*, vol. 12, no. 1, p. 8, Jan. 2020.
- [45] M. Ilić and P. Hofman, "Pros: Can tissue biopsy be replaced by liquid biopsy?" en, *Transl. Lung Cancer Res.*, vol. 5, no. 4, pp. 420–423, Aug. 2016.

- [46] P. Kumar, S. Gupta, and B. C. Das, “Saliva as a potential non-invasive liquid biopsy for early and easy diagnosis/prognosis of head and neck cancer,” en, *Transl. Oncol.*, vol. 40, p. 101 827, Dec. 2023.
- [47] C. K. Chen, J. Liao, M. S. Li, and B. L. Khoo, “Urine biopsy technologies: Cancer and beyond,” en, *Theranostics*, vol. 10, no. 17, pp. 7872–7888, Jun. 2020.
- [48] I. Lopez-Rojo, S. Olmedillas-Lopez, P. Villarejo Campos, V. Dominguez Prieto, J. Barambio Buendia, D. Cortes Guiral, M. Garcia-Arranz, and D. Garcia-Olmo, “Liquid biopsy in peritoneal fluid and plasma as a prognostic factor in advanced colorectal and appendiceal tumors after complete cytoreduction and hyperthermic intraperitoneal chemotherapy,” en, *Ther. Adv. Med. Oncol.*, vol. 12, p. 1 758 835 920 981 351, Dec. 2020.
- [49] R. A. Hickman, A. M. Miller, and M. E. Arcila, “Cerebrospinal fluid: A unique source of circulating tumor DNA with broad clinical applications,” en, *Transl. Oncol.*, vol. 33, p. 101 688, Jul. 2023.
- [50] C. Saitta, I. De Simone, V. Fasulo, M. Corbetta, S. Duga, C. Chiereghin, F. S. Colombo, A. Benetti, R. Contieri, P. P. Avolio, A. Uleri, A. Saita, G. F. Guazzoni, R. Hurle, P. Colombo, N. M. Buffi, P. Casale, G. Lughezzani, R. Asselta, G. Solda, and M. Lazzeri, “Evaluation of semen self-sampling yield predictors and CTC isolation by multi-color flow cytometry for liquid biopsy of localized prostate cancer,” en, *Cancers (Basel)*, vol. 15, no. 10, May 2023.
- [51] J. M. Cameron, A. Sala, G. Antoniou, P. M. Brennan, H. J. Butler, J. J. A. Conn, S. Connal, T. Curran, M. G. Hegarty, R. G. McHardy, D. Orringer, D. S. Palmer, B. R. Smith, and M. J. Baker, “A spectroscopic liquid biopsy for the earlier detection of multiple cancer types,” en, *Br. J. Cancer*, vol. 129, no. 10, pp. 1658–1666, Nov. 2023.
- [52] G. Siravegna, S. Marsoni, S. Siena, and A. Bardelli, “Integrating liquid biopsies into the management of cancer,” en, *Nat. Rev. Clin. Oncol.*, vol. 14, no. 9, pp. 531–548, Sep. 2017.
- [53] E. Kilgour, D. G. Rothwell, G. Brady, and C. Dive, “Liquid biopsy-based biomarkers of treatment response and resistance,” en, *Cancer Cell*, vol. 37, no. 4, pp. 485–495, Apr. 2020.

- [54] C. Alix-Panabières, D. Marchetti, and J. E. Lang, “Liquid biopsy: From concept to clinical application,” en, *Sci. Rep.*, vol. 13, no. 1, Dec. 2023.
- [55] H. Zhou, L. Zhu, J. Song, G. Wang, P. Li, W. Li, P. Luo, X. Sun, J. Wu, Y. Liu, S. Zhu, and Y. Zhang, “Liquid biopsy at the frontier of detection, prognosis and progression monitoring in colorectal cancer,” en, *Mol. Cancer*, vol. 21, no. 1, p. 86, Mar. 2022.
- [56] C. Alix-Panabières and K. Pantel, “Liquid biopsy: From discovery to clinical application,” en, *Cancer Discov.*, vol. 11, no. 4, pp. 858–873, Apr. 2021.
- [57] M. G. Krebs, R. L. Metcalf, L. Carter, G. Brady, F. H. Blackhall, and C. Dive, “Molecular analysis of circulating tumour cells-biology and biomarkers,” en, *Nat. Rev. Clin. Oncol.*, vol. 11, no. 3, pp. 129–144, Mar. 2014.
- [58] P. Gilson, J.-L. Merlin, and A. Harlé, “Deciphering tumour heterogeneity: From tissue to liquid biopsy,” en, *Cancers (Basel)*, vol. 14, no. 6, p. 1384, Mar. 2022.
- [59] H. Schwarzenbach, D. S. B. Hoon, and K. Pantel, “Cell-free nucleic acids as biomarkers in cancer patients,” en, *Nat. Rev. Cancer*, vol. 11, no. 6, pp. 426–437, Jun. 2011.
- [60] C. Huang, Y. R. Neupane, X. C. Lim, R. Shekhani, B. Czarny, M. G. Wacker, G. Pastorin, and J.-W. Wang, “Extracellular vesicles in cardiovascular disease,” in *Advances in Clinical Chemistry*, ser. Advances in clinical chemistry, Elsevier, 2021, pp. 47–95.
- [61] E. Heitzer, I. S. Haque, C. E. S. Roberts, and M. R. Speicher, “Current and future perspectives of liquid biopsies in genomics-driven oncology,” en, *Nat. Rev. Genet.*, vol. 20, no. 2, pp. 71–88, Feb. 2019.
- [62] E. Lefrançais, G. Ortiz-Muñoz, A. Caudrillier, B. Mallavia, F. Liu, D. M. Sayah, E. E. Thornton, M. B. Headley, T. David, S. R. Coughlin, M. F. Krummel, A. D. Leavitt, E. Passegué, and M. R. Looney, “The lung is a site of platelet biogenesis and a reservoir for haematopoietic progenitors,” en, *Nature*, vol. 544, no. 7648, pp. 105–109, Apr. 2017.

- [63] A. L. Palacios-Acedo, D. Mège, L. Crescence, F. Dignat-George, C. Dubois, and L. Panicot-Dubois, “Platelets, thrombo-inflammation, and cancer: Collaborating with the enemy,” en, *Front. Immunol.*, vol. 10, p. 1805, Jul. 2019.
- [64] M. Haemmerle, R. L. Stone, D. G. Menter, V. Afshar-Kharghan, and A. K. Sood, “The platelet lifeline to cancer: Challenges and opportunities,” en, *Cancer Cell*, vol. 33, no. 6, pp. 965–983, Jun. 2018.
- [65] P. Carmeliet and R. K. Jain, “Angiogenesis in cancer and other diseases,” en, *Nature*, vol. 407, no. 6801, pp. 249–257, Sep. 2000.
- [66] H. A. Goubran, T. Burnouf, J. Stakiw, and J. Seghatchian, “Platelet microparticle: A sensitive physiological “fine tuning” balancing factor in health and disease,” en, *Transfus. Apher. Sci.*, vol. 52, no. 1, pp. 12–18, Feb. 2015.
- [67] P. Kanikarla-Marie, M. Lam, D. G. Menter, and S. Kopetz, “Platelets, circulating tumor cells, and the circulome,” en, *Cancer Metastasis Rev.*, vol. 36, no. 2, pp. 235–248, Jun. 2017.
- [68] R. Leblanc and O. Peyruchaud, “Metastasis: New functional implications of platelets and megakaryocytes,” en, *Blood*, vol. 128, no. 1, pp. 24–31, Jul. 2016.
- [69] J. Moss, R. Ben-Ami, E. Shai, O. Gal-Rosenberg, Y. Kalish, A. Klochender, G. Cann, B. Glaser, A. Arad, R. Shemer, and Y. Dor, “Megakaryocyte- and erythroblast-specific cell-free DNA patterns in plasma and platelets reflect thrombopoiesis and erythropoiesis levels,” en, *Nat. Commun.*, vol. 14, no. 1, p. 7542, Nov. 2023.
- [70] L. Plantureux, D. Mège, L. Crescence, F. Dignat-George, C. Dubois, and L. Panicot-Dubois, “Impacts of cancer on platelet production, activation and education and mechanisms of cancer-associated thrombosis,” en, *Cancers (Basel)*, vol. 10, no. 11, p. 441, Nov. 2018.
- [71] L. Plantureux, L. Crescence, F. Dignat-George, L. Panicot-Dubois, and C. Dubois, “Effects of platelets on cancer progression,” *Thrombosis Research*, vol. 164, S40–S47, 2018, Papers and Abstracts of the 9th International Conference on Thrombosis and Hemostasis Issues in Cancer, April 13-15, 2018, Bergamo, Italy, ISSN: 0049-3848. DOI: <https://doi.org/10.1016/j.thromres.2018.01.035>. [Online].

Available: <https://www.sciencedirect.com/science/article/pii/S0049384818300434>.

- [72] S. Sabrkhany, M. J. E. Kuijpers, A. W. Griffioen, and M. G. A. Oude Egbrink, “Platelets: The holy grail in cancer blood biomarker research?” en, *Angiogenesis*, vol. 22, no. 1, pp. 1–2, Feb. 2019.
- [73] S. Sabrkhany, M. J. E. Kuijpers, J. C. Knol, S. W. M. Olde Damink, A.-M. C. Dingemans, H. M. Verheul, S. R. Piersma, T. V. Pham, A. W. Griffioen, M. G. A. oude Egbrink, and C. R. Jimenez, “Exploration of the platelet proteome in patients with early-stage cancer,” en, *J. Proteomics*, vol. 177, pp. 65–74, Apr. 2018.
- [74] M. G. Best, N. Sol, I. Kooi, J. Tannous, B. A. Westerman, F. Rustenburg, P. Schellen, H. Verschueren, E. Post, J. Koster, B. Ylstra, N. Ameziane, J. Dorsman, E. F. Smit, H. M. Verheul, D. P. Noske, J. C. Reijneveld, R. J. A. Nilsson, B. A. Tannous, P. Wesseling, and T. Wurdinger, “RNA-seq of tumor-educated platelets enables blood-based pan-cancer, multiclass, and molecular pathway cancer diagnostics,” en, *Cancer Cell*, vol. 28, no. 5, pp. 666–676, Nov. 2015.
- [75] S. Ding, X. Dong, and X. Song, “Tumor educated platelet: The novel BioSource for cancer detection,” en, *Cancer Cell Int.*, vol. 23, no. 1, p. 91, May 2023.
- [76] S. G. J. G. In ’t Veld, M. Arkani, E. Post, M. Antunes-Ferreira, S. D’Ambrosi, D. C. L. Vessies, L. Vermunt, A. Vancura, M. Muller, A.-L. N. Niemeijer, J. Tannous, L. L. Meijer, T. Y. S. Le Large, G. Mantini, N. E. Wondergem, K. M. Heinhuis, S. van Wilpe, A. J. Smits, E. E. E. Drees, E. Roos, C. E. Leurs, L.-A. Tjon Kon Fat, E. J. van der Lelij, G. Dwarshuis, M. J. Kamphuis, L. E. Visser, R. Harting, A. Gregory, M. W. Schweiger, L. E. Wedekind, J. Ramaker, K. Zwaan, H. Verschueren, I. Bahce, A. J. de Langen, E. F. Smit, M. M. van den Heuvel, K. J. Hartemink, M. J. E. Kuijpers, M. G. A. Oude Egbrink, A. W. Griffioen, R. Rossel, T. J. N. Hiltermann, E. Lee-Lewandrowski, K. B. Lewandrowski, P. C. De Witt Hamer, M. Kouwenhoven, J. C. Reijneveld, W. P. J. Leenders, A. Hoeben, I. M. Verdonck-de Leeuw, C. R. Leemans, R. J. Baatenburg de Jong, C. H. J. Terhaard, R. P. Takes, J. A. Langendijk, S. C. de Jager, A. O. Kraaijeveld, G. Pasterkamp, M. Smits, J. A. Schalken,

- S. Łapińska-Szumczyk, A. Łojkowska, A. J. Żaczek, H. Lokhorst, N. W. C. J. van de Donk, I. Nijhof, H.-J. Prins, J. M. Zijlstra, S. Idema, J. C. Baayen, C. E. Teunissen, J. Killestein, M. G. Besselink, L. Brammen, T. Bachleitner-Hofmann, F. Mateen, J. T. M. Plukker, M. Heger, Q. de Mast, T. Lisman, D. M. Pegtel, H.-J. Bogaard, J. Jassem, A. Supernat, N. Mehra, W. Gerritsen, C. D. de Kroon, C. A. R. Lok, J. M. J. Piek, N. Steeghs, W. J. van Houdt, R. H. Brakenhoff, G. S. Sonke, H. M. Verheul, E. Giovannetti, G. Kazemier, S. Sabrkhany, E. Schuurin, E. A. Sistermans, R. Wolthuis, H. Meijers-Heijboer, J. Dorsman, C. Oudejans, B. Ylstra, B. A. Westerman, D. van den Broek, D. Koppers-Lalic, P. Wesseling, R. J. A. Nilsson, W. P. Vandertop, D. P. Noske, B. A. Tannous, N. Sol, M. G. Best, and T. Wurdinger, “Detection and localization of early- and late-stage cancers using platelet RNA,” en, *Cancer Cell*, vol. 40, no. 9, 999–1009.e6, Sep. 2022.
- [77] S. D’Ambrosi, R. J. Nilsson, and T. Wurdinger, “Platelets and tumor-associated RNA transfer,” en, *Blood*, vol. 137, no. 23, pp. 3181–3191, Jun. 2021.
- [78] R. Hanayama, “Emerging roles of extracellular vesicles in physiology and disease,” en, *J. Biochem.*, vol. 169, no. 2, pp. 135–138, Mar. 2021.
- [79] C. B. Fox, J. Kim, L. V. Le, C. L. Nemeth, H. D. Chirra, and T. A. Desai, “Micro/nanofabricated platforms for oral drug delivery,” en, *J. Control. Release*, vol. 219, pp. 431–444, Dec. 2015.
- [80] J. Liu, Y. Chen, F. Pei, C. Zeng, Y. Yao, W. Liao, and Z. Zhao, “Extracellular vesicles in liquid biopsies: Potential for disease diagnosis,” en, *Biomed Res. Int.*, vol. 2021, p. 6 611 244, Jan. 2021.
- [81] C. P. O’Neill, K. E. Gilligan, and R. M. Dwyer, “Role of extracellular vesicles (EVs) in cell stress response and resistance to cancer therapy,” en, *Cancers (Basel)*, vol. 11, no. 2, p. 136, Jan. 2019.
- [82] B. Irmer, S. Chandrabalan, L. Maas, A. Bleckmann, and K. Menck, “Extracellular vesicles in liquid biopsies as biomarkers for solid tumors,” en, *Cancers (Basel)*, vol. 15, no. 4, Feb. 2023.
- [83] E. Zhou, Y. Li, F. Wu, M. Guo, J. Xu, S. Wang, Q. Tan, P. Ma, S. Song, and Y. Jin, “Circulating extracellular vesicles are effective biomarkers for predicting response to cancer therapy,” en, *EBioMedicine*, vol. 67, no. 103365, p. 103 365, May 2021.

- [84] R. I. Freshney and M. G. Freshney, Eds., *Culture of epithelial cells*, en, ser. Culture of Specialized Cells. Nashville, TN: John Wiley & Sons, Dec. 2002.
- [85] M. T. Barrett, J. Glogovac, L. J. Prevo, B. J. Reid, P. Porter, and P. S. Rabinovitch, “High-quality RNA and DNA from flow cytometrically sorted human epithelial cells and tissues,” en, *Biotechniques*, vol. 32, no. 4, pp. 888–90, 892, 894, 896, Apr. 2002.
- [86] B. Brandt, A. Roetger, S. Heidl, C. Jackisch, R. J. Lelle, G. Assmann, and K. S. Zänker, “Isolation of blood-borne epithelium-derived c-erbB-2 oncoprotein-positive clustered cells from the peripheral blood of breast cancer patients,” en, *Int. J. Cancer*, vol. 76, no. 6, pp. 824–828, Jun. 1998.
- [87] A. H. Talasaz, A. A. Powell, D. E. Huber, J. G. Berbee, K.-H. Roh, W. Yu, W. Xiao, M. M. Davis, R. F. Pease, M. N. Mindrinos, S. S. Jeffrey, and R. W. Davis, “Isolating highly enriched populations of circulating epithelial cells and other rare cells from blood using a magnetic sweeper device,” en, *Proc. Natl. Acad. Sci. U. S. A.*, vol. 106, no. 10, pp. 3970–3975, Mar. 2009.
- [88] K. H. R. Tkaczuk, O. Goloubeva, N. S. Tait, F. Feldman, M. Tan, Z.-P. Lum, S. A. Lesko, D. A. Van Echo, and P. O. P. Ts’o, “The significance of circulating epithelial cells in breast cancer patients by a novel negative selection method,” en, *Breast Cancer Res. Treat.*, vol. 111, no. 2, pp. 355–364, Sep. 2008.
- [89] I. Bhan, K. Mosesso, L. Goyal, J. Philipp, M. Kalinich, J. W. Franses, M. Choz, R. Oklu, M. Toner, S. Maheswaran, D. A. Haber, A. X. Zhu, R. T. Chung, M. Aryee, and D. T. Ting, “Detection and analysis of circulating epithelial cells in liquid biopsies from patients with liver disease,” en, *Gastroenterology*, vol. 155, no. 6, 2016–2018.e11, Dec. 2018.
- [90] K. H. R. Tkaczuk, O. Goloubeva, N. S. Tait, F. Feldman, M. Tan, Z.-P. Lum, S. A. Lesko, D. A. Van Echo, and P. O. P. Ts’o, “The significance of circulating epithelial cells in breast cancer patients by a novel negative selection method,” en, *Breast Cancer Res. Treat.*, vol. 111, no. 2, pp. 355–364, Sep. 2008.

- [91] K. Pantel, E. Denève, D. Nocca, A. Coffy, J.-P. Vendrell, T. Maudelonde, S. Riethdorf, and C. Alix-Panabières, “Circulating epithelial cells in patients with benign colon diseases,” en, *Clin. Chem.*, vol. 58, no. 5, pp. 936–940, May 2012.
- [92] A. Grada and K. Weinbrecht, “Next-generation sequencing: Methodology and application,” en, *J. Invest. Dermatol.*, vol. 133, no. 8, e11, Aug. 2013.
- [93] A. Sanchez-Pla, F. Reverter, M. C. Ruiz de Villa, and M. Comabella, “Transcriptomics: mRNA and alternative splicing,” en, *J. Neuroimmunol.*, vol. 248, no. 1-2, pp. 23–31, Jul. 2012.
- [94] Z. Wang, M. Gerstein, and M. Snyder, “RNA-Seq: A revolutionary tool for transcriptomics,” en, *Nat. Rev. Genet.*, vol. 10, no. 1, pp. 57–63, Jan. 2009.
- [95] K. Athanasopoulou, M. A. Boti, P. G. Adamopoulos, P. C. Skourou, and A. Scorilas, “Third-generation sequencing: The spearhead towards the radical transformation of modern genomics,” en, *Life (Basel)*, vol. 12, no. 1, p. 30, Dec. 2021.
- [96] M. T. Pervez, M. J. U. Hasnain, S. H. Abbas, M. F. Moustafa, N. Aslam, and S. S. M. Shah, “A comprehensive review of performance of next-generation sequencing platforms,” en, *Biomed Res. Int.*, vol. 2022, p. 3 457 806, Sep. 2022.
- [97] J. M. Eastel, K. W. Lam, N. L. Lee, W. Y. Lok, A. H. F. Tsang, X. M. Pei, A. K. C. Chan, W. C. S. Cho, and S. C. C. Wong, “Application of NanoString technologies in companion diagnostic development,” en, *Expert Rev. Mol. Diagn.*, vol. 19, no. 7, pp. 591–598, Jul. 2019.
- [98] W. R. McCombie, J. D. McPherson, and E. R. Mardis, “Next-generation sequencing technologies,” en, *Cold Spring Harb. Perspect. Med.*, vol. 9, no. 11, a036798, Nov. 2019.
- [99] D. Qin, “Next-generation sequencing and its clinical application,” en, *Cancer Biol. Med.*, vol. 16, no. 1, pp. 4–10, Feb. 2019.
- [100] R. Li, Y. Li, X. Fang, H. Yang, J. Wang, K. Kristiansen, and J. Wang, “SNP detection for massively parallel whole-genome resequencing,” en, *Genome Res.*, vol. 19, no. 6, pp. 1124–1132, Jun. 2009.
- [101] F. Ozsolak and P. M. Milos, “RNA sequencing: Advances, challenges and opportunities,” en, *Nat. Rev. Genet.*, vol. 12, no. 2, pp. 87–98, Feb. 2011.

- [102] J. C. Marioni, C. E. Mason, S. M. Mane, M. Stephens, and Y. Gilad, “RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays,” en, *Genome Res.*, vol. 18, no. 9, pp. 1509–1517, Sep. 2008.
- [103] R. Kumar, Y. Ichihashi, S. Kimura, D. H. Chitwood, L. R. Headland, J. Peng, J. N. Maloof, and N. R. Sinha, “A high-throughput method for illumina RNA-Seq library preparation,” en, *Front. Plant Sci.*, vol. 3, p. 202, Aug. 2012.
- [104] *Rna sequencing-definition, principle, steps, types, uses*, Aug. 2023. [Online]. Available: <https://microbenotes.com/rna-sequencing-principle-steps-types-uses/>.
- [105] T. J. Treangen and S. L. Salzberg, “Repetitive DNA and next-generation sequencing: Computational challenges and solutions,” en, *Nat. Rev. Genet.*, vol. 13, no. 1, pp. 36–46, Nov. 2011.
- [106] P. D. Browne, T. K. Nielsen, W. Kot, A. Aggerholm, M. T. P. Gilbert, L. Puetz, M. Rasmussen, A. Zervas, and L. H. Hansen, “GC bias affects genomic and metagenomic reconstructions, underrepresenting GC-poor organisms,” en, *Gigascience*, vol. 9, no. 2, Feb. 2020.
- [107] A. S. Mikheyev and M. M. Y. Tin, “A first look at the oxford nanopore MinION sequencer,” en, *Mol. Ecol. Resour.*, vol. 14, no. 6, pp. 1097–1102, Nov. 2014.
- [108] M. MacKenzie and C. Argyropoulos, “An introduction to nanopore sequencing: Past, present, and future considerations,” en, *Micromachines (Basel)*, vol. 14, no. 2, Feb. 2023.
- [109] R. R. Wick, L. M. Judd, and K. E. Holt, “Performance of neural network basecalling tools for oxford nanopore sequencing,” en, *Genome Biol.*, vol. 20, no. 1, p. 129, Jun. 2019.
- [110] M. Jain, S. Koren, K. H. Miga, J. Quick, A. C. Rand, T. A. Sasani, J. R. Tyson, A. D. Beggs, A. T. Dilthey, I. T. Fiddes, S. Malla, H. Marriott, T. Nieto, J. O’Grady, H. E. Olsen, B. S. Pedersen, A. Rhie, H. Richardson, A. R. Quinlan, T. P. Snutch, L. Tee, B. Paten, A. M. Phillippy, J. T. Simpson, N. J. Loman, and M. Loose, “Nanopore sequencing and assembly of a human genome with ultra-long reads,” en, *Nat. Biotechnol.*, vol. 36, no. 4, pp. 338–345, Apr. 2018.

- [111] P. A. Audano, A. Sulovari, T. A. Graves-Lindsay, S. Cantsilieris, M. Sorensen, A. E. Welch, M. L. Dougherty, B. J. Nelson, A. Shah, S. K. Dutcher, W. C. Warren, V. Magrini, S. D. McGrath, Y. I. Li, R. K. Wilson, and E. E. Eichler, “Characterizing the major structural variant alleles of the human genome,” en, *Cell*, vol. 176, no. 3, 663–675.e19, Jan. 2019.
- [112] L. K. White and J. R. Hesselberth, “Modification mapping by nanopore sequencing,” en, *Front. Genet.*, vol. 13, p. 1037134, Oct. 2022.
- [113] K. H. Miga, S. Koren, A. Rhie, M. R. Vollger, A. Gershman, A. Bzikadze, S. Brooks, E. Howe, D. Porubsky, G. A. Logsdon, V. A. Schneider, T. Potapova, J. Wood, W. Chow, J. Armstrong, J. Fredrickson, E. Pak, K. Tigyi, M. Kremitzki, C. Markovic, V. Maduro, A. Dutra, G. G. Bouffard, A. M. Chang, N. F. Hansen, A. B. Wilfert, F. Thibaud-Nissen, A. D. Schmitt, J.-M. Belton, S. Selvaraj, M. Y. Dennis, D. C. Soto, R. Sahasrabudhe, G. Kaya, J. Quick, N. J. Loman, N. Holmes, M. Loose, U. Surti, R. A. Risques, T. A. Graves Lindsay, R. Fulton, I. Hall, B. Paten, K. Howe, W. Timp, A. Young, J. C. Mullikin, P. A. Pevzner, J. L. Gerton, B. A. Sullivan, E. E. Eichler, and A. M. Phillippy, “Telomere-to-telomere assembly of a complete human X chromosome,” en, *Nature*, vol. 585, no. 7823, pp. 79–84, Sep. 2020.
- [114] S. Hussain, “Native RNA-sequencing throws its hat into the transcriptomics ring,” en, *Trends Biochem. Sci.*, vol. 43, no. 4, pp. 225–227, Apr. 2018.
- [115] M. Polenkowski, A. B. Allister, S. Burbano de Lara, M. Soltau, G. Kendre, and D. D. H. Tran, “Mapping alternative polyadenylation in human cells using direct RNA sequencing technology,” en, *STAR Protoc.*, vol. 4, no. 3, p. 102420, Jul. 2023.
- [116] R. Li, X. Ren, Q. Ding, Y. Bi, D. Xie, and Z. Zhao, “Direct full-length RNA sequencing reveals unexpected transcriptome complexity during *Caenorhabditis elegans* development,” en, *Genome Res.*, vol. 30, no. 2, pp. 287–298, Feb. 2020.

- [117] *Nanopores allow direct sequencing of rna strands, giving full-length reads with low bias*, May 2019. [Online]. Available: <https://nanoporetech.com/resource-centre/nanopores-allow-direct-sequencing-rna-strands-giving-full-length-reads-low-bias-1>.
- [118] J. Lee, I. Sohn, I.-G. Do, K.-M. Kim, S. H. Park, J. O. Park, Y. S. Park, H. Y. Lim, T. S. Sohn, J. M. Bae, M. G. Choi, D. H. Lim, B. H. Min, J. H. Lee, P. L. Rhee, J. J. Kim, D. I. Choi, I. B. Tan, K. Das, P. Tan, S. H. Jung, W. K. Kang, and S. Kim, “Nanostring-based multigene assay to predict recurrence for gastric cancer patients after surgery,” en, *PLoS One*, vol. 9, no. 3, e90133, Mar. 2014.
- [119] R. F. H. Walter, R. Werner, S. Ting, C. Vollbrecht, D. Theegarten, D. C. Christoph, K. W. Schmid, J. Wohlschlaeger, and F. D. Mairinger, “Identification of deregulation of apoptosis and cell cycle in neuroendocrine tumors of the lung via NanoString ncounter expression analysis,” en, *Oncotarget*, vol. 6, no. 28, pp. 24 690–24 698, Sep. 2015.
- [120] Y.-J. Chen, C.-S. Huang, N.-N. Phan, T.-P. Lu, C.-Y. Liu, C.-J. Huang, J.-H. Chiu, L.-M. Tseng, and C.-C. Huang, “Molecular subtyping of breast cancer intrinsic taxonomy with oligonucleotide microarray and NanoString ncounter,” en, *Biosci. Rep.*, vol. 41, no. 8, Aug. 2021.
- [121] S. K. F. Daum, C. Treese, A. Arnold, H. Harloff, H. Lammert, F. Mairinger, M. Hummel, and K. Kleo, “1485P response prediction using NanoString ncounter technology and NGS panel sequencing in neoadjuvant chemotherapy in patients with esophagogastric adenocarcinoma,” en, *Ann. Oncol.*, vol. 31, S923–S924, Sep. 2020.
- [122] *Ncounter® analysis systems for biomarker validation and biomarker development*. [Online]. Available: <https://nanosttring.com/products/ncounter-analysis-system/ncounter-systems-overview/>.
- [123] P. P. Reis, L. Waldron, R. S. Goswami, W. Xu, Y. Xuan, B. Perez-Ordenez, P. Gullane, J. Irish, I. Jurisica, and S. Kamel-Reid, “mRNA transcript quantification in archival samples using multiplexed, color-coded probes,” en, *BMC Biotechnol.*, vol. 11, no. 1, p. 46, May 2011.
- [124] E. Macerola, A. M. Poma, and F. Basolo, “NanoString in the screening of genetic abnormalities associated with thyroid cancer,” en, *Semin. Cancer Biol.*, vol. 79, pp. 132–140, Feb. 2022.

- [125] J. M. Eastel, K. W. Lam, N. L. Lee, W. Y. Lok, A. H. F. Tsang, X. M. Pei, A. K. C. Chan, W. C. S. Cho, and S. C. C. Wong, “Application of NanoString technologies in companion diagnostic development,” en, *Expert Rev. Mol. Diagn.*, vol. 19, no. 7, pp. 591–598, Jul. 2019.
- [126] M. H. Veldman-Jones, R. Brant, C. Rooney, C. Geh, H. Emery, C. G. Harbron, M. Wappett, A. Sharpe, M. Dymond, J. C. Barrett, E. A. Harrington, and G. Marshall, “Evaluating robustness and sensitivity of the NanoString technologies ncounter platform to enable multiplexed gene expression analysis of clinical samples,” en, *Cancer Res.*, vol. 75, no. 13, pp. 2587–2593, Jul. 2015.
- [127] N. Sol, S. G. J. G. In ’t Veld, A. Vancura, M. Tjerkstra, C. Leurs, F. Rustenburg, P. Schellen, H. Verschueren, E. Post, K. Zwaan, J. Ramaker, L. E. Wedekind, J. Tannous, B. Ylstra, J. Killestein, F. Mateen, S. Idema, P. C. de Witt Hamer, A. C. Navis, W. P. J. Leenders, A. Hoeben, B. Moraal, D. P. Noske, W. P. Vandertop, R. J. A. Nilsson, B. A. Tannous, P. Wesseling, J. C. Reijneveld, M. G. Best, and T. Wurdinger, “Tumor-educated platelet RNA for the detection and (pseudo)progression monitoring of glioblastoma,” en, *Cell Rep. Med.*, vol. 1, no. 7, p. 100101, Oct. 2020.
- [128] S. N. Lone, S. Nisar, T. Masoodi, M. Singh, A. Rizwan, S. Hashem, W. El-Rifai, D. Bedognetti, S. K. Batra, M. Haris, A. A. Bhat, and M. A. Macha, “Liquid biopsy: A step closer to transform diagnosis, prognosis and future of cancer treatments,” en, *Mol. Cancer*, vol. 21, no. 1, p. 79, Mar. 2022.
- [129] S. Ben-Aroya and E. Y. Levanon, “A-to-I RNA editing: An overlooked source of cancer mutations,” en, *Cancer Cell*, vol. 33, no. 5, pp. 789–790, May 2018.
- [130] A. Alba-Bernal, R. Lavado-Valenzuela, M. E. Dominguez-Recio, B. Jimenez-Rodriguez, M. I. Queipo-Ortuño, E. Alba, and I. Comino-Mendez, “Challenges and achievements of liquid biopsy technologies employed in early breast cancer,” en, *EBioMedicine*, vol. 62, no. 103100, p. 103100, Dec. 2020.
- [131] A. Bayat, “Science, medicine, and the future: Bioinformatics,” en, *BMJ*, vol. 324, no. 7344, pp. 1018–1022, Apr. 2002.

- [132] S. Ahn, “Introduction to bioinformatics: Sequencing technology,” en, *Asia Pac. Allergy*, vol. 1, no. 2, pp. 93–97, Jul. 2011.
- [133] R. Pereira, J. Oliveira, and M. Sousa, “Bioinformatics and computational tools for next-generation sequencing analysis in clinical genetics,” en, *J. Clin. Med.*, vol. 9, no. 1, p. 132, Jan. 2020.
- [134] B. He, R. Zhu, H. Yang, Q. Lu, W. Wang, L. Song, X. Sun, G. Zhang, S. Li, J. Yang, G. Tian, P. Bing, and J. Lang, “Assessing the impact of data preprocessing on analyzing next generation sequencing data,” en, *Front. Bioeng. Biotechnol.*, vol. 8, p. 817, Jul. 2020.
- [135] M. Ruffalo, T. LaFramboise, and M. Koyutürk, “Comparative analysis of algorithms for next-generation sequencing read alignment,” en, *Bioinformatics*, vol. 27, no. 20, pp. 2790–2796, Oct. 2011.
- [136] A. Metsis, U. Andersson, G. Baurén, P. Ernfors, P. Lönnerberg, A. Montelius, M. Oldin, A. Pihlak, and S. Linnarsson, “Whole-genome expression profiling through fragment display and combinatorial gene identification,” en, *Nucleic Acids Res.*, vol. 32, no. 16, e127, Sep. 2004.
- [137] S. Tang, A. S. Buchman, Y. Wang, D. Avey, J. Xu, S. Tasaki, D. A. Bennett, Q. Zheng, and J. Yang, “Differential gene expression analysis based on linear mixed model corrects false positive inflation for studying quantitative traits,” en, *Sci. Rep.*, vol. 13, no. 1, p. 16 570, Oct. 2023.
- [138] E. E. Baschal, E. D. Larson, T. C. Bootpetch Roberts, S. Pathak, G. Frank, E. Handley, J. Dinwiddie, M. Moloney, P. J. Yoon, S. P. Gubbels, M. A. Scholes, S. P. Cass, H. A. Jenkins, D. N. Frank, I. V. Yang, D. A. Schwartz, V. R. Ramakrishnan, and R. L. P. Santos-Cortez, “Identification of novel genes and biological pathways that overlap in infectious and nonallergic diseases of the upper and lower airways using network analyses,” en, *Front. Genet.*, vol. 10, p. 1352, 2019.
- [139] W. Jiang and L. Chen, “Alternative splicing: Human disease and quantitative analysis from high-throughput sequencing,” en, *Comput. Struct. Biotechnol. J.*, vol. 19, pp. 183–195, 2021.

- [140] C. Heydt, C. B. Wölwer, O. Velazquez Camacho, S. Wagener-Ryczek, R. Pappesch, J. Siemanowski, J. Rehker, F. Haller, A. Agaimy, K. Worm, T. Herold, N. Pfarr, W. Weichert, T. Kirchner, A. Jung, J. Kumbrink, W. Goering, I. Esposito, R. Buettner, A. M. Hillmer, and S. Merkelbach-Bruse, “Detection of gene fusions using targeted next-generation sequencing: A comparative evaluation,” en, *BMC Med. Genomics*, vol. 14, no. 1, p. 62, Feb. 2021.
- [141] S. Wijeratne, M. E. H. Gonzalez, K. Roach, K. E. Miller, K. M. Schieffer, J. R. Fitch, J. Leonard, P. White, B. J. Kelly, C. E. Cottrell, E. R. Mardis, R. K. Wilson, and A. R. Miller, “Full-length isoform concatenation sequencing to resolve cancer transcriptome complexity,” en, *BMC Genomics*, vol. 25, no. 1, Jan. 2024.
- [142] Y. Zhao, K. Wang, W.-L. Wang, T.-T. Yin, W.-Q. Dong, and C.-J. Xu, “A high-throughput SNP discovery strategy for RNA-seq data,” en, *BMC Genomics*, vol. 20, no. 1, p. 160, Feb. 2019.
- [143] A. Leger, P. P. Amaral, L. Pandolfini, C. Capitanichik, F. Capraro, V. Miano, V. Migliori, P. Toolan-Kerr, T. Sideri, A. J. Enright, K. Tzelepis, F. J. van Werven, N. M. Luscombe, I. Barbieri, J. Ule, T. Fitzgerald, E. Birney, T. Leonardi, and T. Kouzarides, “RNA modifications detection by comparative nanopore direct RNA sequencing,” en, *Nat. Commun.*, vol. 12, no. 1, p. 7198, Dec. 2021.
- [144] C.-C. Huang, M. Du, and L. Wang, “Bioinformatics analysis for circulating cell-free DNA in cancer,” en, *Cancers (Basel)*, vol. 11, no. 6, p. 805, Jun. 2019.
- [145] J. Hou, X. Li, and K.-P. Xie, “Coupled liquid biopsy and bioinformatics for pancreatic cancer early detection and precision prognostication,” en, *Mol. Cancer*, vol. 20, no. 1, p. 34, Feb. 2021.
- [146] S.-J. Dawson, D. W. Y. Tsui, M. Murtaza, H. Biggs, O. M. Rueda, S.-F. Chin, M. J. Dunning, D. Gale, T. Forshew, B. Mahler-Araujo, S. Rajan, S. Humphray, J. Becq, D. Halsall, M. Wallis, D. Bentley, C. Caldas, and N. Rosenfeld, “Analysis of circulating tumor DNA to monitor metastatic breast cancer,” en, *N. Engl. J. Med.*, vol. 368, no. 13, pp. 1199–1209, Mar. 2013.

- [147] Y. R. Im, D. W. Y. Tsui, L. A. Diaz Jr, and J. C. M. Wan, “Next-generation liquid biopsies: Embracing data science in oncology,” en, *Trends Cancer*, vol. 7, no. 4, pp. 283–292, Apr. 2021.
- [148] C. Xu and S. A. Jackson, “Machine learning and complex biological data,” en, *Genome Biol.*, vol. 20, no. 1, p. 76, Apr. 2019.
- [149] M. W. Libbrecht and W. S. Noble, “Machine learning applications in genetics and genomics,” en, *Nat. Rev. Genet.*, vol. 16, no. 6, pp. 321–332, Jun. 2015.
- [150] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, “Machine learning applications in cancer prognosis and prediction,” en, *Comput. Struct. Biotechnol. J.*, vol. 13, pp. 8–17, 2015.
- [151] C. Lopez, S. Tucker, T. Salameh, and C. Tucker, “An unsupervised machine learning method for discovering patient clusters based on genetic signatures,” en, *J. Biomed. Inform.*, vol. 85, pp. 30–39, Sep. 2018.
- [152] J. Ko, S. N. Baldassano, P.-L. Loh, K. Kording, B. Litt, and D. Issadore, “Machine learning to detect signatures of disease in liquid biopsies - a user’s guide,” en, *Lab Chip*, vol. 18, no. 3, pp. 395–405, Jan. 2018.
- [153] H.-J. Kwon, U.-H. Park, C. J. Goh, D. Park, Y. G. Lim, I. K. Lee, W.-J. Do, K. J. Lee, H. Kim, S.-Y. Yun, J. Joo, N. Y. Min, S. Lee, S.-W. Um, and M.-S. Lee, “Enhancing lung cancer classification through integration of liquid biopsy multi-omics data with machine learning techniques,” en, *Cancers (Basel)*, vol. 15, no. 18, Sep. 2023.
- [154] J. Ko, S. N. Baldassano, P.-L. Loh, K. Kording, B. Litt, and D. Issadore, “Machine learning to detect signatures of disease in liquid biopsies - a user’s guide,” en, *Lab Chip*, vol. 18, no. 3, pp. 395–405, Jan. 2018.
- [155] S. Giannoukakos, S. D’Ambrosi, D. Koppers-Lalic, C. Gomez-Martin, A. Fernandez, and M. Hackenberg, “Assessing the complementary information from an increased number of biologically relevant features in liquid biopsy-derived RNA-Seq data,” en, *Heliyon*, vol. 10, no. 6, e27360, Mar. 2024.

- [156] C. Pedraz-Valdunciel, S. Giannoukakos, N. Potie, A. Gimenez-Capitan, C.-Y. Huang, M. Hackenberg, A. Fernandez-Hilario, J. Bracht, M. Filipiska, E. Aldeguer, S. Rodriguez, T. G. Bivona, S. Warren, C. Aguado, M. Ito, A. Aguilar-Hernandez, M. A. Molina-Vila, and R. Rosell, “Digital multiplexed analysis of circular RNAs in FFPE and fresh non-small cell lung cancer specimens,” en, *Mol. Oncol.*, vol. 16, no. 12, pp. 2367–2383, Jun. 2022.
- [157] D. Fortunato, S. Giannoukakos, A. Gimenez-Capitan, M. Hackenberg, M. A. Molina-Vila, and N. Zarovni, “Selective isolation of extracellular vesicles from minimally processed human plasma as a translational strategy for liquid biopsies,” en, *Biomark. Res.*, vol. 10, no. 1, p. 57, Aug. 2022.
- [158] C. Pedraz-Valdunciel, S. Giannoukakos, A. Gimenez-Capitan, D. Fortunato, M. Filipiska, J. Bertran-Alamillo, J. W. P. Bracht, A. Drozdowskyj, J. Valarezo, N. Zarovni, A. Fernandez-Hilario, M. Hackenberg, A. Aguilar-Hernandez, M. A. Molina-Vila, and R. Rosell, “Multiplex analysis of CircRNAs from plasma extracellular vesicle-enriched samples for the detection of early-stage non-small cell lung cancer,” en, *Pharmaceutics*, vol. 14, no. 10, p. 2034, Sep. 2022.
- [159] S. D’Ambrosi, S. Giannoukakos, M. Antunes-Ferreira, C. Pedraz-Valdunciel, J. W. P. Bracht, N. Potie, A. Gimenez-Capitan, M. Hackenberg, A. Fernandez Hilario, M. A. Molina-Vila, R. Rosell, T. Würdinger, and D. Koppers-Lalic, “Combinatorial blood platelets-derived circRNA and mRNA signature for early-stage lung cancer detection,” en, *Int. J. Mol. Sci.*, vol. 24, no. 5, Mar. 2023.
- [160] C. Pedraz-Valdunciel, M. Ito, S. Giannoukakos, A. Gimenez-Capitan, M. A. Molina-Vila, and R. Rosell, “Brief report: Circular runt-related transcription factor (circRUNX1) as potential biomarker for cancer recurrence in EGFR mutation-positive surgically resected NSCLC,” en, *JTO Clin. Res. Rep.*, no. 100604, p. 100 604, Nov. 2023.
- [161] [Online]. Available: <https://github.com/s-andrews/FastQC>.
- [162] B. Bushnell, J. Rood, and E. Singer, “BBMerge – accurate paired shotgun read merging via overlap,” en, *PLoS One*, vol. 12, no. 10, e0185056, Oct. 2017.

- [163] P. Ewels, M. Magnusson, S. Lundin, and M. Käller, “MultiQC: Summarize analysis results for multiple tools and samples in a single report,” *Bioinformatics*, vol. 32, no. 19, pp. 3047–3048, Oct. 2016.
- [164] A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras, “STAR: Ultrafast universal RNA-seq aligner,” en, *Bioinformatics*, vol. 29, no. 1, pp. 15–21, Jan. 2013.
- [165] A. Frankish, M. Diekhans, I. Jungreis, J. Lagarde, J. E. Loveland, J. M. Mudge, C. Sisu, J. C. Wright, J. Armstrong, I. Barnes, A. Berry, A. Bignell, C. Boix, S. Carbonell Sala, F. Cunningham, T. Di Domenico, S. Donaldson, I. T. Fiddes, C. Garcia Giron, J. M. Gonzalez, T. Grego, M. Hardy, T. Hourlier, K. L. Howe, T. Hunt, O. G. Izuogu, R. Johnson, F. J. Martin, L. Martinez, S. Mohanan, P. Muir, F. C. P. Navarro, A. Parker, B. Pei, F. Pozo, F. C. Riera, M. Ruffier, B. M. Schmitt, E. Stapleton, M.-M. Suner, I. Sycheva, B. Uszczyńska-Ratajczak, M. Y. Wolf, J. Xu, Y. T. Yang, A. Yates, D. Zerbino, Y. Zhang, J. S. Choudhary, M. Gerstein, R. Guigo, T. J. P. Hubbard, M. Kellis, B. Paten, M. L. Tress, and P. Flicek, “GENCODE 2021,” en, *Nucleic Acids Res.*, vol. 49, no. D1, pp. D916–D923, Jan. 2021.
- [166] S. Uhrig, J. Ellermann, T. Walther, P. Burkhardt, M. Fröhlich, B. Hutter, U. H. Toprak, O. Neumann, A. Stenzinger, C. Scholl, S. Fröhling, and B. Brors, “Accurate and efficient detection of gene fusions from RNA sequencing data,” en, *Genome Res.*, vol. 31, no. 3, pp. 448–460, Mar. 2021.
- [167] By Broad Institute, *Picard Tools - By Broad Institute*, Dec. 2023. [Online]. Available: <http://broadinstitute.github.io/picard>.
- [168] L. Wang, S. Wang, and W. Li, “RSeQC: Quality control of RNA-seq experiments,” en, *Bioinformatics*, vol. 28, no. 16, pp. 2184–2185, Aug. 2012.
- [169] R. Patro, G. Duggal, M. I. Love, R. A. Irizarry, and C. Kingsford, “Salmon provides fast and bias-aware quantification of transcript expression,” en, *Nat. Methods*, vol. 14, no. 4, pp. 417–419, Apr. 2017.
- [170] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and 1000 Genome Project Data Processing Subgroup, “The sequence Alignment/Map format and SAMtools,” en, *Bioinformatics*, vol. 25, no. 16, pp. 2078–2079, Aug. 2009.

- [171] A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, and M. A. DePristo, “The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data,” en, *Genome Res.*, vol. 20, no. 9, pp. 1297–1303, Sep. 2010.
- [172] P. Danecek, J. K. Bonfield, J. Liddle, J. Marshall, V. Ohan, M. O. Pollard, A. Whitwham, T. Keane, S. A. McCarthy, R. M. Davies, and H. Li, “Twelve years of SAMtools and BCFtools,” en, *Gigascience*, vol. 10, no. 2, Feb. 2021.
- [173] S. T. Sherry, M. H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, and K. Sirotkin, “dbSNP: The NCBI database of genetic variation,” en, *Nucleic Acids Res.*, vol. 29, no. 1, pp. 308–311, Jan. 2001.
- [174] E. Picardi and G. Pesole, “REDIttools: High-throughput RNA editing detection made easy,” en, *Bioinformatics*, vol. 29, no. 14, pp. 1813–1814, Jul. 2013.
- [175] [Online]. Available: https://community.nanoporetech.com/docs/prepare/library_prep_protocols/Guppy-protocol/v/gpb_2003_v1_revax_14dec2018/guppy-software-overview.
- [176] W. De Coster and R. Rademakers, “NanoPack2: Population-scale evaluation of long-read sequencing data,” en, *Bioinformatics*, vol. 39, no. 5, May 2023.
- [177] K. Sahlin and P. Medvedev, “Error correction enables use of oxford nanopore technology for reference-free transcriptome analysis,” en, *Nat. Commun.*, vol. 12, no. 1, p. 2, Jan. 2021.
- [178] H. Li, “Minimap2: Pairwise alignment for nucleotide sequences,” en, *Bioinformatics*, vol. 34, no. 18, pp. 3094–3100, Sep. 2018.
- [179] A. Leger and T. Leonardi, “pycoQC, interactive quality control for oxford nanopore sequencing,” *J. Open Source Softw.*, vol. 4, no. 34, p. 1236, Feb. 2019.
- [180] [Online]. Available: <https://rseqc.sourceforge.net/>.

- [181] D. Wyman, G. Balderrama-Gutierrez, F. Reese, S. Jiang, S. Rahmanian, S. Forner, D. Matheos, W. Zeng, B. Williams, D. Trout, W. England, S.-H. Chu, R. C. Spitale, A. J. Tenner, B. J. Wold, and A. Mortazavi, “A technology-agnostic long-read analysis pipeline for transcriptome discovery and quantification,” Jun. 2019.
- [182] D. Wyman and A. Mortazavi, “TranscriptClean: Variant-aware correction of indels, mismatches and splice junctions in long-read transcripts,” en, *Bioinformatics*, vol. 35, no. 2, pp. 340–342, Jan. 2019.
- [183] S. Noguchi, T. Arakawa, S. Fukuda, M. Furuno, A. Hasegawa, F. Hori, S. Ishikawa-Kato, K. Kaida, A. Kaiho, M. Kanamori-Katayama, T. Kawashima, M. Kojima, A. Kubosaki, R.-I. Manabe, M. Murata, S. Nagao-Sato, K. Nakazato, N. Ninomiya, H. Nishiyori-Sueki, S. Noma, E. Saijyo, A. Saka, M. Sakai, C. Simon, N. Suzuki, M. Tagami, S. Watanabe, S. Yoshida, P. Arner, R. A. Axton, M. Babina, J. K. Baillie, T. C. Barnett, A. G. Beckhouse, A. Blumenthal, B. Bodega, A. Bonetti, J. Briggs, F. Brombacher, A. J. Carlisle, H. C. Clevers, C. A. Davis, M. Detmar, T. Dohi, A. S. B. Edge, M. Edinger, A. Ehrlund, K. Ekwall, M. Endoh, H. Enomoto, A. Eslami, M. Fagiolini, L. Fairbairn, M. C. Farach-Carson, G. J. Faulkner, C. Ferrai, M. E. Fisher, L. M. Forrester, R. Fujita, J.-I. Furusawa, T. B. Geijtenbeek, T. Gingeras, D. Goldowitz, S. Guhl, R. Guler, S. Gustincich, T. J. Ha, M. Hamaguchi, M. Hara, Y. Hasegawa, M. Herlyn, P. Heutink, K. J. Hitchens, D. A. Hume, T. Ikawa, Y. Ishizu, C. Kai, H. Kawamoto, Y. I. Kawamura, J. S. Kempfle, T. J. Kenna, J. Kere, L. M. Khachigian, T. Kitamura, S. Klein, S. P. Klinken, A. J. Knox, S. Kojima, H. Koseki, S. Koyasu, W. Lee, A. Lennartsson, A. Mackay-sim, N. Mejhert, Y. Mizuno, H. Morikawa, M. Morimoto, K. Moro, K. J. Morris, H. Motohashi, C. L. Mummery, Y. Nakachi, F. Nakahara, T. Nakamura, Y. Nakamura, T. Nozaki, S. Ogishima, N. Ohkura, H. Ohno, M. Ohshima, M. Okada-Hatakeyama, Y. Okazaki, V. Orlando, D. A. Ovchinnikov, R. Passier, M. Patrikakis, A. Pombo, S. Pradhan-Bhatt, X.-Y. Qin, M. Rehli, P. Rizzu, S. Roy, A. Sajantila, S. Sakaguchi, H. Sato, H. Satoh, S. Savvi, A. Saxena, C. Schmidl, C. Schneider, G. G. Schulze-Tanzil, A. Schwegmann, G. Sheng, J. W. Shin, D. Sugiyama, T. Sugiyama, K. M. Summers, N. Takahashi, J. Takai, H. Tanaka, H. Tatsukawa, A. Tomoiu, H. Toyoda, M. van de Wetering, L. M. van den Berg,

- R. Verardo, D. Vijayan, C. A. Wells, L. N. Winteringham, E. Wolvetang, Y. Yamaguchi, M. Yamamoto, C. Yanagi-Mizuochi, M. Yoneda, Y. Yonekura, P. G. Zhang, S. Zucchelli, I. Abugessaisa, E. Arner, J. Harshbarger, A. Kondo, T. Lassmann, M. Lizio, S. Sahin, T. Sengstag, J. Severin, H. Shimoji, M. Suzuki, H. Suzuki, J. Kawai, N. Kondo, M. Itoh, C. O. Daub, T. Kasukawa, H. Kawaji, P. Carninci, A. R. R. Forrest, and Y. Hayashizaki, “FANTOM5 CAGE profiles of human and mouse samples,” en, *Sci. Data*, vol. 4, no. 1, p. 170 112, Aug. 2017.
- [184] M. I. Love, W. Huber, and S. Anders, “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2,” en, *Genome Biol.*, vol. 15, no. 12, p. 550, 2014.
- [185] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, “Edger: A bioconductor package for differential expression analysis of digital gene expression data,” en, *Bioinformatics*, vol. 26, no. 1, pp. 139–140, Jan. 2010.
- [186] U. Raudvere, L. Kolberg, I. Kuzmin, T. Arak, P. Adler, H. Peterson, and J. Vilo, “G:profiler: A web server for functional enrichment analysis and conversions of gene lists (2019 update),” en, *Nucleic Acids Res.*, vol. 47, no. W1, W191–W198, Jul. 2019.
- [187] L. Kolberg, U. Raudvere, I. Kuzmin, J. Vilo, and H. Peterson, “Gprofiler2 – an R package for gene list functional enrichment analysis and namespace conversion toolset g:profiler,” en, *F1000Res.*, vol. 9, p. 709, Jul. 2020.
- [188] G. Yu, L.-G. Wang, Y. Han, and Q.-Y. He, “Clusterprofiler: An R package for comparing biological themes among gene clusters,” en, *OMICS*, vol. 16, no. 5, pp. 284–287, May 2012.
- [189] K. Vitting-Seerup and A. Sandelin, “IsoformSwitchAnalyzeR: Analysis of changes in genome-wide patterns of alternative splicing and its functional consequences,” en, *Bioinformatics*, vol. 35, no. 21, pp. 4469–4471, Nov. 2019.
- [190] J. Mistry, S. Chuguransky, L. Williams, M. Qureshi, G. A. Salazar, E. L. L. Sonnhammer, S. C. E. Tosatto, L. Paladin, S. Raj, L. J. Richardson, R. D. Finn, and A. Bateman, “Pfam: The protein families database in 2021,” en, *Nucleic Acids Res.*, vol. 49, no. D1, pp. D412–D419, Jan. 2021.

- [191] B. Meszaros, G. Erdős, and Z. Dosztanyi, “IUPred2A: Context-dependent prediction of protein disorder as a function of redox state and protein binding,” en, *Nucleic Acids Res.*, vol. 46, no. W1, W329–W337, Jul. 2018.
- [192] F. Teufel, J. J. Almagro Armenteros, A. R. Johansen, M. H. Gislason, S. I. Pihl, K. D. Tsirigos, O. Winther, S. Brunak, G. von Heijne, and H. Nielsen, “SignalP 6.0 predicts all five types of signal peptides using protein language models,” en, *Nat. Biotechnol.*, vol. 40, no. 7, pp. 1023–1025, Jul. 2022.
- [193] Y.-J. Kang, D.-C. Yang, L. Kong, M. Hou, Y.-Q. Meng, L. Wei, and G. Gao, “CPC2: A fast and accurate coding potential calculator based on sequence intrinsic features,” *Nucleic Acids Res.*, vol. 45, no. W1, W12–W16, Jul. 2017.
- [194] D. M. Bryant, K. Johnson, T. DiTommaso, T. Tickle, M. B. Couger, D. Payzin-Dogru, T. J. Lee, N. D. Leigh, T.-H. Kuo, F. G. Davis, J. Bateman, S. Bryant, A. R. Guzikowski, S. L. Tsai, S. Coyne, W. W. Ye, R. M. Freeman Jr, L. Peshkin, C. J. Tabin, A. Regev, B. J. Haas, and J. L. Whited, “A tissue-mapped axolotl DE novo transcriptome enables identification of limb regeneration factors,” en, *Cell Rep.*, vol. 18, no. 3, pp. 762–776, Jan. 2017.
- [195] [Online]. Available: <https://github.com/jts/nanopolish>.
- [196] N. J. Loman, J. Quick, and J. T. Simpson, “A complete bacterial genome assembled de novo using only nanopore sequencing data,” en, *Nat. Methods*, vol. 12, no. 8, pp. 733–735, Aug. 2015.
- [197] M. H. Celik and A. Mortazavi, “Analysis of alternative polyadenylation from long-read or short-read RNA-seq with LAPA,” Nov. 2022.
- [198] P. N. Pratanwanich, F. Yao, Y. Chen, C. W. Q. Koh, Y. K. Wan, C. Hendra, P. Poon, Y. T. Goh, P. M. L. Yap, J. Y. Chooi, W. J. Chng, S. B. Ng, A. Thiery, W. S. S. Goh, and J. Göke, “Identification of differential RNA modifications from nanopore direct RNA sequencing with xpore,” en, *Nat. Biotechnol.*, vol. 39, no. 11, pp. 1394–1402, Nov. 2021.
- [199] J. Vandesompele, K. De Preter, F. Pattyn, B. Poppe, N. Van Roy, A. De Paepe, and F. Speleman, *Genome Biol*, vol. 3, no. 7, research0034.1, 2002.

- [200] M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, and G. K. Smyth, “Limma powers differential expression analyses for RNA-sequencing and microarray studies,” en, *Nucleic Acids Res.*, vol. 43, no. 7, e47, Apr. 2015.
- [201] D. Risso, J. Ngai, T. P. Speed, and S. Dudoit, “Normalization of RNA-seq data using factor analysis of control genes or samples,” en, *Nat. Biotechnol.*, vol. 32, no. 9, pp. 896–902, Sep. 2014.
- [202] F. Ferrara, S. Zoupanou, E. Primiceri, Z. Ali, and M. S. Chiriaco, “Beyond liquid biopsy: Toward non-invasive assays for distanced cancer diagnostics in pandemics,” en, *Biosens. Bioelectron.*, vol. 196, no. 113698, p. 113 698, Jan. 2022.
- [203] S. Bratulic, F. Gatto, and J. Nielsen, “The translational status of cancer liquid biopsies,” en, *Regen. Eng. Transl. Med.*, vol. 7, no. 3, pp. 312–352, Sep. 2021.
- [204] A. Krishnan and S. Thomas, “Toward platelet transcriptomics in cancer diagnosis, prognosis and therapy,” en, *Br. J. Cancer*, vol. 126, no. 3, pp. 316–322, Feb. 2022.
- [205] B. M. Hussen, S. T. Abdullah, A. Salihi, D. K. Sabir, K. R. Sidiq, M. F. Rasul, H. J. Hidayat, S. Ghafouri-Fard, M. Taheri, and E. Jamali, “The emerging roles of NGS in clinical oncology and personalized medicine,” en, *Pathol. Res. Pract.*, vol. 230, no. 153760, p. 153 760, Feb. 2022.
- [206] E. Kilgour, D. G. Rothwell, G. Brady, and C. Dive, “Liquid biopsy-based biomarkers of treatment response and resistance,” en, *Cancer Cell*, vol. 37, no. 4, pp. 485–495, Apr. 2020.
- [207] E. Sanchez-Herrero, R. Serna-Blasco, V. Ivanchuk, R. Garcia-Campelo, M. Domine Gomez, J. M. Sanchez, B. Massuti, N. Reguart, C. Camps, S. Sanz-Moreno, S. Calabuig-Fariñas, E. Jantus-Lewintre, M. Arnal, D. Fernandez-Orth, V. Calvo, V. Gonzalez-Rumayor, M. Provencio, and A. Romero, “NGS-based liquid biopsy profiling identifies mechanisms of resistance to ALK inhibitors: A step toward personalized NSCLC treatment,” en, *Mol. Oncol.*, vol. 15, no. 9, pp. 2363–2376, Sep. 2021.
- [208] D. Shyr and Q. Liu, “Next generation sequencing in cancer research and clinical application,” en, *Biol. Proced. Online*, vol. 15, no. 1, p. 4, Feb. 2013.

- [209] M. C. Liefwaard, K. S. Moore, L. Mulder, D. van den Broek, J. Wesseling, G. S. Sonke, L. F. A. Wessels, M. Rookus, and E. H. Lips, “Tumour-educated platelets for breast cancer detection: Biological and technical insights,” en, *Br. J. Cancer*, vol. 128, no. 8, pp. 1572–1581, Apr. 2023.
- [210] S. Yu, Y. Li, Z. Liao, Z. Wang, Z. Wang, Y. Li, L. Qian, J. Zhao, H. Zong, B. Kang, W.-B. Zou, K. Chen, X. He, Z. Meng, Z. Chen, S. Huang, and P. Wang, “Plasma extracellular vesicle long RNA profiling identifies a diagnostic signature for the detection of pancreatic ductal adenocarcinoma,” en, *Gut*, vol. 69, no. 3, pp. 540–550, Mar. 2020.
- [211] M. Antunes-Ferreira, S. D’Ambrosi, M. Arkani, E. Post, S. G. J. G. In ’t Veld, J. Ramaker, K. Zwaan, E. D. Kucukguzel, L. E. Wedekind, A. W. Griffioen, M. Oude Egbrink, M. J. E. Kuijpers, D. van den Broek, D. P. Noske, K. J. Hartemink, S. Sabrkhany, I. Bahce, N. Sol, H.-J. Bogaard, D. Koppers-Lalic, M. G. Best, and T. Wurdinger, “Tumor-educated platelet blood tests for Non-Small cell lung cancer detection and management,” en, *Sci. Rep.*, vol. 13, no. 1, p. 9359, Jun. 2023.
- [212] K. Swanson, E. Wu, A. Zhang, A. A. Alizadeh, and J. Zou, “From patterns to patients: Advances in clinical machine learning for cancer diagnosis, prognosis, and treatment,” en, *Cell*, vol. 186, no. 8, pp. 1772–1791, Apr. 2023.
- [213] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
- [214] K. Katz, O. Shutov, R. Lapoint, M. Kimelman, J. R. Brister, and C. O’Sullivan, “The sequence read archive: A decade more of explosive growth,” en, *Nucleic Acids Res.*, vol. 50, no. D1, pp. D387–D390, Jan. 2022.
- [215] Uhrig Sebastian, *Arriba Software*, Feb. 2023. [Online]. Available: <https://github.com/suhrig/arriba>.

- [216] G. A. Van der Auwera, M. O. Carneiro, C. Hartl, R. Poplin, G. del Angel, A. Levy-Moonshine, T. Jordan, K. Shakir, D. Roazen, J. Thibault, E. Banks, K. V. Garimella, D. Altschuler, S. Gabriel, and M. A. DePristo, “From FastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline,” in *Current Protocols in Bioinformatics*. Hoboken, NJ, USA: John Wiley & Sons, Inc., Oct. 2013, pp. 11.10.1–11.10.33.
- [217] Calzolari Manuel, *sklearn-genetic*, Aug. 2023. [Online]. Available: <https://github.com/manuel-calzolari/sklearn-genetic>.
- [218] M. G. Best, N. Sol, S. G. J. G. In ‘t Veld, A. Vancura, M. Muller, A.-L. N. Niemeijer, A. V. Fejes, L.-A. Tjon Kon Fat, A. E. Huis In ‘t Veld, C. Leurs, T. Y. Le Large, L. L. Meijer, I. E. Kooi, F. Rustenburg, P. Schellen, H. Verschueren, E. Post, L. E. Wedekind, J. Bracht, M. Esenkbrink, L. Wils, F. Favaro, J. D. Schoonhoven, J. Tannous, H. Meijers-Heijboer, G. Kazemier, E. Giovannetti, J. C. Reijneveld, S. Idema, J. Killestein, M. Heger, S. C. de Jager, R. T. Urbanus, I. E. Hoefler, G. Pasterkamp, C. Mannhalter, J. Gomez-Arroyo, H.-J. Bogaard, D. P. Noske, W. P. Vandertop, D. van den Broek, B. Ylstra, R. J. A. Nilsson, P. Wesseling, N. Karachaliou, R. Rosell, E. Lee-Lewandrowski, K. B. Lewandrowski, B. A. Tannous, A. J. de Langen, E. F. Smit, M. M. van den Heuvel, and T. Wurdinger, “Swarm intelligence-enhanced detection of non-small-cell lung cancer using tumor-educated platelets,” en, *Cancer Cell*, vol. 32, no. 2, 238–252.e9, Aug. 2017.
- [219] S. Li, Y. Li, B. Chen, J. Zhao, S. Yu, Y. Tang, Q. Zheng, Y. Li, P. Wang, X. He, and S. Huang, “exoRBase: A database of circRNA, lncRNA and mRNA in human blood exosomes,” en, *Nucleic Acids Res.*, vol. 46, no. D1, pp. D106–D112, Jan. 2018.
- [220] Y. Li, Q. Zheng, C. Bao, S. Li, W. Guo, J. Zhao, D. Chen, J. Gu, X. He, and S. Huang, “Circular RNA is enriched and stable in exosomes: A promising biomarker for cancer diagnosis,” en, *Cell Res.*, vol. 25, no. 8, pp. 981–984, Aug. 2015.
- [221] T. Liu, X. Wang, W. Guo, F. Shao, Z. Li, Y. Zhou, Z. Zhao, L. Xue, X. Feng, Y. Li, F. Tan, K. Zhang, Q. Xue, S. Gao, Y. Gao, and J. He, “RNA sequencing of tumor-educated platelets reveals a three-gene

- diagnostic signature in esophageal squamous cell carcinoma,” en, *Front. Oncol.*, vol. 12, p. 824354, May 2022.
- [222] L. Xu, X. Li, X. Li, X. Wang, Q. Ma, D. She, X. Lu, J. Zhang, Q. Yang, S. Lei, L. Wang, and Z. Wang, “RNA profiling of blood platelets noninvasively differentiates colorectal cancer from healthy donors and noncancerous intestinal diseases: A retrospective cohort study,” en, *Genome Med.*, vol. 14, no. 1, p. 26, Mar. 2022.
- [223] C. Scheepbouwer, M. Hackenberg, M. A. J. van Eijndhoven, A. Gerber, M. Pegtel, and C. Gomez-Martin, “NORMSEQ: A tool for evaluation, selection and visualization of RNA-Seq normalization methods,” en, *Nucleic Acids Res.*, vol. 51, no. W1, W372–W378, Jul. 2023.
- [224] C. Evans, J. Hardin, and D. M. Stoebel, “Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions,” en, *Brief. Bioinform.*, vol. 19, no. 5, pp. 776–792, Sep. 2018.
- [225] K. C. Thandra, A. Barsouk, K. Saginala, J. S. Aluru, and A. Barsouk, “Epidemiology of lung cancer,” en, *Contemp. Oncol. (Pozn.)*, vol. 25, no. 1, pp. 45–52, Feb. 2021.
- [226] A. Leiter, R. R. Veluswamy, and J. P. Wisnivesky, “The global burden of lung cancer: Current status and future trends,” en, *Nat. Rev. Clin. Oncol.*, vol. 20, no. 9, pp. 624–639, Sep. 2023.
- [227] J. Liu, W. Lee, Z. Jiang, Z. Chen, S. Jhunjhunwala, P. M. Haverty, F. Gnad, Y. Guan, H. N. Gilbert, J. Stinson, C. Klijn, J. Guillory, D. Bhatt, S. Vartanian, K. Walter, J. Chan, T. Holcomb, P. Dijkgraaf, S. Johnson, J. Koeman, J. D. Minna, A. F. Gazdar, H. M. Stern, K. P. Hoefflich, T. D. Wu, J. Settleman, F. J. de Sauvage, R. C. Gentleman, R. M. Neve, D. Stokoe, Z. Modrusan, S. Seshagiri, D. S. Shames, and Z. Zhang, “Genome and transcriptome sequencing of lung cancers reveal diverse mutational and splicing events,” en, *Genome Res.*, vol. 22, no. 12, pp. 2315–2327, Dec. 2012.
- [228] C. Martinez-Ruiz, J. R. M. Black, C. Puttick, M. S. Hill, J. Demeulemeester, E. Larose Cadieux, K. Thol, T. P. Jones, S. Veeriah, C. Naceur-Lombardelli, A. Toncheva, P. Prymas, A. Rowan, S. Ward, L. Cubitt, F. Athanasopoulou, O. Pich, T. Karasaki, D. A. Moore, R. Salgado, E. Colliver, C. Castignani, M. Dietzen, A. Huebner, M. Al Bakir, M. Tanic, T. B. K. Watkins, E. L. Lim, A. M. Al-Rashed,

- D. Lang, J. Clements, D. E. Cook, R. Rosenthal, G. A. Wilson, A. M. Frankell, S. de Carne Trecesson, P. East, N. Kanu, K. Litchfield, N. J. Birkbak, A. Hackshaw, S. Beck, P. Van Loo, M. Jamal-Hanjani, TRACERx Consortium, C. Swanton, and N. McGranahan, “Genomic-transcriptomic evolution in lung cancer and metastasis,” en, *Nature*, vol. 616, no. 7957, pp. 543–552, Apr. 2023.
- [229] Y. Tian, Q. Li, Z. Yang, S. Zhang, J. Xu, Z. Wang, H. Bai, J. Duan, B. Zheng, W. Li, Y. Cui, X. Wang, R. Wan, K. Fei, J. Zhong, S. Gao, J. He, C. M. Gay, J. Zhang, J. Wang, and F. Tang, “Single-cell transcriptomic profiling reveals the tumor heterogeneity of small-cell lung cancer,” en, *Signal Transduct. Target. Ther.*, vol. 7, no. 1, p. 346, Oct. 2022.
- [230] M. Jain, R. Abu-Shumays, H. E. Olsen, and M. Akeson, “Advances in nanopore direct RNA sequencing,” en, *Nat. Methods*, vol. 19, no. 10, pp. 1160–1164, Oct. 2022.
- [231] *Mirvana™ mirna isolation kit, with phenol*. [Online]. Available: <https://www.thermofisher.com/order/catalog/product/AM1560>.
- [232] V. Newman, B. Moore, H. Sparrow, and E. Perry, “The ensembl genome browser: Strategies for accessing eukaryotic genome data,” in *Eukaryotic Genomic Databases: Methods and Protocols*, M. Kollmar, Ed. New York, NY: Springer New York, 2018, pp. 115–139, ISBN: 978-1-4939-7737-6. DOI: 10.1007/978-1-4939-7737-6_6. [Online]. Available: https://doi.org/10.1007/978-1-4939-7737-6_6.
- [233] G. Yu, L.-G. Wang, Y. Han, and Q.-Y. He, “ClusterProfiler: An R package for comparing biological themes among gene clusters,” en, *OMICS*, vol. 16, no. 5, pp. 284–287, May 2012.
- [234] [Online]. Available: <https://github.com/YuLab-SMU/enrichplot>.
- [235] W. J. Kent, C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler, Haussler, and David, “The human genome browser at UCSC,” en, *Genome Res.*, vol. 12, no. 6, pp. 996–1006, Jun. 2002.
- [236] A. R. Quinlan and I. M. Hall, “BEDTools: A flexible suite of utilities for comparing genomic features,” en, *Bioinformatics*, vol. 26, no. 6, pp. 841–842, Mar. 2010.

- [237] [Online]. Available:
<https://github.com/nanoporetech/pipeline-polya-diff>.
- [238] A. M. Young, S. Van Buren, and N. U. Rashid, “Differential transcript usage analysis incorporating quantification uncertainty via compositional measurement error regression modeling,” en, *Biostatistics*, Apr. 2023.
- [239] W. Jiang and L. Chen, “Alternative splicing: Human disease and quantitative analysis from high-throughput sequencing,” en, *Comput. Struct. Biotechnol. J.*, vol. 19, pp. 183–195, 2021.
- [240] R. Batra, M. Manchanda, and M. S. Swanson, “Global insights into alternative polyadenylation regulation,” en, *RNA Biol.*, vol. 12, no. 6, pp. 597–602, 2015.
- [241] H. Gujar, D. J. Weisenberger, and G. Liang, “The roles of human DNA methyltransferases and their isoforms in shaping the epigenome,” en, *Genes (Basel)*, vol. 10, no. 2, p. 172, Feb. 2019.
- [242] H. Sudo, A. B. Tsuji, A. Sugyo, M. Abe, O. Hino, and T. Saga, “AHNAK is highly expressed and plays a key role in cell migration and invasion in mesothelioma,” en, *Int. J. Oncol.*, vol. 44, no. 2, pp. 530–538, Feb. 2014.
- [243] M. Sohn, S. Shin, J.-Y. Yoo, Y. Goh, I. H. Lee, and Y. S. Bae, “Ahnak promotes tumor metastasis through transforming growth factor- β -mediated epithelial-mesenchymal transition,” en, *Sci. Rep.*, vol. 8, no. 1, p. 14379, Sep. 2018.
- [244] J. Gleeson, A. Leger, Y. D. J. Prawer, T. A. Lane, P. J. Harrison, W. Haerty, and M. B. Clark, “Accurate expression quantification from nanopore direct RNA sequencing with NanoCount,” en, *Nucleic Acids Res.*, vol. 50, no. 4, e19, Feb. 2022.
- [245] C. Sonesson, Y. Yao, A. Bratus-Neuenschwander, A. Patrignani, M. D. Robinson, and S. Hussain, “A comprehensive examination of nanopore native RNA sequencing for characterization of complex transcriptomes,” en, *Nat. Commun.*, vol. 10, no. 1, p. 3359, Jul. 2019.
- [246] D. R. Garalde, E. A. Snell, D. Jachimowicz, B. Sipos, J. H. Lloyd, M. Bruce, N. Pantic, T. Admassu, P. James, A. Warland, M. Jordan, J. Ciccone, S. Serra, J. Keenan, S. Martin, L. McNeill, E. J. Wallace, L. Jayasinghe, C. Wright, J. Blasco, S. Young, D. Brocklebank, S. Juul, J. Clarke, A. J. Heron, and D. J. Turner, “Highly parallel direct RNA

- sequencing on an array of nanopores,” en, *Nat. Methods*, vol. 15, no. 3, pp. 201–206, Mar. 2018.
- [247] P. Uapinyoying, J. Goecks, S. M. Knoblach, K. Panchapakesan, C. G. Bonnemann, T. A. Partridge, J. K. Jaiswal, and E. P. Hoffman, “A long-read RNA-seq approach to identify novel transcripts of very large genes,” en, *Genome Res.*, vol. 30, no. 6, pp. 885–897, Jun. 2020.
- [248] R. E. Workman, A. D. Tang, P. S. Tang, M. Jain, J. R. Tyson, R. Razaghi, P. C. Zuzarte, T. Gilpatrick, A. Payne, J. Quick, N. Sadowski, N. Holmes, J. G. de Jesus, K. L. Jones, C. M. Soulette, T. P. Snutch, N. Loman, B. Paten, M. Loose, J. T. Simpson, H. E. Olsen, A. N. Brooks, M. Akeson, and W. Timp, “Nanopore native RNA sequencing of a human poly(a) transcriptome,” en, *Nat. Methods*, vol. 16, no. 12, pp. 1297–1305, Dec. 2019.
- [249] Y. Chen, N. M. Davidson, Y. K. Wan, H. Patel, F. Yao, H. M. Low, C. Hendra, L. Watten, A. Sim, C. Sawyer, V. Iakovleva, P. L. Lee, L. Xin, H. E. V. Ng, J. M. Loo, X. Ong, H. Q. A. Ng, J. Wang, W. Q. C. Koh, S. Y. P. Poon, D. Stanojevic, H.-D. Tran, K. H. E. Lim, S. Y. Toh, P. A. Ewels, H.-H. Ng, N. G. Iyer, A. Thiery, W. J. Chng, L. Chen, R. DasGupta, M. Sikic, Y.-S. Chan, B. O. P. Tan, Y. Wan, W. L. Tam, Q. Yu, C. C. Khor, T. Wüstefeld, P. N. Pratanwanich, M. I. Love, W. S. S. Goh, S. B. Ng, A. Oshlack, J. Göke, and SG-NEx consortium, “A systematic benchmark of nanopore long read RNA sequencing for transcript level analysis in human cell lines,” Apr. 2021.
- [250] D. A. Glinos, G. Garborcauskas, P. Hoffman, N. Ehsan, L. Jiang, A. Gokden, X. Dai, F. Aguet, K. L. Brown, K. Garimella, T. Bowers, M. Costello, K. Ardlie, R. Jian, N. R. Tucker, P. T. Ellinor, E. D. Harrington, H. Tang, M. Snyder, S. Juul, P. Mohammadi, D. G. MacArthur, T. Lappalainen, and B. B. Cummings, “Transcriptome variation in human tissues revealed by long-read sequencing,” en, *Nature*, vol. 608, no. 7922, pp. 353–359, Aug. 2022.
- [251] M. S. Bang, K. Kang, J.-J. Lee, Y.-J. Lee, J. E. Choi, J. Y. Ban, and C.-H. Oh, “Transcriptome analysis of non-small cell lung cancer and genetically matched adjacent normal tissues identifies novel prognostic marker genes,” en, *Genes Genomics*, vol. 39, no. 3, pp. 277–284, Mar. 2017.

- [252] M. Pertea, G. M. Pertea, C. M. Antonescu, T.-C. Chang, J. T. Mendell, and S. L. Salzberg, “StringTie enables improved reconstruction of a transcriptome from RNA-seq reads,” en, *Nat. Biotechnol.*, vol. 33, no. 3, pp. 290–295, Mar. 2015.
- [253] A. Bankevich, S. Nurk, D. Antipov, A. A. Gurevich, M. Dvorkin, A. S. Kulikov, V. M. Lesin, S. I. Nikolenko, S. Pham, A. D. Prjibelski, A. V. Pyshkin, A. V. Sirotkin, N. Vyahhi, G. Tesler, M. A. Alekseyev, and P. A. Pevzner, “SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing,” en, *J. Comput. Biol.*, vol. 19, no. 5, pp. 455–477, May 2012.
- [254] P. Jenjaroenpun, T. Wongsurawat, T. D. Wadley, T. M. Wassenaar, J. Liu, Q. Dai, V. Wanchai, N. S. Akel, A. Jamshidi-Parsian, A. T. Franco, G. Boysen, M. L. Jennings, D. W. Ussery, C. He, and I. Nookaew, “Decoding the epitranscriptional landscape from native RNA sequences,” en, *Nucleic Acids Res.*, vol. 49, no. 2, e7, Jan. 2021.
- [255] P. Gkogkolou and M. Böhm, “Advanced glycation end products: Key players in skin aging?” en, *Dermatoendocrinol.*, vol. 4, no. 3, pp. 259–270, Jul. 2012.
- [256] S. F. Yan, R. Ramasamy, and A. M. Schmidt, “Receptor for AGE (RAGE) and its ligands-cast into leading roles in diabetes and the inflammatory response,” en, *J. Mol. Med.*, vol. 87, no. 3, pp. 235–247, Mar. 2009.
- [257] K. Prasad, “AGE-RAGE stress: A changing landscape in pathology and treatment of alzheimer’s disease,” en, *Mol. Cell. Biochem.*, vol. 459, no. 1-2, pp. 95–112, Sep. 2019.
- [258] E. A. Oczypok, T. N. Perkins, and T. D. Oury, “All the “RAGE” in lung disease: The receptor for advanced glycation endproducts (RAGE) is a major mediator of pulmonary inflammatory responses,” en, *Paediatr. Respir. Rev.*, vol. 23, pp. 40–49, Jun. 2017.
- [259] Q. Wang, W. Zhu, G. Xiao, M. Ding, J. Chang, and H. Liao, “Effect of AGER on the biological behavior of non-small cell lung cancer H1299 cells,” en, *Mol. Med. Rep.*, vol. 22, no. 2, pp. 810–818, Aug. 2020.
- [260] W. Zhang, J. Fan, Q. Chen, C. Lei, B. Qiao, and Q. Liu, “SPP1 and AGER as potential prognostic biomarkers for lung adenocarcinoma,” *Oncol. Lett.*, Mar. 2018.

- [261] M.-C. Chen, K.-C. Chen, G.-C. Chang, H. Lin, C.-C. Wu, W.-H. Kao, C.-L. J. Teng, S.-L. Hsu, and T.-Y. Yang, "Rage acts as an oncogenic role and promotes the metastasis of human lung cancer," *Cell Death and Disease*, vol. 11, no. 4, Apr. 2020, ISSN: 2041-4889. DOI: 10.1038/s41419-020-2432-1. [Online]. Available: <http://dx.doi.org/10.1038/s41419-020-2432-1>.
- [262] Y. Wang, J. Liu, B. O. Huang, Y.-M. Xu, J. Li, L.-F. Huang, J. Lin, J. Zhang, Q.-H. Min, W.-M. Yang, and X.-Z. Wang, "Mechanism of alternative splicing and its regulation," en, *Biomed. Rep.*, vol. 3, no. 2, pp. 152–158, Mar. 2015.
- [263] A. Reyes and W. Huber, "Alternative start and termination sites of transcription drive most transcript isoform differences across human tissues," en, *Nucleic Acids Res.*, vol. 46, no. 2, pp. 582–592, Jan. 2018.
- [264] Y. Wang, J. Zhang, Y.-J. Li, N.-N. Yu, W.-T. Liu, J.-Z. Liang, W. W. Xu, Z.-H. Sun, B. Li, and Q.-Y. He, "MEST promotes lung cancer invasion and metastasis by interacting with VCP to activate NF- κ B signaling," en, *J. Exp. Clin. Cancer Res.*, vol. 40, no. 1, p. 301, Sep. 2021.
- [265] H. Nakanishi, T. Suda, M. Katoh, A. Watanabe, T. Igishi, M. Kodani, S. Matsumoto, M. Nakamoto, Y. Shigeoka, T. Okabe, M. Oshimura, and E. Shimizu, "Loss of imprinting of PEG1/MEST in lung cancer cell lines," en, *Oncol. Rep.*, vol. 12, no. 6, pp. 1273–1278, Dec. 2004.
- [266] M. S. Kim, H. S. Lee, Y. J. Kim, D. Y. Lee, S. G. Kang, and W. Jin, "MEST induces twist-1-mediated EMT through STAT3 activation in breast cancers," en, *Cell Death Differ.*, vol. 26, no. 12, pp. 2594–2606, Dec. 2019.
- [267] G. Blobel and B. Dobberstein, "Transfer of proteins across membranes. i. presence of proteolytically processed and unprocessed nascent immunoglobulin light chains on membrane-bound ribosomes of murine myeloma," en, *J. Cell Biol.*, vol. 67, no. 3, pp. 835–851, Dec. 1975.
- [268] B. Tian, J. Hu, H. Zhang, and C. S. Lutz, "A large-scale analysis of mRNA polyadenylation of human and mouse genes," en, *Nucleic Acids Res.*, vol. 33, no. 1, pp. 201–212, Jan. 2005.
- [269] H. Chang, J. Lim, M. Ha, and V. N. Kim, "TAIL-seq: Genome-wide determination of poly(a) tail length and 3' end modifications," en, *Mol. Cell*, vol. 53, no. 6, pp. 1044–1052, Mar. 2014.

- [270] I. S. Harris, A. E. Treloar, S. Inoue, M. Sasaki, C. Gorrini, K. C. Lee, K. Y. Yung, D. Brenner, C. B. Knobbe-Thomsen, M. A. Cox, A. Elia, T. Berger, D. W. Cescon, A. Adeoye, A. Brüstle, S. D. Molyneux, J. M. Mason, W. Y. Li, K. Yamamoto, A. Wakeham, H. K. Berman, R. Khokha, S. J. Done, T. J. Kavanagh, C.-W. Lam, and T. W. Mak, “Glutathione and thioredoxin antioxidant pathways synergize to drive cancer initiation and progression,” en, *Cancer Cell*, vol. 27, no. 2, pp. 211–222, Feb. 2015.
- [271] E. Sparrow and M. D. Bodman-Smith, “Granulysin: The attractive side of a natural born killer,” en, *Immunol. Lett.*, vol. 217, pp. 126–132, Jan. 2020.
- [272] A. Kishi, Y. Takamori, K. Ogawa, S. Takano, S. Tomita, M. Tanigawa, M. Niman, T. Kishida, and S. Fujita, “Differential expression of granulysin and perforin by NK cells in cancer patients and correlation of impaired granulysin expression with progression of cancer,” en, *Cancer Immunol. Immunother.*, vol. 50, no. 11, pp. 604–614, Jan. 2002.
- [273] F. Perros, S. Cohen-Kaminsky, N. Gambaryan, B. Girerd, N. Raymond, I. Klingelschmitt, A. Huertas, O. Mercier, E. Fadel, G. Simonneau, M. Humbert, P. Dorfmueller, and D. Montani, “Cytotoxic cells and granulysin in pulmonary arterial hypertension and pulmonary veno-occlusive disease,” en, *Am. J. Respir. Crit. Care Med.*, vol. 187, no. 2, pp. 189–196, Jan. 2013.
- [274] C. Carbone, G. Piro, V. Merz, F. Simionato, R. Santoro, C. Zecchetto, G. Tortora, and D. Melisi, “Angiopoietin-Like proteins in angiogenesis, inflammation and cancer,” *Int. J. Mol. Sci.*, vol. 19, no. 2, p. 431, Feb. 2018.
- [275] M. Endo, M. Nakano, T. Kadomatsu, S. Fukuhara, H. Kuroda, S. Mikami, T. Hato, J. Aoi, H. Horiguchi, K. Miyata, H. Odagiri, T. Masuda, M. Harada, H. Horio, T. Hishima, H. Nomori, T. Ito, Y. Yamamoto, T. Minami, S. Okada, T. Takahashi, N. Mochizuki, H. Iwase, and Y. Oike, “Tumor cell-derived angiopoietin-like protein ANGPTL2 is a critical driver of metastasis,” en, *Cancer Res.*, vol. 72, no. 7, pp. 1784–1794, Apr. 2012.

- [276] K. Meng, S. Lu, Y.-Y. Li, L.-L. Hu, J. Zhang, Y. Cao, Y. Wang, C. Z. Zhang, and Q.-Y. He, “LINC00493-encoded microprotein SMIM26 exerts anti-metastatic activity in renal cell carcinoma,” en, *EMBO Rep.*, vol. 24, no. 6, e56282, Jun. 2023.
- [277] B. Chen, J.-H. Cha, M. Yan, N. Cao, P. Ye, X. Yan, and W.-H. Yang, “ATXN7L3B promotes hepatocellular carcinoma stemness and is downregulated by metformin,” en, *Biochem. Biophys. Res. Commun.*, vol. 573, pp. 1–8, Oct. 2021.
- [278] E. Y. Leberfarb, A. O. Degtyareva, I. I. Brusentsov, V. N. Maximov, M. I. Voevoda, A. I. Autenshlus, D. V. Morozov, A. V. Sokolov, and T. I. Merkulova, “Potential regulatory SNPs in the ATXN7L3B and KRT15 genes are associated with gender-specific colorectal cancer risk,” en, *Per. Med.*, vol. 17, no. 1, pp. 43–54, Jan. 2020.
- [279] X. Yang, Y. Zhang, Z. Xue, Y. Hu, W. Zhou, Z. Xue, X. Liu, G. Liu, W. Li, X. Liu, X. Li, M. Han, and J. Wang, “TRIM56 promotes malignant progression of glioblastoma by stabilizing cIAP1 protein,” en, *J. Exp. Clin. Cancer Res.*, vol. 41, no. 1, p. 336, Dec. 2022.
- [280] Y. Chen, J. Zhao, D. Li, J. Hao, P. He, H. Wang, and M. Zhang, “TRIM56 suppresses multiple myeloma progression by activating TLR3/TRIF signaling,” en, *Yonsei Med. J.*, vol. 59, no. 1, p. 43, 2018.
- [281] K. Lu, Y. Sui, and L. Fu, “Identification of TRIM56 as a potential biomarker for lung adenocarcinoma,” en, *Cancer Manag. Res.*, vol. 13, pp. 2201–2213, Mar. 2021.
- [282] P. Ershov, E. Yablokov, Y. Mezentsev, and A. Ivanov, “Uncharacterized proteins CxORF_x: Subinteractome analysis and prognostic significance in cancers,” en, *Int. J. Mol. Sci.*, vol. 24, no. 12, Jun. 2023.
- [283] Jan. 2024. [Online]. Available: <https://www.ncbi.nlm.nih.gov/gene/27086#:~:text=The%20FoxP1%20gene%20regulates%20lung,in%20Scarring%20After%20Glaucoma%20Surgery..>
- [284] W. Shu, H. Yang, L. Zhang, M. M. Lu, and E. E. Morrissey, “Characterization of a new subfamily of winged-helix/forkhead (fox) genes that are expressed in the lung and act as transcriptional repressors,” en, *J. Biol. Chem.*, vol. 276, no. 29, pp. 27 488–27 497, Jul. 2001.

- [285] J. Xiao, B. He, Y. Zou, X. Chen, X. Lu, M. Xie, W. Li, S. He, S. You, and Q. Chen, “Prognostic value of decreased FOXP1 protein expression in various tumors: A systematic review and meta-analysis,” en, *Sci. Rep.*, vol. 6, no. 1, Jul. 2016.
- [286] J. Akaishi, M. Onda, J. Okamoto, S. Miyamoto, M. Nagahama, K. Ito, A. Yoshida, and K. Shimizu, “Down-regulation of transcription elongation factor a (SII) like 4 (TCEAL4) in anaplastic thyroid cancer,” en, *BMC Cancer*, vol. 6, no. 1, p. 260, Nov. 2006.
- [287] Y. Sun and J. Zhao, “Transcription elongation factor a (SII)-like (TCEAL) gene family member-TCEAL2: A novel prognostic marker in pan-cancer,” en, *Cancer Inform.*, vol. 21, p. 11 769 351 221 126 285, Sep. 2022.
- [288] C.-Y. Huang, Y.-M. Chen, J.-J. Zhao, Y.-B. Chen, S.-S. Jiang, S.-M. Yan, B.-W. Zhao, K. Pan, D.-D. Wang, L. Lv, Y.-F. Li, W. Wang, Z.-W. Zhou, and J.-C. Xia, “Decreased expression of transcription elongation factor a-like 7 is associated with gastric adenocarcinoma prognosis,” en, *PLoS One*, vol. 8, no. 1, e54671, Jan. 2013.
- [289] S.-F. Zhang, X.-Y. Wang, Z.-Q. Fu, Q.-H. Peng, J.-Y. Zhang, F. Ye, Y.-F. Fu, C.-Y. Zhou, W.-G. Lu, X.-D. Cheng, and X. Xie, “TXNDC17 promotes paclitaxel resistance via inducing autophagy in ovarian cancer,” en, *Autophagy*, vol. 11, no. 2, pp. 225–238, 2015.
- [290] F. Wang, X. Chen, X. Yu, and Q. Lin, “Degradation of CCNB1 mediated by APC11 through UBA52 ubiquitination promotes cell cycle progression and proliferation of non-small cell lung cancer cells,” en, *Am. J. Transl. Res.*, vol. 11, no. 11, pp. 7166–7185, Nov. 2019.
- [291] Y. Liu, M. Zhao, and H. Qu, “A database of lung cancer-related genes for the identification of subtype-specific prognostic biomarkers,” en, *Biology (Basel)*, vol. 12, no. 3, Feb. 2023.
- [292] S. Wei, C. Hao, X. Li, H. Zhao, J. Chen, and Q. Zhou, “Effects of BTG2 on proliferation inhibition and anti-invasion in human lung cancer cells,” en, *Tumour Biol.*, vol. 33, no. 4, pp. 1223–1230, Aug. 2012.
- [293] B. Mao, Z. Zhang, and G. Wang, “BTG2: A rising star of tumor suppressors (review),” en, *Int. J. Oncol.*, vol. 46, no. 2, pp. 459–464, Feb. 2015.

- [294] L. Yuniati, B. Scheijen, L. T. van der Meer, and F. N. van Leeuwen, "Tumor suppressors BTG1 and BTG2: Beyond growth control," en, *J. Cell. Physiol.*, vol. 234, no. 5, pp. 5379–5389, May 2019.
- [295] H.-S. Kim, R. G. Rosenfeld, and Y. Oh, "Biological roles of insulin-like growth factor binding proteins (IGFBPs)," en, *Exp. Mol. Med.*, vol. 29, no. 2, pp. 85–96, Jun. 1997.
- [296] X. Cai, L. Wang, X. Wang, and F. Hou, "Silence of IGFBP7 suppresses apoptosis and epithelial mesenchymal transformation of high glucose induced-podocytes," en, *Exp. Ther. Med.*, vol. 16, no. 2, pp. 1095–1102, Aug. 2018.
- [297] Y. Chen, M. Pacyna-Gengelbach, F. Ye, T. Knösel, P. Lund, N. Deutschmann, K. Schlüns, W. F. M. A. Kotb, C. Sers, H. Yasumoto, T. Usui, and I. Petersen, "Insulin-like growth factor binding protein-related protein 1 (IGFBP-rP1) has potential tumour-suppressive activity in human lung cancer," en, *J. Pathol.*, vol. 211, no. 4, pp. 431–438, Mar. 2007.
- [298] J. Okamura, Y. Huang, D. Moon, M. Brait, X. Chang, and M. S. Kim, "Downregulation of insulin-like growth factor-binding protein 7 in cisplatin-resistant non-small cell lung cancer," en, *Cancer Biol. Ther.*, vol. 13, no. 3, pp. 148–155, Feb. 2012.
- [299] W. Zhao, J. Wang, B. Zhu, Y. Duan, F. Chen, W. Nian, J. Sun, B. Zhang, Z. Tong, and Z. Chen, "IGFBP7 functions as a potential lymphangiogenesis inducer in non-small cell lung carcinoma," en, *Oncol. Rep.*, vol. 35, no. 3, pp. 1483–1492, Mar. 2016.
- [300] L. Jin, F. Shen, M. Weinfeld, and C. Sergi, "Insulin growth factor binding protein 7 (IGFBP7)-related cancer and IGFBP3 and IGFBP7 crosstalk," en, *Front. Oncol.*, vol. 10, p. 727, May 2020.
- [301] A. Hall, "Rho family GTPases," en, *Biochem. Soc. Trans.*, vol. 40, no. 6, pp. 1378–1382, Dec. 2012.
- [302] R. Rathinam, "Role of rho GTPases and their regulators in cancer progression," en, *Front. Biosci.*, vol. 16, no. 1, p. 2561, 2011.
- [303] A. P. Wheeler and A. J. Ridley, "Why three rho proteins? RhoA, RhoB, RhoC, and cell motility," en, *Exp. Cell Res.*, vol. 301, no. 1, pp. 43–49, Nov. 2004.

- [304] Y. Shikada, I. Yoshino, T. Okamoto, S. Fukuyama, T. Kameyama, and Y. Maehara, “Higher Expression of RhoC Is Related to Invasiveness in Non-Small Cell Lung Carcinoma,” *Clinical Cancer Research*, vol. 9, no. 14, pp. 5282–5286, Nov. 2003, ISSN: 1078-0432. eprint: <https://aacrjournals.org/clincancerres/article-pdf/9/14/5282/2087768/df1403005282.pdf>.
- [305] G. Sita, A. Graziosi, P. Hrelia, and F. Morroni, “Sulforaphane causes cell cycle arrest and apoptosis in human glioblastoma U87MG and U373MG cell lines under hypoxic conditions,” en, *Int. J. Mol. Sci.*, vol. 22, no. 20, p. 11 201, Oct. 2021.
- [306] A. Morán, T. Fernández-Marcelo, J. Carro, C. De Juan, I. Pascua, J. Head, A. Gómez, F. Hernando, A.-J. Torres, M. Benito, and P. Iniesta, “Methylation profiling in non-small cell lung cancer: Clinical implications,” en, *Int. J. Oncol.*, vol. 40, no. 3, pp. 739–746, Mar. 2012.
- [307] J. Y. Kim, M.-J. Kim, J. S. Lee, J. Son, D.-H. Kim, J. S. Lee, S.-K. Jeong, E. Chun, and K.-Y. Lee, “Stratifin (SFN) regulates lung cancer progression via nucleating the Vps34-BECN1-TRAF6 complex for autophagy induction,” en, *Clin. Transl. Med.*, vol. 12, no. 6, e896, Jun. 2022.
- [308] R. E. Husni, A. Shiba-Ishii, T. Nakagawa, T. Dai, Y. Kim, J. Hong, S. Sakashita, N. Sakamoto, Y. Sato, and M. Noguchi, “DNA hypomethylation-related overexpression of SFN, GORASP2 and ZYG11A is a novel prognostic biomarker for early stage lung adenocarcinoma,” en, *Oncotarget*, vol. 10, no. 17, pp. 1625–1636, Feb. 2019.
- [309] *RNA-based regulation in human health and disease*, en, ser. Translational Epigenetics. San Diego, CA: Academic Press, Aug. 2020.
- [310] L. Romão, “MRNA metabolism in health and disease,” en, *Biomedicines*, vol. 10, no. 9, p. 2262, Sep. 2022.
- [311] S. Anders and W. Huber, “Differential expression analysis for sequence count data,” en, *Genome Biol.*, vol. 11, no. 10, R106, Oct. 2010.
- [312] M. J. Heller, “DNA microarray technology: Devices, systems, and applications,” en, *Annu. Rev. Biomed. Eng.*, vol. 4, no. 1, pp. 129–153, Mar. 2002.

- [313] N. Majumdar, T. Wessel, and J. Marks, “Digital PCR modeling for maximal sensitivity, dynamic range and measurement precision,” en, *PLoS One*, vol. 10, no. 3, e0118833, Mar. 2015.
- [314] R. Hitzemann, D. Bottomly, P. Darakjian, N. Walter, O. Iancu, R. Searles, B. Wilmot, and S. McWeeney, “Genes, behavior and next-generation RNA sequencing,” en, *Genes Brain Behav.*, vol. 12, no. 1, pp. 1–12, Feb. 2013.
- [315] C. D. Brumbaugh, H. J. Kim, M. Giovacchini, and N. Pourmand, “NanoStriDE: Normalization and differential expression analysis of NanoString ncounter data,” en, *BMC Bioinformatics*, vol. 12, no. 1, p. 479, Dec. 2011.
- [316] M. M. Kulkarni, “Digital multiplexed gene expression analysis using the NanoString ncounter system,” in *Current Protocols in Molecular Biology*. Hoboken, NJ, USA: John Wiley & Sons, Inc., Apr. 2011.
- [317] H.-F. Tsang, V. W. Xue, S.-P. Koh, Y.-M. Chiu, L. P.-W. Ng, and S.-C. C. Wong, “NanoString, a novel digital color-coded barcode technology: Current and future applications in molecular diagnostics,” en, *Expert Rev. Mol. Diagn.*, vol. 17, no. 1, pp. 95–103, Jan. 2017.
- [318] J. P. Bruce, A. B. Y. Hui, W. Shi, B. Perez-Ordóñez, I. Weinreb, W. Xu, B. Haibe-Kains, D. M. Waggott, P. C. Boutros, B. O’Sullivan, J. Waldron, S. H. Huang, E. X. Chen, R. Gilbert, and F.-F. Liu, “Identification of a microRNA signature associated with risk of distant metastasis in nasopharyngeal carcinoma,” en, *Oncotarget*, vol. 6, no. 6, pp. 4537–4550, Feb. 2015.
- [319] C. A. Class, C. J. Lukan, C. A. Bristow, and K.-A. Do, “Easy NanoString ncounter data analysis with the NanoTube,” en, *Bioinformatics*, vol. 39, no. 1, Jan. 2023.
- [320] P. T. Inc. (2015). Collaborative data science, [Online]. Available: <https://plot.ly>.
- [321] L. C. Gandolfo and T. P. Speed, “RLE plots: Visualizing unwanted variation in high dimensional data,” en, *PLoS One*, vol. 13, no. 2, e0191629, Feb. 2018.

- [322] Y. S. Low, C. Blöcker, J. R. McPherson, S. A. Tang, Y. Y. Cheng, J. Y. S. Wong, C. Chua, T. K. H. Lim, C. L. Tang, M. H. Chew, P. Tan, I. B. Tan, S. G. Rozen, and P. Y. Cheah, “A formalin-fixed paraffin-embedded (FFPE)-based prognostic signature to predict metastasis in clinically low risk stage I/II microsatellite stable colorectal cancer,” *Cancer Lett.*, vol. 403, pp. 13–20, Sep. 2017.
- [323] I. J. Park, Y. S. Yu, B. Mustafa, J. Y. Park, Y. B. Seo, G.-D. Kim, J. Kim, C. M. Kim, H. D. Noh, S.-M. Hong, Y. W. Kim, M.-J. Kim, A. A. Ansari, L. Buonaguro, S.-M. Ahn, and C.-S. Yu, “A nine-gene signature for predicting the response to preoperative chemoradiotherapy in patients with locally advanced rectal cancer,” en, *Cancers (Basel)*, vol. 12, no. 4, p. 800, Mar. 2020.
- [324] [Online]. Available: <https://www.python.org/>.
- [325] [Online]. Available: <https://numpy.org/>.
- [326] [Online]. Available: <https://pandas.pydata.org/>.
- [327] [Online]. Available: <https://biopython.org/>.
- [328] [Online]. Available: <https://pypi.org/project/fast-ml/>.
- [329] [Online]. Available: <https://matplotlib.org/>.
- [330] [Online]. Available: <https://seaborn.pydata.org/>.
- [331] [Online]. Available: <https://github.com/ewels/MultiQC>.
- [332] [Online]. Available: <https://sourceforge.net/projects/bbmap/>.
- [333] [Online]. Available: <https://github.com/alexdobin/STAR>.
- [334] [Online]. Available: <https://combine-lab.github.io/salmon/>.
- [335] [Online]. Available: <https://github.com/suhrig/arriba>.
- [336] [Online]. Available: <https://github.com/broadinstitute/picard>.
- [337] [Online]. Available: <https://github.com/broadinstitute/gatk>.
- [338] [Online]. Available: <https://github.com/samtools/bcftools>.
- [339] [Online]. Available: <https://github.com/BioinfoUNIBA/REDItools>.
- [340] [Online]. Available: <https://github.com/samtools/samtools>.
- [341] [Online]. Available: <https://www.r-project.org/>.
- [342] [Online]. Available: <https://CRAN.R-project.org/package=dplyr>.

- [343] [Online]. Available: <https://CRAN.R-project.org/package=reshape>.
- [344] [Online]. Available:
<https://CRAN.R-project.org/package=data.table>.
- [345] [Online]. Available: <https://bioconductor.org/packages/release/bioc/html/edgeR.html>.
- [346] [Online]. Available: <https://bioconductor.org/packages/release/bioc/html/DESeq2.html>.
- [347] [Online]. Available: <https://bioconductor.org/packages/release/bioc/html/RUVSeq.html>.
- [348] [Online]. Available: <https://CRAN.R-project.org/package=ggplot2>.
- [349] [Online]. Available:
<https://CRAN.R-project.org/package=RColorBrewer>.
- [350] [Online]. Available: <https://bioconductor.org/packages/release/bioc/html/BiocParallel.html>.
- [351] [Online]. Available: <https://CRAN.R-project.org/package=tidyverse>.
- [352] [Online]. Available:
<https://cran.r-project.org/package=gprofiler2>.
- [353] M. Tardaguila, L. de la Fuente, C. Marti, C. Pereira, F. J. Pardo-Palacios, H. del Risco, M. Ferrell, M. Mellado, M. Macchietto, K. Verheggen, M. Edelmann, I. Ezkurdia, J. Vazquez, M. Tress, A. Mortazavi, L. Martens, S. Rodriguez-Navarro, V. Moreno-Manzano, and A. Conesa, “SQANTI: Extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification,” en, *Genome Res.*, vol. 28, no. 3, pp. 396–411, Mar. 2018.
- [354] G. K. Geiss, R. E. Bumgarner, B. Birditt, T. Dahl, N. Dowidar, D. L. Dunaway, H. P. Fell, S. Ferree, R. D. George, T. Grogan, J. J. James, M. Maysuria, J. D. Mitton, P. Oliveri, J. L. Osborn, T. Peng, A. L. Ratcliffe, P. J. Webster, E. H. Davidson, L. Hood, and K. Dimitrov, “Direct multiplexed measurement of gene expression with color-coded probe pairs,” en, *Nat. Biotechnol.*, vol. 26, no. 3, pp. 317–325, Mar. 2008.

Appendix

This appendix provides the additional content that supplements the main body of the thesis. It includes detailed figures, tables, and other relevant material pertinent to each chapter.

A.1 Assessing the complementary information of biologically relevant features in lbRNA-Seq data (Supplementary Materials)

A.1.1 Detailed analysis applied to the individual datasets

All datasets adhered to standardised principles of analysis as detailed in the Materials and Methods Section. While we fine-tuned specific ML parameters for each biofeature type, we consistently applied these parameters across all datasets sharing the same biofeature. This approach aimed to create a versatile yet biofeature-specific prediction model suitable for application to diverse datasets.

Each of the six datasets followed identical bioinformatics principles and analyses, outlined in Figure 4.2 A-G and detailed in Materials and Methods Sections 4.2.1 - 4.2.8. In the ML analysis for each dataset, a uniform scheme was applied, illustrated in Figure 4.2 E-G and described in Materials and Methods Section 4.2.9. In datasets with an independent validation set (NSCLC, CRC, and PDAC), the primary dataset underwent feature selection and training, with the independent validation set serving as the test set. For the remaining datasets (GBM, ESCC, and HCC), a random 70-30% split was used, with 70% for feature selection and training and the remaining 30% as the test set.

The Genetic Algorithm (GA) was employed across all datasets for feature selection, utilising stratified 5-fold cross-validation and empirical parameters: `n_population = 130`, `n_generations = 130`, `scoring = "accuracy"`, and `max_features = 50`. Each biofeature type utilised a distinct base estimator, selected based on its suitability for the specific data type. For Gene Expression data, the GA used the `RandomForestClassifier` as the base estimator, with the parameter `class_weight = 'balanced'`. Isoform Expression employed the GA with the `SVC` as the base estimator and parameters `C = 10`, `kernel = 'linear'`, `class_weight = 'balanced'`, and `probability = True`.

For the other biofeatures (FoCT, Gene Fusion, RNA editing, and SNVs), the GA used the `LogisticRegression` base estimator, each with distinct parameters:

- **FoCT:** `LogisticRegression` with parameters
`solver = 'liblinear', C = 1, class_weight = 'balanced'`.
- **Gene Fusion:** `LogisticRegression` with parameters
`solver = 'liblinear', class_weight = 'balanced', C = 0.1`.
- **RNA Editing:** `LogisticRegression` with parameters
`solver = 'liblinear', class_weight = 'balanced', C = 1`.
- **SNVs:** `LogisticRegression` with parameters
`solver = 'liblinear', class_weight = 'balanced', C = 5`.

Regarding classification models, different classifiers were employed for each biofeature:

- **Gene Expression and Isoform Expression:** `AdaBoost` algorithm with the `ExtraTreesClassifier` as the base estimator.
- **Other Biofeatures:** `LogisticRegression` classifier using the same base estimator and parameters employed in the GA for feature selection. For example, in FoCT, `LogisticRegression` was used with parameters `solver = 'liblinear', C = 1, class_weight = 'balanced'`.

Importantly, the specifics and individual parameters tuned for each biofeature are embedded in the `classification.py` script available on our GitHub page. Additionally, it's noteworthy that, to account for skewed class balance in some datasets, the parameter `class_weight = 'balanced'` was utilised across all algorithms.

A.1.2 Comprehensive elucidation of the Decision Output module

To gain further insight into how each biofeature contributes to the ultimate decision output, consider the presented example in Figure 4.2, illustrating the calculation of the final decision output for an imaginary sample through Soft Voting. As detailed in Materials and Methods Section 4.2.9 and Results Section 4.3.2 (Decision Output module), Soft Voting entails averaging all probabilities obtained from individual biofeatures.

Taking the example in Figure 4.2, denoted as sample N, Gene Expression analysis yields a probability of 0.21, classifying sample N as Non-Cancer based on the standard classification threshold of 0.5. Gene Fusion output probability is 0.74, classifying it as Cancer. RNA editing assigns a probability of 0.88, also classifying it as Cancer. SNV profiling gives a probability of 0.48, indicating Non-Cancer. Isoform Expression profiling assigns a probability of 0.9, classifying it as Cancer. Fraction of Canonical Transcript (FoCT) profiling yields a probability of 0.15, categorising it as Non-Cancer.

To determine the final decision for this imaginary sample N, the average probability is computed by summing all individual probabilities and dividing by the total number of biofeatures. In this example,

$$\text{AverageProbability} = \frac{0.21 + 0.74 + 0.88 + 0.48 + 0.90 + 0.15}{6} = 0.56 \quad (1)$$

As 0.56 exceeds the standard classification threshold of 0.5, Soft Voting classifies sample N as Cancer. It's worth noting that, in this theoretical example, the true label for sample N is indeed Cancer, demonstrating the correct classification by the Soft Voting ensemble learning approach compared to the misclassification by standard Gene Expression analysis.

A.1.3 Normalisation and Benchmarking of Gene and Isoform Expression Matrices

Gene-level expression data were initially obtained using the STAR aligner, while transcript-level expression data were obtained using STAR (with `quantMode` parameter set to `TranscriptomeSAM`) output in combination with the Salmon in alignment-based mode. Count matrices need to undergo normalisation to account for variations in read yield and technical artefacts. In this section, we meticulously benchmarked diverse normalisation methods for both gene and isoform expression matrices. The same methods were applied across all six datasets. Specifically, eight widely recognised normalisation techniques were tested, including TMM and TMMwsp from the `edgeR` (v3.28.1) package, RLE from the `DESeq2` (v1.26.0) package, RUV from the `RUVseq` (v1.20.0) package, Full Quantile (FQ) normalisation from the `preprocessCore` (v1.48.0) package, and `edgeR`'s Upper Quartile (UQ), CPM, and RPKM. To maintain methodological robustness and minimise the influence of machine learning models, six diverse classifiers were employed: `AdaBoost`, `KNN`, `linearSV`, `LogisticRegression`, `NaiveBayes`, and `RandomForest`. The evaluation of performance involved a consistent stratified 5-fold cross-validation across all datasets, with the average AUC serving as the metric (Figure 4.3 and Figure 2).

Preceding any normalisation or benchmarking of gene or isoform expression matrices, Principal Component Analysis (PCA) was executed to detect potential batch effects. Subsequent PCAs were conducted post-normalisation as well to ensure the efficacy of the applied normalisation techniques.

For the utilisation of the `RUVSeq` package, the `RUVg` function was adopted. Stable genes were identified using TMM and RLE normalisation, selecting non-differentially expressed genes with an adjusted p-value greater than 0.1. The common genes (intersection) from both TMM and RLE non-DE gene sets were used as reference genes, with a K value of 2 in the `RUVg` function.

A.1.4 Implementation of different ensemble learning techniques

We explored four distinct ensemble classification techniques to optimise the combination of the diverse biofeature types derived from each dataset. Soft voting, a technique suitable for probabilistic classifiers, was implemented by averaging the output probabilities generated by individual biofeature type models. In contrast, majority voting, also known as hard voting, made final decisions based on the majority class prediction from the different biofeature types. Finally, stacking, a more advanced approach, involved training a meta-learner, in our case, the GaussianNB classifier, to intelligently integrate predictions from the base models during training, facilitating more sophisticated decision-making and improving overall performance.

A.1.5 Calculation of Accuracy and Misclassification rate

For accuracy assessment, we employed the "accuracy_score" metric from the sklearn Python toolkit. To elaborate on the calculation, accuracy is determined as the ratio of correctly predicted instances to the total number of instances in the dataset, expressed mathematically as follows:

$$Accuracy = \frac{NumberofCorrectPredictions}{TotalNumberofInstances} \quad (2)$$

The misclassification rate, alternatively calculated as the percentage of misclassified samples over the total number of instances, can be expressed using the formula:

$$MisclassificationRate(\%) = \left(\frac{NumberofMisclassifiedSamples}{TotalNumberofInstances} \right) * 100 \quad (3)$$

Figure 3 presents the confusion matrices depicting the prediction outcomes for the test sets of the six datasets. These matrices provide a comprehensive summary of the ensemble learning (soft voting) predictions in matrix form, detailing the number of instances that were correctly and incorrectly classified.

A.1.6 Interpretation of the Biological Significance of Selected Features in the NSCLC Dataset

In the intricate landscape of the human cellular milieu, a consortium of genes congregated, each with a distinct function and purpose. Among the notable inhabitants were HLA-A and HLA-B, pivotal in antigen presentation, serving as sentinels at the cellular threshold.

Deep within this molecular community, NT5C3AP1 emerged as a conductor of gene orchestration, regulating the nuanced cadence of gene expression, harmonising their synchronised activities.

- B2M faithfully accompanied HLA-A and HLA-B, engaged in the meticulous process of antigen presentation, thereby fortifying the cell's defence mechanisms.
- CALM3, serving as a calcium signalling modulator, acted as a cellular pacemaker, orchestrating intracellular events with precision.
- CNST emerged as the architectural engineer, meticulously crafting intricate cellular structures, which facilitated the orchestration of molecular transactions.
- PLEK emerged as a versatile multitasker, adroitly coordinating multifaceted cellular processes, adapting to the ever-shifting cellular demands.
- HLA-DRB1 assumed the role of a molecular diplomat, mediating interactions between the immune system and the cell, ensuring judicious immune responses.
- MTFR1L, functioning as an adenosine triphosphate (ATP) producer, served as the cellular powerhouse, fuelling diverse cellular activities.
- FERMT3 stood as the molecular anchor, fortifying cellular interactions with the extracellular matrix, ensuring cellular adhesion and tissue stability.
- RTN4, akin to a cellular architect, sculpted the endoplasmic reticulum into intricate configurations, a testament to its artistic prowess.
- PIP4K2A functioned as the biochemical artisan, synthesising vital signalling molecules, catalysing fundamental cellular responses.

Together, this constellation of genes engendered a dynamic and symbiotic cellular environment, a symphony of molecular interactions contributing to the cellular homeostasis and, consequently, the overall well-being of the human organism.

A.1.7 Supplementary Figures

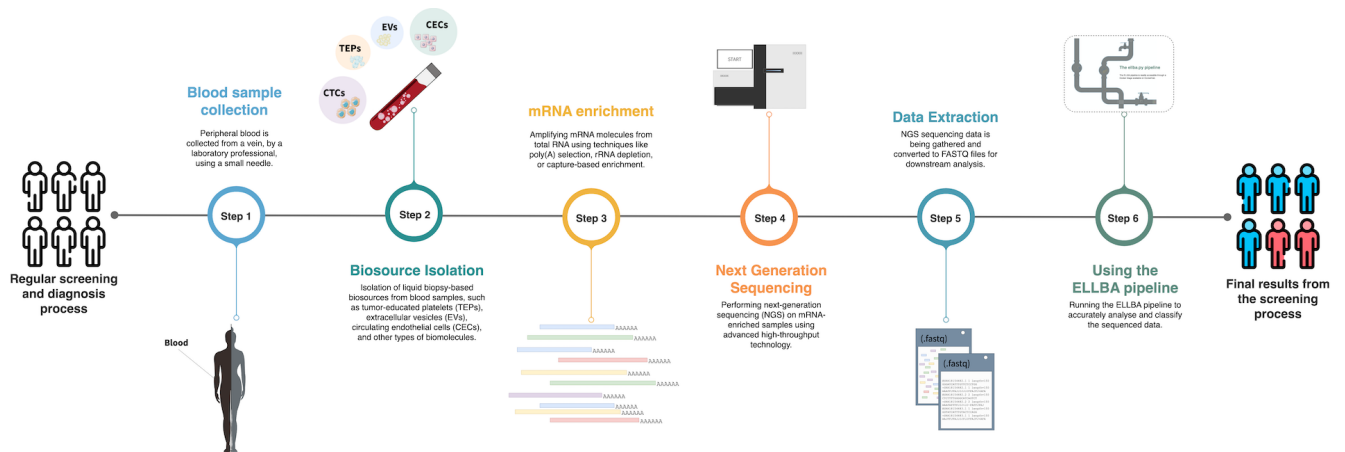


FIGURE 1: **ELBA Methodology Integration into Clinical Screening.** This schematic illustrates the integration of the ELBA methodology into the route clinical screening process. When an individual undergoes regular screening at a healthcare facility, a blood sample is collected, and one of the blood-based biosources is extracted. Subsequently, the content of this biosource can be sequenced, and the resulting raw data serves as input for the ELBA methodology. The ELBA pipeline efficiently processes the raw data and provides a final prediction regarding whether the individual may potentially develop the disease or not, aiding in early disease detection during clinical screening.

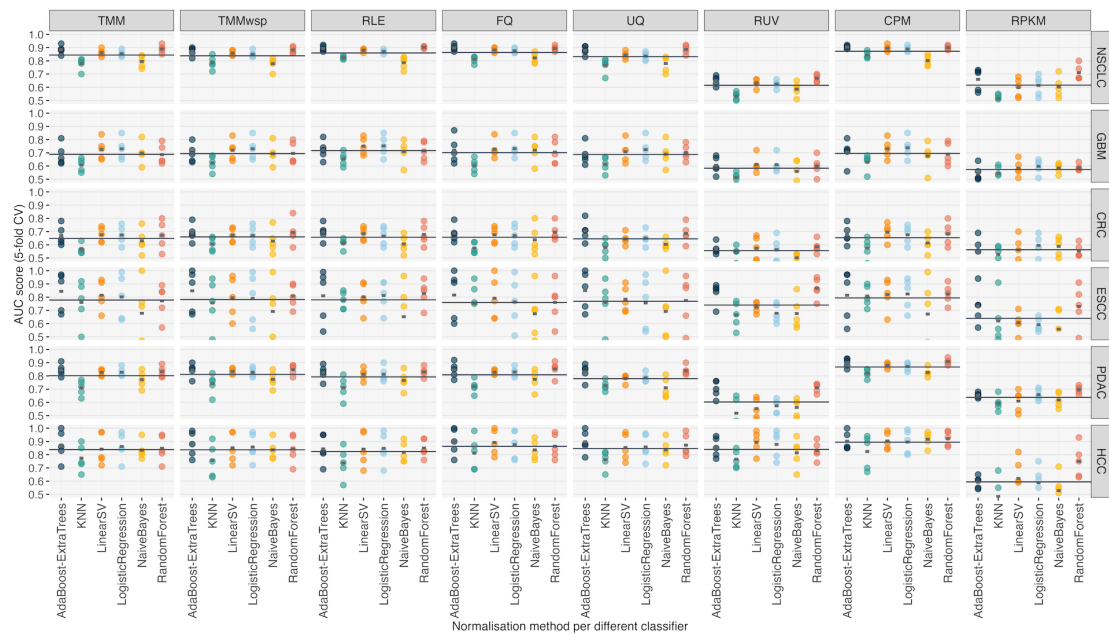


FIGURE 2: Overview of Different Normalisation Methods Applied to Isoform Expression. A comprehensive evaluation of eight normalisation techniques was conducted across all six datasets. Each column corresponds to a specific normalisation method, and each row pertains to a different dataset utilised. The x-axis illustrates the diverse models employed within each normalisation method, while the y-axis portrays the mean AUC score achieved through 5-fold CV. Each model is denoted by dots representing the AUC for each CV fold. Furthermore, a dashed line signifies the mean AUC across the 5 folds, while a solid line represents the mean AUC across all models.

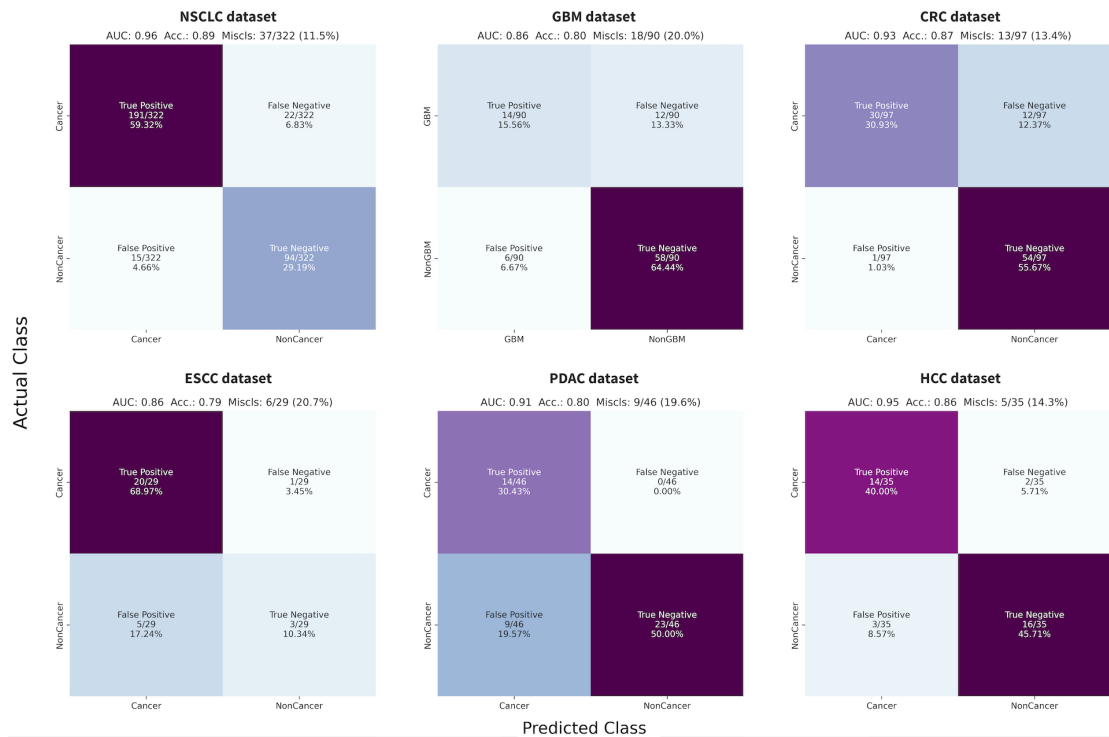


FIGURE 3: **Confusion Matrices for the Six Analysed Datasets.** Each confusion matrix illustrates the prediction outcomes for the respective test set of each dataset. Dataset names are specified in the title of each matrix. The x-axis denotes the predicted classes, and the y-axis represents the actual classes. The colour intensity within each quadrant of the confusion matrix corresponds to the percentage of instances for each class. Darker colours indicate a higher percentage, while lighter colours denote a lower class percentage, offering a visual representation of the soft voting classification performance.

A.1.8 Supplementary Tables

TABLE 1: List of software and packages used in the ELLBA software.

Software/Package	Version	Usage	Source
Python	v3.8.18	-	[324]
numpy	v1.23.5	Data manipulation	[325]
pandas	v1.5.2	Data manipulation	[326]
Bio	v1.79	Reading fastq file containing adapter sequences	[327]
scikit-learn	v1.2.0	Machine learning workflow	[213]
genetic_selection	v0.5.1	Utilisation of the Genetic algorithm for feature selection	[217]
fast_ml	v3.68	Detecting quasi-constant features	[328]
matplotlib	v3.6.3	Data visualisation	[329]
seaborn	v0.12.2	Data visualisation	[330]
FastQC	v0.11.9	Quality control checks on the raw sequence data	[161]
MultiQC	v1.13	Summary QC report	[331]
BBDuk	v38.18	Performing adapter removal and quality trimming	[332]
STAR	v2.7.6a	Aligning the RNA-Seq data against the reference genome	[333]
Salmon	v1.9.0	Quantifying the expression of transcripts using the RNA-seq data	[334]
Arriba	v2.1.0	Detection of gene fusions from the RNA-Seq data	[335]
RSeQC	v3.0.0	Quality control metrics on the aligned data	[180]
Picard	v2.23.3	Quality control metrics on the aligned data	[336]
GATK	v4.1.9.0	Preprocessing the aligned BAM files for variant calling	[337]
BCFtools	v1.7	Calling SNPs from the aligned BAM files	[338]
REDItools	v2.0	RNA editing profiling in the RNA-Seq data	[339]
samtools	v1.7	Indexing and deduplication on the aligned data	[340]
R	v3.6.3	-	[341]
dplyr	v1.0.7	expression filtering, alternative isoform expression, gene fusion filtering	[342]
reshape	v0.8.8	expression filtering, alternative isoform expression, gene fusion filtering	[343]
data.table	v1.14.0	expression filtering, alternative isoform expression, gene fusion filtering	[344]
edgeR	v3.28.1	expression filtering	[345]
DESeq2	v1.26.0	expression filtering	[346]
RUVSeq	v1.20.0	expression filtering	[347]
ggplot2	v3.3.5	expression filtering	[348]
RColorBrewer	v1.1.2	expression filtering	[349]
BiocParallel	v1.20.1	expression filtering	[350]
tidyverse	v1.3.1	gene fusion filtering	[351]
gprofiler2	v0.2.1	gene enrichment analysis	[352]

TABLE 2: Detailed machine learning configuration.

Biofeature Matrix	Biofeature Pre-processing	Feature Selection		Classifier	
		Algorithm	Base Estimator	Algorithm	Base Estimator
Gene expression	MinMaxScaler	Genetic Algorithm	RandomForestClassifier	AdaBoostClassifier	ExtraTreesClassifier
Isoform expression	MinMaxScaler	Genetic Algorithm	SVC	AdaBoostClassifier	ExtraTreesClassifier
FoCT	StandardScaler	Genetic Algorithm	LogisticRegression	LogisticRegression	-
Gene fusion	-	Genetic Algorithm	LogisticRegression	LogisticRegression	-
RNA editing	-	Genetic Algorithm	LogisticRegression	LogisticRegression	-
SNV	-	Genetic Algorithm	LogisticRegression	LogisticRegression	-

TABLE 3: Detailed overview of the publicly available datasets utilised in the study

Liquid Biopsy Biosource	TEPs	TEPs	TEPs	TEPs	EVs	CECs
Dataset	NSCLC	GBM	CRC	ESCC	PDAC	HCC
Cancer Type	Non-Small Cell Lung Cancer	Glioblastoma	Colorectal Cancer	Esophageal Squamous-Cell Carcinomas	Pancreatic Ductal Adenocarcinoma	Hepatocellular Carcinoma
Publication (DOI)	http://dx.doi.org/10.1016/j.ccell.2017.07.004	https://doi.org/10.1016/j.xcrm.2020.100101	https://doi.org/10.1186/s13073-022-01033-x	https://doi.org/10.3389/fonc.2022.824354	http://dx.doi.org/10.1136/gut.jnl-2019-318860	https://doi.org/10.1053/j.gastro.2018.09.020
BioProject Accession	PRJNA353588	PRJNA659491	PRJNA737596	PRJNA810728	PRJNA552230	PRJNA483004
Data Accession	GSE89843	GSE156902	NA	GSE197514	GSE133684	GSE117623
Sequencing Type	SE100 mRNA-Seq	SE100 mRNA-Seq	PE100 mRNA-Seq	PE250 mRNA-Seq	PE150 exLR-Seq	PE100 scRNA-Seq
Comparison	402 NSCLC vs. 377 non-Cancer	88 GBM vs. 212 nonGBM	132 CRC vs. 190 non-Cancer	71 ESCC vs. 25 non-Cancer	284 PDAC vs. 117 non-Cancer	52 HCC vs. 64 non-Cancer
Type of Samples with Abbreviations	NSCLC vs. Healthy Control (HC), MS, PH, EP, NSA, CP, UAP, SAP	GBM vs. BrM and MS	CRC vs. HC, CD, UC, Pl, Ad	ESCC vs. HC	PDAC vs. HC	HCC vs. CLD
Detailed Comparison	402 NSCLC vs. 234 HC, 58 MS, 34 PH, 21 EP, 13 NSA, 6 CP, 6 UAP, 5 SAP	88 GBM vs. 126 BrM and 86 MS	132 CRC vs. 59 Ad, 48 Pl, 40 CD, 22 UC, 21 HC	71 ESCC vs. 25 HC	284 PDAC vs. 117 HC	52 HCC vs. 64 CLD
Total Number of Samples	779	300	322	96	401	116
External Validation Set	YES	NO	YES	NO	YES	NO
Publication (DOI)	https://doi.org/10.1016/j.ccell.2022.08.006	NA	http://dx.doi.org/10.1016/j.ccell.2015.09.018	NA	https://doi.org/10.1093/nar/gkx891 & https://doi.org/10.1038/cr.2015.82	NA
BioProject Accession	PRJNA761450	NA	PRJNA281708	NA	PRJNA391134/PRJNA391134	NA
Data Accession	GSE183635	NA	GSE68086	NA	GSE100232/GSE100206	NA
Detailed Comparison	213 NSCLC vs. 75 HC, 20 EP, 13 PD, 1 MS	NA	42 CRC vs. 55 HC	NA	14 PDAC vs. 32 HC	NA
Total Number of Samples	322	NA	97	NA	46	NA

TABLE 4: Number of extracted features per dataset prior any filtering steps

	NSCLC	CRC	PDAC	HCC	GBM	ESCC
	SE100 mRNA-Seq	PE100 mRNA-Seq	PE150 exLR-Seq	PE100 scRNA-Seq	SE100 mRNA-Seq	PE250 mRNA-Seq
	TEPs	TEPs	EVs	CECs	TEPs	TEPs
Gene expression	17818	13356	14081	11587	19820	11633
Isoform expression	23095	30493	33979	8147	30171	34735
FoCT	1410	2300	3088	2478	2994	6283
Gene fusion	1783	1407	3444	173	1011	218
RNA editing	486	1674	1129	403	531	1360
SNV	367	1230	383	92	350	338
	External Validation Sets					
	SE100 mRNA-Seq	SE100 mRNA-Seq	PE150 exLR-Seq			
Gene expression	23812	17270	14368			
Isoform expression	33129	25521	33862			
FoCT	2947	3820	9258			
Gene fusion	711	855	1798			
RNA editing	1353	359	2040			
SNV	463	466	800			

TABLE 5: Summary of features selected for each dataset using the Genetic Algorithm.

NSCLC DATASET	GBM DATASET	CRC DATASET	ESCC DATASET	PDAC DATASET	HCC DATASET
Gene expression	Gene expression	Gene expression	Gene expression	Gene expression	Gene expression
ENSG00000210196.2	ENSG00000185436.12	ENSG00000110090.13	ENSG00000162407.9	ENSG00000007341.19	ENSG00000049249.9
ENSG00000140022.13	ENSG00000189280.3	ENSG00000163154.6	ENSG00000162607.13	ENSG00000145375.9	ENSG00000197056.11
ENSG00000165879.9	ENSG00000235927.4	ENSG00000050393.11	ENSG00000163785.13	ENSG00000186132.15	ENSG00000162367.11
ENSG00000257027.1	ENSG00000265491.5	ENSG00000258422.5	ENSG00000236548.2	ENSG00000104972.15	ENSG00000143013.13
ENSG00000131504.17	ENSG00000114978.18	ENSG00000198185.11	ENSG00000133112.17	ENSG00000128626.11	ENSG00000265491.5
ENSG00000177272.9	ENSG00000152127.9	ENSG00000272799.1	ENSG00000136146.15	ENSG00000115271.11	ENSG00000117533.15
ENSG00000126267.11	ENSG00000171596.7	ENSG00000269335.5	ENSG00000100575.14	ENSG00000236875.3	ENSG00000003509.16
ENSG00000166716.10	ENSG00000144681.10	ENSG00000140931.20	ENSG00000182809.11	ENSG00000033100.16	ENSG00000115310.18
ENSG00000175727.14	ENSG00000164086.10	ENSG00000184319.16	ENSG00000083457.12	ENSG00000035664.11	ENSG00000153250.20
ENSG00000111644.8	ENSG00000181004.10	ENSG00000090530.10	ENSG00000102362.15	ENSG00000252316.1	ENSG00000115806.13
ENSG00000103254.10	ENSG00000064652.11	ENSG00000106443.17	Isoform expression	ENSG00000114019.14	ENSG00000068745.15
ENSG00000168310.11	ENSG00000145990.11	ENSG00000184349.13	ENST00000491035.5	ENSG00000163554.15	ENSG00000145191.15
ENSG00000125971.16	ENSG00000130340.16	ENSG00000182195.9	ENST00000428056.6	ENSG00000131876.17	ENSG00000226435.10
ENSG00000155749.12	ENSG00000230487.8	ENSG00000228474.6	ENST00000430705.5	ENSG00000156738.18	ENSG00000053900.11
ENSG00000100889.12	ENSG00000106772.18	ENSG00000229117.9	ENST00000571428.5	ENSG00000177096.9	ENSG00000205302.7
ENSG00000257261.6	ENSG00000160446.19	ENSG00000136003.16	ENST00000564286.1	ENSG00000071626.17	ENSG00000119048.8
ENSG00000183878.15	ENSG00000197746.14	ENSG00000187951.11	ENST00000565777.1	ENSG00000204371.11	ENSG00000177683.14
ENSG00000092841.19	ENSG00000184743.13	ENSG00000186665.9	ENST00000457662.2	ENSG00000167037.18	ENSG00000121022.14
ENSG00000163374.19	ENSG00000214530.9	ENSG00000134222.16	ENST00000450573.5	ENSG00000167664.8	ENSG00000187210.14
ENSG00000114544.16	ENSG00000254659.3	ENSG00000143079.15	ENST00000393263.7	ENSG00000066629.18	ENSG00000168209.5
ENSG00000068976.14	ENSG00000281026.1	Isoform expression	ENST00000543258.1	ENSG00000161692.18	ENSG00000174456.15
ENSG00000285774.3	ENSG00000259330.3	ENST00000417615.1	ENST00000533239.1	ENSG00000105784.15	ENSG00000185787.15
ENSG00000155091.15	ENSG00000140400.17	ENST00000442956.1	ENST00000510824.5	ENSG00000167315.18	ENSG00000132507.18
ENSG00000198959.12	ENSG00000102981.9	ENST00000469375.1	ENST00000376582.7	ENSG00000024048.10	ENSG00000274180.1
ENSG00000235527.7	ENSG00000126653.18	ENST00000227348.9	ENST00000295984.7	ENSG00000154473.18	Isoform expression
ENSG00000100263.14	ENSG00000167664.8	ENST00000507146.5	ENST00000379907.9	ENSG00000144401.14	ENSG00000347635.9
ENSG00000103316.12	ENSG00000073050.12	ENST00000371856.6	ENST00000871059.7	ENSG00000198169.9	ENST00000263246.8
ENSG00000125354.23	ENSG00000183963.18	ENST00000467310.1	ENST00000298198.5	ENSG00000151702.17	ENST00000372409.8
ENSG00000114544.16	ENSG00000068366.20	ENST00000668018.1	ENST00000432042.5	ENSG00000165983.14	ENST00000356487.11
ENSG00000110628.16	ENSG00000198840.2	ENST00000372661.6	ENST00000538533.5	ENSG00000169221.14	ENST00000361574.10
ENSG00000105135.16	Isoform expression	FoCT	ENST00000496497.1	Isoform expression	ENST00000329138.9
ENSG00000042753.12	ENST00000062252.5	ENSG00000171490.13	ENST00000539528.5	ENST00000514985.6	ENST00000375215.3
ENSG00000148842.18	ENST00000317991.9	ENSG00000161203.13	ENST00000396151.7	ENST00000392221.5	ENST00000560044.5
ENSG00000130714.17	ENST00000436233.9	ENSG00000198836.10	ENST00000554076.5	ENST00000257497.11	ENST00000249636.11
ENSG00000132879.14	ENST00000469257.2	ENSG00000159840.16	ENST00000588995.5	ENST00000442394.5	ENST00000256383.10
ENSG00000261115.6	ENST00000556083.1	ENSG00000103512.15	ENST00000520339.5	ENST00000399702.5	ENST00000354586.5
ENSG00000225648.5	ENST00000437406.1	ENSG00000108846.16	ENST00000389858.4	ENST00000370331.3	ENST00000298125.7
ENSG00000117505.16	ENST00000513321.1	ENSG00000092199.17	ENST00000443698.5	ENST00000465399.5	ENST00000256015.5
ENSG00000224114.1	ENST00000606369.5	ENSG00000100380.14	ENST00000535933.5	ENST00000265769.9	ENST00000537653.5
Isoform expression	ENST00000653163.1	ENSG00000061676.15	ENST00000521644.5	ENST00000306730.8	ENST00000540163.5
ENST00000394547.7	ENST00000378495.7	ENSG00000125037.12	ENST00000458549.7	ENST00000546010.6	ENST00000514325.1
ENST00000482727.1	ENST00000497464.5	ENSG00000166337.10	ENST00000418688.5	ENST00000444448.6	ENST00000637561.1
ENST00000335968.8	ENST00000396894.8	ENSG00000005020.13	ENST00000392486.3	ENST00000376007.8	ENST00000295888.9
ENST00000261250.8	ENST00000614713.4	ENSG00000101079.21	ENST00000391865.7	ENST00000368309.4	ENST00000330953.6
ENST00000290716.13	ENST00000537020.5	ENSG00000073578.17	FoCT	ENST00000372922.8	ENST00000308488.10
ENST00000323374.8	ENST00000563856.1	ENSG00000127483.19	ENSG00000068323.17	ENST00000512986.5	ENST00000447998.7
ENST00000676278.1	ENST00000288599.9	ENSG00000124795.17	ENSG00000077721.16	ENST00000370192.8	ENST00000372663.9
ENST00000542854.5	ENST00000661691.1	ENSG00000091157.13	ENSG00000101337.16	ENST00000288840.10	ENST00000292357.7
ENST00000559302.1	ENST00000489673.1	ENSG00000164985.15	ENSG00000103121.9	ENST00000438527.7	ENST00000443035.8
ENST00000442128.2	ENST00000460569.1	ENSG00000100836.10	ENSG00000106462.11	ENST00000357137.9	ENST00000337859.11
ENST00000460495.5	ENST00000665523.1	ENSG00000049323.16	ENSG00000106477.20	ENST00000314622.9	ENST00000311893.14
ENST00000409028.8	ENST00000338961.11	ENSG00000043462.13	ENSG00000106772.18	ENST00000510585.3	ENST00000504430.5
ENST00000543139.1	ENST00000372667.9	ENSG00000129473.10	ENSG00000107862.5	ENST00000622512.1	ENST00000368599.4
ENST00000221130.11	ENST00000669077.1	ENSG00000234456.8	ENSG00000112234.9	ENST00000577246.5	ENST00000313349.3
ENST00000568399.1	ENST00000585931.5	ENSG00000131389.17	ENSG00000115271.11	ENST0000030085.13	ENST00000552459.2
ENST00000273432.8	ENST00000532444.5	ENSG00000164305.19	ENSG00000119636.16	ENST00000622683.5	ENST00000526627.1
ENST00000442834.6	FoCT	ENSG00000006125.18	ENSG00000121766.16	ENST00000606059.1	ENST00000356245.8
ENST00000591399.5	ENSG00000028116.18	ENSG00000094975.14	ENSG00000124635.9	ENST00000488782.1	ENST00000518219.5
ENST00000502665.1	ENSG00000066294.15	ENSG00000174695.10	ENSG00000126524.10	FoCT	ENST00000602637.1
ENST00000399494.5	ENSG00000099624.8	ENSG00000257267.3	ENSG00000127526.15	ENSG00000150316.12	ENST00000568588.5
ENST00000366769.7	ENSG00000100934.15	ENSG00000126581.13	ENSG00000128513.16	ENSG00000083168.11	ENST00000644876.2
ENST00000370277.5	ENSG00000102362.15	ENSG00000172053.18	ENSG00000130830.15	ENSG00000108061.12	ENST00000322428.10
ENST00000458001.2	ENSG00000104824.17	ENSG00000122779.18	ENSG00000132823.11	ENSG00000153071.15	ENSG00000446751.6
ENST00000507142.6	ENSG00000104964.15	ENSG00000110917.8	ENSG00000133703.13	ENSG00000113648.16	ENST00000536173.5
ENST00000343537.12	ENSG00000105372.8	ENSG00000223482.8	ENSG00000136279.21	ENSG00000179833.4	ENST00000265462.9
ENST00000496144.5	ENSG00000117523.16	ENSG00000136104.21	ENSG00000137275.14	ENSG00000144028.15	FoCT
ENST00000356674.7	ENSG00000128731.18	ENSG00000184009.12	ENSG00000139687.16	ENSG00000085224.23	ENSG00000069329.18

NSCLC DATASET	GRM DATASET	CRG DATASET	ESCC DATASET	PDAC DATASET	HCC DATASET
ENST000001037.9	ENSG00000134548.11	Gene fusion	ENSG00000158856.18	ENSG00000136536.15	ENSG00000082175.15
ENST00000379666.7	ENSG00000135018.14	ENSG00000143353.12	ENSG00000167258.15	ENSG00000140105.18	ENSG00000082515.18
ENST00000383463.9	ENSG00000136280.17	ENSG00000059377.18	ENSG0000016770.11	ENSG00000213741.11	ENSG00000100934.10
ENST00000361131.5	ENSG00000136754.17	ENSG00000225630.1	ENSG00000168894.10	ENSG00000156261.13	ENSG00000104529.17
ENST00000409939.8	ENSG00000136929.13	RNA editing	ENSG00000169313.10	ENSG00000135999.12	ENSG00000125991.20
ENST00000510223.1	ENSG00000136938.9	chr1.10283837.AC	chr1.10283837.AC	ENSG00000108671.14	ENSG00000134321.13
ENST00000557682.6	ENSG00000143409.15	chr4.73096373.CG	chr4.73096373.CG	ENSG00000178537.10	ENSG00000140538.16
ENST00000336095.10	ENSG00000143418.20	chr1.246667186.TC	ENSG00000180694.14	ENSG00000075785.14	ENSG00000143458.12
FoCT	ENSG00000145495.16	chr17.55419885.CT	ENSG00000181929.13	ENSG00000119392.15	ENSG00000145287.11
ENSG00000125037.12	ENSG00000147459.18	chr19.21572595.CA	ENSG00000196405.13	ENSG00000152558.15	ENSG00000145495.16
ENSG00000120866.11	ENSG00000149932.17	chr9.110244140.AG	ENSG00000197386.12	ENSG00000102734.13	ENSG00000179833.4
ENSG00000158796.17	ENSG00000160014.17	chr2.65270218.AG	ENSG00000263001.6	ENSG00000137309.20	ENSG00000184009.12
ENSG00000158405.5	ENSG00000180964.17	chr20.32703359.CT	Gene fusion	ENSG00000162889.11	ENSG00000186687.16
ENSG00000172466.16	ENSG00000180964.17	chr9.75146941.AG	ENSG00000113638.14	ENSG00000108563.14	ENSG00000109649.14
ENSG00000051382.9	Gene fusion	chr3.69170657.TA	RNA editing	ENSG0000021150.10	ENSG00000106776.10
ENSG000000198160.14	ENSG00000146416.19	chr14.56393712.AG	chr1.20650728.AT	ENSG00000085733.16	ENSG00000197102.12
ENSG00000166710.20	RNA editing	chr2.190203335.AT	chr1.22091783.GT	ENSG00000138107.13	ENSG00000250241.6
ENSG00000138434.17	chr1.246666314.TC	chr11.104948600.TC	chr1.86125458.AG	ENSG00000135968.21	ENSG00000256269.10
ENSG00000142599.19	chr2.214780740.CG	chr6.36741140.AG	chr1.247864917.GA	ENSG00000106211.10	ENSG00000267002.4
ENSG00000170802.16	chr3.93470433.AT	chr1.140072285.AC	chr3.56593570.AG	ENSG00000142230.13	ENSG00000273153.6
ENSG00000133633.18	chr3.93470433.AT	chr1.116917885.CT	chr1.13538671.CG	ENSG00000100162.22	Gene fusion
ENSG00000102054.18	chr4.117574884.TA	chr3.134960179.TC	chr6.170318526.TC	ENSG00000104216.16	NA
ENSG00000184178.16	chr3.31270252.TC	chr9.132901261.CG	chr7.5641153.CA	ENSG00000102316.17	RNA editing
ENSG00000189861.10	chr6.31356739.CA	chr11.22821609.AT	chr7.130047528.CG	ENSG00000105723.12	chr4.49710871.GT
ENSG00000160014.17	chr6.31356751.GT	chr10.22536667.CT	chr11.44619242.TC	ENSG00000135968.21	chr4.49711528.AG
ENSG00000166337.10	chr6.36741288.CG	chr22.39018761.CT	chr12.60192277.TC	ENSG00000216490.4	chr6.31270025.AG
ENSG00000137488.17	chr6.103967528.CA	chr17.77094744.AG	chr16.66579622.GA	ENSG00000151726.14	chr1.1319903.AG
ENSG00000151141.18	chr1.2732983.TC	chr20.51101314.AG	chr3.15659382.AG	ENSG00000105162.13	chr1.18214740.TC
ENSG00000189403.15	chr10.171816550.CT	chr17.50795126.CA	chrM.4769.AG	ENSG00000048649.13	chr11.5227013.AG
ENSG00000116688.18	chr12.30543572.TC	chr19.1106616.TC	chr2.3498938.AG	ENSG00000165168.8	chr10.1086353.CA
ENSG00000078596.11	chr16.89561665.GA	chr15.40036464.GA	chr8.106697739.TG	ENSO00000197894.11	KI270466.1.972.GA
ENSG00000188677.15	chr17.47313196.AG	chr1.40067594.TG	chr11.320606.GT	ENSG0000009995.19	chr1.192812042.CG
ENSG00000100387.9	chr17.47316373.CA	chr17.17264850.GA	chr13.21374438.TC	ENSG00000130955.17	chr6.29945489.GT
ENSG00000028363.1	chr19.1038222.CG	chr6.2991110.CT	chr14.50118530.CT	ENSG00000085872.15	chr6.29945567.GT
ENSG00000108061.12	chr17.4732983.TC	chr17.4732983.TC	chr19.5659382.AG	ENSG00000156136.10	chr1.18214740.TC
ENSG00000128294.16	chr19.353034.CT	chr6.31356827.TC	chr21.8214740.TC	Gene fusion	chrM.8364.AG
ENSG00000160789.21	chr19.55013640.AC	chr2.241352963.CT	chr2.411012075.AT	ENSG0000005189.16	chrM.8701.AG
ENSG00000143162.9	chr21.46664908.CG	chr6.31356374.CT	chr8.70227307.CG	RNA editing	KI270466.1.407.AG
ENSG00000163041.12	chr22.206239041.AG	chr22.20623936.GA	chr16.24092867.CT	KI270467.1.1753.TC	chr17.7014479.TC
ENSG00000108960.9	chrX.132074568.TA	chrX.116495595.GT	chr6.29942827.CG	chr1.166909861.GT	chr3.93470432.GA
ENSG00000162852.14	chr16.106998043.TC	chrX.152972245.GT	chr1.171587026.CC	chr2.68396142.TC	chr17.77486061.AG
ENSG00000248343.6	ENSG00000248343.6	SNV	chr5.180071596.CT	chr17.77486061.AG	chr1.33463690.CG
ENSG00000109614.18	chr3.123706855.GA	chr6.2994058.C.CC	chr18.11851811.CA	chr8.89600109.CT	chr11.54171512.CG
ENSG00000105372.8	chr5.15166342.GC	chr2.105894020.C.CACAC,CAC	chr16.56362518.CA	chr8.89600109.CT	chrM.1610.AG
ENSG00000103512.15	chr6.27122863.TC	chr8.51819594.TAAGATAA.CCT.CT	chr19.23832768.CT	chr16.400399.CT	KI270467.1.2133.TC
ENSG00000132824.14	chr6.29942953.TA	chr11.66276271.CCTT.C	chr17.3660774.GA	chr6.32660181.TC	KI270467.1.3101.AT
ENSG00000112335.15	chr6.149811183.AT	chr12.62885493.G.GG	chr2.201477679.GT	chr5.17354196.CG	chr11.5248354.TG
ENSG00000101096.20	chr7.13516461.GA	chr11.164223524.ACA	SNV	chr8.26655797.GT	chr5.49659878.CG
ENSG00000245532.9	chr10.22534868.CT	chr17.75319194.CT.C	chr1.13779904.T.TCCT.TTCT	chr2.214780740.CG	chr6.29944350.AG
ENSG00000104323.18	chr3.123706855.GA	chr20.51100907.G.CTG	chr1.50226707.CTCCCTT.C	chr6.29943462.TC	chr1.3793969.GA
ENSG00000113282.14	chr12.6537943.TC	chr2.230543149.AAA	chr1.223801167.A.T	chr2.37341798.CT	chr6.29942940.AG
ENSG00000206172.4	chr14.91378946.AG	chr14.51255393.GT.TTTT.TT.G	chr1.247878164.AAAATAAA.AAAAT.A	chr2.44065442.GT	chr6.31271180.AG
ENSG00000185418.16	chr17.68532637.CG	chr2.74154536.AAG.A	chr2.69514621.CTGT.C	chr6.31271153.AC	chr6.31356818.CG
ENSG00000180354.16	chr22.32479203.GA	chr16.9111907.AT.A	chr2.108771700.GA	chrM.16362.TC	chr6.31357115.CG
ENSG00000088726.16	chr22.39961075.AG	chr1.211314005.GT.F	chr3.183379296.C.T	chr7.135164528.TA	chr6.31357118.CG
Gene fusion	chrM.1610.AT	chr3.157261131.T.TT	chr4.117574910.A.G	chr19.46001278.TA	chr6.32578849.CG
ENSG00000167996.16	ENSG00000167996.16	chr3.30095841.CA	chr1.11753770.GT	chr6.29942953.TG	chr10.1086353.CA
RNA editing	chr9.12145206.CT	chr2.230543149.AAA	chr4.119628904.GG.C	chr8.5589027.GA	chr10.41859053.AG
chr11.48186831.AT	chrX.152972245.GT	chr11.118228725.AT.A	chr4.139730430.TCTGT.G	chr9.21699386.CT	chr16.87403158.AG
chr4.70585580.CT	chr2.68396015.TC	chr1.110672471.G.GGG	chr5.56882021.TCAACAACAA.TCAACAAT	chr6.31271273.GA	chr6.31271273.GA
chr21.10325858.GA	chr19.56379264.AG	chr17.66450796.T.TTT	chr5.137942873.A.C	chr12.48940979.CG	chr1.16162436.CT
chr4.117574857.CT	chr5.139330067.GA	chr14.54685187.ACA	chr6.350941.A.AA	chr10.110881295.GC	chr22.32479203.GA
chr8.30069055.AG	chr1.15630331.TG	chr15.28755013.CATT.C	chr6.29943338.G.C.A	chr4.9083845.TC	SNV
chr16.11672238.CG	chrM.11467.AG	chr7.0924377.A.GAA	chr6.31161865.A.C	chr4.10583558.CT	chr2.88836337.G.C
chr2.31267850.CG	chr13.18212056.AG	chr17.4034660.G.GG	chr6.31217273.GA.T	chr1.80576832.TC	chr2.105836337.A.AGCA
chr6.29943451.AT	chr1.150150179.AG	chr18.9400306.AA	chr6.31356423.G.C.A	chrY.10198498.TC	chr4.49153203.G.C
chrM.11719.GA	SNV	chr19.48210101.A.AGAAAA	chr6.80069652.A.C	chr19.574709.AT	chr4.73986201.A.G.T
chr6.33086636.TG	chr1.25832412.C.CTC	chr12.11893361.A.AAGAA	chr6.80070027.T.C	chr1.15807671.AG	chr5.49601888.A.G.C.T
chr1.151034063.CA	chr1.51840391.ATCT.A	chr14.102081436.A.ACA	chr6.159680958.A.AA.AAA	chr2.19740802.AG	chr6.29887984.A.G.T
chr19.41353016.GA	chr1.110659238.T.TTA	chr9.122847654.A.ATA	chr9.33294904.A.AA	chr6.11380752.TC	chr6.29886499.C.G.T
chr7.70374333.CG	chr1.117627844.CT.C	chr10.22536965.CCA.A	chr9.97927168.TACACAGC.TACACAC.T	chr11.12840090.A.G	chr6.29942940.A.G
chr12.10864693.TC	chr5.150965488.AAGG.A	chr10.460402912.T.TCT	chr11.6390705.TGCTGGC.T	chr1.30572118.CT	chr6.29942982.T.A
chr19.56379298.AG	chr3.121631459.A.ACA	chr1.110659238.T.TTA	chr11.27368542.T.C	chr9.13302018.GT	chr6.31271132.G.A.C
chr1.47180174.GT	chr3.127736188.GT.G	chr19.8438980.ACA	chr1.134970131.T.C	chr16.2765236.AC	chr6.31356825.T.G.A.C
chr7.66956001.TC	chr4.73984853.AGT.A	chr15.51911003.C.CC	chr12.93083679.A.G	chr6.31271165.CT	chr6.31356889.A.T.C
chr19.55013640.AC	chr4.73984853.AGT.A	chr1.73984853.AGT.A	chr14.54957029.A.T	chr12.118246461.GA	chr6.89080840.CTCA.C
chr2.120294133.TA	chr5.113020369.CTA.C	chr22.38484092.GA.G	chr15.92712026.T.C	chr7.139565716.CT	chr8.51818304.T.C.A
chr12.120764861.AG	chr6.29945567.G.T.A.C	chr1.24789638.ATTGTGGC.A	chr16.21501823.TGCCCGCC.CT.CGCC.TGCCGCC	chr14.35292469.CG	chr11.9608471.T.TT
chr3.31356739.CG	chr3.30095730.A.G	chr7.7289521.AAA	chr16.71922908.AATGCC.A	chr6.29943495.TG	chr12.15902993.TT
chr4.117574799.AG	chr6.30000746.C.T	chr3.5225273.C.CC	chr17.18805689.CTT.C	SNV	chr14.102081436.A.ACA
chr22.36295561.AC	chr7.23311727.CATT.C	chr12.96267832.TAAAAG.T	chr17.49424742.A.G	chr17.109714333.TT	chr16.2771987.TC.T
chr17.3896624.GC	chr7.134969636.TTTTA.T	chr2.191834914.G.GG	chr17.56856667.T.G	chr12.7920898.GT	chr17.21976479.G.A.T
chr21.39397364.AG	chr8.109532238.CCA.C	chr3.3130033.C.TGACTCC	chr19.5260688.GT.G	chr1.152978558.TG.T	KI270467.1.3305.A.C
chr6.32443258.AG	chr9.34252468.GCTAA.G	chr17.41692548.GATT.G	chr22.21805846.T.A	chr3.50258896.AG	
chr1.173488460.AG	chr9.35729074.T.TCT		chrX.39787945.A.A	chr2.34972902.CAGAT.C	
chr1.18956451.AG	chr10.74119159.A.ATA			chr5.139369499.C.CATT	
chr11.64853046.GA	chr11.10509421.C			chr19.2426665.AAT.A	
chr14.22551978.CG	chr12.10699507.T.C.A			chr9.34088024.G.GAAAGA	
chr2.3611718.TA	chr12.53041516.T.TTTT.TTTT			chr14.104795689.C.CC	
chr2.68396142.TC	chr14.5628337.T.TT			chr21.37367357.G.GG	
chr16.84158.AG	chr15.43988968.T.A.C			chr16.11678744.T.TTCTGA	
chr12.21532174.AC	chr16.682287.C.C			chr15.64500167.G.C	
chr11.48170123.AG	chr17.47263339.TC.T			chr2.59322186.A.C	
chr2.64338218.TG	chr18.5290626.TG.T			chr9.13443842.CGCTG.C	
chr6.31271324.GT	chr18.35306564.T.TAGTC.TAGGC			chr9.117180684.G.GG	
chr6.31356323.GT	chr22.18730667.C.G.A			chr10.110286187.T.TT	
chr14.56303317.TG				chr14.35401929.TTTC.T	
chr11.50997091.AG				chr12.48938679.G.GAG	
chr14.81278392.AG				chr7.29929051.CTGT.C	
chr1.160996644.TC				chr1.25382412.C.CTC	
chr11.843019.CT				chr2.101697932.G.CCCG	
chr10.120763363.TC				chr20.5610946.TC.T	
chr1.160998043.TC				chr7.150627451.C.CC	
chr9.100451665.AG				chr10.225	

NSCLC DATASET	GBM DATASET	CRC DATASET	ESCC DATASET	PDAC DATASET	HCC DATASET
chr2.190203041.T.TCT					
chr19.40796645.GC.G					
chr11.64223524.AC.A					
chr11.58775233.A.ACA					
chr6.31356751.G.T,A					
chr18.9400306.A.AA					
chr4.49153253.G.C					
chr1.205769847.AC.A					
chr3.88055009.ATC.A					
chr6.31271273.G.A,T					
chr17.16441919.CA.C					
chr18.3248250.A.AA					
chr14.104714369.A.G,C					
chr2.54972426.G.GTTG					
chr5.49659077.T.A					
chr20.25297834.G.GTGGG					
chr11.102398666.AG.A					
chr18.79903076.C.CTC					
chr18.5290626.TG.T					
chr7.152842573.T.TT					

TABLE 6: Machine learning output stats per dataset

DATASETS	NSCLC	GBM	CRC	ESCC	PDAC	HCC
Liquid Biopsy Biosource	tumour educated platelets	tumour educated platelets	tumour educated platelets	tumour educated platelets	extracellular vesicles	circulating epithelial cells
External Validation Set	YES	NO	YES	NO	YES	NO
Gene expression # of selected features	39	30	20	10	30	24
Gene expression AUC	0.91	0.79	0.86	0.71	0.94	0.84
Gene expression Accuracy	0.84	0.79	0.76	0.79	0.74	0.77
Gene expression Balanced Accuracy	0.84	0.70	0.75	0.66	0.81	0.78
Gene expression F1 score	0.88	0.58	0.69	0.87	0.70	0.78
Gene expression Average Precision score	0.95	0.68	0.84	0.87	0.87	0.78
Gene expression misclassified samples	50/322 (15.5%)	19/90 (21.1%)	23/97 (23.7%)	6/29 (20.7%)	12/46 (26.1%)	8/35 (22.9%)
Isoform expression # of selected features	35	25	9	33	27	40
Isoform expression AUC	0.89	0.77	0.84	0.85	0.92	0.88
Isoform expression Accuracy	0.78	0.76	0.80	0.79	0.74	0.80
Isoform expression Balanced Accuracy	0.78	0.62	0.80	0.62	0.81	0.81
Isoform expression F1 score	0.83	0.42	0.77	0.88	0.70	0.81
Isoform expression Average Precision score	0.94	0.57	0.76	0.92	0.85	0.83
Isoform expression misclassified samples	70/322 (21.7%)	22/90 (24.4%)	19/97 (19.6%)	6/29 (20.7%)	12/46 (26.1%)	7/35 (20.0%)
FoCT # of selected features	46	24	36	34	43	21
FoCT AUC	0.85	0.69	0.82	0.75	0.65	0.84
FoCT Accuracy	0.74	0.73	0.81	0.76	0.63	0.80
FoCT Balanced Accuracy	0.78	0.71	0.80	0.72	0.61	0.80
FoCT F1 score	0.77	0.59	0.76	0.83	0.48	0.79
FoCT Average Precision score	0.92	0.47	0.78	0.88	0.49	0.81
FoCT misclassified samples	83/322 (25.8%)	24/90 (26.7%)	18/97 (18.6%)	7/29 (24.1%)	17/46 (37.0%)	7/35 (20.0%)
Gene fusion # of selected features	1	1	3	1	1	NA
Gene fusion AUC	0.47	0.52	0.56	0.48	0.56	NA
Gene fusion Accuracy	0.34	0.69	0.63	0.69	0.41	NA
Gene fusion Balanced Accuracy	0.47	0.52	0.59	0.48	0.56	NA
Gene fusion F1 score	0.15	0.18	0.38	0.82	0.49	NA
Gene fusion Average Precision score	0.65	0.30	0.51	0.71	0.33	NA
Gene fusion misclassified samples	211/322 (65.5%)	28/90 (31.1%)	36/97 (37.1%)	9/29 (31.0%)	27/46 (58.7%)	NA
RNA editing # of selected features	47	49	36	32	41	39
RNA editing AUC	0.83	0.77	0.76	0.73	0.82	0.94
RNA editing Accuracy	0.78	0.72	0.70	0.69	0.72	0.86
RNA editing Balanced Accuracy	0.75	0.70	0.69	0.63	0.76	0.86
RNA editing F1 score	0.83	0.58	0.64	0.78	0.65	0.86
RNA editing Average Precision score	0.90	0.49	0.76	0.89	0.62	0.91
RNA editing misclassified samples	72/322 (22.4%)	25/90 (27.8%)	29/97 (29.9%)	9/29 (31.0%)	13/46 (28.3%)	5/35 (14.3%)
SNV # of selected features	49	29	42	37	48	20
SNV AUC	0.72	0.64	0.66	0.65	0.71	0.87
SNV Accuracy	0.70	0.62	0.71	0.72	0.70	0.83
SNV Balanced Accuracy	0.68	0.61	0.70	0.58	0.70	0.83
SNV F1 score	0.76	0.47	0.65	0.83	0.59	0.81
SNV Average Precision score	0.83	0.37	0.60	0.84	0.50	0.85
SNV misclassified samples	98/322 (30.4%)	34/90 (37.8%)	28/97 (28.9%)	8/29 (27.6%)	14/46 (30.4%)	6/35 (17.1%)
Ensemble Learning (MajorityVoting) # of selected features	216*	157*	143*	146*	189*	144*
Ensemble Learning (MajorityVoting) AUC	0.87	0.73	0.85	0.69	0.82	0.86
Ensemble Learning (MajorityVoting) Accuracy	0.86	0.82	0.87	0.83	0.78	0.86
Ensemble Learning (MajorityVoting) Balanced Accuracy	0.87	0.73	0.85	0.69	0.82	0.86
Ensemble Learning (MajorityVoting) F1 score	0.89	0.62	0.83	0.89	0.72	0.86
Ensemble Learning (MajorityVoting) Average Precision score	NA	NA	NA	NA	NA	NA
Ensemble Learning (MajorityVoting) misclassified samples	44/322 (13.7%)	16/90 (17.8%)	13/97 (13.4%)	5/29 (17.2%)	10/46 (21.7%)	5/35 (14.3%)
Ensemble Learning (SoftVoting) # of selected features	216*	157*	143*	146*	189*	144*
Ensemble Learning (SoftVoting) AUC	0.95	0.87	0.94	0.85	0.92	0.95
Ensemble Learning (SoftVoting) Accuracy	0.88	0.80	0.87	0.79	0.78	0.91
Ensemble Learning (SoftVoting) Balanced Accuracy	0.89	0.71	0.85	0.66	0.84	0.92
Ensemble Learning (SoftVoting) F1 score	0.90	0.59	0.82	0.87	0.74	0.91
Ensemble Learning (SoftVoting) Average Precision score	0.98	0.74	0.94	0.93	0.84	0.94
Ensemble Learning (SoftVoting) misclassified samples	39/322 (12.1%)	18/90 (20.0%)	13/97 (13.4%)	6/29 (20.7%)	10/46 (21.7%)	3/35 (8.6%)
Ensemble Learning (Stacking) # of selected features	216*	157*	143*	146*	189*	144*
Ensemble Learning (Stacking) AUC	0.93	0.88	0.91	0.81	0.98	0.99
Ensemble Learning (Stacking) Accuracy	0.85	0.82	0.85	0.79	0.78	0.91
Ensemble Learning (Stacking) Balanced Accuracy	0.84	0.70	0.85	0.74	0.84	0.92
Ensemble Learning (Stacking) F1 score	0.89	0.80	0.83	0.86	0.74	0.91
Ensemble Learning (Stacking) Average Precision score	0.97	0.76	0.91	0.94	0.95	0.99
Ensemble Learning (Stacking) misclassified samples	48/322 (14.9%)	16/90 (17.8%)	15/97 (15.5%)	6/29 (20.7)	10/46 (21.7%)	3/35 (8.6%)

TABLE 7: Table of genes exhibiting recurring appearances exceeding a threshold of two occurrences across all datasets.

GeneID	Gene Name	Gene Type	Occur.	Description
ENSG00000234745.11	HLA-B	protein coding	6	major histocompatibility complex, class I, B
ENSG00000206503.13	HLA-A	protein coding	5	major histocompatibility complex, class I, A
ENSG00000213492.2	NT5C3AP1	transcribed processed pseudo-gene	3	NT5C3A pseudo-gene 1
ENSG00000166710.20	B2M	protein coding	3	beta-2-microglobulin
ENSG00000160014.17	CALM3	protein coding	3	calmodulin 3
ENSG00000162852.14	CNST	protein coding	3	consortin, connexin sorting protein
ENSG00000115956.10	PLEK	protein coding	3	pleckstrin
ENSG00000196126.11	HLA-DRB1	protein coding	3	major histocompatibility complex, class II, DR beta 1

GeneID	Gene Name	Gene Type	Occur.	Description
ENSG00000117640.18	MTFR1L	protein coding	3	mitochondrial fission regulator 1 like
ENSG00000149781.12	FERMT3	protein coding	3	FERM domain containing kindlin 3
ENSG00000115310.18	RTN4	protein coding	3	reticulon 4
ENSG00000150867.14	PIP4K2A	protein coding	3	phosphatidylinositol-5-phosphate 4-kinase type 2 alpha
ENSG00000285774.3	AL133444.1	lncRNA	2	NA
ENSG00000240225.10	ZNF542P	transcribed unprocessed pseudo-gene	2	zinc finger protein 542, pseudogene
ENSG00000158769.18	F11R	protein coding	2	F11 receptor
ENSG00000177272.9	KCNA3	protein coding	2	potassium voltage-gated channel subfamily A member 3
ENSG00000125354.23	SEPTIN6	protein coding	2	septin 6

GeneID	Gene Name	Gene Type	Occur.	Description
ENSG00000143514.17	TP53BP2	protein coding	2	tumor protein p53 binding protein 2
ENSG00000169221.14	TBC1D10B	protein coding	2	TBC1 domain family member 10B
ENSG00000272888.7	LINC01578	lncRNA	2	NA
ENSG00000125037.12	EMC3	protein coding	2	ER membrane protein complex subunit 3
ENSG00000133639.6	BTG1	protein coding	2	BTG anti-proliferation factor 1
ENSG00000166337.10	TAF10	protein coding	2	TATA-box binding protein associated factor 10
ENSG00000108061.12	SHOC2	protein coding	2	SHOC2 leucine rich repeat scaffold protein
ENSG00000108960.9	MMD	protein coding	2	monocyte to macrophage differentiation associated

GeneID	Gene Name	Gene Type	Occur.	Description
ENSG00000248334.6	WHAMMP2	transcribed unpro- cessed pseudo- gene	2	WHAMM pseudo- gene 2
ENSG00000105372.8	RPS19	protein coding	2	ribosomal protein S19
ENSG00000103512.15	NOMO1	protein coding	2	NODAL modulator 1
ENSG00000101096.20	NFATC2	protein coding	2	nuclear factor of activated T cells 2
ENSG00000204323.5	SMIM5	protein coding	2	small inte- gral mem- brane pro- tein 5
ENSG00000250334.6	LINC00989	lncRNA	2	long inter- genic non- protein coding RNA 989
ENSG00000184602.6	SNN	protein coding	2	stannin
ENSG00000198886.2	MT-ND4	protein coding	2	mitochondrially encoded NADH dehydro- genase 4
ENSG00000267265.5	AC011476.3	lncRNA	2	NA

GeneID	Gene Name	Gene Type	Occur.	Description
ENSG00000100345.22	MYH9	protein coding	2	myosin heavy chain 9
ENSG00000134548.11	SPX	protein coding	2	spexin hor- mone
ENSG00000143409.15	MINDY1	protein coding	2	MINDY lysine 48 deubiq- uitinase 1
ENSG00000086232.13	EIF2AK1	protein coding	2	eukaryotic translation initiation factor 2 al- pha kinase 1
ENSG00000244041.7	LINC01011	lncRNA	2	long inter- genic non- protein coding RNA 1011
ENSG00000204623.9	ZNRD1ASP	transcribed unitary pseudo- gene	2	NA
ENSG00000205038.12	PKHD1L1	protein coding	2	PKHD1 like 1
ENSG00000136754.17	ABI1	protein coding	2	abl in- teractor 1

GeneID	Gene Name	Gene Type	Occur.	Description
ENSG00000180353.11	HCLS1	protein coding	2	hematopoietic cell-specific Lyn substrate 1
ENSG00000164091.12	WDR82	protein coding	2	WD repeat domain 82
ENSG00000063046.18	EIF4B	protein coding	2	eukaryotic translation initiation factor 4B
ENSG00000185245.8	GP1BA	protein coding	2	glycoprotein Ib platelet subunit alpha
ENSG00000179632.10	MAF1	protein coding	2	NA
ENSG00000023734.11	STRAP	protein coding	2	serine/threonine kinase receptor associated protein
ENSG00000187699.10	C2orf88	protein coding	2	NA
ENSG00000128791.12	TWSG1	protein coding	2	twisted gastrulation BMP signaling modulator 1

GeneID	Gene Name	Gene Type	Occur.	Description
ENSG00000117280.13	RAB29	protein coding	2	RAB29, member RAS oncogene family
ENSG00000152558.15	TMEM123	protein coding	2	transmembrane protein 123
ENSG00000198081.11	ZBTB14	protein coding	2	zinc finger and BTB domain containing 14
ENSG00000133627.18	ACTR3B	protein coding	2	actin related protein 3B
ENSG00000160446.19	ZDHHC12	protein coding	2	zinc finger DHHC-type palmitoyl-transferase 12
ENSG00000265491.5	RNF115	protein coding	2	ring finger protein 115
ENSG00000114978.18	MOB1A	protein coding	2	MOB kinase activator 1A
ENSG00000106772.18	PRUNE2	protein coding	2	prune homolog 2 with BCH domain

GeneID	Gene Name	Gene Type	Occur.	Description
ENSG00000167664.8	TMIGD2	protein coding	2	transmembrane and immunoglobulin domain containing 2
ENSG00000129657.16	SEC14L1	protein coding	2	SEC14 like lipid binding 1
ENSG00000173812.11	EIF1	protein coding	2	eukaryotic translation initiation factor 1
ENSG00000182149.21	IST1	protein coding	2	IST1 factor associated with ESCRT-III
ENSG00000100934.15	SEC23A	protein coding	2	SEC23 homolog A, COPII coat complex component
ENSG00000102362.15	SYTL4	protein coding	2	synaptotagmin like 4
ENSG00000117523.16	PRRC2C	protein coding	2	proline rich coiled-coil 2C
ENSG00000136929.13	HEMGN	protein coding	2	hemogen

GeneID	Gene Name	Gene Type	Occur.	Description
ENSG00000145495.16	MARCHF6	protein coding	2	membrane associated ring-CH-type finger 6
ENSG00000138376.11	BARD1	protein coding	2	BRCA1 associated RING domain 1
ENSG00000124772.12	CPNE5	protein coding	2	copine 5
ENSG00000064666.15	CNN2	protein coding	2	calponin 2
ENSG00000182551.14	ADI1	protein coding	2	acireductone dioxygenase 1
ENSG00000146859.6	TMEM140	protein coding	2	transmembrane protein 140
ENSG00000263465.4	SRSF8	protein coding	2	serine and arginine rich splicing factor 8
ENSG00000100225.18	FBXO7	protein coding	2	F-box protein 7
ENSG00000210077.1	MT-TV	Mt tRNA	2	mitochondrially encoded tRNA valine
ENSG00000147394.18	ZNF185	protein coding	2	zinc finger protein 185 with LIM domain

GeneID	Gene Name	Gene Type	Occur.	Description
ENSG00000279516.2	FAM230C	lncRNA	2	family with sequence similarity 230 member C
ENSG00000122786.20	CALD1	protein coding	2	caldesmon 1
ENSG00000257267.3	ZNF271P	transcribed unitary pseudo-gene	2	zinc finger protein 271, pseudogene
ENSG00000136003.16	ISCU	protein coding	2	iron-sulfur cluster assembly enzyme
ENSG00000011275.19	RNF216	protein coding	2	ring finger protein 216
ENSG00000223482.8	NUTM2A-AS1	lncRNA	2	NUTM2A antisense RNA 1
ENSG00000184009.12	ACTG1	protein coding	2	actin gamma 1
ENSG00000141030.13	COPS3	protein coding	2	COP9 signalosome subunit 3
ENSG00000071051.14	NCK2	protein coding	2	NCK adaptor protein 2
ENSG00000168300.14	PCMTD1	protein coding	2	protein-L-isoaspartate
ENSG00000177885.15	GRB2	protein coding	2	growth factor receptor bound protein 2

GeneID	Gene Name	Gene Type	Occur.	Description
ENSG00000080824.19	HSP90AA1	protein coding	2	heat shock protein 90 alpha fam- ily class A member 1
ENSG00000110367.13	DDX6	protein coding	2	DEAD- box he- licase 6
ENSG00000115271.11	GCA	protein coding	2	grancalcin
ENSG00000223361.5	FTH1P10	transcribed processed pseudo- gene	2	ferritin heavy chain 1 pseudo- gene 10
ENSG00000142089.16	IFITM3	protein coding	2	interferon induced trans- membrane protein 3
ENSG00000179833.4	SERTAD2	protein coding	2	SERTA domain containing 2
ENSG00000048649.13	RSF1	protein coding	2	remodeling and spac- ing factor 1

A.2 Expanding the landscape of cancer transcriptome by native RNA sequencing of NSCLC tissue samples (Supplementary Materials)

A.2.1 TALON transcript classification categories.

In our study, we utilised the TALON software for transcript categorisation, which incorporates the classification system originally developed by the SQANTI software [353]. This approach allowed us to accurately categorise transcripts based on their alignment with established transcript models (Figure 4). Transcripts perfectly matching known models at splice junctions are classified as 'known', with a degree of flexibility allowed at their 5' and 3' ends to account for minor variations.

When a transcript partially matches a known model, particularly if it has novel potential start or end points, it falls under the category of an 'incomplete splice match' (ISM). TALON further refines this category into prefix ISMs, which align with the beginning (5' end) of an existing transcript model, and suffix ISMs, which align with the end (3' end). This nuanced classification of ISMs is vital, as it helps in evaluating the integrity and completeness of the transcripts.

The 'novel in catalog' (NIC) category includes transcripts that form new connections between known splice donors and acceptors. This novel rearrangement of existing exons expands our understanding of the plasticity and diversity of gene expression. In contrast, 'novel not in catalog' (NNC) transcripts are characterised by having at least one novel splice site, signifying the presence of new exon boundaries and potentially undiscovered aspects of gene structure.

We also take into account 'genomic' transcripts, which typically represent either incomplete overlap with known genes or DNA contamination.

Another intriguing category is 'antisense' transcripts, which are defined by their overlap with known genes but in the opposite orientation. This phenomenon provides a unique window into the complexities of genomic regulation and expression patterns. Lastly, transcripts that do not align with any known gene structure are classified as 'intergenic'. This category is particularly interesting as it may point to novel gene discovery or unexplored genomic regions, offering opportunities for new insights into genomic function and organisation.

A.2.2 Supplementary Figures

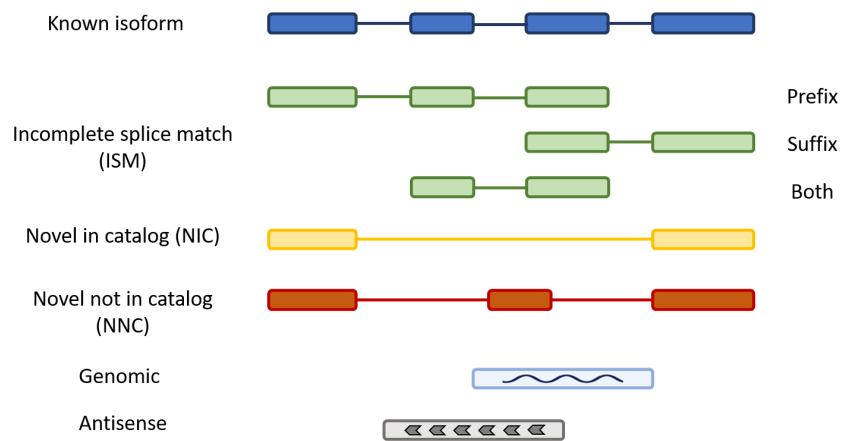
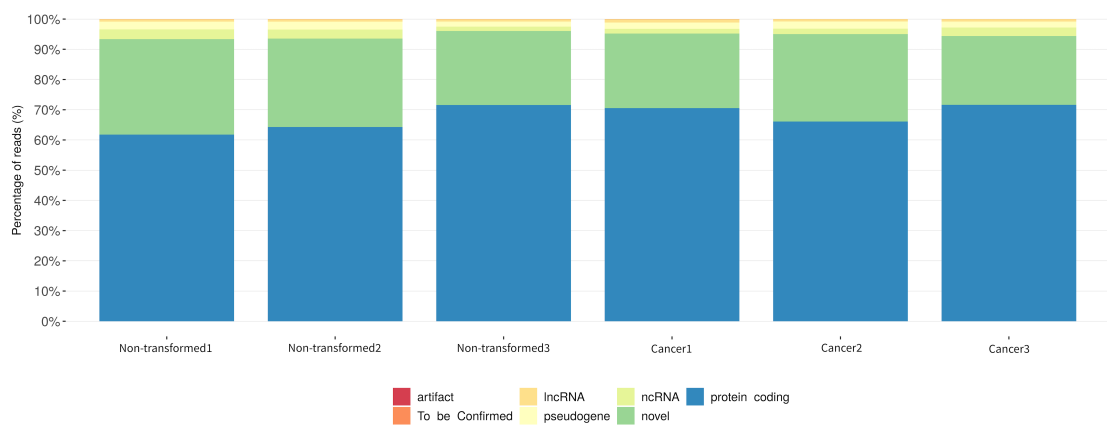


FIGURE 4: **Classification of Transcripts Based on TALON Analysis.** The schematic illustrates various categories of transcripts identified: Known isoforms (blue) represent exact matches to established transcript models; Incomplete splice match (ISM) transcripts (green) are partially matching sequences, further divided into prefix (aligning at the start), suffix (aligning at the end), or both; Novel in catalog (NIC) transcripts (yellow) connect known splice sites in new configurations; Novel not in catalog (NNC) transcripts (red) contain at least one novel splice junction; Genomic transcripts (light blue) are typically excluded due to partial overlap or potential DNA contamination; Antisense transcripts (grey patterned) are transcribed in the opposite direction to known genes.



*Stats based on the filtered expression matrix

FIGURE 5: Transcript Type Composition Across Samples. The bar plot summarises the transcript composition for non-transformed and cancer samples based on the reference genome annotation and TALON analysis. It displays the proportion of reads categorised as protein-coding, non-coding RNAs (ncRNAs), long non-coding RNAs (lncRNAs), pseudogenes, novel transcripts, artifacts, and those requiring confirmation. Each bar represents a different sample, allowing for a comparative view of transcript types between non-transformed and cancerous tissues. Please note that the statistics are derived from the filtered expression matrix.



Percentage of Known and novel category composition for each sample.

FIGURE 6: Distribution of Transcript Categories Across Samples. The circular bar chart depicts the proportion of known and novel transcript categories for each sample, including cancerous and non-transformed tissues. Each segment represents a category such as known transcripts, ISM Prefix, ISM Suffix, ISM Both, NIC, NNC, Antisense, and Intergenic, with the length of the bar corresponding to the percentage composition of that category within the sample. The chart provides an overview of the transcriptomic landscape between the different sample types.

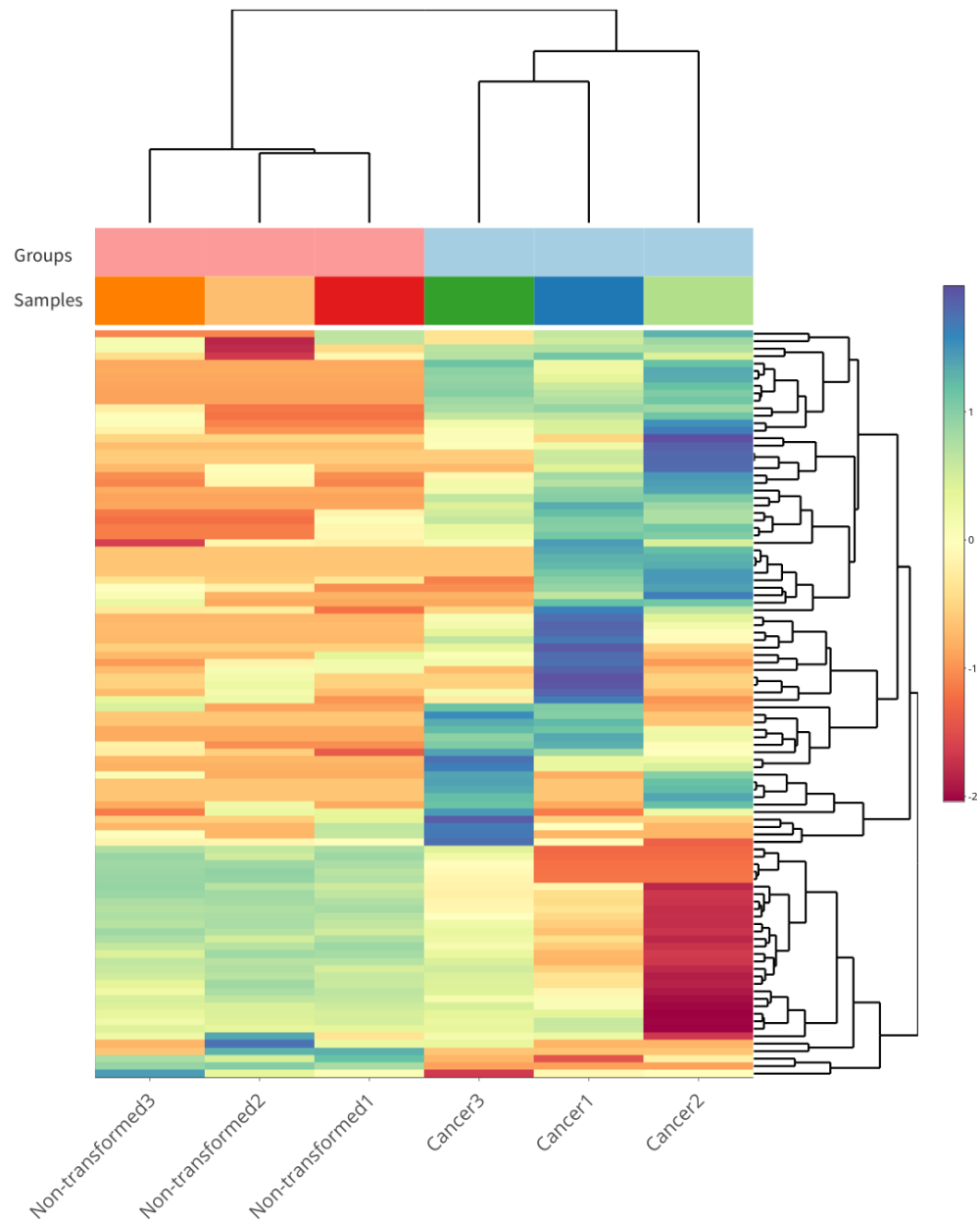


FIGURE 7: **Heatmap of Gene Expression Variability in NSCLC.** This heatmap showcases the top 100 most variable genes across non-transformed and cancerous lung tissue samples. Each row represents a gene, and each column corresponds to a sample. The colour gradient from blue to red indicates the expression level from low to high, standardised across samples. Hierarchical clustering on both genes and samples illustrates the relative similarity of expression patterns, with the dendrogram on the top and left reflecting the clustering results.

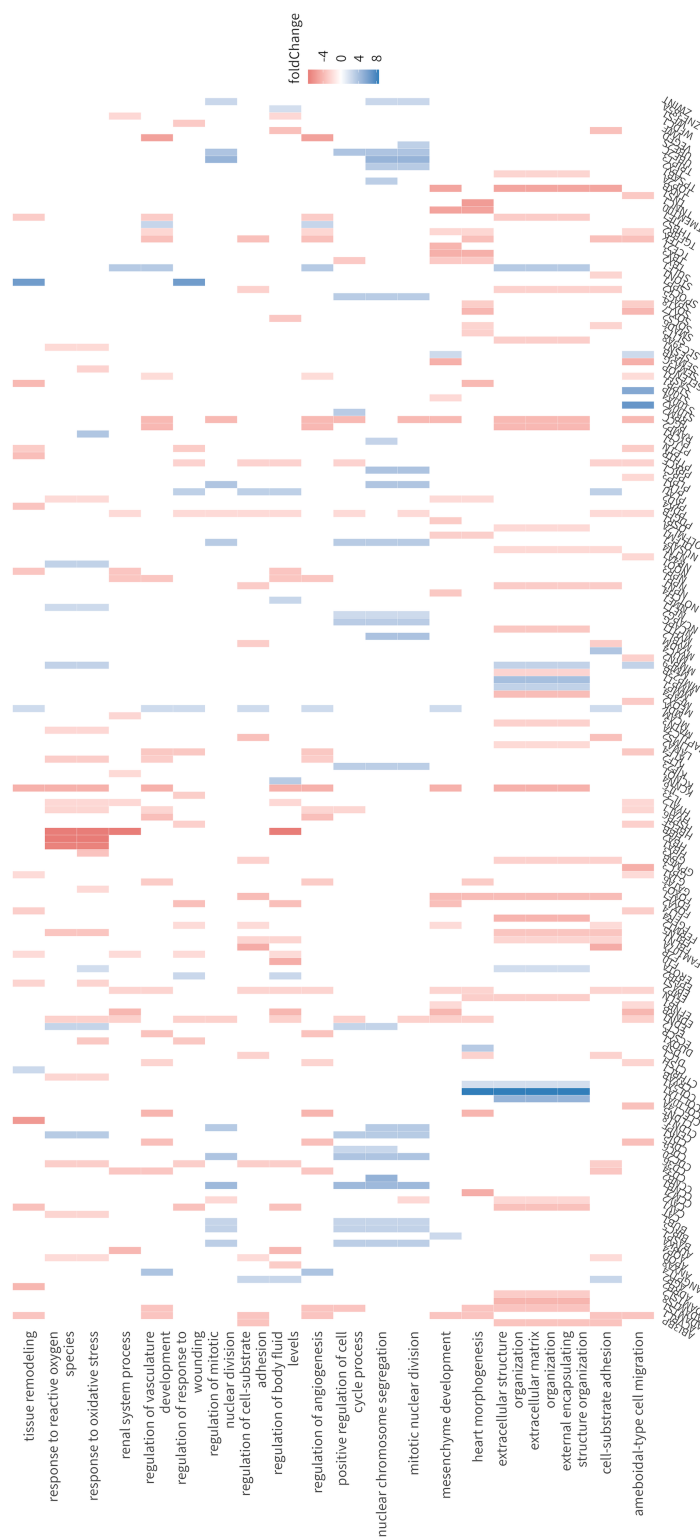


FIGURE 8: Heatplot Visualisation. The heatplot presents a visual mapping between genes and biological concepts through a heatmap representation. In instances where the network of gene-concept connections is exceedingly intricate, particularly when numerous significant terms are involved, the heatplot provides a streamlined perspective. This streamlining aids in the clearer identification of expression trends.



FIGURE 9: Gene Ontology and Pathway Enrichment Analysis for DE Novel Transcripts. (A) The top panel shows the enrichment analysis of up-regulated novel transcripts, identifying significant biological processes, molecular functions, and cellular components, with the degree of enrichment indicated by the negative log₁₀ p-value. (B) The bottom panel presents a similar enrichment analysis for down-regulated novel transcripts, highlighting differentially involved pathways and biological terms. The color coding corresponds to various categories of gene ontology and pathways, while the size of the dots represents the magnitude of enrichment.

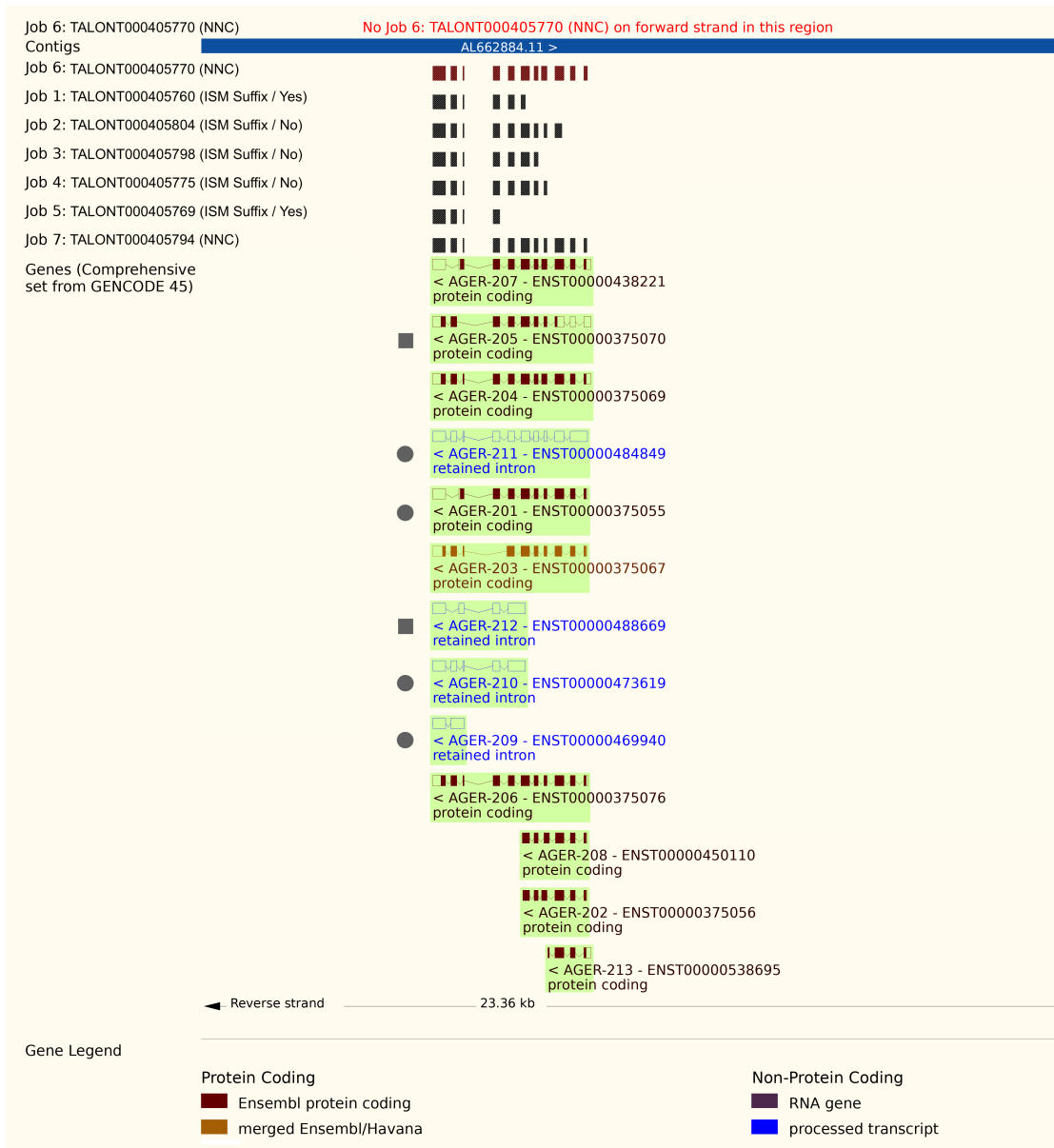


FIGURE 10: Detailed visualisation of all captured isoforms of the AGER gene. The figure provides a comprehensive genomic landscape of the AGER gene, outlining both novel and known isoforms. On the left, novel isoforms are cataloged with their respective names, detailing the type of isoform (NNC or ISM Suffix), and an indication of CAGE support for their 5' ends (Yes or No). The central part of the figure graphically maps out the structure of these novel isoforms, juxtaposed with the lower-positioned known isoforms of AGER. Grey circles adjacent to four of these isoforms signify those that were quantified and met the quality and filtering standards of the study, indicating their presence at detectable levels. Conversely, grey squares next to other isoforms represent those that, despite being identified, did not pass the filtering step due to insufficient read counts. The absence of markers alongside the rest of the isoforms suggests that they were not detected in the analysis. The figure, taken from Ensembl Genome Browser, is colour-coded to distinguish between different genomic elements: protein-coding sequences, RNA genes, and processed transcripts are each depicted with a distinct colour as explained in the gene legend.

A.2.3 Supplementary Tables

TABLE 8: Significantly Differentially Expressed Genes in Non-Transformed vs. Cancer samples.

GeneID	GeneName	Non-transformed1	Non-transformed2	Non-transformed3	Cancer1	Cancer2	Cancer3	log2FC	pvalue	padj
ENSG00000123243.15	ITIH5	74.58	97.88	64.23	1.74	3.13	1.89	-4.97695	1.0802847797811e-06	0.00162
ENSG00000179776.19	CDH5	247.48	244.7	354.21	26.03	22.98	20.82	-3.60007	1.2925018333233e-06	0.00162
ENSG00000266964.6	FXYD1	142.39	139.09	79.79	5.21	2.09	2.84	-5.15928	1.49914890928915e-06	0.00162
ENSG00000131477.11	RAMP2	267.83	342.58	288.04	10.41	20.89	22.71	-4.00628	1.52880812007726e-06	0.00162
ENSG00000144655.15	CSRNP1	125.44	149.4	140.13	12.15	10.45	13.25	-3.51295	1.57074249682202e-06	0.00162
ENSG00000168477.19	TNXB	264.43	200.91	340.59	8.68	3.13	11.36	-5.09092	1.71422773163568e-06	0.00162
ENSG00000176435.7	CLEC14A	240.7	242.12	241.33	19.09	15.67	27.45	-3.51891	1.79741899545996e-06	0.00162
ENSG00000136160.17	EDNRB	67.8	87.58	108.99	5.21	5.22	4.73	-4.08962	1.88451811915116e-06	0.00162
ENSG00000127920.6	GNG11	461.07	329.7	326.96	32.921	41.79	36.91	-3.30921	2.25181018919064e-06	0.00162
ENSG00000118526.7	TCF21	176.29	182.88	229.65	6.94	7.31	16.09	-4.20997	2.30568312910799e-06	0.00162
ENSG0000022267.19	FHL1	132.22	180.31	338.64	10.41	8.36	7.57	-4.63146	2.84920388336413e-06	0.00168
ENSG00000133401.16	PDZD2	50.85	56.67	58.39	1.74	1.04	2.84	-4.71715	2.8640536752753e-06	0.00168
ENSG00000114854.8	TNNC1	240.7	139.09	99.26	5.21	1.04	3.79	-5.5672	3.7373565382948e-06	0.00194
ENSG00000110799.14	VWF	864.5	752.13	794.06	78.09	43.88	86.13	-3.53223	4.53310933330303e-06	0.00194
ENSG00000165072.10	MAMDC2	67.8	105.61	62.28	3.47	1.04	3.79	-4.75788	4.56563531003521e-06	0.00194
ENSG00000162545.6	CAMK2N1	155.95	236.97	210.19	22.56	21.94	13.25	-3.39215	4.66078695474228e-06	0.00194
ENSG00000152583.13	SPARCL1	779.74	837.13	1078.2	74.61	87.76	114.52	-3.27687	4.71714123516946e-06	0.00194
ENSG00000131634.14	TMEM204	179.68	193.18	159.59	27.76	18.81	19.88	-3.00931	4.9626833328531e-06	0.00194
ENSG00000137309.20	HMGGA1	13.56	33.49	31.14	275.9	279.99	303.81	3.4145	6.3003136886366e-06	0.00213
ENSG00000189129.14	PLAC9	294.95	105.61	181	8.68	8.36	12.3	-4.25648	6.18283375365173e-06	0.00213
ENSG00000108622.11	ICAM2	132.22	146.82	147.91	10.41	19.85	17.04	-3.12983	6.85508585211539e-06	0.00213
ENSG00000261371.6	PECAM1	789.91	703.19	895.26	126.67	95.07	101.27	-2.88717	6.97655577314883e-06	0.00213
ENSG00000172005.11	MAL	47.46	64.39	44.76	3.47	3.13	3.79	-3.84373	6.97830627073732e-06	0.00213
ENSG00000094963.14	PMO2	206.8	309.25	216.03	13.88	11.49	25.55	-3.97597	7.8896434393924e-06	0.00222
ENSG0000017469.13	CAVIN1	833.99	852.59	739.56	72.88	112.83	117.36	-2.99335	8.05047199656218e-06	0.00222
ENSG00000072163.20	LIMS2	125.44	136.52	99.26	15.62	9.4	6.63	-3.54648	8.62914478689307e-06	0.00222
ENSG00000102760.13	RGCC	840.77	989.1	1354.57	85.03	34.48	82.34	-3.98179	8.79358610900897e-06	0.00222
ENSG00000112782.19	CLIC5	44.07	38.64	44.76	3.47	2.09	2.84	-3.87139	9.14554206139505e-06	0.00222
ENSG00000186994.12	KANK3	61.02	110.76	81.74	5.21	4.18	8.52	-3.75772	9.31850758019253e-06	0.00222
ENSG00000151433.15	ROBO4	115.27	79.85	204.35	5.21	5.22	9.46	-4.25965	9.71666630485621e-06	0.00222
ENSG00000139567.13	ACVRL1	108.49	175.15	145.97	15.62	9.4	17.98	-3.31107	9.8902184570104e-06	0.00222
ENSG00000197253.13	TPSB2	237.31	190.61	159.59	24.29	30.3	32.18	-2.73594	1.02344005589891e-05	0.00222
ENSG00000160867.15	FGFR4	71.19	175.15	122.61	5.21	4.18	9.46	-4.22111	1.04232755343995e-05	0.00222
ENSG00000171967.12	CYBRD1	400.04	401.82	247.17	50.32	55.37	45.43	-2.78837	1.19486254835826e-05	0.00245
ENSG00000141934.10	PLP2	10.17	12.88	7.78	203.02	100.29	119.25	3.76424	1.25058047961542e-05	0.00245
ENSG0000026524.3	GDF10	64.41	77.27	38.92	1.74	0	1.89	-5.45941	1.26900938884472e-05	0.00245
ENSG00000103241.7	FOXP1	67.8	87.58	48.66	6.94	3.13	4.73	-3.78609	1.28960470266239e-05	0.00245
ENSG0000017434.10	CA4	105.1	193.18	173.21	0	0	4.73	-6.25889	1.34947781376951e-05	0.0025
ENSG00000249751.4	ECSCR	125.44	66.97	136.24	12.15	8.36	10.41	-3.40601	1.41397829174042e-05	0.00251
ENSG00000126218.12	F10	33.9	41.21	25.3	1.74	1.04	0.95	-4.61162	1.44147470621778e-05	0.00251
ENSG00000174059.17	CD34	271.22	133.94	272.47	26.03	28.21	18.93	-3.20741	1.48490940342111e-05	0.00251
ENSG00000136732.16	GYPC	183.07	167.43	229.65	36.44	26.12	35.02	-2.57302	1.55361581309665e-05	0.00251
ENSG0000010391.7	SEMA3G	23.73	46.36	42.82	1.74	1.04	1.89	-4.46418	1.55450113355675e-05	0.00251
ENSG00000095370.20	SH2D3C	61.02	85	77.85	10.41	9.4	11.36	-2.82594	1.5711616943636e-05	0.00251
ENSG00000122679.8	RAMP3	139	123.64	108.99	6.94	8.36	17.98	-3.40385	1.60246676975908e-05	0.00251
ENSG00000136826.15	KLF4	67.8	110.76	118.72	13.88	11.49	9.46	-3.10671	1.6650820471791e-05	0.00255
ENSG00000149564.12	ESAM	271.22	180.31	326.96	31.23	26.12	43.54	-2.93601	1.75177802392567e-05	0.00262
ENSG00000119147.10	ECRCG4	47.46	54.09	29.19	1.74	0	0	-6.18806	1.88902274884368e-05	0.00267
ENSG00000135111.16	TBX3	23.73	28.33	36.98	1.74	1.04	0.95	-4.47626	1.88941194869137e-05	0.00267
ENSG00000175899.15	A2M	4458.1	4427.79	4595.02	402.57	306.11	654.95	-3.30428	1.89809592890626e-05	0.00267
ENSG00000115306.16	SPTBN1	718.72	631.07	770.17	93.7	126.41	141.02	-2.54854	2.04831035840247e-05	0.00278
ENSG0000014092.15	FBLN5	139	133.94	247.17	17.35	11.49	23.66	-3.29565	2.05277634515236e-05	0.00278
ENSG00000161835.11	TAMALIN	61.02	110.76	60.33	5.21	6.27	9.46	-3.39234	2.24152990185216e-05	0.00291
ENSG00000169583.13	CLIC3	372.92	216.37	188.78	15.62	19.85	35.97	-3.40009	2.25039510696925e-05	0.00291
ENSG00000101331.17	CCM2L	54.24	33.49	38.92	0	2.09	1.89	-4.61066	2.27488691828954e-05	0.00291
ENSG00000090006.18	LTBP4	250.87	479.1	593.6	46.85	61.64	47.32	-3.08257	2.35323485149012e-05	0.00296
ENSG00000172236.18	TPSAB1	410.21	499.7	537.16	90.23	80.44	108.84	-2.36967	2.61926294991684e-05	0.00317
ENSG00000143416.21	SELENBP1	386.48	522.89	266.63	48.59	47.01	70.04	-2.81683	2.73563542195187e-05	0.00317
ENSG00000135052.16	GOLM1	3.39	2.58	5.84	279.37	80.44	108.84	5.16495	2.74355958138746e-05	0.00317
ENSG00000163751.4	CPA3	115.27	115.91	122.61	20.82	14.63	24.61	-2.54786	2.78605502664497e-05	0.00317
ENSG00000161281.11	COX7A1	159.34	123.64	110.93	24.29	12.54	10.41	-3.09261	2.78862552173955e-05	0.00317
ENSG00000161940.11	BCL6B	57.63	46.36	91.47	3.47	7.31	2.84	-3.77863	2.81061843192507e-05	0.00317
ENSG00000131097.7	HIGD1B	118.66	100.46	66.17	6.94	2.09	9.46	-3.89406	2.85953088050129e-05	0.00317
ENSG00000170323.9	FABP4	145.78	247.28	231.6	5.21	0	9.46	-5.34252	2.9716917300021e-05	0.00317
ENSG00000066056.14	TIE1	88.14	151.97	101.2	10.41	16.72	17.98	-2.87585	3.04063137692989e-05	0.00317
ENSG00000117399.14	CDC20	3.39	5.15	3.89	59	77.31	40.7	3.75706	3.05125216832121e-05	0.00317
ENSG00000239672.8	NME1	50.85	25.76	35.03	216.9	197.45	258.38	2.60452	3.06253768864212e-05	0.00317
ENSG00000168309.18	FAM107A	33.9	74.7	105.1	1.74	1.04	4.73	-4.67377	3.06384698267443e-05	0.00317
ENSG00000135218.19	CD36	190.61	138.18	138.18	20.82	21.94	19.88	-2.74093	3.27553730951531e-05	0.00329
ENSG0000009953.10	MMP11	3.39	2.58	1.95	57.26	53.28	32.18	4.12347	3.32423282276861e-05	0.00329
ENSG00000163815.6	CLEC3B	623.8	762.43	463.2	15.62	6.27	43.54	-4.78117	3.41035345594833e-05	0.00329
ENSG00000151718.16	WWC2	128.83	90.15	101.2	10.41	16.72	19.88	-2.71432	3.48470920458422e-05	0.00329
ENSG00000165197.5	VEGFD	98.32	77.27	73.96	0	0	4.73	-5.32387	3.49201450820416e-05	0.00329
ENSG00000131055.5	COX4I2	108.49	123.64	79.79	5.21	3.13	13.25	-3.75673	3.63158732123318e-05	0.00329
ENSG00000049283.19	EPN3	0	0	1.95	46.85	42.83	32.18	5.30727	3.63241367222502e-05	0.00329
ENSG00000179820.16	MYADM	342.41	198.34	375.62	45.12	29.25	53.95	-2.83154	3.65919997812524e-05	0.00329
ENSG00000131471.7	AOC3	145.78	177.73	219.92	1.74	6.27	17.98	-4.25126	3.67198431865422e-05	0.00329
ENSG00000120833.14	SOC2S	44.07	72.12	60.33	6.94	4.18	8.52	-3.13965	3.6750479848823e-05	0.00329
ENSG00000248290.1	TNXA	54.24	25.76	31.14	1.74	0	0.95	-5.21921	3.74627483389712e-05	0.00329
ENSG00000174640.15	SLC20A1	155.95	128.79	157.64	17.35	7.31	23.66	-3.17819	3.75334490211692e-05	0.00329
ENSG00000137033.12	IL33	54.24	66.97	93.42	8.68	12.54	9.46	-2.78343	3.78438290770046e-05	0.00329
ENSG00000166292.12	TMEM100	50.85	144.24	36.98	1.74	1.04	2.84	-5.19148	3.858581449520897e-05	0.00333
ENSG00000120913.24	PDLIM2	254.26	154.55	210.19	38.18	41.79	46.38	-2.27981	4.26652068226739e-05	0.00351
ENSG00000077152.12	UBE2T	0	2.58	3.89	48.59	45.97	44.48	4.12366	4.29825282033531e-05	0.00351
ENSG00000184113.10	CLDN5	915.35	780.46	1693.21	27.76	44.92	107.9	-4.21054	4.33066499987319e-05	0.0035

GeneID	GeneName	Non-transformed1	Non-transformed2	Non-transformed3	Cancer1	Cancer2	Cancer3	log2FC	pvalue	padj
ENSG00000105124.19	SVEP1	88.14	61.82	79.79	8.68	14.63	15.14	-2.52153	5.9138933181863e-05	0.00433
ENSG000000668001.14	HYAL2	403.43	383.79	463.2	71.14	89.85	114.52	-2.17531	6.11278932105998e-05	0.00443
ENSG00000079308.20	TNS1	552.6	543.49	544.94	112.79	45.97	39.75	-3.0632	6.27310440772698e-05	0.0045
ENSG00000152137.8	HSPB8	67.8	92.73	64.23	17.35	10.45	10.41	-2.57928	6.3314830610254e-05	0.0045
ENSG00000167588.13	GPD1	27.12	43.79	81.74	3.47	0	0.95	-5.24205	6.67999796625049e-05	0.0047
ENSG00000118689.5	FOXO3	118.66	157.12	181	32.97	34.48	37.86	-2.43952	6.83210105485297e-05	0.00476
ENSG00000137648.19	TMPRSS4	3.39	7.73	0	140.55	62.68	102.22	4.8116	7.0874237400107e-05	0.00487
ENSG00000108691.10	CCL2	216.97	149.4	165.43	29.5	44.92	22.71	-2.43952	7.12857163214947e-05	0.00487
ENSG00000011201.12	ANOS1	47.46	48.94	66.17	5.21	1.04	6.63	-3.62526	7.25778220125299e-05	0.00491
ENSG00000077942.19	FBLN1	620.41	880.92	628.63	111.05	113.88	183.61	-2.37777	7.76185308646109e-05	0.0052
ENSG00000182010.11	RTKN2	64.41	56.67	48.66	8.68	1.04	1.89	-4.02583	8.07903300856279e-05	0.00536
ENSG00000134057.15	CNCB1	0	2.58	7.78	100.64	130.59	53	4.48221	8.2057084571839e-05	0.0054
ENSG00000043591.6	ADRB1	57.63	33.49	48.66	6.94	1.04	1.89	-3.94688	8.5511975733188e-05	0.00557
ENSG00000163513.19	TGFBFR2	515.31	486.82	422.33	91.97	71.04	134.4	-2.25461	8.832432427986e-05	0.0057
ENSG00000101057.16	MYBL2	6.78	7.73	5.84	81.56	169.25	51.11	3.86911	9.07366003740505e-05	0.00572
ENSG00000134531.10	EMP1	491.58	345.16	628.63	78.09	100.29	117.36	-2.30332	9.09173924876446e-05	0.00572
ENSG00000050555.19	LAMC3	44.07	28.33	93.42	5.21	2.09	3.79	-3.9281	9.10903894013967e-05	0.00572
ENSG00000148671.14	ADIRF	772.96	1138.35	274.42	27.76	59.55	15.14	-4.398	9.23366189188803e-05	0.00575
ENSG00000169418.10	NPR1	81.36	46.36	79.79	6.94	3.13	11.36	-3.21949	9.3727646719748e-05	0.00578
ENSG00000066405.13	CLDN18	596.67	718.65	517.69	8.68	0	23.66	-5.76105	9.52673556689956e-05	0.00578
ENSG00000103811.18	CTSH	2166.33	1313.65	1298.13	242.93	267.45	366.28	-2.44386	9.536432864277e-05	0.00578
ENSG00000121068.14	TBX2	50.85	87.58	134.29	12.15	7.31	14.2	-3.02081	9.80522483879181e-05	0.00581
ENSG00000064205.11	CEN35	115.27	167.43	138.18	26.03	5.22	8.52	-3.49359	9.95351985501858e-05	0.00581
ENSG00000090530.10	P3H2	33.9	61.82	44.76	8.68	6.27	6.63	-2.70999	9.95622746167827e-05	0.00581
ENSG00000109846.9	CRYAB	98.32	100.46	93.42	27.76	14.63	17.04	-2.31875	9.96174672312315e-05	0.00581
ENSG00000164855.16	TMEM184A	0	2.58	1.95	26.03	38.66	31.23	4.12508	0.000101993692317128	0.00581
ENSG00000025423.11	HSD17B6	67.8	175.15	50.6	8.68	3.13	7.57	-3.91949	0.000102563617769164	0.00581
ENSG00000168743.13	NPNT	132.22	234.4	188.78	22.56	13.58	39.75	-2.85284	0.000103109463388304	0.00581
ENSG00000019102.12	VSG2	115.27	118.49	52.55	6.94	1.04	9.46	-3.98487	0.000103701721320499	0.00581
ENSG00000149557.14	FEZ1	30.51	33.49	50.6	5.21	6.27	4.73	-2.80032	0.000104115288125862	0.00581
ENSG00000103710.11	RASL12	33.9	33.49	31.14	3.47	4.18	5.68	-2.78641	0.0001061727829529789	0.00588
ENSG00000130052.14	STARSD8	71.19	41.21	62.28	5.21	8.36	11.36	-2.72107	0.000107426684069823	0.0059
ENSG00000182481.10	PKNA2	44.07	18.03	40.87	189.14	165.07	182.67	2.37994	0.000113647355546468	0.00611
ENSG0000013726.17	FXYD6	44.07	72.12	108.99	13.88	11.49	9.46	-2.72136	0.00011467773756534	0.00611
ENSG00000211445.13	GPX3	684.82	1437.29	1337.05	81.56	61.64	192.13	-3.36012	0.00011478168904544	0.00611
ENSG00000149435.12	CGTLC1	189.85	115.91	44.76	3.47	0	5.68	-5.13956	0.000115203127151703	0.00611
ENSG00000158764.7	ITLN2	37.29	61.82	27.25	0	0	2.84	-4.97878	0.000115451699934627	0.00611
ENSG00000116690.13	PRG4	37.29	23.18	48.66	1.74	0	2.84	-4.1118	0.000116505929114497	0.00612
ENSG00000197766.9	CFD	447.51	775.31	749.29	90.23	21.94	76.66	-3.39218	0.000117921788601278	0.00615
ENSG00000172899.16	EGFL7	233.92	128.91	235.49	50.32	67.91	34.07	-2.80032	0.000121814891106156	0.0063
ENSG00000080546.14	SESN1	84.75	74.7	107.04	19.09	18.81	23.66	-2.10188	0.000124496560245546	0.00639
ENSG00000147526.20	TACC1	33.9	56.67	97.31	10.41	5.22	5.68	-3.20598	0.000127546426307059	0.0065
ENSG00000154175.18	ABIBBP	50.85	36.06	83.69	1.74	6.27	7.57	-3.32055	0.000129060079906087	0.00653
ENSG00000037280.16	FLT4	54.24	48.94	64.23	6.94	12.54	4.73	-2.75873	0.000129872122197986	0.00653
ENSG00000267107.9	PCAT19	111.88	38.64	83.69	6.94	1.04	7.57	-3.88581	0.00013411636241015	0.00669
ENSG00000154783.12	FGD5	67.8	72.12	60.33	6.94	8.36	17.04	-2.55448	0.000136264866476324	0.00673
ENSG00000187479.9	C1orf96	555.99	589.86	568.3	64.2	153.58	62.47	-2.60547	0.000136699019223205	0.00673
ENSG00000162407.9	PLPP3	183.07	170	235.49	60.73	29.25	37.86	-2.2202	0.000141137168645647	0.0069
ENSG00000136244.12	IL6	88.14	59.24	344.48	5.21	9.4	7.57	-4.41806	0.000145852613306988	0.00708
ENSG00000185112.6	FAM43A	30.51	30.91	42.82	5.68	5.21	6.27	5.68	0.000148470113181642	0.00715
ENSG00000170890.14	PLA2G1B	30.51	43.79	27.25	0	0	2.84	-4.67062	0.000150345487617745	0.00715
ENSG00000143867.7	OSR1	40.68	38.64	35.03	6.94	2.09	5.68	-2.95996	0.000151185993528652	0.00715
ENSG00000169252.6	ADRB2	54.24	51.52	36.98	0	1.04	5.68	-4.07454	0.000151487872347923	0.00715
ENSG00000164736.6	SOX17	37.29	46.36	38.92	0	5.22	0.95	-4.05257	0.000155275032557775	0.00723
ENSG00000188783.6	PRELP	281.39	378.64	295.83	72.88	40.74	18.93	-2.87542	0.000156812930005163	0.00726
ENSG00000127329.16	PTPRB	64.41	56.67	101.2	14.63	13.25	13.25	-2.75997	0.000160165810705687	0.0073
ENSG00000258947.8	TUBB3	0	0	1.95	38.18	91.94	312.33	7.18715	0.000160921729214482	0.0073
ENSG00000114698.15	PLSCR4	30.51	33.49	52.55	3.47	7.31	2.84	-3.0455	0.00016367813623899	0.00738
ENSG00000101955.15	SRPX	33.9	28.33	52.55	5.21	3.13	6.63	-2.90897	0.000165247127650637	0.0074
ENSG00000113389.16	NPR3	44.07	51.52	97.31	1.74	9.4	7.57	-3.25242	0.000167027156674871	0.00744
ENSG00000071539.14	TRIP13	3.39	0	27.76	30.3	29.34	4.81269	0.000169149654573882	0.00748	
ENSG00000157456.8	CCNB2	0	2.58	1.95	34.7	100.29	34.07	4.94057	0.000172594168701237	0.00759
ENSG00000120156.22	TEK	44.07	90.15	33.09	3.47	2.09	7.57	-3.55385	0.000179938111775761	0.00781
ENSG00000116016.14	EPAS1	223.75	427.58	576.08	95.44	67.91	62.47	-2.45067	0.00018338567842987	0.00784
ENSG00000155066.16	PROM2	3.39	5.15	1.95	24.29	55.37	41.64	3.52992	0.000183553168338735	0.00784
ENSG00000076604.16	TRAF4	13.56	18.03	29.19	126.67	81.49	114.52	2.34681	0.00018549858224899	0.00784
ENSG00000104951.16	IL41	3.39	5.15	1.95	34.7	49.1	26.5	3.37404	0.000186917900859362	0.00784
ENSG00000125798.15	FOXA2	74.58	56.67	29.19	1.74	2.09	7.57	-3.60437	0.000187149960752946	0.00784
ENSG00000173269.14	MMRN2	57.63	51.52	87.58	17.35	7.31	8.52	-2.63213	0.000187325960561824	0.00784
ENSG00000117394.24	SLC2A1	33.9	12.88	11.68	199.55	344.76	98.43	3.52123	0.000189420390023744	0.00789
ENSG00000107317.13	PTGDS	1271.32	589.86	737.62	83.29	77.31	190.24	-2.88021	0.000191909108208154	0.0079
ENSG00000154065.17	ANKRD29	40.68	25.76	38.92	6.94	4.18	5.68	-2.6385	0.000191954011404678	0.0079
ENSG00000121691.7	CAT	210.19	329.7	235.55	50.32	28.21	70.98	-2.36533	0.000193448287127396	0.00791
ENSG00000104413.18	ESR1	33.9	33.49	19.46	124.94	128.5	105.06	2.06269	0.000195172866176738	0.00794
ENSG00000163072.16	NOSTRIN	57.63	64.39	31.14	6.94	4.18	9.46	-2.83156	0.00019848651280312	0.00798
ENSG00000130558.20	OLFM1	44.07	33.49	33.09	5.21	3.13	7.57	-2.72225	0.000202170963963547	0.00807
ENSG00000157778.9	PSMG3	20.34	12.88	11.68	57.26	71.04	79.5	2.2459	0.000204116679539269	0.00807
ENSG00000244734.4	HBB	13682.82	816.53	1605.63	32.97	11.49	54.89	-7.32763	0.000206768667328785	0.00813
ENSG00000143554.14	SLC27A3	111.88	131.37	108.99	24.29	20.89	38.8	-2.04843	0.000211291515360515	0.00824
ENSG0										

GeneID	GeneName	Non-transformed1	Non-transformed2	Non-transformed3	Cancer1	Cancer2	Cancer3	log2FC	pvalue	padj
ENSG00000175063.17	UBE2C	3.39	0	19.46	109.32	146.26	106.95	3.77647	0.000391780516549564	0.01198
ENSG00000088325.16	TPX2	6.78	0	5.84	39.91	51.19	38.8	3.31992	0.000393585689395967	0.01199
ENSG00000168734.14	PKIG	216.97	164.85	184.89	65.94	44.92	25.55	-2.0697	0.000396675391374251	0.01199
ENSG0000011732.14	PRKCE	33.9	36.06	58.39	10.41	4.18	8.52	-2.51149	0.000398916389211107	0.01199
ENSG00000123358.20	NRA1	467.85	244.7	412.6	111.05	86.71	48.27	-2.19883	0.000402144440393727	0.01199
ENSG00000160013.9	PTGIR	33.9	41.21	50.6	8.68	9.4	2.84	-2.62135	0.000428293710489134	0.01249
ENSG0000012822.16	CALCOYCO1	47.46	82.43	83.69	22.56	12.54	17.04	-2.06437	0.000430428131650812	0.01249
ENSG00000114378.17	HYAL1	44.07	41.21	40.87	3.47	10.45	9.46	-2.32564	0.000433916831064276	0.01249
ENSG00000214274.10	ANG	30.51	46.36	21.41	5.21	6.27	3.79	-2.63808	0.000434708986935007	0.01249
ENSG00000167641.11	PPP1R14A	328.85	249.85	219.92	48.59	6.27	35.97	-3.1574	0.000435023137610683	0.01249
ENSG00000060718.22	COL1A1	0	0	0	83.29	138.05	22.71	8.55587	0.000437401998847496	0.01251
ENSG00000108846.16	ABC3	23.73	7.73	31.14	85.03	133.73	180.77	2.64793	0.00043937800542024	0.01259
ENSG00000204301.6	NOTCH4	30.51	25.76	77.85	8.68	4.18	3.79	-3.10371	0.000454665678995647	0.01285
ENSG00000106511.6	MEOX2	30.51	28.33	25.3	3.47	1.04	5.68	-2.99365	0.000457367275385792	0.01287
ENSG00000142973.15	CYP4B1	257.65	370.91	527.43	5.21	2.09	41.64	-4.47763	0.000462761381640299	0.01297
ENSG00000096696.15	DSP	3.39	2.58	3.89	27.76	21.94	56.79	3.35566	0.000466947357697231	0.01304
ENSG00000105894.13	PTN	47.46	23.18	48.66	8.68	2.09	5.68	-2.9021	0.000476503789268776	0.01332
ENSG00000204839.9	MROH6	6.78	5.15	7.78	52.06	27.16	39.75	2.52485	0.000478518704147849	0.01332
ENSG00000169679.15	BUB1	6.78	2.58	3.89	26.03	34.48	25.55	2.71653	0.000489050430788811	0.01345
ENSG00000145247.12	OClAD2	50.85	59.24	33.09	138.82	182.83	336.94	2.21459	0.000501095891115319	0.01369
ENSG00000152661.9	GJA1	271.22	154.55	107.04	39.91	42.83	46.38	-2.02423	0.000503593517830506	0.01369
ENSG00000154330.13	PGM5	81.36	23.18	54.49	10.41	2.09	2.84	-3.48475	0.000504118365472816	0.01369
ENSG00000185033.15	SEMA4B	27.12	23.18	21.41	71.14	188.05	96.54	2.32166	0.000512647931323454	0.01387
ENSG00000176788.9	BASP1	23.73	25.76	52.55	282.84	118.05	155.22	2.39514	0.00052343890308744	0.01411
ENSG00000135063.20	FAM189A2	111.88	59.24	52.55	15.62	2.09	9.46	-3.08144	0.000539063410524255	0.01447
ENSG00000165507.9	DEPP1	464.46	486.82	794.06	119.73	121.19	32.18	-2.68286	0.000545420183420848	0.01459
ENSG00000134020.8	PBP4	352.58	406.98	62.28	1.74	1.04	19.88	-5.0133	0.000549910344216643	0.01465
ENSG00000196754.13	S100A2	6.78	18.03	5.84	46.85	846.23	2651.02	6.83577	0.000552207169711753	0.01466
ENSG00000113368.12	LMNB1	6.78	0	7.78	65.94	65.82	33.13	3.40806	0.000561568466139153	0.01485
ENSG00000184347.15	SLIT3	6.78	43.79	46.71	8.68	3.13	13.25	-2.60591	0.00056815178468187	0.01492
ENSG00000128050.9	PAICS	23.73	20.61	36.98	97.17	153.58	85.18	2.01508	0.00057444599888173	0.01498
ENSG00000110492.17	MDK	111.88	69.55	89.53	229.05	402.22	697.54	2.29916	0.000574905932381426	0.01498
ENSG00000046653.15	PGM6B	23.73	25.76	25.3	6.94	3.13	2.84	-2.59396	0.000580916590044097	0.01508
ENSG00000123500.10	COL10A1	6.78	0	93.7	47.01	3.01	27.45	4.84693	0.000601733252457035	0.01545
ENSG00000203883.7	SOX18	179.68	123.64	36.98	17.35	15.67	15.14	-2.78927	0.000610119595211273	0.01548
ENSG00000183048.12	SLC25A10	13.56	10.3	0	98.91	65.82	72.88	3.44941	0.000611718651326430	0.01548
ENSG00000168497.5	CAVIN2	494.97	461.07	648.09	12.15	108.65	34.07	-3.34658	0.0006118792414400732	0.01548
ENSG00000145113.22	MUC4	6.78	7.73	1.95	38.18	125.37	38.8	3.68203	0.000615447986666009	0.01552
ENSG00000105974.13	CAV1	416.99	613.04	805.73	10.41	107.61	43.54	-3.48084	0.000622807340387699	0.01556
ENSG00000276409.5	CCL14	189.85	126.21	169.32	12.15	5.22	38.8	-3.05489	0.000633830107806697	0.01581
ENSG00000064989.13	CALCRL	23.73	74.7	35.03	8.68	5.22	7.57	-2.64165	0.000642802324644593	0.01589
ENSG00000106541.12	AGR2	128.83	131.37	52.55	1018.58	573.56	329.37	2.62647	0.000644112466369289	0.01589
ENSG00000156804.7	FBXO32	13.56	7.73	11.68	71.14	40.74	162.79	3.05692	0.00064467201000542	0.01589
ENSG00000213853.10	EMP2	84.75	72.12	73.96	24.29	5.22	14.2	-2.48624	0.000647442853579544	0.01589
ENSG00000099994.11	SUSD2	793.3	759.86	286.09	6.94	7.31	69.09	-4.40318	0.000652987261649602	0.0159
ENSG00000121075.11	TBX4	44.07	61.82	40.87	5.21	0	7.57	-3.46797	0.00066024180918025	0.01602
ENSG00000100311.17	PDGFB	44.07	28.33	48.66	6.94	11.49	9.46	-2.0686	0.000671591805965627	0.01623
ENSG00000109501.15	WFS1	206.8	177.73	163.48	45.12	16.72	57.73	-2.19067	0.000682350961247071	0.01632
ENSG00000162772.17	ATF3	115.27	95.3	38.92	15.62	14.63	19.88	-2.26978	0.00068236902058147	0.01632
ENSG00000263155.6	MYZAP	30.51	61.82	23.35	0	3.13	5.68	-3.44154	0.000684407378011321	0.01632
ENSG00000154734.16	ADAMTS1	33.9	77.27	237.44	10.41	10.45	14.2	-3.30796	0.000693533406741153	0.01645
ENSG00000156298.13	TSPAN7	162.73	133.94	114.83	12.15	45.97	17.04	-2.40876	0.000694622508818291	0.01645
ENSG0000007402.12	CACNA2D2	50.85	54.09	42.82	0	0	7.57	-3.96319	0.000726488125754367	0.01709
ENSG00000124785.9	NRN1	37.29	36.06	23.35	3.47	0	4.73	-3.46116	0.000731913867701464	0.01716
ENSG00000165495.16	PKNOX2	27.12	23.18	23.35	0	4.18	2.84	-3.11713	0.0007399589900012	0.01719
ENSG00000129235.11	TXNDC17	84.75	54.09	58.39	180.46	234.02	527.18	2.26654	0.000748758956440722	0.0173
ENSG00000076382.17	SPAG5	0	2.58	7.78	59	32.39	33.13	3.27139	0.000750862055540509	0.0173
ENSG00000104267.10	CA2	44.07	59.24	130.4	13.88	15.67	18.93	-2.26887	0.000754885484611115	0.0173
ENSG00000105339.11	DENND3	30.51	56.67	81.74	8.68	12.54	15.14	-2.19501	0.000780157984045838	0.01774
ENSG00000159713.11	TPPP3	105.1	231.82	206.3	46.85	10.45	8.52	-3.13585	0.000781248713892611	0.01774
ENSG00000131153.9	GINS2	6.78	2.58	1.95	22.56	22.98	43.54	3.07489	0.00078281816863825	0.01774
ENSG00000074527.13	NTN4	40.68	79.85	36.98	8.68	11.49	14.2	-2.14503	0.000818970054046801	0.01832
ENSG00000050165.19	DKK3	98.32	128.79	239.38	52.06	33.43	27.45	-2.0741	0.00081934942322209	0.01832
ENSG00000040776.13	HSPB6	30.51	46.36	81.74	6.94	0	6.63	-3.52629	0.00082123313510041	0.01832
ENSG00000206384.11	COL6A6	54.24	30.91	62.28	0	2.09	9.46	-3.42884	0.000823083727260575	0.01832
ENSG00000253159.3	PCDHGA12	189.85	162.27	130.4	19.09	58.5	39.75	-2.00557	0.000829258563240205	0.0184
ENSG00000145506.14	NKD2	33.9	41.21	64.23	5.21	5.22	14.2	-2.4366	0.000831732357561767	0.0184
ENSG00000213088.12	ACKR1	250.87	90.15	108.99	24.29	8.36	33.13	-2.75009	0.000845136961748023	0.01858
ENSG00000101460.13	MAP1LC3A	88.14	118.49	70.06	24.29	26.12	8.52	-2.24227	0.000849532494892884	0.01859
ENSG00000035664.11	DAPK2	37.29	38.64	19.46	0	4.18	4.73	-3.1386	0.000861195714942911	0.01875
ENSG00000289027.1	ENSG00000289027	23.73	25.76	21.41	6.94	2.09	2.84	-2.64799	0.000864922699892406	0.01875
ENSG00000260244.1	ENSG00000260244	54.24	41.21	29.19	6.94	5.22	12.3	-2.26263	0.00086880087045447	0.01875
ENSG00000110328.6	GALNT18	101.71	59.24	147.91	29.5	7.31	17.04	-2.57064	0.00087969143869395	0.0191
ENSG00000120279.7	MYCT1	37.29	33.09	33.09	0	6.27	6.63	-2.89517	0.000917934084811662	0.01957
ENSG00000143344.16	RGL1	101.71	54.09	66.17	8.68	15.67	24.61	-2.10386	0.000931862201121816	0.01979
ENSG00000135363.12	LMO2	94.93	30.91	56.44	12.15	12.54	14.2	-2.18087	0.00093417584090438	0.01979
ENSG00000181019.13	NQO1	91.54	23.18	33.09	489.33	412.67	152.38	2.8687	0.000943264166638715	0.01992
ENSG0000009635.14	MAOB	23.73	59.24	46.71	1.74	9.4	8.52	-2.60709	0.000959887067043369	0.02008
ENSG00000196154.12	S100A4	1400.15	1839.12	998.41	510.16	196.41	265.95	-2.12543	0.00096528432778087	0.02009
ENSG00000164078.14	MST1R	6.78	2.58	3.89	20.82	32.39	23.66	2.56214	0.000975281332026521	0.02024
ENSG00000115380.20	EFEMP1	291.56	218.94	435.95	83.29	36.57	102.22	-2.09309	0.00100070284726992	0.02064
ENSG00000182253.15	SYNM	44.07	41.21	58.39	5.21	13.58	1.89	-2.7604	0.00101754829942945	0.02078
ENSG00000243244.7	STON1	44.07	25.76	19.46	3.47	6.27	5.68	-2.3929	0.00101913418812375	0.02078
ENSG00000078596.11	ITM2A	88.14	69.55	134.29	3.47	27.16	8.52	-2.82718	0.00103887030727112	0.02101
ENSG00000235750.10	KIAA0040	64.41	97.88	95.36	8.68	22.98	28.39	-2.05191	0.00104105973196684	0.02101
ENSG00000140284.11	SLC27A2	0	5.15	0	24.29	27.16	23.66	3.74892	0.00104237405521901	0.02101
ENSG00000138821.14	SL									

GeneID	GeneName	Non-transformed1	Non-transformed2	Non-transformed3	Cancer1	Cancer2	Cancer3	log2FC	pvalue	padj
ENSG00000108773.11	KAT2A	16.95	5.15	5.84	45.12	44.92	33.13	2.21991	0.00154563749830781	0.02652
ENSG00000142235.13	LMTK3	0	5.15	3.89	31.23	73.13	18.93	3.56836	0.0155420949079188	0.0266
ENSG00000137804.14	NUSAP1	3.39	0	11.68	62.47	82.53	28.39	3.30525	0.0158961689867837	0.02688
ENSG00000125740.14	FOSB	98.32	30.91	95.36	3.47	1.04	16.09	-3.31012	0.0159570442457345	0.02689
ENSG00000137834.15	SMAD6	30.51	64.39	62.28	13.88	3.13	13.25	-2.41357	0.0160001961184412	0.02689
ENSG00000089685.15	BIRC5	13.56	0	19.46	86.76	88.8	70.98	2.82965	0.0160176344752592	0.02689
ENSG00000159166.15	LAD1	10.17	23.18	13.62	29.5	188.05	171.31	3.0323	0.0161227977644899	0.02701
ENSG00000154146.13	NRGN	176.29	170	97.31	8.68	8.36	45.43	-2.75346	0.0163999281368624	0.02737
ENSG00000189409.14	MMP23B	88.14	30.91	85.63	10.41	6.27	20.82	-2.39759	0.0166150098900585	0.02757
ENSG00000183615.6	FAM167B	23.73	28.33	21.41	8.68	4.18	4.73	-2.09683	0.0169829423594756	0.02809
ENSG00000166979.13	EVA1C	74.58	51.52	62.28	5.21	24.03	8.52	-2.24104	0.0170996490350427	0.02809
ENSG00000113070.8	HBEGF	33.9	28.33	97.31	3.47	7.31	13.25	-2.65963	0.0173856285465668	0.02846
ENSG00000108823.17	SCCA	47.46	30.91	29.19	8.68	0	3.79	-3.21023	0.0173932879347856	0.02846
ENSG00000176945.18	MUC20	16.95	5.15	1.95	64.2	48.06	33.13	-2.73737	0.0174388578176778	0.02846
ENSG00000176971.4	FIBIN	54.24	177.73	101.2	10.41	31.34	5.68	-2.75643	0.01784026368852581	0.02892
ENSG00000113555.6	PCDH12	50.85	25.76	44.76	15.62	5.22	6.63	-2.21032	0.0180020240292081	0.02905
ENSG00000277196.4	ENSG00000277196	47.46	141.67	105.1	15.62	4.18	26.5	-2.65546	0.018586393982631	0.02972
ENSG00000182575.8	NXP3	23.73	33.49	35.03	8.68	1.04	0	-3.54773	0.019561058113156	0.03051
ENSG00000124664.11	SPDEF	6.78	7.73	1.95	29.5	19.85	44.48	2.57443	0.0200704375667513	0.03103
ENSG00000164741.15	DLG1	33.9	48.94	95.36	10.41	3.13	17.04	-2.5416	0.0020293932599295	0.03124
ENSG00000163687.14	DNASE1L3	50.85	41.21	62.28	3.47	1.04	15.14	-3.84961	0.00204366367646121	0.03132
ENSG00000079462.8	PAFAH1B3	13.56	25.76	11.68	183.93	59.55	54.89	2.52271	0.0021339403409577	0.03244
ENSG00000143324.14	XPR1	10.17	46.36	15.57	67.67	136.86	173.2	2.3754	0.00213960483965614	0.03244
ENSG00000190776.6	EPN1	61.02	136.52	85.63	12.15	37.61	16.09	-2.07298	0.00215016655260699	0.03253
ENSG00000171885.18	AQP4	193.24	92.73	118.72	0	0	20.82	-4.07213	0.00216471850014662	0.03268
ENSG00000164611.13	PTTG1	23.73	7.73	9.73	102.38	135.82	30.29	-2.76018	0.0021759320000989	0.03271
ENSG00000170276.6	HSPD2	27.12	30.91	23.35	6.94	1.04	6.63	-2.47534	0.00218300386156849	0.03271
ENSG00000103175.11	WFDRC1	105.1	33.49	48.66	19.09	4.18	2.84	-2.96389	0.0022581919963024	0.03334
ENSG00000134201.12	GSTM5	23.73	56.67	35.03	10.41	1.04	0	-3.6423	0.00233201653130657	0.03397
ENSG00000181104.7	F2R	88.14	54.09	192.68	36.44	24.03	22.71	-2.03985	0.00233706328148451	0.03397
ENSG00000107742.14	SPOCK2	71.19	72.12	142.07	1.74	18.81	24.61	-2.57571	0.00239006219024359	0.03424
ENSG00000183019.8	MCEMP1	91.54	159.7	175.16	10.41	1.04	34.07	-3.17378	0.00239240502941133	0.03424
ENSG00000276850.5	ENSG00000276850	23.73	28.33	19.46	0	6.27	1.89	-2.89035	0.0023961495737097	0.03424
ENSG00000179639.11	FCER1A	61.02	23.18	29.19	1.74	2.09	10.41	-2.77707	0.0024350817073987	0.03461
ENSG00000053918.18	KCNQ1	47.46	85	50.6	8.68	7.31	24.61	-2.10947	0.00245827740801852	0.0348
ENSG00000137857.18	DUOX1	47.46	97.88	23.35	1.74	2.09	13.25	-3.12009	0.00247746072840289	0.03489
ENSG00000077348.10	EXOSC5	6.78	18.03	7.78	29.5	56.42	45.43	2.00137	0.00258369794065159	0.03599
ENSG00000105538.10	RASIP1	101.71	146.82	124.56	45.12	6.27	6.63	-2.79965	0.00261298349535683	0.03633
ENSG00000105388.17	CEACAM5	13.56	25.76	3.89	414.23	106.56	87.07	6.66014	0.00270216331263857	0.03713
ENSG00000104783.15	KCNM4	13.56	0	11.68	24.29	128.5	95.59	3.30209	0.00282752471788112	0.03818
ENSG00000114346.14	ECT2	0	2.58	7.78	24.29	38.66	20.82	2.7312	0.00293354051208189	0.03923
ENSG00000094804.12	CDC6	0	7.73	3.89	24.29	30.3	20.82	2.54656	0.00308055014603133	0.04036
ENSG00000132182.13	NUA2	6.78	23.18	19.46	76.35	102.38	37.86	2.06224	0.00311633292166927	0.04052
ENSG00000109805.10	NCAPG	0	0	7.78	26.03	47.01	21.77	3.22083	0.00319913271564115	0.04145
ENSG00000076706.17	MCAM	172.9	59.24	325.02	53.79	22.98	37.86	-2.304	0.00323293193796431	0.04173
ENSG00000137807.16	KIF23	3.39	0	5.84	19.09	50.15	18.93	3.10085	0.00333147291646722	0.04259
ENSG00000166123.14	GPT2	0	0	7.78	248.14	19.85	44.48	4.90141	0.00333562949945208	0.04259
ENSG00000173848.19	NET1	6.78	28.33	25.3	50.32	82.53	188.34	2.35643	0.00337166877090381	0.04283
ENSG00000156802.13	ATAD2	6.78	10.3	13.62	55.53	84.62	21.77	2.31561	0.00341048431713184	0.04307
ENSG00000170312.17	CDK1	3.39	0	17.52	50.32	152.53	40.7	3.33485	0.00341951462878605	0.04311
ENSG00000106992.19	AK1	244.09	226.67	126.5	31.23	21.94	88.02	-2.05076	0.00349241770913444	0.04379
ENSG00000133216.17	EPHB2	3.39	2.58	13.62	53.79	20.89	53.95	2.52376	0.00350510697702527	0.04379
ENSG00000049540.19	ELN	91.54	64.39	247.17	12.15	6.27	39.75	-2.75667	0.00350864891699886	0.04379
ENSG00000161642.18	ZNF385A	3.39	10.3	19.46	57.26	63.73	32.18	2.06616	0.00360410051144547	0.04464
ENSG00000140525.20	FANCI	3.39	5.15	9.73	31.23	71.04	18.93	2.59921	0.00362235639330198	0.04471
ENSG00000105971.15	CAV2	193.24	149.4	311.39	13.88	85.67	41.64	-2.18521	0.00365757594469946	0.04497
ENSG0000007933.13	FMO3	44.07	38.64	38.92	19.09	3.13	5.68	-2.24409	0.00381473140376998	0.04607
ENSG00000288825.1	HZAC18	27.12	20.61	11.68	93.7	34.48	121.15	2.21032	0.00381650050823728	0.04607
ENSG00000122952.17	ZWINT	13.56	0	5.84	38.18	34.48	33.13	2.52065	0.00381800936202196	0.04607
ENSG00000106483.12	SFRP4	47.46	20.61	29.19	147.49	268.5	53	2.29004	0.00382448159190042	0.04607
ENSG00000133026.14	MYH10	27.12	64.39	159.59	24.29	16.72	11.36	-3.25666	0.00384784586007348	0.04619
ENSG00000164466.13	SFXN1	10.17	0	13.62	46.85	42.83	43.54	2.40801	0.00385540430686522	0.04621
ENSG00000186340.17	THBS2	0	12.88	31.14	102.38	148.35	62.47	2.67269	0.00392068952556599	0.04636
ENSG00000131747.15	TOP2A	0	0	25.3	0	93.7	79.4	3.10967	0.0041093496829532	0.04794
ENSG00000117791.16	MTARC2	50.85	41.21	42.82	0	10.45	12.3	-2.39542	0.00420930097443666	0.04839
ENSG00000078401.7	EDN1	81.36	48.94	157.64	8.68	4.18	32.18	-2.62276	0.0042468878313613	0.04866

TABLE 9: Significantly Differentially Expressed Transcripts in Non-Transcribed vs. Cancer.

TranscriptID	TranscriptName	GeneName	TranscriptNovelty	ISM	subtype	Non-transformed1	Non-transformed2	Non-transformed3	Cancer1	Cancer2	Cancer3	log2FC	pvalue	padj
ENST0000031325.0	FXYD1-201	FXYD1	Known	None	None	81.97	77.15	74.21	2.49	4.11	1.34	-5.37101	3.04495907107231e-06	0.00712
TALONT000370816	TALONT000370816	SPARCL1	ISM	Both	Both	14.07	160.73	105.61	7.48	4.22	5.37	-4.52643	5.06162528408178e-06	0.00712
TALONT000381758	TALONT000381758	ADH1B	ISM	Suffix	ISM	50.98	83.58	85.63	0	1.34	-6.52699	8.48520813447456e-06	0.00712	
TALONT000405760	TALONT000405760	AGER	ISM	Suffix	ISM	280.39	122.15	105.61	2.49	0	1.34	-6.93911	9.15638700163578e-06	0.00712
ENST0000041122.5	CAV1-206	CAV1	Known	None	None	154.43	96.44	154.43	4.49	5.63	5.37	-4.49456	1.077045187034e-05	0.00712
ENST0000046990.5	AGER-209	AGER	Known	None	None	169.93	96.44	59.94	0	1.34	-7.06846	1.0873568150175e-05	0.00712	
ENST0000023297.8	TNXC1-201	TNXC1	Known	None	None	131.7	70.72	85.63	4.99	1.41	1.34	-5.25543	1.20711037446709e-05	0.00712
TALONT000279133	TALONT000279133	PDZD2	ISM	Both	Both	50.98	41.79	59.94	2.49	1.41	1.34	-4.74943	1.67829122926905e-05	0.00712
TALONT000432337	TALONT000432337	TCF21	ISM	Suffix	ISM	437.58	143.61	402.4	32.42	29.54	5.37	-3.22166	1.74048829012709e-05	0.00712
ENST0000030900.9	CA4-201	CA4	Known	None	None	67.97	135.01	171.26	0	2.68	-6.5671	1.7414531823665e-05	0.00712	
ENST0000024861.6	GNGL1-201	GNGL1	Known	None	None	57.77	411.47	479.52	47.38	56.27	5.37	-3.2248	1.79365247700305e-05	0.00712
ENST0000037655.6	AGER-201	AGER	Known	None	None	169.93	109.3	174.11	0	4.03	-6.35894	1.8034063959825e-05	0.00712	
ENST00000367382.5	TCF21-201	TCF21	Known	None	None	152.94	186.45	276.87	7.48	8.44	17.45	-4.13931	2.04868617740536e-05	0.00712
TALONT000370868	TALONT000370868	SPARCL1	ISM	Suffix	ISM	80.72	70.72	79.92	2.49	4.22	6.71	-3.94986	2.0560020893226e-05	0.00712
TALONT000457959	TALONT000457959	CDH5	ISM	Both	Both	186.93	202.52	254.03	22.44	18.29	10.74	-3.65285	2.1682166029696e-05	0.00712
ENST0000034231.3	CLEC4E-201	CLEC4E	Known	None	None	301.63	302.17	353.93	27.43	21.1	38.93	-3.43697	2.16858087263953e-05	0.00712
ENST0000029790.4	VEGFD-201	VEGFD	ISM	Both	Both	67.97	67.51	88.48	0	2.68	-6.82054	2.21700473255458e-05	0.00712	
TALONT000484277	TALONT000484277	CAV1	ISM	Suffix	ISM	484.27	893.65	794.29	7.48	109.72	110.79	-3.24297	3.89635437786804e-05	0.00712
ENST0000037263.4	PLAC9-201	PLAC9	Known	None	None	305.88	112.51	245.47	9.97	11.25	14.77	-4.15294	2.42855985715012e-05	0.00712
TALONT000284485	TALONT000284485	RGCC	ISM	Suffix	ISM	33.99	61.08	68.5	0	1.34	-6.10639	2.4342054034920e-05	0.00712	
ENST0000025796.10	RAMP2-201	RAMP2	Known	None	None	212.42	273.24	322.54	9.97	18.29	26.85	-3.82965	2.50772139411763e-05	0.00712
TALONT000405804	TALONT000405804	AGER	ISM	Suffix	ISM	61.08	77.07	0	0	-8.38128	2.7037573605239e-05	0.00712		
ENST0000018305.10	CLDN18-201	CLDN18	Known	None	None	144.44	241.31	305.41	2.49	0	6.71	-6.07306	2.8513376714060e-05	0.00712
TALONT000342272	TALONT000342272	A2M	ISM	Suffix	ISM	152.94	163.94	154.13	9.97	15.47	20.14	-3.31281	2.9077018648721e-05	0.00712
ENST00000376075.4	COX4I2-201	COX4I2	Known	None	None	80.72	61.08	51.38	2.49	4.22	5.37	-3.83154	3.01630433749054e-05	0.00712
TALONT000272771	TALONT000272771	VWF	ISM	Both	Both	161.44	54.45	122.74	2.49	1.41	4.03	-4.69889	3.0283758127093e-05	0.00712
ENST0000064607.2	EDNRB-207	EDNRB	Known	None	None	50.98	61.08	88.48	2.49	5.63	2.68	-4.0785	3.09966433769643e-05	0.00712
TALONT000284236	TALONT000284236	RGCC	ISM	Suffix	ISM	76.47	41.79	85.63	4.99	1.41	1.34	-4.77161	3.12680828845103e-05	0.00712
TALONT000370809	TALONT000370809	SPARCL1	ISM	Suffix	ISM	195.42	231.45	339.66	19.95	25.32	30.88	-3.30655	3.63327170540417e-05	0.00712
TALONT000342325	TALONT000342325	A2M	ISM	Suffix	ISM	271.89	398.61	356.79	32.42	26.73	48.33	-3.24297	3.89635437786804e-05	0.00712
ENST0000027153.10	CSRP1-201	CSRP1	Known	None	None	128.41	112.51	128.41	9.97	9.85	14.77	-3.85301	3.853013470674e-05	0.00712
TALONT000341267	TALONT000341267	CAMK2N1	ISM	Both	Both	93.46	173.59	77.07	9.97	8.44	6.71	-3.76161	4.0278386327713e-05	0.00712
ENST000004848.9.5	AGER-211	AGER	Known	None	None	63.72	64.29	40.2	0	2.68	-6.40779	4.18895036292003e-05	0.00809	
ENST00000379359.4	RGCC-201	RGCC	Known	None	None	815.68	1022.23	1675.48	112.22	37.98	92.63	-3.85867	4.38266158314143e-05	0.00822
ENST00000360997.7	FAM107A-201	FAM107A	Known	None	None	161.44	122.74	141.41	2.49	1.41	4.03	-4.69889	4.56422834126236e-05	0.00712
ENST00000372013.8	ADRF-201	ADRF	Known	None	None	926.13	1356.55	382.48	37.41	58.96	17.45	-5.59392	4.635614713377e-05	0.00823
TALONT000405770	TALONT000405770	AGER	NKC	None	None	89.21	86.79	57.09	2.49	0	4.03	-4.99902	4.8684766425576e-05	0.00839
ENST00000238044.8	ECRG4-201	ECRG4	Known	None	None	46.73	64.29	39.96	2.49	0	0	-5.29993	5.11866711769586e-05	0.00859
ENST00000373800.1	DES-201	DES	Known	None	None	161.44	77.15	108.46	19.97	1.41	4.03	-4.29428	5.35695614126236e-05	0.00909
ENST00000296130.5	CLEC4B-201	CLEC4B	Known	None	None	256.53	684.7	562.3	19.95	8.44	49.67	-4.5021	5.4434375067537e-05	0.00868
ENST0000056626.2	TMEM204-201	TMEM204	Known	None	None	512.42	231.45	202.66	37.41	29.1	25.1	-2.90188	5.7139254881472e-05	0.00889
TALONT000541095	TALONT000541095	PECAM1	ISM	Both	Both	420.58	356.82	336.81	67.33	42.2	44.3	-3.89071	6.4249936084576e-05	0.00909
ENST00000473019.5	AGER-210	AGER	Known	None	None	169.93	141.08	62.79	2.49	0	5.37	-4.11881	6.40956201747745e-05	0.00909
TALONT000284459	TALONT000284459	RGCC	ISM	Suffix	ISM	190.21	61.08	108.46	4.99	1.41	8.05	-4.17422	6.5675878127093e-05	0.00909
ENST0000049662.6	ICAM2-203	ICAM2	Known	None	None	76.47	57.86	125.59	2.49	8.44	5.37	-3.85642	6.7627289700589e-05	0.00909
ENST0000042486.3	TMEM100-201	TMEM100	Known	None	None	59.48	15.43	37.11	2.49	1.41	2.68	-5.11005	6.8133331697459e-05	0.00909
TALONT000398208	TALONT000398208	TNXC1	ISM	Suffix	ISM	46.73	41.79	39.96	2.49	0	1.34	-4.97057	6.86866047143318e-05	0.00909
ENST0000022987.8	COX7A1-201	COX7A1	Known	None	None	161.44	125.59	125.59	24.94	18.28	13.42	-2.9558	7.062050073067216e-05	0.00909
TALONT000405794	TALONT000405794	AGER	NKC	None	None	45	45.67	0	0	2.68	-5.37887	7.1109248248134e-05	0.00909	
TALONT000577197	TALONT000577197	ESAM	ISM	Both	Both	118.95	48.22	105.65	7.48	7.03	8.05	-3.54996	7.15107215267446e-05	0.00909
TALONT000342292	TALONT000342292	A2M	ISM	Suffix	ISM	429.08	265.41	340.74	3.99	15.47	40.27	-3.43181	7.43044828635414e-05	0.00909
ENST00000592696	TALONT000592696	FMO2	Known	None	None	53.99	73.92	79.92	2.49	1.41	5.37	-4.19883	7.48540705223906e-05	0.00909
ENST0000051403.1	TALONT00051403.1	SELENO1P	ISM	Suffix	ISM	72.22	99.65	77.07	9.97	9.85	14.77	-2.92688	7.5821422642345e-05	0.00909
ENST0000053245.1	TALONT00053245.1	LTBP4	ISM	Suffix	ISM	106.21	173.59	137.01	17.48	18.29	5.37	-3.68389	7.6918238202634e-05	0.00909
TALONT000363255	TALONT000363255	SPTBN1	ISM	Both	Both	492.81	392.18	411.02	59.85	81.59	67.12	-3.26289	8.08138644378332e-05	0.00924
TALONT000370874	TALONT000370874	SPARCL1	ISM	Suffix	ISM	72.22	67.07	77.07	2.49	5.63	9.4	-4.34462	8.1749005127529e-05	0.00924
TALONT000342299	TALONT000342299	A2M	ISM	Suffix	ISM	190.21	45.67	148.95	4.99	1.41	4.03	-3.68492	8.26324717327912e-05	0.00924
TALONT000342279	TALONT000342279	A2M	ISM	Suffix	ISM	637.25	511.12	488.09	64.84	26.73	76.52	-3.28028	8.4336538558541e-05	0.00927
ENST00000403084.1	CLDN5-201	CLDN5	Known	None	None	1134.3	922.58	2369.07	39.9	60.49	149.01	-4.13455	8.80574765704648e-05	0.00952
TALONT000405798	TALONT000405798	AGER	ISM	Suffix	ISM	114.7	80.36	48.52	0	4.03	-6.54475	8.98251877675312e-05	0.00955	
TALONT000342476	TALONT000342476	ADM	Known	None	None	97.71	61.08	108.46	19.97	1.41	4.03	-3.51119	9.5042801142721e-05	0.00991
ENST0000018155.3	ECSCR-201	ECSCR	Known	None	None	67.97	45	88.48	7.48	2.81	6.71	-3.55674	9.770319239205e-05	0.00991
ENST00000495709.1	A2M-207	A2M	Known	None	None	692.48	472.54	174.11	37.41	16.88	28.19	-4.02405	9.79317012842498e-05	0.00991
TALONT000381824	TALONT000381824	ADH1B	ISM	Suffix	ISM	70.72	80.36	105.61	4.99	0	5.37	-4.60962	0.00010428892737634	0.00991
TALONT000484788	TALONT000484788	GF3X	ISM	Suffix	ISM	50.98	73.92	85.63	4.99	9.85	8.05	-3.11885	0.00010724830617433	0.00991
TALONT000392670	TALONT000392670	BAG1	ISM	Both										

TranscriptID	TranscriptName	GeneName	TranscriptNovelty	ISM_subtype	Non-transformed1	Non-transformed2	Non-transformed3	Cancer1	Cancer2	Cancer3	log2FC	pvalue	padj	
ENST0000029254.9	GYPC-201	GYPC	Known	None	208.17	176.8	282.58	49.87	29.54	48.33	-2.8622	0.00053572767922283	0.01554	
ENST0000059886.6	ADRB2-201	ADRB2	Known	None	67.97	64.29	54.23	0.0	6.71	-4.4321	0.00036143534060399	0.00198		
ENST0000019837.8	ADRB2-201	ADRB2	Known	None	45.87	106.08	85.63	24.94	14.77	-2.3460	0.003026705101777	0.01677		
ENST0000058825.9	GPX3-201	GPX3	Known	None	675.48	1424.06	1655.5	92.27	54.86	228.87	-3.2149	0.0003829826670074179	0.01627	
ENST0000057264.4	PLAU-201	PLAU	Known	None	0	6.14	8.56	40.98	14.16	154.38	4.0696	0.000387474833879046	0.01645	
TALON000272688	TALON000272688	VWF	ISM	Suffix	72.22	54.05	47.67	7.48	0	-3.0923	0.000389524820516822	0.01645		
TALON000442920	TALON000442920	KLF2	ISM	Suffix	263.4	292.53	154.11	42.39	9.85	33.56	-1.0003	0.000389321570197995	0.01648	
ENST00000504613	ADRH1-201	ADRH1	Known	None	543.26	616.53	7.48	0	25.51	-5.5301	0.00039044600000777	0.01648		
TALON000368386	TALON000368386	MGP	ISM	Suffix	63.72	41.79	62.79	12.47	2.81	5.37	-3.1097	0.00039636171774667	0.01648	
TALON000398833	TALON000398833	SUSD2	ISM	Suffix	101.96	106.08	79.92	0	1.41	12.08	-1.16503	0.000403876369417235	0.01648	
ENST0000028807.3	PSMG3-202	PSMG3	Known	None	16.99	16.07	14.27	69.82	78.78	91.29	2.33685	0.00040400675873901	0.01648	
ENST0000029280.8	PSPI-203	PSPI	Known	None	42.8	28.54	100.95	84.4	242.98	4.49811	0.000404229910492331	0.01648		
ENST0000057988.6	ICAM2-208	ICAM2	Known	None	38.23	54.65	51.38	4.99	8.44	9.4	-2.57572	0.00041990717877996	0.0166	
ENST00000248071.6	KLF2-201	KLF2	Known	None	195.42	282.88	291.14	54.86	30.95	61.75	-2.83611	0.0004278025465702	0.01686	
ENST0000023021.10	V5IC2-201	V5IC2	Known	None	46.73	70.72	59.14	7.48	0	44.89	-3.83039	0.0004303862353072805	0.01686	
TALON000342285	A2M	A2M	ISM	Suffix	28.23	28.54	27.09	2.49	1.41	6.71	-3.4559	0.0004311630311847	0.01686	
ENST00000282470.11	SPARCL1-201	SPARCL1	Known	None	271.89	270.23	676.47	44.89	43.61	69.81	-2.88467	0.00043158514945586	0.01686	
TALON000285163	TALON000285163	NPR 3.00	ISM	Both	50.98	64.29	97.05	2.49	12.66	9.1	-2.99137	0.0004323803807416	0.01686	
ENST0000047526.1	HHB-203	HHB	Known	None	1584.62	99.65	111.32	2.49	0	5.37	-7.66498	0.000434817528414375	0.01686	
ENST0000050690.10	SCL2A10-202	SCL2A10	Known	None	28.23	28.54	28.54	44.89	26.73	48.33	4.85339	0.000435299191023231	0.01686	
TALON000305387	TALON000305387	SELENOP	ISM	Suffix	38.23	34.57	38.23	7.48	2.81	2.68	-1.50709	0.000438069711127103	0.01688	
TALON000370826	TALON000370826	SPARCL1	ISM	Suffix	38.23	34.57	38.23	7.48	2.81	2.68	-1.50709	0.000438069711127103	0.01688	
ENST0000039318.8	PPP1R14B-201	PPP1R14B	Known	None	114.7	57.86	154.13	526.17	745.55	475.23	2.36898	0.00044098843240005	0.01688	
ENST0000012142.1	TPS21-203	TPS21	Known	None	84.97	106.08	142.72	19.95	23.91	100.95	84.4	242.98	0.01688	
ENST0000056455.9	UBE2C-204	UBE2C	Known	None	0	0	14.27	99.75	135.04	118.14	4.38855	0.000450431265856589	0.01689	
ENST00000566274	TALON000566274	HSPD1	ISM	Suffix	8.5	6.43	5.71	57.36	50.64	36.25	2.77881	0.000457062487565343	0.01704	
TALON000342344	TALON000342344	A2M	ISM	Suffix	55.23	125.37	142.72	14.96	9.85	21.48	-2.78853	0.000460517408105707	0.01717	
TALON000378666	TALON000378666	TAC1C1	ISM	Suffix	28.23	28.54	28.23	2.48	-0.4149	0.0004640623588363	0.0004640623588363	0.01717		
TALON000304214	TALON000304214	TGFBF2	ISM	Suffix	33.99	61.08	39.96	7.48	7.03	6.71	-2.64082	0.0004890852368997	0.01717	
TALON000260071	TALON000260071	IL3	ISM	Both	42.48	57.86	57.86	2.49	11.25	8.05	-2.80072	0.0004819658571779	0.01758	
ENST0000071625.8	PTGDS-202	PTGDS	Known	None	955.87	504.69	842.02	72.32	80.18	196	-2.71459	0.0004852533389709	0.01758	
ENST0000027738.1.1	CFD	CFD	ISM	Suffix	105.42	802.71	802.71	62.44	62.44	62.44	-3.67881	0.00049247409369575	0.01758	
TALON000541111	TALON000541111	PECAM1	ISM	Suffix	42.48	41.79	34.25	7.48	5.63	2.68	-2.91552	0.00050421752880978	0.01806	
ENST0000068708.9	S10A2-202	S10A2	Known	None	0	6.43	0	32.42	694.91	2390.91	8.70477	0.0005076835237041	0.01806	
ENST0000019083.3	CN3-201	CN3	Known	None	55.23	99.65	119.88	19.95	4.22	5.37	-3.3922	0.0005129910236013	0.01816	
TALON000383844	TALON000383844	GAPDH	ISM	Suffix	106.21	106.21	59.94	0	6.71	0.0005129910236013	0.0005129910236013	0.01816		
TALON000401055	TALON000401055	ENSG00000285043	ISM	Suffix	16.99	16.99	25.69	0.0	97.25	84.4	103.37	2.66057	0.000519662895468326	0.01816
TALON000342494	TALON000342494	A2M	ISM	Suffix	38.23	23.14	137.01	12.47	4.42	5.37	-3.21237	0.000521109282411	0.01816	
TALON000543063	TALON000543063	FOXO3	ISM	Both	140.19	186.45	219.78	37.41	42.2	46.99	-2.10077	0.00052409981485708	0.01816	
TALON000272778	TALON000272778	VWF	ISM	Suffix	105.42	131.8	185.53	4.95	4.95	5.63	15.42	-2.22414	0.0005249823583636	0.01816
TALON000428747	TALON000428747	NR4A1	ISM	Suffix	140.19	67.51	111.32	24.94	11.25	6.71	-2.95016	0.000534139213223637	0.01816	
ENST0000060687.2	CRYAB-217	CRYAB	Known	None	97.71	86.79	91.34	17.46	4.22	17.45	-2.82531	0.000534964389171	0.01816	
ENST0000029759.6	FAM143A-201	FAM143A	Known	None	38.23	38.67	62.79	7.48	8.44	8.05	-2.35551	0.00053533470640084	0.01816	
ENST0000022804.1	CCL2-201	CCL2	Known	None	131.8	185.53	21.48	27.48	21.48	21.48	-2.88109	0.00054731151679009	0.01834	
TALON000342419	TALON000342419	A2M	ISM	Suffix	59.48	65.05	49.84	5.99	5.63	20.14	-3.28904	0.00054817916109903	0.01834	
TALON000272826	TALON000272826	VWF	ISM	Suffix	55.23	48.22	51.38	4.99	4.22	8.05	-2.83463	0.00054944240991849	0.01834	
TALON000343369	TALON000343369	PHRC2	ISM	Both	29.74	45	28.54	7.48	2.81	6.71	-2.8967	0.00055653053898939	0.01834	
ENST0000027897.1	EPAS1-206	EPAS1	Known	None	288.89	572.19	185.53	30.88	19.95	11.25	30.88	0.00056189584633166	0.01848	
ENST0000066465.5	EPAS1-206	EPAS1	Known	None	59.48	93.22	117.03	19.95	7.03	16.11	-2.67397	0.00056236708132827	0.01848	
ENST00000578482.7	TPSANT-202	TPSANT	Known	None	63.72	106.08	99.9	12.47	21.1	13.42	-2.30652	0.0005703277208971	0.01848	
ENST00000320623.10	NQO1-201	NQO1	Known	None	63.72	16.07	25.69	466.32	355.9	161.09	3.2427	0.000592503778711409	0.01918	
TALON000300064	TALON000300064	EPAS1	ISM	Suffix	105.42	105.42	231.54	231.54	231.54	231.54	2.30451	0.0006025135925921048	0.01918	
TALON000365688	TALON000365688	EMP1	ISM	Both	475.81	276.45	596.55	7.98	88.62	115.45	-2.24225	0.000612789533291048	0.01958	
ENST00000342265	TALON000342265	A2M	ISM	Suffix	89.21	93.22	114.13	7.48	8.44	26.85	-2.90959	0.00061914229673729	0.01958	
ENST00000395659.9	MDR-203	MDR	Known	None	55.23	41.79	77.07	216.95	306.66	675.25	2.77616	0.00061944381158362	0.01958	
ENST00000514958.5	SELENOP-214	SELENOP	Known	None	106.21	106.21	57.86	62.44	128.88	128.88	0.00062186236943166	0.00062186236943166	0.01958	
ENST0000055426.8	EFEMP1-201	EFEMP1	Known	None	42.48	57.86	125.59	14.96	5.63	8.05	-3.0405	0.000623490859804908	0.01958	
TALON000446389	TALON000446389	CTSH	ISM	Suffix	118.95	83.58	59.94	23.91	13.42	-2.25363	0.000641022379240551	0.02		
ENST00000256044.8	PEBP4-201	PEBP4	Known	None	225.16	305.38	68.5	2.49	0	14.77	-4.95618	0.00064313886680679	0.02	
ENST0000028591.4	CCL2-201	CCL2	Known	None	140.19	140.19	145.66	145.66	145.66	145.66	0.00064313886680679	0.00064313886680679	0.02	
TALON000284944	TALON000284944	SFTPC	NIC	None	29.74	295.74	0	0	8.05	-5.60331	0.0006612846209003164	0.02031		
TALON000595680	TALON000595680	UBE2T	ISM	Suffix	0	0	2.85	29.92	26.73	28.19	43.7135	0.000665495768121044	0.02031	
ENST0000038922.9	ACVR1L-201	ACVR1L	Known	None	29.74	48.22	11.32	2.49	56.3	0	-3.88392	0.0006671975543269	0.02031	
ENST00000516512.3	PKP3-201	PKP3	Known	None	8.5	16.07	11.42	68.47	91.44	68.47	2.65071	0.00067055884633166	0.0205	
ENST0000041473.7	HMG1A1-204	HMG1A1	Known	None	0	3.21	14.27	82.29	90.03	63.1	3.54636	0.00069250762758536	0.02083	
TALON000387890	TALON000387890	OLEML2A	ISM	Both	46.73	45	59.94	7.48	11.25	12.08	-2.23858	0.00069601661147523	0.02083	
ENST00000579818.4	MTM1-201	MTM1	Known	None	72.22	456.47	137.01	9.97	19.69	2.68	-4.33142	0.0007064307473332	0.02085	
ENST00000467371.1	SPRY1-201	SPRY1	Known	None	12.74	11.42	11.42	7.47	11.25	5.37	-2.6737	0.0007064307473332	0.02085	

TranscriptID	TranscriptName	GeneName	TranscriptNovelty	ISM	subtype	Non-transformed1	Non-transformed2	Non-transformed3	Cancer1	Cancer2	Cancer3	log2FC	pvalue	padj	
ENST00000569419.1	ENST00000569419	ENSG00000290244	Known	Noae		67.97	51.43	42.81	9.97	7.03	17.45	-2.1744	0.001708818998286	0.0341	
ENST0000055321.4	SUSD2-201	SUSD2	Known	Noae		118.95	228.23	122.74	0	4.22	26.85	-3.76747	0.0017364684409096	0.0341	
ENST00000512576.3	PAICS-207	PAICS	Known	Noae		4.25	9.64	11.42	49.87	40.79	38.93	-2.25466	0.0017403307805423	0.0341	
ENST00000394825.6	NR4A1-204	NR4A1	Known	Noae		106.21	96.44	205.51	32.42	37.98	14.77	-2.26694	0.00174338340706192	0.0341	
ENST00000301242.9	PPP1R14A-201	PPP1R14A	Known	Noae		305.88	206.38	285.43	57.36	5.63	44.3	-3.00628	0.0017687874978768	0.03434	
ENST0000066609.6	BOP1-205	BOP1	Known	Noae		12.74	6.43	14.27	62.34	45.01	45.64	-2.15664	0.0017708453061991	0.03434	
ENST0000069550.10	DKC1-201	DKC1	Known	Noae		4.25	12.86	5.71	42.39	35.17	49.67	-2.40953	0.00177195138282908	0.03434	
TALNT0000354811	VSIR	VSIR	ISM	Suffix		29.74	38.57	34.25	7.48	7.03	8.05	-2.14885	0.0017964993705204	0.03447	
TALNT0000403503	FBLN1	FBLN1	ISM	Suffix		42.48	54.05	57.09	14.96	2.81	10.74	-2.49047	0.0017965213040409	0.03447	
TALNT0000301880	GAPPD	GAPPD	ISM	Suffix		101.96	48.22	85.63	403.38	530.33	189.29	-2.53007	0.0017973847396571	0.03447	
TALNT0000412156	ENG	ENG	ISM	Both		237.12	225.02	91.34	42.39	45.01	49.99	-2.24384	0.0018055381898484	0.03447	
ENST0000068584.1	H2AW-202	H2AW	Known	Noae		4.25	12.86	2.85	74.81	35.17	40.27	-2.82775	0.001805875130479	0.03447	
TALNT0000468842	PTGDS	PTGDS	ISM	Suffix		186.93	54.05	74.21	7.48	2.81	21.48	-3.21869	0.001818258282892636	0.03461	
ENST00000327858.11	FBLN1-202	FBLN1	Known	Noae		67.97	215.38	199.8	34.91	32.35	38.93	-2.1841	0.001818740940573538	0.03514	
ENST00000342058.9	FBLN5-202	FBLN5	Known	Noae		46.73	54.05	185.53	14.96	4.22	16.11	-3.04113	0.0018039743188092	0.03571	
TALNT0000334128	SFTPC	SFTPC	ISM	Suffix		858.16	668.63	516.63	4.99	0	57.73	-4.95395	0.00181886691754406	0.03571	
TALNT0000277204	BASP1	BASP1	ISM	Both		29.74	28.93	71.36	339.14	201.37	23.5005	0.00194420571905841	0.03629		
ENST00000360000.8	TCEAL8-201	TCEAL8	Known	Noae		33.99	32.15	28.54	7.48	7.03	5.37	-2.22131	0.00196300915289152	0.03649	
TALNT0000338367	SFTPC	SFTPC	ISM	Suffix		271.89	498.26	259.74	0	0	30.88	-4.91439	0.001967051818762	0.03731	
TALNT000520690	COL1A1	COL1A1	ISM	Suffix		12.74	3.21	5.71	47.38	45.01	30.88	-2.53519	0.00204286548629531	0.03765	
TALNT000320183	ABCA3	ABCA3	ISM	Suffix		29.74	28.93	42.81	4.99	1.41	8.05	-2.75133	0.0020302975985663	0.03765	
TALNT000559875	COL3A1	COL3A1	ISM	Suffix		4.25	6.43	14.27	39.9	88.62	61.75	-3.27872	0.00210562931347862	0.03858	
TALNT000333335	SFTPC	SFTPC	ISM	Suffix		543.79	1086.53	570.86	2.49	0	56.38	-5.14298	0.0021267235988667	0.03885	
TALNT000560154	SASH1	SASH1	ISM	Both		118.95	51.43	117.03	27.43	15.47	25.51	-2.06953	0.0021435058631863	0.03905	
TALNT000446397	CTSH	CTSH	ISM	Both		280.39	109.3	45.67	9.97	25.32	20.14	-2.89982	0.0021606142709299	0.03925	
ENST00000473336.5	RAB25-206	RAB25	Known	Noae		8.5	3.21	8.56	32.42	50.64	30.88	-2.46365	0.00229042189672303	0.03986	
TALNT0000446415	CTSH	CTSH	ISM	Suffix		135.95	64.29	42.81	14.96	14.07	20.14	-2.25451	0.002213761876176	0.03997	
ENST00000288666	SFTPC	SFTPC	ISM	Suffix		586.27	1375.84	804.91	4.99	0	69.81	-5.14617	0.0022454261815419	0.04025	
ENST00000494979.1	MARCO-203	MARCO	Known	Noae		72.22	51.43	37.11	7.48	0	9.4	-3.21663	0.0022576836481217	0.04025	
ENST0000052332.1	PDIA4-204	PDIA4	Known	Noae		25.49	16.07	59.94	182.04	112.54	265.81	-2.43412	0.00226209643412058	0.04025	
ENST0000022226.1	PDLIM2-220	PDLIM2	Known	Noae		106.21	48.22	34.25	9.97	15.47	10.74	-2.30736	0.0022829835967487	0.04025	
ENST0000045674.2	RPST-210	RPST	Known	Noae		12.74	19.29	19.98	42.39	150.52	84.57	-2.38958	0.00228385854922995	0.04055	
TALNT0000344786	P2R	P2R	ISM	Both		106.21	61.08	196.95	34.91	25.32	17.45	-2.24915	0.0022982771563654	0.04055	
TALNT0000553671	OLFML3	OLFML3	ISM	Both		123.2	106.08	74.21	19.95	37.98	14.77	-2.03202	0.00230042189672303	0.04055	
ENST00000318602.12	A2M-201	A2M	Known	Noae		212.42	469.23	1233.06	49.87	46.42	139.62	-3.01048	0.0023127301612384	0.04055	
TALNT0000300689	GAPDH	GAPDH	ISM	Suffix		106.21	77.15	77.07	311.71	682.25	202.71	-2.20209	0.0023151145994507	0.04055	
ENST00000677950.1	NDUFB9-214	NDUFB9	Known	Noae		8.5	6.43	14.27	94.76	36.57	44.3	-2.50069	0.0023405177918506	0.04073	
ENST0000038931.8	JUP-203	JUP	Known	Noae		12.74	14.27	14.27	59.85	43.61	48.33	-2.2943	0.0023385078364559	0.04073	
TALNT0000357780	EPCAM	EPCAM	ISM	Suffix		12.74	16.07	2.85	57.36	43.61	71.15	-2.45755	0.002346752306648	0.04073	
ENST0000039261.6	MGP-203	MGP	Known	Noae		1091.82	1170.11	2874.29	426.42	68.93	241.64	-2.8053	0.0023723666452191	0.04094	
TALNT0000520842	COL1A1	COL1A1	ISM	Suffix		8.5	3.21	14.27	32.42	80.18	48.33	-2.58412	0.0023750531563854	0.04094	
TALNT0000429047	NR4A1	NR4A1	ISM	Suffix		93.46	35.36	42.81	14.96	12.66	6.71	-2.3109	0.00240696926781907	0.04125	
ENST00000407067.1	MDK	MDK	Known	Noae		16.99	3.21	8.56	29.92	66.12	123.51	-2.88663	0.00241332112587266	0.04145	
TALNT0000329693	COL1A1	COL1A1	ISM	Suffix		50.98	48.22	48.52	4.99	2.81	16.11	-2.52195	0.00242474188787026	0.04175	
TALNT0000305366	SELENOP	SELENOP	ISM	Suffix		50.98	48.22	48.52	4.99	2.81	16.11	-2.52195	0.00242474188787026	0.04175	
TALNT0000285220	SFTPC	SFTPC	ISM	Suffix		50.98	138.23	117.03	0	0	16.11	-4.02967	0.002429036127508964	0.04205	
TALNT000477908	SYN1	SYN1	ISM	Both		50.98	48.22	74.21	7.48	18.29	2.68	-2.6142	0.0025254229088117	0.04205	
ENST0000023957.9	EEF1B2-201	EEF1B2	Known	Noae		16.99	6.43	14.27	47.38	67.52	45.64	-2.09462	0.002526932295754	0.04205	
ENST00000366615.10	COA6-203	COA6	Known	Noae		8.5	6.43	22.83	64.84	99.88	46.99	-2.41404	0.00255374651672733	0.04205	
ENST00000264522.14	UBE2S-201	UBE2S	Known	Noae		16.99	19.29	54.23	271.81	381.22	69.81	-2.9537	0.00255666630014067	0.04213	
TALNT0000594340	HGL1	HGL1	ISM	Both		106.21	48.22	48.52	4.99	15.47	17.45	-2.30561	0.00262206247297537	0.04299	
TALNT0000264176	ENSG00000250171		ISM	Both		39.48	131.8	34.25	4.99	9.85	18.79	-2.64746	0.00262206247297537	0.04299	
ENST0000004982.6	HSPB6-201	HSPB6	Known	Noae		106.21	48.22	114.17	9.97	0	9.4	-3.39734	0.00262668152784391	0.04301	
TALNT0000560257	COL3A1	COL3A1	ISM	Suffix		21.24	6.43	34.25	104.74	133.64	76.52	-2.31635	0.00269998877051479	0.04402	
ENST00000372898.7	ITIH2A-201	ITIH2A	Known	Noae		59.48	48.22	142.72	4.99	22.51	4.03	-2.91565	0.0027398112363545	0.04461	
TALNT0000571132	RAN	RAN	ISM	Suffix		9.64	9.64	11.42	37.41	59.08	49.67	-2.64884	0.0027398112363545	0.04461	
ENST0000030411.5	CAVIN2-201	CAVIN2	Known	Noae		148.69	138.23	488.09	2.49	46.42	22.82	-3.37025	0.002807675682142	0.04509	
ENST00000318627.4	FIBIN-201	FIBIN	Known	Noae		67.97	221.81	148.42	-2.72074	14.96	42.2	8.05	-2.72074	0.0028418093552992	0.04553
TALNT0000329262	ACR3	ACR3	ISM	Suffix		67.97	38.57	31.4	2.49	1.41	12.08	-2.92484	0.00287281291608854	0.04579	
ENST00000372838.9	CERCAM-201	CERCAM	Known	Noae		8.5	6.43	11.42	37.41	36.57	37.59	-2.63417	0.0028828308099388	0.04608	
TALNT0000402367	CD52	CD52	ISM	Both		144.44	93.22	68.5	37.41	12.66	22.82	-2.08864	0.002998581535123745	0.04717	
ENST00000678361.1	SFTPD-203	SFTPD	Known	Noae		131.7	196.09	48.52	0	1.41	21.48	-3.84609	0.00303581835123745	0.04717	
ENST00000438072.7	SFRP4-201	SFRP4	Known	Noae		21.24	12.86	28.54	104.74	222.26	48.33	-2.96187	0.0031161909061364	0.04834	
ENST00000328895.9	GKN2-201	GKN2	Known	Noae		72.22	28.93	28.93	0	0	12.08	-3.78014	0.0031380172800276	0.04903	
TALNT0000490585	FGC	FGC	ISM	Suffix		93.46	28.93	28.54	0	0	8.05	-3.8643	0.00326072861799318	0.04908	

TABLE 11: Significantly Alternative Polyadenylated Genes in Non-Transformed vs. Cancer samples.

Region	GeneID	GeneName	DeltaUsage	OddsRatio	pvalue	padj	log10FDR	Significance
chr10:103882537-	ENS00000107900.12	STN1	-0.396190476190476	0	3.04925021939776e-13	4.7088328510331e-11	10.327086722249	Down
chr10:11334357+	ENS00000048740.19	CELF2	-0.473306600663657	0.141935483870968	9.73756238539081e-10	9.55312149785576e-08	7.01985469880478	Down
chr10:48435875+	ENS00000107643.17	MAPK8	-0.375727513227513	0.188888888888889	0.00153888607329007	0.0170421925168205	1.76847453300225	Down
chr10:62193195-	ENS00000182010.11	RTKN2	-0.603302227475703	0	0.000120801032553038	0.00252471130441049	2.5978827539467	Down
chr10:76058784+	ENS00000148655.15	LRMDA	-0.302910052910053	0	0.00100039095738565	0.0126243312451738	1.89879161869234	Down
chr11:6532092+	ENS00000132286.12	TBM10B	-0.451159762150474	0.109903381642512	7.64186942717331e-07	4.206136443515764e-05	4.37611664413779	Down
chr11:6532092+	ENS00000149798.5	CDC42EP2	-0.493055555555555	0	2.70840880072203e-06	0.00122746853202288	3.91098963284572	Down
chr11:751095+	ENS00000185801.1	OLFML1	-0.522222222222222	0	0.00562568034595777	0.0424702079332334	1.37191561207078	Down
chr11:94419865-	ENS00000029222.13	MRE11	-0.46031746031746	0	4.49647091285339e-05	0.00115728614019036	2.93655924801314	Down
chr12:104103450+	ENS00000111727.12	HCFPC2	-0.681481481481482	0	0.002253141214239613	0.0045009479435909	2.3469600983243	Down
chr12:104350283+	ENS00000198431.18	TXNRD1	-0.31552538243658	0.224477205657881	5.08909069770954e-47	4.7236642586887e-44	43.3257209780739	Down
chr12:106510104+	ENS0000013503.10	POLRB	-0.513888888888889	0	0.0068406309892378	0.048936582555335	1.30986013119271	Down
chr12:1103747318+	ENS0000017427.18	ATP2A2	-0.342322175203610	0.197570352145164	1.26233432969979e-17	3.39567934689243e-15	14.469073268056	Down
chr12:12829890+	ENS00000178878.13	ALPOLD1	-0.57246417027356	0.0061614935528582	3.09704878454395e-40	1.84473498673657e-37	36.7340660154478	Down
chr12:56591797-	ENS00000076067.13	RBMS2	-0.424832451499118	0.170940170940171	0.00018900293653783	0.00359851256657282	2.44887697555956	Down
chr12:69579182-	ENS00000166225.9	FRS2	-0.327777777777778	0	0.00123259901946366	0.014735057386333	1.83942662624761	Down
chr13:77895486-	ENS00000136160.17	EDNRB	-0.325293917399181	0	0.00378591032294561	0.0318253086522615	1.4727273752383	Down
chr14:21089542-	ENS00000165891.10	ARHGEF40	-0.324786324786325	0	0.00207171945647192	0.021457387362538	1.677477139719159	Down
chr14:63293218-	ENS00000126785.13	RHOJ	-0.498274672187716	0	0.000597829344454064	0.00867008504939554	2.06197664228134	Down
chr14:73885463-	ENS00000140043.12	PTGR2	-0.483333333333333	0	0.00271907057224076	0.0257370368902723	1.58944145492572	Down
chr14:91969516-	ENS00000100815.12	TRIP11	-0.448244348244348	0.1125	0.000180587243180794	0.00346185920893022	2.4668881494703	Down
chr15:42141201-	ENS00000168907.13	PLA2G4F	-0.427868427868428	0	0.00227080205463458	0.022603507617216	1.64582413011418	Down
chr15:4526179-	ENS0000013770.14	CTDSP12	-0.386640211640212	0.192307692307692	0.0048586595136135	0.038165001102275	1.41833472075100	Down
chr15:62163535-	ENS00000105502.4	C2CD4B	-0.453846153846154	0	0.00682016384888857	0.048943600136676	1.91029734586417	Down
chr15:65105453-	ENS00000166855.9	CLPX	-0.315600675194598	0.241935483870968	0.00220809897148944	0.02216444045057	1.65434256285007	Down
chr15:81309088-	ENS00000172349.18	IL16	-0.379292929292929	0.200716845878136	0.00100068149095334	0.0126243312451738	1.89879161869234	Down
chr16:84662309-	ENS00000135686.13	KLHL36	-0.346089885109849	0.20176018451415	0.000357251877255063	0.00588033023651995	2.2305982834696	Down
chr16:37518463-	ENS00000207506.5	SYNRG	-0.43097423307557	0.135802469135802	0.0020910543072911	0.00385781014790942	1.3651951483939	Down
chr17:40092625-	ENS00000126351.13	THRA	-0.33296066252588	0.240740740740741	0.00105162474786889	0.013064991696295	1.8838086196068	Down
chr17:4271432-	ENS00000132388.13	UBE2G1	-0.341941913498585	0.298759305210918	0.00132947873399765	0.0152706931237425	1.81614125023991	Down
chr17:45150649-	ENS00000186834.4	HEXIM1	-0.332152667021088	0.228310502283105	6.36463339985261e-05	0.00154286854622599	1.87167017520021	Down
chr17:5383719+	ENS00000209275.17	RABEP1	-0.300798480311665	0.282250864516129	0.00550622161039333	0.041825046969047	1.3676356269239	Down
chr17:54961942-	ENS00000166260.13	COX11	-0.401008820919833	0.110478741211918	1.70500180198444e-08	1.36711634872579e-06	5.86419452313564	Down
chr17:57927576-	ENS00000136451.9	VEZF1	-0.373940603184563	0.196663740122915	5.6958908495453e-07	3.23121109041334e-05	4.4906346692163	Down
chr17:60043200-	ENS00000068097.17	HEATR6	-0.464285714285714	0	0.00271692254614455	0.0257370368902723	1.58944145492572	Down
chr17:65635540-	ENS00000154240.17	CEP112	-0.485185185185185	0	0.00271197344661	0.0257283038420115	1.5895884403978	Down
chr17:76679570-	ENS00000182534.14	MXRA7	-0.317085060767243	0.175384615384615	6.3811345021643e-17	1.478276167043e-14	18.330244219534	Down
chr17:82244365-	ENS00000141551.15	CSNK1D	-0.397591958513335	0.1475038925652	4.72793414282241e-23	1.71418447030118e-20	19.7659402437172	Down
chr18:42081396-	ENS00000078142.14	PIK3C3	-0.306397306397306	0	0.0002272923842327	0.0041192709132968	2.38517964493929	Down
chr18:45850646-	ENS00000152223.16	EPG6	-0.333333333333333	0	0.000820969042003765	0.010971251348181	1.95974383526763	Down
chr19:46758478-	ENS00000181027.11	FKRP	-0.398412698412698	0	0.00115501159756839	0.0139078921007394	1.85673868728956	Down
chr19:51642578-	ENS00000254415.4	SIGLEC14	-0.483333333333333	0	0.00191184649610678	0.019535518536110	1.6999797860362	Down
chr19:57623353-	ENS00000213762.12	ZNF314	-0.533333333333333	0	0.00457665903890159	0.036483685970746	1.43786915200102	Down
chr19:10381596+	ENS00000054523.20	KIF1B	-0.364171122994563	0.240196078431373	0.00220480933781949	0.02216444045057	1.621644256285007	Down
chr19:109736975-	ENS00000134202.12	GSTM3	-0.362851326503809	0	2.61447631579145e-08	2.00019431168669e-06	5.6989272113895	Down
chr19:153860109-	ENS00000143614.10	GATAD2B	-0.380532212885154	0.233766233766234	0.00471645745816524	0.0372448283557196	1.4289402284862	Down
chr19:243501764-	ENS00000117020.19	AKT3	-0.317566854568543	0.283766233766234	0.00132859719264444	0.0152706931237425	1.81614125023991	Down
chr19:26798294-	ENS00000060642.12	PIGV	-0.366666666666667	0	0.00532350532350534	0.049812612194601	1.38741460902938	Down
chr19:28500362-	ENS00000210438.13	PHACTR4	-0.462301587301587	0	5.39262071015353e-05	0.00135448988258947	2.8692423459684	Down
chr19:4464503-	ENS00000126106.14	TMEM53	-0.509357309357509	0.0777777777777778	6.2013868634567e-06	0.0022983718019403	3.682857971932609	Down
chr19:54217020+	ENS00000134748.13	PRPF38A	-0.304248837582171	0.157894736842105	0.000148893984746739	0.00297039937512693	2.5271851514298	Down
chr19:89506572-	ENS00000191747.15	LRRCSB	-0.5	0	7.56286632633766e-05	0.00174217520152845	2.75890817238751	Down
chr20:20389559-	ENS00000188559.16	RALGAPA2	-0.362433862433862	0	1.69874111466074e-05	0.000534558571892677	3.4270470194981	Down
chr20:33077559-	ENS0000010147.12	PXMP4	-0.309868421052632	0.279411764705882	0.00224504572269151	0.022408817742808	1.6939471158562	Down
chr20:6040777-	ENS00000125872.9	LRRN4	-0.477777777777778	0	6.60763843002512e-05	0.00156537207101305	2.805284419273	Down
chr20:62294884+	ENS00000130703.17	OSBPL2	-0.3921455093869732	0.183270676691729	0.000560642909258046	0.00827481476290212	1.28822171940721	Down
chr20:62945640+	ENS00000101193.8	GIDS	-0.307617333933124	0.309855160246861	0.000137982630998563	0.00280643209730979	2.55184546123708	Down
chr20:825714739-	ENS00000154721.15	JAM2	-0.554858934169279	0	0.0014543703856756	0.0138623030898619	1.8581364149082	Down
chr22:17729221-	ENS00000099968.18	BCL2L3	-0.310612535612536	0.3	0.00112868367359156	0.0137002811558661	1.86327052020211	Down
chr22:28757500-	ENS0000010209.11	HSCB	-0.33149831649832	0	9.77437497314947e-05	0.00215062039316869	2.66743624033366	Down
chr22:30027851-	ENS00000100330.16	MTMR3	-0.388888888888889	0	0.00010179738779051	0.0022922985258771	2.65184514969757	Down
chr22:102715842-	ENS00000135953.11	MIRD9	-0.472222222222222	0	0.00166034155597723	0.0179812834224599	1.7511793152603	Down
chr22:131483974-	ENS00000173272.16	MZT2A	-0.306878306878307	0	3.57037898712289e-06	0.00155881624992763	3.80720507553965	Down
chr22:16552631-	ENS00000198782.11	CYRIA	-0.3384767400319	0.148809523809524	0.00352278635158831	0.0306964633081452	1.51291165886276	Down
chr22:187345104-	ENS00000064989.13	CALCR1	-0.325269416573764	0.209660842754368	0.000322380393780579	0.0054202003743598	2.2659908121967	Down
chr22:201623802-	ENS00000157555.20	TMEM237	-0.409722222222222	0.138888888888889	0.00116801262735775	0.0139943352004832	1.8540472728243	Down
chr22:201700553-	ENS0000003093.16	ALS2	-0.477777777777778	0	0.00023692025960859	0.0042487698989553	2.37173677919954	Down
chr22:236125247-	ENS00000119772.9	AGAP1	-0.462962962962963	0	0.0011333914559721	0.0137374292897549	1.86209452999326	Down
chr22:25232976-	ENS00000119772.9	DNMT3A	-0.45	0	3.07583040430393e-05	0.00084096226606424	3.0752247992867	Down
chr22:5366980-	ENS00000115239.24	ASB3	-0.416666666666667	0	0.000109813317360488	0.00234803398325412	2.62929622180563	Down
chr22:61049256+	ENS00000162928.9	PEX13	-0.316382361025218	0.20889748549323	0.00188367135682237	0.0179335872418866	1.70479395997102	Down
chr22:7044093+	ENS00000148181.15	RNF144A	-0.564814814814815	0	0.000628516386967805	0.0090317965590024	2.0442258534527	Down
chr22:8506288-	ENS00000168887.11	C2orf68	-0.40592380952381	0.157248157248157	0.00243726822259918	0.023990394710241	1.62344057071038	Down
chr3:101822337-	ENS0000014815.17	NXPE3	-0.619047619047619	0	0.000964891885607994	0.0123408488252838	1.90685496773724	Down
chr3:138494346-	ENS00000114107.9	CEP70	-0.355555555555556	0	0.00203003254022748	0.0207710936846098	1.68254063544921	Down
chr3:141952496-	ENS00000114126.18	TDP2	-0.387761069340017	0.126315789473684	0.000300525147176618	0.00515654156853048	2.2876417913149	Down
chr3:149321032-	ENS00000163762.7	TM4						

Region	GeneID	GeneName	DeltaUsage	OddsRatio	pvalue	padj	log10FDR	Significance
chr8:23841929-	ENSG00000159167.12	STC1	-0.458333333333333	0	0.00458818374149134	0.0365433278130815	1.439719190621582	Down
chr8:56217465+	ENSG0000017091.18	CHCHD7	-0.35808166296341	0.1596283783787878	5.91878113733145e-06	0.000224348708655486	6.64907642590087	Down
chr8:56301300-	ENSG00000170786.13	SDR16C5	-0.329471189463742	0.308510638297872	2.0584866506743e-05	0.000613061434990047	3.212496002571	Down
chr8:90623545-	ENSG00000180694.14	TMEM64	-0.555258467023173	0	7.6667794770309e-07	4.20613643515764e-05	4.37611664413779	Down
chr9:120869579-	ENSG00000119403.15	PHF19	-0.350584809971096	0.2047764227644228	0.000277284110546032	0.00483738944626456	3.5358894722903	Down
chr9:131093049-	ENSG0000050555.19	LAMC3	-0.835714285714286	0	4.18163586824146e-06	0.000193887952668017	7.3124917520365	Down
chr9:136724035-	ENSG00000165716.11	DIPK1B	-0.341736694677871	0	1.55646843756139e-07	1.02199923628539e-05	4.99054942873865	Down
chr9:27329130-	ENSG00000120162.10	MOB3B	-0.397222222222222	0.20066889632107	0.00200440655250167	0.0266550474476769	1.685497428608024	Down
chr9:35847234+	ENSG00000137103.20	TMEM58B	-0.452380952380952	0	0.00152949170425933	0.016982640769887	1.76997883739054	Down
chr9:5776554+	ENSG0000017036.12	RIC1	-0.333333333333333	0	0.0028966076957396	0.02695849506113	1.56930435564152	Down
chrX:129645809-	ENSG00000171388.12	APLN	-0.572567783094099	0	0.000443942672308382	0.00683385101707461	2.16533449336981	Down
chrX:16321849-	ENSG00000129680.16	MAP7D3	-0.5	0	0.00329835082458771	0.0293856276989711	1.53186502799577	Down
chrX:45148373-	ENSG00000147113.17	DIPK2B	-0.307777777777778	0	0.00485573213320732	0.0381650011012275	1.41833472075109	Down
chrX:48903613-	ENSG00000102100.16	SLC35A2	-0.408208020050125	0	9.83298500637906e-12	1.20584208776757e-09	8.91870956193175	Down
chrX:73820659-	ENSG00000229807.13	XIST	-0.581087258994236	0.0368421052631579	2.59350253849879e-17	6.3609463731004e-15	14.1964782657919	Down
chr10:128096662-	ENSG00000148773.14	MKI67	0.687275469628411	Inf	0.000567045433857053	0.0083103547854727	2.08038043493021	Up
chr10:70112273-	ENSG0000042286.15	AIFM2	0.344821824676634	4.40322580645161	0.00189578213520708	0.0198355423155481	1.2575592115755	Up
chr10:76884877-	ENSG00000156113.24	KCNMA1	0.575084175084175	Inf	0.00215404131126744	0.021772884783748	1.66208594655748	Up
chr10:90900761+	ENSG00000148688.14	PPP30	0.324186133783657	Inf	0.000461495199372243	0.00713990439251416	2.146370630414	Up
chr10:99711222+	ENSG00000198018.7	RNTD7	0.65826023918129	Inf	0.00163083541184905	0.0178013457080637	1.6450167584462	Up
chr11:124146901+	ENSG00000110002.16	VWA5A	0.390579710144928	6.03529411764706	5.55517892907725e-06	0.000216470266773716	3.66460174763807	Up
chr11:128965060-	ENSG00000134909.19	ARHGAP32	0.554292929292929	Inf	0.00275363527270991	0.025975751741095	1.64711664413779	Up
chr11:129659509+	ENSG00000151715.8	TMEM45B	0.578703703703704	Inf	0.0025309267801823	0.024501045473431	1.61081538364956	Up
chr11:129850950+	ENSG00000132286.12	TMM10B	0.451159762150474	9.0989010989011	7.64186942717331e-07	4.20613643515764e-05	4.37611664413779	Up
chr12:104265892+	ENSG00000198431.18	TXNRD1	0.315525582543658	4.45479529678425	5.09890069770954e-06	4.72366425868887e-44	4.3257290780739	Up
chr12:108522688-	ENSG0000075856.12	SART3	0.415032679738562	Inf	1.08525152606995e-06	5.6561952974358e-05	4.2474756030748	Up
chr12:10908016+	ENSG00000135093.13	USP30	0.4	Inf	0.00546678800850816	0.0416324613275567	1.80569791222349	Up
chr12:110351090+	ENSG00000174437.18	ATP2A2	0.36161971373239	6.04545454545455	1.28494466401415e-20	4.28606142128559e-18	17.3679416096044	Up
chr12:12791463+	ENSG00000178873.18	ALP01D	0.526354587622193	12.2131332082552	6.36368460103367e-34	3.53778439253465e-31	30.4512686335257	Up
chr12:27235092+	ENSG00000211455.8	STK38L	0.396828792849332	Inf	0.000238095953560561	0.00426069132347965	2.37051992828079	Up
chr12:42081847-	ENSG00000152333.11	GXYLT1	0.645833333333333	Inf	0.000150228406842159	0.0029898678247439	2.43244798571125	Up
chr12:42159034-	ENSG0000015153.15	YAF2	0.646428571428571	Inf	0.00568990042674255	0.042746017710456	1.36910433864184	Up
chr12:63809107+	ENSG00000118600.13	RXYLT1	0.484848484848485	Inf	0.000473702720054314	0.0073016749266714	2.136557412767	Up
chr12:6439167+	ENSG00000215039.9	CD27-AS1	0.785873440285205	Inf	0.00050394255759658	0.0076546028921632	2.6170773484364	Up
chr12:67626663+	ENSG00000127334.11	DYRK2	0.513428600385122	Inf	0.00115007936309247	0.013880180376053	1.85760489129341	Up
chr13:10720771+	ENSG0000015487.13	ING1	0.312665112665112	3.69886363636364	0.00675513577608903	0.048687188622996	1.3585630248779	Up
chr13:79311833-	ENSG00000139746.16	RBM26	0.643035430335433	Inf	6.47826254548304e-06	0.00023729696717336	3.62470781200806	Up
chr14:103701543+	ENSG00000126214.22	KLC1	0.401481481481481	Inf	7.57631706235504e-06	1.08929151694791e-10	9.96285587850806	Up
chr14:45203190-	ENSG00000129534.14	MIS18BP1	0.341099785035712	Inf	0.00689857344654995	0.049237571703337	1.30770332606572	Up
chr14:461795467+	ENSG0000023608.5	SNAPC1	0.535612535612536	Inf	0.00600577590232419	0.0445174802217612	1.35143492561672	Up
chr14:91965995-	ENSG00000100815.12	TRIP11	0.567291967291967	12.6984126984127	3.6962097255011e-06	0.00015887986031494	3.7899311506765	Up
chr15:42569990+	ENSG00000137814.12	HAUS2	0.6	Inf	0.002711977344661	0.0257283038420115	1.5898884403978	Up
chr15:52307285-	ENSG00000197355.9	MYO5A	0.49702380958231	Inf	0.000452744668747545	0.00705689307044071	1.5138646316235	Up
chr15:65148221-	ENSG00000166855.9	CLPX	0.31560067154958	4.13333333333333	0.0022080987148944	0.0221644744045057	1.6543426285007	Up
chr15:68301711-	ENSG00000137809.17	ITGA11	0.538766788766789	Inf	0.00515797120349809	0.041869301069378	1.3781402888797	Up
chr15:84873475+	ENSG00000136383.7	ALPK3	0.366666666666667	Inf	0.00634844719173024	0.0462759625278308	1.33646453927029	Up
chr16:16149565-	ENSG0000091262.17	ABCC6	0.533333333333333	Inf	0.0014197874181189	0.0160864229343661	1.7935405172792	Up
chr16:84464179+	ENSG00000064270.13	ATP2C2	0.472222222222222	Inf	0.01163067859936	0.01163067859936	1.93515428457395	Up
chr17:39533549-	ENSG00000167258.15	CDK12	0.413215488215488	7.41818181818182	0.0025696510030446	0.0427282130288557	1.60690211556524	Up
chr17:45152096+	ENSG00000186834.4	HEXIM1	0.332152667021008	4.38	6.3646333998526e-05	0.0015428685442259	2.81167107520021	Up
chr17:54961206-	ENSG00000166260.13	COX11	0.309417974490277	6.4207792207922	4.57674649504787e-06	0.00018842860203775	3.74263317387393	Up
chr17:57971554-	ENSG00000136451.9	VEZF1	0.373940603184564	5.08482142857143	5.69598309498454e-07	3.2321109041334e-05	4.9685460926163	Up
chr17:62615478+	ENSG00000146872.19	TLK2	0.423313492063492	10.125	0.0003768551537095	0.006090306633014	2.21356126891699	Up
chr17:64039152-	ENSG00000178607.17	ERN1.00	0.350869157451635	Inf	0.00065855976152731	0.00929226709200176	2.03187831556167	Up
chr17:82242666-	ENSG00000141551.13	CNKN1D	0.397591958531335	6.7794821045687	4.72793414282421e-23	1.71418447030418e-20	19.76592444371762	Up
chr19:18670497-	ENSG00000167487.12	KLHL26	0.322222222222222	Inf	0.00447548484600417	0.0359894582071834	1.44382469125767	Up
chr1:100026887+	ENSG00000117620.15	SLC35A3	0.308030661222151	4.0199335481728	0.00294361792659778	0.0272137803657416	1.56521112459502	Up
chr1:15806072-	ENSG00000143365.20	RORC	0.41	Inf	0.00364005221903507	0.0311668780776911	1.50630669793282	Up
chr1:155614933+	ENSG00000125459.18	MSTO1	0.501798941798942	Inf	1.52529750754275e-05	0.00048997205756554	3.3098039655918	Up
chr1:185119056-	ENSG0000021486.12	TRMT1L	0.535817805383023	Inf	0.000894828836172977	0.0116049419359976	1.935350281455	Up
chr1:200644046-	ENSG00000118197.14	DDX59	0.312091503267942	Inf	0.0016665655714825	0.0180251817510496	1.74412034751949	Up
chr1:212737758-	ENSG00000117697.15	NSL1	0.301709401709402	Inf	0.0020846952671216	0.02117451367268	1.67418606286019	Up
chr1:24349724-	ENSG00000117020.19	AKT3	0.317568542658543	3.52402745995423	0.0013285971926444	0.0152706931237425	1.81641215023991	Up
chr1:3812088-	ENSG00000116198.14	CEP104	0.63403264032634	Inf	0.000561318768919449	0.00827481476290212	2.08224171940721	Up
chr1:44850840-	ENSG00000070785.17	EIF2B3	0.401171963038089	Inf	3.85548320233942e-05	0.00101103378692974	2.9952343082722	Up
chr2:151354263-	ENSG00000085832.17	EPS15	0.309513783707332	4.0790816265306	4.60233468419951e-05	0.00117726591814539	1.92912542862704	Up
chr20:33704113-	ENSG0000010417.12	PXMP4	0.309864821052362	3.57894736842105	0.0022450457226915	0.022408817742808	1.6493473135862	Up
chr20:4140246-	ENSG00000124177.16	CHD6	0.564089635854342	Inf	0.00227146429563769	0.0226035092617216	1.64582243011418	Up
chr20:4343602+	ENSG00000124193.16	SRSF6	0.308662811454147	4.59421312632322	1.19904498584546e-06	6.0831415660989e-05	4.21987207651557	Up
chr20:435481-	ENSG0000012875.15	TBC1D20	0.444420396944438	7	5.65978112938571e-07	3.2312109041334e-05	4.906866962163	Up
chr20:62296181+	ENSG00000130703.17	OSBPL2	0.392145593869732	5.45641025641026	0.000560642909258047	0.00827481476290212	2.08224171940721	Up
chr20:62948473+	ENSG00000101193.8	GID8	0.30761733933123	3.22733516483516	0.000137982630998562	0.0020643209730979	2.55184546123708	Up
chr21:36294								

Region	GeneID	GeneName	DeltaUsage	OddsRatio	pvalue	padj	log10FDR	Significance
chr7:139043521:-	ENSG00000105939.14	ZC3HAV1	0.302915204018769	3.7958041958042	0.000171920153265355	0.00332830757995834	2.47777654593114	Up
chr7:23175417:+	ENSG00000122550.18	KLHL7	0.350271155095717	6.30172413793104	0.00118887338016077	0.0141510535276042	1.84921122627446	Up
chr7:44826795:-	ENSG00000105968.19	H2AZ2	0.380381568445851	6.615	1.291850060611e-25	6.33690450319712e-23	22.1981228376194	Up
chr7:4771436:+	ENSG00000164916.11	FOXK1	0.396491228070175	Inf	0.00192600288879416	0.0200331965639888	1.69824974769401	Up
chr7:55956500:+	ENSG00000239789.6	MRPS17	0.313506659368728	5.41369047619048	0.000113717744958315	0.00241713024276084	2.61669994777671	Up
chr7:99576451:+	ENSG00000197343.11	ZNF655	0.320672797420855	5.34146341463415	1.39854233666402e-05	0.000453791616554134	3.3431435316285	Up
chr8:140520158:-	ENSG00000123908.12	AGO2	0.408068783068783	Inf	0.0026662462666575	0.0254975087349676	1.59350225076256	Up
chr8:38262766:-	ENSG00000085788.14	DDHD2	0.305555555555556	Inf	0.000314826632002528	0.00534692318588407	2.27189605494703	Up
chr8:56218615:+	ENSG00000170791.18	CHCHD7	0.35808166296341	6.26455026455026	5.91878113733142e-06	0.000224348708655486	3.64907642590987	Up
chr8:93729249:-	ENSG00000188343.14	CIBAR1	0.458333333333333	Inf	0.00181572994590491	0.0191905855752865	1.71691177314178	Up
chr9:120855698:-	ENSG00000119403.15	PHF19	0.350584898971996	4.8833746898263	0.000277284105446032	0.00483738944626456	2.31538894722903	Up
chr9:134159964:+	ENSG00000196363.10	WDR5	0.301888162672476	Inf	0.000261772050281448	0.0046248244222394	2.33490475025468	Up
chr9:14615533:-	ENSG00000175893.12	ZDHHHC21	0.769919590643275	Inf	0.00052882072972608	0.00789759978116473	2.10250487835279	Up
chr9:91213826:-	ENSG00000148090.12	AUH	0.314814814814815	Inf	0.0059517222057272	0.0441953798729794	1.35462312885329	Up
chr9:98732009:-	ENSG00000165138.18	ANKS6	0.316666666666667	Inf	0.00465815406392657	0.0368542189175367	1.43351278873498	Up
chrX:40626922:-	ENSG00000185753.13	CXorf38	0.301851851851852	Inf	0.003605325224613	0.0310908035471611	1.50736805339623	Up
chrX:73821670:-	ENSG00000229807.13	XIST	0.581087258994236	27.1428571428571	2.59350253849879e-17	6.3609463731004e-15	14.1964782657919	Up

A.3 NanoInsights: A Web Platform for Advanced NanoString nCounter Data Analysis (Supplementary Materials)

A.3.1 NanoString nCounter Technology: A Multiplexed Approach for RNA Target Analysis

The NanoString nCounter technology is a powerful and versatile platform used for the analysis of RNA targets. This technology relies on the parallel hybridisation of complementary probes to RNA molecules of interest. Specifically, for each RNA target, two distinct probes are meticulously designed: the reporter probe and the capture probe.

The reporter probe plays a pivotal role in this technology, as it contains a unique and distinctive fluorescent molecular barcode that is specific to the RNA target under investigation. This molecular barcode serves as a molecular signature, allowing for the precise identification of the target of interest. On the other hand, the capture probes are biotinylated, enabling the immobilisation of the hybridisation complexes formed between the target RNA and the specific probes.

The process of hybridisation takes place in a controlled environment, typically overnight, within a thermocycler equipped with a programmable heated lid. Following hybridisation, excess unbound probes are removed, and the resulting hybridisation complexes are immobilised and aligned on a streptavidin-coated surface.

A unique feature of this technology is the structure of the fluorescent molecular barcodes, which consist of seven individual light signals. This innovative design allows for the multiplexing of up to 800 different RNA targets within a single experiment. The resulting data are presented in the form of a matrix of digital numbers, representing the count of each barcode in each sample investigated [354].

A.3.2 Supplementary Figures

Guidance and Insights
Video Tutorial
Run our Example

NanoInsights Roadmap

Uploading your Data

NanoInsights streamlines the data submission process, accommodating nCounter raw data in the RCC format. You can effortlessly upload RCC files individually or opt for a more organised approach by consolidating them into a single zipped file (.zip) or a tape archive (.tar.gz) for submission.

In addition to the raw data, we require a clinical sample information table to accompany your data for comprehensive analysis. Ensure the clinical information table is submitted in a comma-delimited format, compatible with both .csv and .txt formats. The table should include, at a minimum, the essential headers: "Filename" and "Condition", with the exact format (first letter of each header capitalised). Feel free to enrich the table with additional clinical information tailored to your needs. For your convenience, an example clinical data file, aligning with the required input format, is available for reference [here](#).

FIGURE 11: **NanoInsights Web Service Interface and How to Use Guide.** Screenshot of the NanoInsights web service interface showcasing the 'How to use?' tab. This tab provides comprehensive guidance on navigating and utilising the website efficiently. It includes detailed explanations of each accessible variable that users can adjust to tailor their experience. The guide offers insights into the functionality of the website, enabling users to make informed decisions and optimise their interactions with the platform.

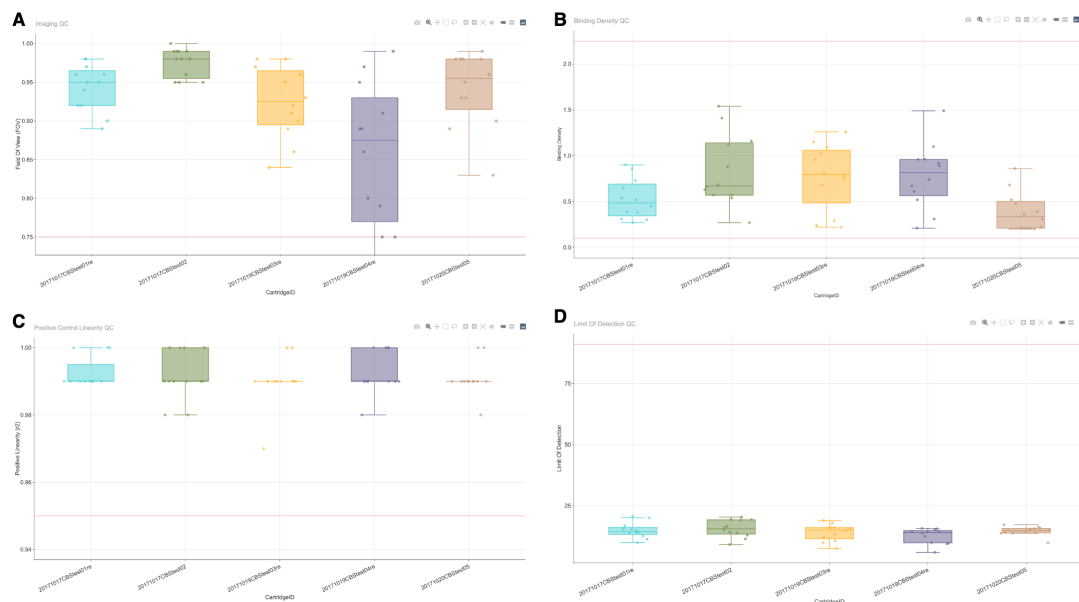


FIGURE 12: Standard NanoString QC Metrics for Gene Expression Analysis (Training Set). This multi-panel figure showcases the standard QC metrics for a NanoString gene expression analysis. **(A)** The Imaging QC boxplot displays the field of view uniformity across different cartridges. The red line in 0.75 represents the minimum limit. Sample below this limit will be potential outliers. **(B)** The Binding Density QC boxplot shows the level of image saturation. Samples outside the red lines (indicating the low and upper limit) would be potential outliers. **(C)** The Positive Control Linearity QC boxplot shows the linearity of the assay across a range of positive control concentrations. These positive controls are typically used to measure the efficiency of the hybridization reaction. Sample with a linearity below 0.95 (red line), would be indicated as potential outliers. **(D)** The Limit of Detection QC boxplot indicates the sensitivity of the assay in detecting low-abundance targets. Each panel delineates the QC metrics across various cartridge IDs, providing a comprehensive overview of the assay's performance and reliability.

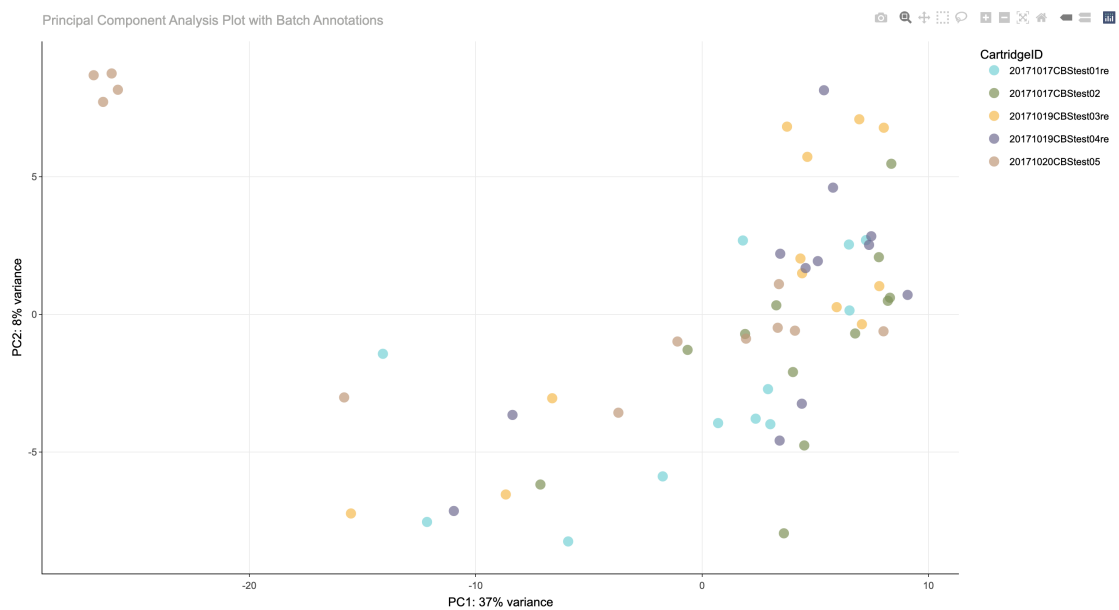


FIGURE 13: **PCA Plot Highlighting Batch Variation in the Training Set.** This Principal Component Analysis plot visualises the variance in gene expression data across different batches, represented by CartridgeIDs. Each point corresponds to a sample, with its position reflecting the sample's score on the first two principal components that together explain 45% of the variance (PC1: 37%, PC2: 8%). The colour coding indicates the batch each sample belongs to, providing a clear illustration of batch effects within the dataset.

A.3.3 Supplementary Tables

TABLE 12: Comprehensive Inventory of Packages Employed in NanoInsights.

Package	Version	Usage
Python	v3.11.4	-
fast_ml	v3.68	Removing quasi-constant features
matplotlib	v3.8.2	Machine Learning visualisations
numpy	v1.26.0	Data manipulation
pandas	v2.1.1	Data manipulation
plotly	v5.18.0	Interactive visualisations
scipy.stats	v1.12.0	Linear Regression analysis
seaborn	v0.13.0	Machine Learning visualisations
sklearn	v1.3.2	Machine Learning
R	v4.3.1	-
ctrlGene	v1.0.1	Assessing the Stability of Candidate Housekeeping Genes
DESeq2	v1.42.0	Exploratory Analysis
edgeR	v4.0.7	Exploratory Analysis and filtering
ggdendro	v0.1.23	Visualisations
ggplot2	v3.4.4	Visualisations
heatmaply	v1.5.0	Visualisations
htmltools	v0.5.7	Interactive visualisations
limma	v3.58.1	Differential Expression
NanoStringClustR	v0.1.1	Normalisation
NanoTube	v1.8.0	Normalisation
optparse	v1.7.3	Input arguments
pheatmap	v1.0.12	Visualisations
RColorBrewer	v1.1-3	Visualisations
reshape2	v1.4.4	Data manipulation
reticulate	v1.34.0	Interoperability between R and Python
RUVSeq	v1.36.0	Normalisation
tidyverse	v2.0.0	Data manipulation

TABLE 13: Gene Features Selected for Predictive Modelling in Chemoradiotherapy Response.

Selected Genes		
CACNA1D	GAS1	NKD1
CACNA2D1	ID1	PIK3CA
CAMK2B	IL2RB	SFN
CD40	IRS1	SMAD9
ETV7	MAPK8IP1	STAT1
FAS	MAPK8IP2	TNFSF10
FGFR3	MDM2	TSPAN7
FGFR4	MLLT4	WNT11
FUT8	MYD88	
FZD7	NGFR	

TABLE 14: Comparative Performance Metrics of the Utilised Classifier Models.

Metric	Abbrev.	Random Forest	Extra Trees	Logistic Regression
True Positive	TP	25.0	0.0	20.0
False Positive	FP	22.0	1.0	15.0
True Negative	TN	40.0	61.0	47.0
False Negative	FN	9.0	34.0	14.0
Population	Population	96.0	96.0	96.0
Accuracy	Accuracy	0.68	0.64	0.7
Balanced Accuracy	BA	0.69	0.49	0.67
False Positive Rate	FPR	0.35	0.02	0.24
False Negative Rate	FNR	0.26	1.0	0.41
True Negative Rate	TNR	0.65	0.98	0.76
Negative Predictive Value	NPV	0.82	0.64	0.77
False Discovery Rate	FDR	0.47	1.0	0.43
True Positive Rate	TPR	0.74	0.0	0.59
Positive Predictive Value	PPV	0.53	0.0	0.57
F1 score	F1	0.62	0.0	0.58
F2 score	F2	0.68	0.0	0.58
Cohen Kappa Metric	Cohen Kappa	0.35	-0.02	0.34
Matthews Correlation Coefficient	MCC	0.36	-0.08	0.34
ROC AUC score	ROC AUC	0.7	0.69	0.7
Average precision	Avg. Precision	0.54	0.52	0.55
Log loss	Log Loss	0.65	0.62	0.83
Brier score	BS	0.23	0.22	0.24
Negative Likelihood Ratios	LR-	0.41	1.02	0.54
Positive Likelihood Ratios	LR+	2.07	0.0	2.43
Diagnostic Odds Ratio	DOR	5.05	0.0	4.48