A BIOINFORMATICS APPROACH FOR IDENTICAL SALIVARY BIOMARKER DISCOVERY

Project Report submitted to
Shri Ramdeobaba College of Engineering & Management, Nagpur in
partial fulfillment of requirement for the award of degree of

Bachelor of Engineering

In

BIOMEDICAL ENGINEERING

By

Krunal Parate

Under the guidance of

Dr. Arpita Parakh

RCOEM

Shri Ramdeobaba College of Engineering and Management, Nagpur

BIOMEDICAL ENGINEERING

Shri Ramdeobaba College of Engineering & Management, Nagpur 440013

(An Autonomous Institute affiliated to Rashtrasant Tukdoji Maharaj, Nagpur University Nagpur)

April 2025

SHRI RAMDEOBABA COLLEGE OF ENGINEERING & MANAGEMENT, NAGPUR

(An Autonomous Institute affiliated to Rashtrasant Tukdoji Maharaj Nagpur University Nagpur) Department of Biomedical Engineering

CERTIFICATE

This is to certify that the thesis titled **"A Bioinformatics Approach for Identical Salivary Biomarker Discovery"** is a bonafide work of Krunal Parate submitted to the Rashtrasant Tukdoji Maharaj Nagpur University, Nagpur in partial fulfillment of the award of Bachelor of Engineering in Biomedical Engineering has been carried out at the Department of Biomedical Engineering, Shri Ramdeobaba College of Engineering and Mangement, Nagpur during the academic year 2024-2025.

Date:10/05/2025
Place: Nagpur

Dr. Arpita Parakh
Research Guide

Dr. Nitin Narkhede
H.O.D

Dr. M.B. Chandak
Principal

DECLARATION

I, hereby declare that the thesis titled "**A Bioinformatics Approach for Identical Salivary Biomarker Discovery**" submitted herein, has been carried out in the Department of Biomedical Engineering of Shri Ramdeobaba College of Engineering & Management, Nagpur. The work is original and has not been submitted earlier as a whole or part for the award of any degree / diploma at this or any other institution / University.

Date:10/05/2025

Place:- Nagpur

Krunal Parate

Roll No. 30

Approval Sheet

This thesis/dissertation/report entitled **A Bioinformatics Approach for Identical Salivary Biomarker Discovery** by Krunal Parate is approved for the degree of Bachelor of Engineering in Biomedical Engineering.

Name & signature of Supervisor(s) Examiner(s)

Parate 10/5/2025

Name & signature of External.

A. B. Lambat 10/5/2025

Name & signature RRC Members

Name & signature of HOD

Date: 10/05/25

Place: Nagpur

ACKNOWLEDGEMENT

I would like to express my deepest gratitude to my guide, Dr. Arpita Parakh, for his invaluable guidance, constant encouragement, and insightful feedback throughout this research. His expertise and mentorship were instrumental in shaping this work. I would also like to thank Dr. Nitin Narkhede, HoD, Biomedical Engineering Dept., RCOEM, and Dr. M.B. Chandak, Principal, for their academic support and encouragement during my research.

ABSTRACT

Saliva has emerged as a highly promising bio fluid in the realm of clinical diagnostics due to its non-invasive, easy-to-collect nature and its rich content of biological markers. Unlike blood or tissue biopsies, saliva collection does not require trained personnel or invasive procedures, making it ideal for frequent monitoring and early detection of diseases. Over recent years, technological advancements in molecular biology, proteomics, and genomics have facilitated the identification of numerous biomarkers in saliva, such as DNA, RNA, proteins, metabolites, and microbiota. These discoveries have positioned saliva as a reliable alternative for diagnosing a wide range of conditions including infectious diseases, autoimmune disorders, metabolic syndromes, and various cancers. The integration of saliva-based testing with point-of-care devices is accelerating its translation into clinical practice, offering rapid, cost-effective, and patient-friendly diagnostics.

The diagnostic potential of saliva is rooted in its molecular composition, which reflects the physiological and pathological state of the body. Salivary biomarkers have shown considerable promise in detecting oral diseases like periodontitis and oral squamous cell carcinoma, as well as systemic diseases such as diabetes, cardiovascular disorders, and neurodegenerative diseases. For example, changes in salivary glucose, cortisol, and certain cytokines have been linked to disease progression and response to treatment. Moreover, recent studies have demonstrated the utility of exosomal RNA and microRNA in saliva for cancer diagnostics, particularly in detecting early-stage tumors with high sensitivity and specificity. These findings not only highlight the breadth of salivary biomarkers but also reinforce the clinical relevance of saliva as a mirror of systemic health.

Despite its potential, several challenges must be addressed before salivary diagnostics can become routine in clinical settings. Issues such as standardization of collection methods, variability in biomarker expression, and sensitivity of detection techniques remain significant. Moreover, environmental and individual factors like circadian rhythms, hydration, and oral hygiene can influence salivary composition. Nevertheless, continued

research and the development of robust analytical platforms, such as lab-on-a-chip technologies and biosensors, are paving the way for overcoming these limitations. As multidisciplinary collaborations between clinicians, biotechnologists, and engineers grow, saliva-based diagnostics is poised to transform healthcare by enabling early disease detection, continuous monitoring, and personalized medicine in a non-invasive and accessible manner.

Table of Contents

LIST OF FIGURES

LIST OF TABLE

CHAPTER 1
INTRODUCTION

The landscape of disease diagnosis is undergoing a paradigm shift from invasive and expensive techniques to non-invasive, rapid, and cost-effective approaches. Among these, the use of biomarker measurable indicators of normal or pathological biological processes has become a cornerstone of modern diagnostics and personalized medicine. While blood and tissue biopsies remain standard biological sources, the emergence of saliva as a diagnostic fluid is gaining momentum due to its accessibility, patient compliance, and molecular richness. This thesis explores the utility of saliva in identifying biomarkers that are commonly expressed in cancer, control (healthy individuals), and chronic disease conditions, emphasizing its potential as a universal, non-invasive tool for disease detection and monitoring.

1.1 Background on Cancer, Control, and Chronic Conditions

Cancer remains one of the leading causes of mortality worldwide, characterized by uncontrolled cellular proliferation and the disruption of normal tissue function. Despite advancements in imaging and tissue-based diagnostics, early-stage cancers often go undetected due to a lack of visible symptoms and accessible screening methods. In parallel, chronic diseases such as cardiovascular disorders, diabetes, and autoimmune condition are on the rise, contributing significantly to global morbidity and healthcare costs. These conditions are often progressive, necessitating long-term monitoring to manage symptoms and prevent complications.

In contrast, the control group, representing healthy individuals, serves as a crucial benchmark in biomarker discovery studies. By comparing biomarker expression profiles across cancer, chronic, and control groups, researchers can identify disease-specific changes, differentiate between overlapping conditions, and even uncover early molecular signs of disease onset. Integrating data across these groups provides a comprehensive picture of disease dynamics and lays the foundation for developing pan-disease diagnostic strategies.

1.2 Importance of Biomarkers

Biomarkers are fundamental to the practice of precision medicine, providing critical information that guides early detection, diagnosis, prognosis, and therapeutic decision-making. These biological indicators ranging from genes and proteins to metabolites and nucleic acids allow clinicians to identify disease states even before symptoms appear, offering a proactive approach to healthcare. In the field of oncology, for example, biomarkers such as HER2 in breast cancer, BRCA1/2 mutations in hereditary cancers, and PSA for prostate cancer have significantly improved patient stratification and treatment outcomes. These markers not only assist in diagnosing malignancies but also help determine the most effective therapeutic strategies, such as targeted treatments or immunotherapies.

In chronic disease management, biomarkers are equally indispensable. Conditions like diabetes, cardiovascular disease, and autoimmune disorders depend heavily on reliable biomarkers to monitor disease progression and response to treatment. For instance, HbA1c is routinely used to assess long-term glucose control in diabetic patients, while C-reactive protein (CRP) serves as a marker of systemic inflammation in both cardiovascular and autoimmune conditions. Similarly, troponins are vital for detecting heart muscle damage during myocardial infarction. These biomarkers facilitate timely intervention, minimize complications, and improve patient outcomes by enabling more accurate clinical decisions.

Despite their value, traditional biomarkers often require invasive sampling methods such as blood draws or tissue biopsies, which limit their utility for frequent monitoring, especially in resource-limited or high-risk populations. These procedures can be painful, costly, and logistically challenging. As a result, there is growing interest in identifying biomarkers that can be obtained through non-invasive means. Saliva has emerged as a promising alternative due to its ease of collection, minimal discomfort, and reduced risk of infection. It contains a wide array of biomolecules DNA, RNA, proteins, enzymes, and metabolites—that reflect both oral and systemic health. The ability to collect saliva samples repeatedly, even in home settings, makes it an ideal medium for large-scale screening and continuous monitoring. Thus, saliva-based biomarkers represent a transformative opportunity to enhance accessibility, compliance, and cost-effectiveness in diagnostic and prognostic testing across various diseases.

1.3 Why We Use Saliva Samples

Saliva is an increasingly valuable diagnostic medium owing to its rich molecular composition and non-invasive collection method. This complex biofluid comprises enzymes, proteins, hormones, DNA, RNA, metabolites, and diverse microbiota. These components mirror both local (oral) and systemic physiological states, making saliva a highly informative biological sample. Remarkably, it contains over 30% of the proteins found in blood plasma, affirming its potential for systemic disease detection. Unlike blood or tissue biopsies, saliva can be collected easily, safely, and repeatedly without the need for trained personnel. This simplicity makes it especially suitable for large-scale public health screenings, pediatric and geriatric populations, and longitudinal patient monitoring.

Advancements in molecular technologies have further enhanced the relevance of saliva in diagnostic research. Salivary transcriptomics and proteomics have uncovered numerous biomarkers associated with various pathological conditions, including cancer, autoimmune diseases, infectious illnesses, and metabolic disorders. RNA transcripts and specific protein markers present in saliva have been shown to correlate with disease states, allowing early detection and real-time health monitoring. Additionally, saliva offers insight into the oral microbiome—a dynamic ecosystem now recognized for its contribution to systemic inflammation and tumor development. This dual representation of host and microbial factors positions saliva as a comprehensive diagnostic tool.

Cutting-edge analytical platforms such as next-generation sequencing (NGS), real-time quantitative PCR (RT-qPCR), and microfluidic lab-on-a-chip systems have made it feasible to evaluate salivary biomarkers with high sensitivity and specificity. These technologies enable the rapid detection of molecular changes using small volumes of saliva, enhancing its clinical utility. As a result, saliva-based diagnostics are rapidly progressing toward implementation in routine clinical practice, offering a patient-friendly, cost-effective alternative to traditional diagnostic approaches.

CHAPTER 2
LITERATURE REVIEW

2.1 Traditional Biomarkers for Cancer

Traditional cancer biomarkers are molecular, biochemical, or cellular indicators present in tissues, blood, or other bodily fluids that provide insights into the presence and progression of malignancies. They serve various clinical purposes, including screening, diagnosis, prognosis, therapy selection, and monitoring. Commonly utilized biomarkers in clinical oncology include:

i. **Prostate-Specific Antigen (PSA)**: PSA is a glycoprotein produced by prostate cells and used as a biomarker for prostate cancer detection and monitoring. Elevated levels may suggest malignancy but can also result from benign conditions like BPH or prostatitis, leading to false positives. While not highly specific, PSA remains valuable when combined with digital rectal exams and imaging. Advances like PSA velocity and free/total PSA ratios help improve its diagnostic precision.

ii. **Carcinoembryonic Antigen (CEA)**: CEA is a glycoprotein primarily used to monitor colorectal cancer recurrence and treatment response. Though elevated in other cancers and some benign conditions (e.g., liver disease, smoking), it lacks specificity for early diagnosis. CEA is best used in patients with known malignancies for tracking disease progression over time, especially when combined with imaging and other clinical indicators.

iii. **Cancer Antigen 125 (CA-125)**: CA-125 is a glycoprotein used mainly to monitor ovarian cancer treatment and recurrence. It can also be elevated in benign conditions like menstruation, pregnancy, and endometriosis, limiting its specificity. Therefore, it's not suitable for general screening. However, when combined with imaging or used in high-risk individuals, CA-125 improves diagnostic accuracy and remains crucial in ovarian cancer management.

iv. **Alpha-Fetoprotein (AFP)**: AFP is a fetal plasma protein that serves as a tumor marker for hepatocellular carcinoma and testicular germ cell tumors. High AFP levels suggest liver cancer, especially in hepatitis or cirrhosis patients, but may also be elevated in benign liver conditions. AFP is widely used in cancer surveillance, particularly in high-risk populations, and aids in monitoring treatment response and disease recurrence.

v. **HER2/neu and BRCA1/2**: HER2 is overexpressed in ~20–25% of breast cancers and guides targeted treatment with agents like trastuzumab. BRCA1/2 mutations impair DNA repair and significantly increase hereditary breast and ovarian cancer risk. BRCA testing informs prevention strategies and treatment decisions, including the use of PARP inhibitors. Together, HER2 and BRCA integrate molecular diagnostics into precision oncology for improved patient outcomes.

While these biomarkers have significantly improved clinical oncology, their application is often constrained by invasiveness, variability, and limited predictive power across all cancer types.

2.2 Limitations of Current Biomarkers

Despite their clinical utility, traditional cancer biomarkers have several limitations that hinder their widespread effectiveness:

i. **Invasiveness**: Traditional biomarker testing typically relies on blood draws or tissue biopsies, which can be uncomfortable, expensive, and carry a risk of complications. These invasive methods are impractical for frequent monitoring and may be unsuitable for certain populations, such as children, the elderly, or those in remote areas.

ii. **Lack of Sensitivity and Specificity**: Many conventional biomarkers are not disease-specific and can be elevated in benign or unrelated conditions, leading to false positives or inconclusive results. This reduces diagnostic accuracy and may cause unnecessary anxiety or interventions, highlighting the need for more precise and disease-specific diagnostic tools.

iii. **Tumor Heterogeneity**: Cancers are genetically diverse, and a single biomarker often fails to represent the full molecular landscape of a tumor. Intratumoral and intertumoral heterogeneity can compromise the reliability of biomarker-based diagnostics, resulting in suboptimal treatment decisions and reduced clinical outcomes.

iv. **Delayed Detection**: Several biomarkers only become detectable during advanced stages of disease, when treatment options may be limited and less effective. This delay undermines the goal of early detection, which is crucial for improving survival rates and enabling timely therapeutic interventions.

v. **Cost and Infrastructure**: Advanced biomarker testing methods such as genomic sequencing or multiplex assays require specialized equipment and trained personnel, making them costly and inaccessible in low-resource or rural settings. This limits widespread implementation and contributes to healthcare disparities, particularly in underserved populations.

These challenges underscore the need for new biomarker platforms that are non-invasive, cost-effective, sensitive, and applicable across diverse patient populations. Saliva, with its diverse molecular content and ease of collection, presents an attractive solution for overcoming these limitations, which is the focus of the subsequent chapters.

CHAPTER 3
SALIVA AS A DIAGNOSTICS

3.1 Overview of Saliva Diagnostics

Saliva, a clear biofluid secreted by salivary glands, plays a critical role in oral and systemic health. Composed of water, electrolytes, proteins, nucleic acids, and various metabolites, saliva reflects both local (oral) and systemic physiological states. In recent years, it has emerged as a powerful non-invasive diagnostic medium, capable of mirroring the molecular signatures of diseases such as cancer, diabetes, cardiovascular disorders, and autoimmune conditions. Its diagnostic relevance stems from its accessibility, ease of collection, and the presence of biomolecules like DNA, RNA, proteins, and metabolites that parallel blood-based indicators of disease

3.2 Advantages Over Other Biological Fluids

Saliva stands out as a diagnostic fluid due to its unique combination of accessibility, safety, and bio molecular richness. Compared to more traditional biological samples like blood, serum, urine, or tissue biopsies, saliva offers a number of important clinical and practical advantages:

1. Non-Invasiveness

Saliva collection is entirely non-invasive, requiring no needles, incisions, or penetration of body tissues. This characteristic greatly reduces patient anxiety and discomfort, making the process far more agreeable, especially for sensitive populations such as children, the elderly, or those with needle phobia. Moreover, the non-invasive nature ensures minimal risk of infection or procedural complications, which is particularly beneficial in outpatient or remote diagnostic settings.

2. Ease of Collection and Storage

Saliva can be collected quickly and painlessly by passive drooling, spitting, or using absorbent swabs. This process does not require trained medical personnel, which makes saliva sampling highly scalable and ideal for large epidemiological studies or routine screening. It also supports self-collection, enabling decentralized and home-based diagnostics. Furthermore, modern stabilizing agents now allow for ambient

temperature storage of saliva, which simplifies transport and reduces reliance on cold chain logistics.

3. Cost-Effectiveness

Compared to blood sampling—which involves sterile syringes, phlebotomy professionals, and cold storage—saliva-based diagnostics are significantly less expensive. Collection tools are simpler, and because samples can be obtained without clinical supervision, labour and facility costs are reduced. This makes saliva an economically attractive option for mass screening programs, especially in resource-limited settings or for diseases that require repeated monitoring over time.

4. Safety and Patient Compliance

Saliva poses a lower biohazard risk than blood or tissue, as it typically contains fewer infectious agents and less cellular debris. This minimizes risks to healthcare providers and laboratory personnel. The non-invasive nature and ease of collection also enhance patient compliance, promoting participation in longitudinal studies and follow-up testing. These features are crucial for managing chronic diseases where regular monitoring is essential.

5. Reflective of Systemic Conditions

Though produced locally in the salivary glands, saliva contains a rich array of systemically derived biomarkers. These include hormones (like cortisol and testosterone), antibodies (such as IgA and IgG), nucleic acids (DNA, RNA, microRNAs), and metabolic by-products that closely parallel those found in blood. Because of the extensive vascularization of salivary glands and their passive exchange with blood plasma, changes in systemic health such as inflammation, infection, cancer, or metabolic imbalance are often mirrored in salivary composition. This makes saliva a reliable and informative medium for monitoring both oral and systemic health conditions.

3.3 Major Components and Functional Properties

Saliva is a complex biological fluid composed of various molecular constituents that reflect both the local oral environment and systemic physiological conditions. The diversity and abundance of these components enable saliva to perform essential physiological functions while also serving as a rich source of biomarkers for diagnostic applications.

1. Proteins and Enzymes

Saliva contains a wide array of proteins and enzymes that contribute to oral health, digestion, and immune defense. Key examples include:

i. **Salivary Amylase**: This enzyme initiates the digestion of dietary starches into simpler sugars. It also reflects sympathetic nervous system activity and can serve as a stress biomarker.
ii. **Mucins**: High molecular weight glycoproteins that help in lubricating oral surfaces, facilitating speech, chewing, and swallowing. Altered mucin levels can indicate mucosal inflammation or salivary gland dysfunction.
iii. **Lactoferrin**: An iron-binding protein with potent antimicrobial activity, playing a role in inhibiting bacterial and fungal growth in the oral cavity.
iv. **Defensins and Histatins**: These are small cationic peptides with antimicrobial and antifungal properties. They also aid in wound healing and maintaining the balance of the oral microbiota.

Changes in the levels or activity of these proteins may signal the presence of infections, autoimmune diseases (e.g., Sjögren's syndrome), or neoplastic processes (e.g., oral squamous cell carcinoma).

2. DNA and RNA

Saliva contains both extracellular and intracellular nucleic acids derived from epithelial cells, leukocytes, and exosomes:

i. **Human Genomic DNA**: Saliva provides high-quality DNA suitable for genotyping, SNP analysis, and epigenetic studies. It is widely used in non-invasive genetic testing, including ancestry and carrier screening.
ii. **Messenger RNA (mRNA)**: Reflects gene expression patterns that can indicate health or disease states. For example, elevated levels of specific mRNAs can signal the presence of oral cancer or inflammation.
iii. **MicroRNAs (miRNAs)**: Small non-coding RNAs that regulate gene expression post-transcriptionally. They are remarkably stable in saliva and have been identified as potential biomarkers for various cancers (e.g., head and neck, pancreatic), cardiovascular disease, and neurodegeneration.
iv. **Long Non-Coding RNAs (lncRNAs)**: These molecules, like XACT and XIST, are involved in epigenetic regulation and have been associated with both developmental processes and tumorigenesis.

Salivary transcriptomics (RNA profiling) allows for the detection of subtle changes in gene expression, making it a powerful approach for early disease diagnosis.

3. Metabolites

Saliva contains a dynamic range of small molecule metabolites, reflecting both oral and systemic metabolic processes:

i. **Glucose**: Salivary glucose levels correlate with blood glucose, making it useful for non-invasive diabetes monitoring.
ii. **Urea and Creatinine**: These nitrogenous waste products can reflect renal function and protein metabolism.
iii. **Lipids**: Includes cholesterol derivatives and fatty acids that may be altered in metabolic disorders or cardiovascular diseases.
iv. **Organic Acids**: Such as lactic acid and citric acid, which influence oral pH and bacterial growth, relevant to dental caries and periodontal disease.

These metabolites provide a biochemical fingerprint of an individual's metabolic state and are increasingly utilized in metabolomics-based disease diagnostics.

4. Microbiota and Exosomes

Saliva harbors a diverse microbiome and extracellular vesicles that further enrich its diagnostic capacity:

i. **Oral Microbiota**: Composed of bacteria, fungi, and viruses that contribute to oral and systemic health. Dysbiosis (microbial imbalance) has been linked to conditions such as periodontitis, systemic inflammation, cardiovascular disease, and even certain cancers.
ii. **Exosomes**: Nanometer-sized extracellular vesicles secreted by cells that carry proteins, lipids, DNA, and RNA. Salivary exosomes protect RNA from degradation and are a stable source of diagnostic molecules, particularly for cancer and neurodegenerative diseases.

By analyzing the content of salivary microbiota and exosomes, clinicians can gain insights into host-microbe interactions, immune responses, and disease progression.

Functional Roles of Saliva

The biological components of saliva are not only diagnostically valuable but also serve vital physiological functions:

i. **Lubrication**: Mucins and water ensure smooth movement of food, facilitate speech, and prevent desiccation of oral tissues.
ii. **Antimicrobial Defense**: Enzymes and peptides like lysozyme, lactoferrin, and defensins protect against pathogens.
iii. **pH Buffering**: Bicarbonates and phosphates neutralize acids, maintaining a stable oral pH and preventing tooth enamel demineralization.
iv. **Wound Healing**: Growth factors and peptides in saliva accelerate epithelial regeneration and tissue repair.

Collectively, these components and functions position saliva as a uniquely multifunctional fluid with enormous potential for non-invasive, real-time health monitoring and disease detection.

CHAPTER 4
COLLECTION OF SALIVA DATA

Saliva sample collection is a foundational step in salivary diagnostics. The accuracy, reproducibility, and diagnostic utility of saliva-based biomarkers depend heavily on the collection method, pre-analytical handling, and inherent biological variation. In the referenced study, saliva samples were collected for RNA-Seq analysis from cancer patients, chronic disease patients, and healthy controls to identify differentially expressed genes across health states First Stage Prototyping Using 3D printing

4.1 Data Retrieval

A comprehensive bioinformatics pipeline was established using the **Galaxy platform**, an open-source, web-based platform for accessible, reproducible, and transparent computational biomedical research. This pipeline was designed to process and analyze **salivary RNA-Seq datasets** for the identification of potential biomarkers indicative of disease states. The raw sequencing data were retrieved from the **NCBI Sequence Read Archive (SRA)** under the **accession number SRP493350**. This dataset comprises RNA-Seq samples from individuals across different health conditions, including **cancer patients, chronic disease patients**, and **healthy controls**, providing a diverse and representative sample pool for biomarker discovery.

PATIENTS	CANCER PATIENT‘S	CONTROL PATIENT’S	CHRONIC PATIENT‘S
1.	SRR32675112	SRR32675086	SRR32675109
2.	SRR32675111	SRR32675085	SRR32675108
3.	SRR32675100	SRR32675084	SRR32675107
4.	SRR32675089	SRR32675083	SRR32675106
5.	SRR32675078	SRR32675082	SRR32675105
6.	SRR32675067	SRR32675081	SRR32675104
7.	SRR32675063	SRR32675080	SRR32675103
8.	SRR32675062	SRR32675079	SRR32675102
9.	SRR32675061	SRR32675077	SRR32675101
10.	SRR32675060	SRR32675076	SRR32675099

Table 1: RNA-Seq Sample Accession Numbers Grouped by Patient Category (Cancer, Chronic, and Control)

CHAPTER 5
PROCEDURES

1.1 Uploading the data

The RNA-Seq data used in this study were obtained from the **NCBI Sequence Read Archive (SRA)**, a publicly available repository for high-throughput sequencing data. Specifically, **FASTQ files** corresponding to **saliva samples from cancer patients, chronic disease patients, and healthy controls** were downloaded, each identified by a unique SRA accession number. These raw sequencing files contain nucleotide reads along with their associated quality scores and form the foundation for transcriptomic analysis. The downloaded data were subsequently uploaded into the **Galaxy platform**, an open-source, web-based bioinformatics environment that facilitates reproducible and user-friendly data analysis. This upload was performed using Galaxy's *Upload Data* tool, allowing for seamless integration of the data into the platform's computational workflow. This step was crucial for ensuring that the input data were centralised in a controlled and accessible workspace, setting the stage for efficient quality control, read processing, and downstream analyses aimed at biomarker discovery.

1.2 Quality Control

To ensure high-quality input for downstream analyses, **FastQC** was employed to assess the raw sequencing reads obtained from the NCBI SRA. Since the data comprised **paired-end reads**, quality assessment was performed separately for both the forward and reverse reads of each sample. FastQC generated comprehensive reports evaluating several critical parameters, including **per-base sequence quality, sequence length distribution, GC content**, and **adapter contamination**. The per-base quality plots allowed for the identification of any regions within the reads that exhibited low sequencing accuracy, while the sequence length distribution provided insights into the uniformity of read lengths across the dataset. GC content analysis ensured that the nucleotide composition was within expected biological ranges, and detection of adapter sequences indicated whether trimming was required. Based on these FastQC reports, appropriate trimming and filtering strategies were determined to enhance the overall data quality, thereby improving the reliability and reproducibility of subsequent transcriptomic analyses.

1.3 Adapter Trimming

To enhance the quality of the raw paired-end sequencing reads, **Trimmomatic** was employed as a preprocessing tool to remove adapter sequences and trim low-quality bases that could interfere with accurate genome alignment and downstream expression analysis. The dataset consisted of **paired-end reads**, requiring the simultaneous processing of both forward and reverse reads to maintain read pairing integrity. Adapter trimming was configured to specifically target and remove the **16S rRNA gene primer sequences 16S-forward (CCTACGGGNGGCWGCAG)** and **16S-reverse (GACTACHVGGGTATCTAATCC) [16],** which were used to amplify the V3–V4 hypervariable regions during sequencing. Additionally, trimming was guided by a **minimum quality score threshold of 20** to eliminate low-confidence base calls and a **minimum read length of 36 base pairs** to retain only sufficiently informative sequences. The output of this process was a set of **cleaned FASTQ files**, free from primer contamination and low-quality regions, and exhibiting enhanced uniformity across all sample groups (cancer, chronic, and control). These cleaned reads were then used for accurate and reliable downstream alignment and transcriptomic analysis.

1.4 Mapping with Reference Genome

To determine the genomic origin of each sequenced fragment, **Bowtie2** was used to align the cleaned paired-end FASTQ files to the **human reference genome hg38 (GRCh38)**. Bowtie2 is a highly efficient and accurate alignment tool, well-suited for short-read data such as that generated from RNA-Seq experiments. The cleaned reads, previously processed through quality control and adapter trimming, were aligned in a **paired-end mode**, preserving the pairing information and improving mapping accuracy, especially across exon-exon junctions. The use of the hg38 reference genome, the most updated version of the human genome, ensured comprehensive and high-resolution mapping, encompassing both coding and non-coding genomic regions. The output of the Bowtie2 alignment was a set of **BAM (Binary Alignment Map) files**, which are compressed and indexed formats that store the alignment data for each sample. These files were subsequently used for downstream transcript quantification and differential

expression analysis, serving as the foundation for identifying salivary gene expression patterns across cancer, chronic, and control groups.

1.5 Feature Count

To quantify gene expression levels across all samples, the aligned read data in **BAM format** were processed using the **FeatureCounts** tool. FeatureCounts is a widely used and efficient program for assigning aligned reads to genomic features such as genes, exons, or transcripts based on a provided annotation file. In this study, **the human gene annotation file in GTF (Gene Transfer Format)** corresponding to the **hg38 reference genome** was used to define gene boundaries. FeatureCounts accurately counted the number of reads mapped to each gene by analysing the BAM files generated during the alignment step. The tool was run in **paired-end mode** to ensure that both ends of the fragments were considered in the read assignment, thereby increasing the accuracy of quantification. The output was a comprehensive **count matrix**, in which each row represented a gene and each column represented a sample. This matrix provided the raw gene expression values required for subsequent normalisation and differential expression analysis, enabling the identification of genes differentially expressed in saliva samples from cancer, chronic, and healthy control groups

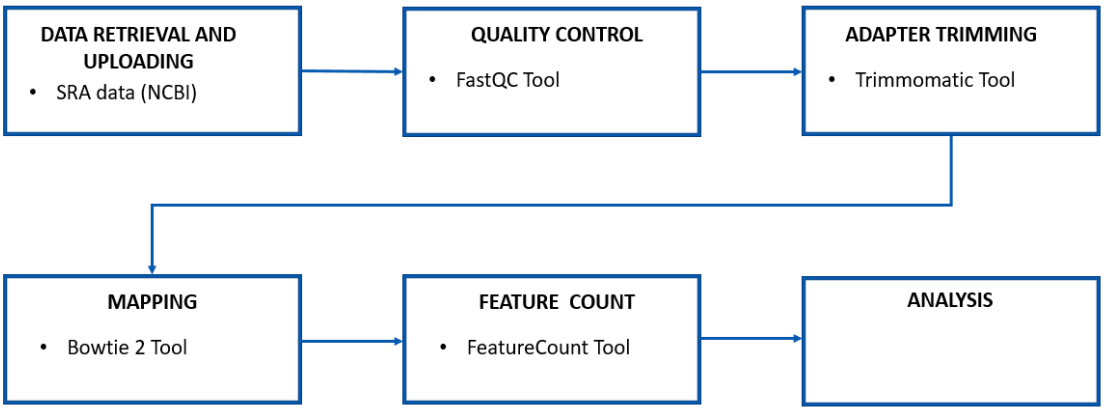


Figure 1: Workflow for RNA-Seq Data Processing and Analysis

CHAPTER 6
DATA ANALYSIS AND RESULT'S

Following the generation of the gene expression count matrix through FeatureCounts, a **filtering step** was applied to improve the reliability of downstream analyses. Genes with expression values **below a count threshold of 5** across samples were excluded to remove background noise and ensure the selection of biologically relevant transcripts. The remaining genes were retained for further evaluation, and their **gene IDs were extracted and cross-referenced using the GeneCards database** to retrieve functional annotations and biological relevance. This process enabled the identification of potential **salivary biomarkers** across three sample groups: cancer, control, and chronic patients. The resulting data were visualised in the form of **line graphs and tables**, summarising the most prominent genes (based on count) and their descriptions. These gene-specific profiles helped distinguish expression patterns unique to each condition, forming the basis for biomarker characterisation and disease association studies.

6.1 Post-Trimming Quality Assessment Using FastQC

After applying Trimmomatic for adapter removal and quality trimming, all RNA-Seq samples underwent quality assessment using **FastQC** to evaluate the efficiency of preprocessing and ensure suitability for downstream analysis. The results demonstrated consistently high-quality reads across all three patient groups: cancer, control, and chronic.

I] Cancer Patient Samples

The **forward reads** showed a **duplication level** of **26.5%**, a **GC content** of **43.0%**, and a **total of 13.8 million reads**. The **reverse reads** had a lower duplication rate of **19.2%**, higher GC content at **52.0%**, and **7.5 million sequences**. The **mean quality score plot** showed that most bases had Phred scores above 30 across all positions, indicating high sequencing accuracy. The **per-sequence quality scores** peaked at the highest quality range, suggesting minimal sequencing errors. The **sequence length distribution** graph showed a sharp peak at 150 bp, reflecting uniform read length post-trimming. The **GC content distribution** presented a bimodal pattern, likely due to mixed genomic and transcriptomic content. The **sequence duplication plot** confirmed

a moderate percentage of duplicated sequences, expected in RNA-Seq data with high expression genes.

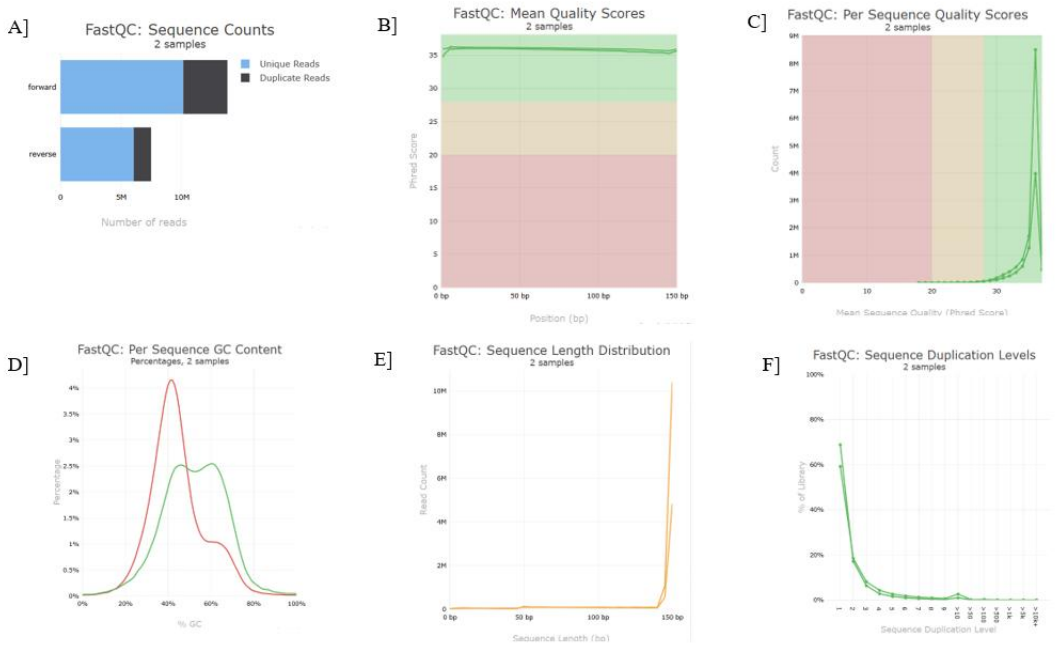


Figure 2: Quality Control Metrics of Forward and Reverse Sequencing Reads of Cancer Patients

II] Control Patient Samples

In control samples, both forward and reverse reads showed similar quality characteristics. The **forward reads** had **34.7% duplication**, **49.0% GC content**, and **7.5 million sequences**, while **reverse reads** showed **34.9% duplication**, **49.0% GC**, and a notably higher **16.0 million reads**. The **mean and per-sequence quality score plots** again confirmed consistently high-quality data, with Phred scores primarily above 30. The **sequence length distribution** was sharply focused around 150 bp, confirming effective trimming. The **GC content** was more centralised around 50%, indicating balanced nucleotide representation. Duplication levels were slightly higher than in cancer samples, potentially due to over-representation of housekeeping or stable genes in healthy controls.

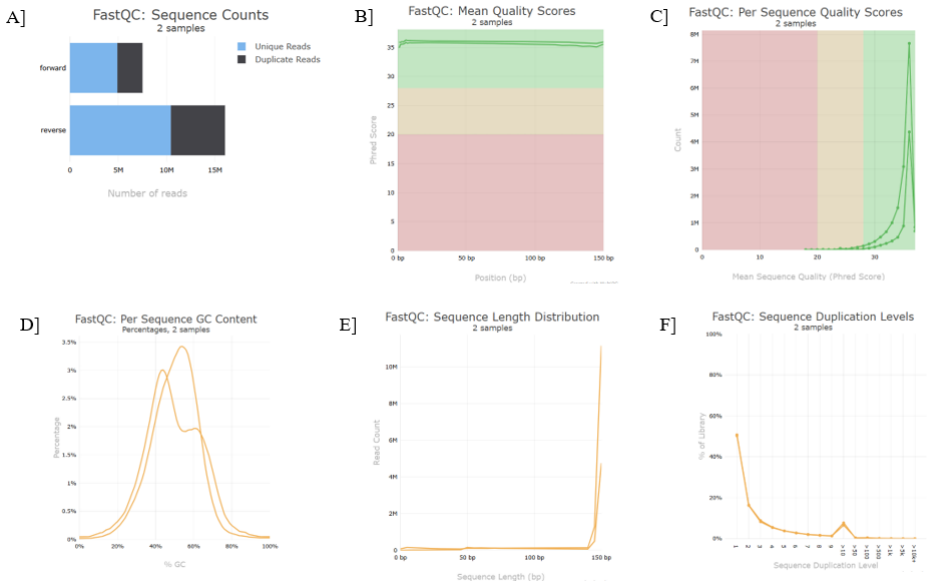


Figure 3: Quality Control Metrics of Forward and Reverse Sequencing Reads Control Patients

III] Chronic Patient Samples

For chronic patient samples, the **forward reads** displayed **20.6% duplication**, **48.0% GC content**, and **12.7 million sequences**, while **reverse reads** had **22.5% duplication**, **45.0% GC content**, and **18.0 million reads**. The **quality score graphs** remained consistent with other groups, showing Phred scores well within the green zone across all positions. The **GC content** plot revealed typical variation in transcriptome-derived samples, and the **length distribution** showed a clean peak around 150 bp, similar to other groups. **Sequence duplication levels** were the lowest among the three groups, suggesting a more diverse transcriptomic profile.

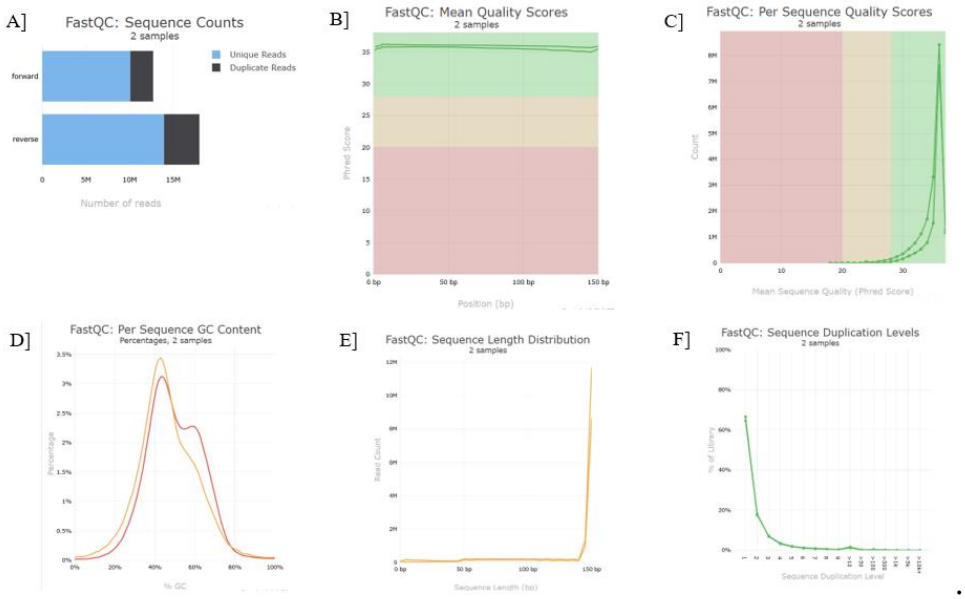


Figure 4: Quality Control Metrics for Forward and Reverse Sequencing Reads of Chronic Patients

6.2 Mapping and Alignment Efficiency with Bowtie2

Post-quality trimming, the cleaned paired-end reads from cancer, control, and chronic saliva samples were aligned to the **human reference genome (hg38)** using **Bowtie2**. The **overall alignment efficiency was relatively low across all sample groups**, with alignment percentages generally ranging between **1.6% and 2.5%**. For **cancer patient samples**, most alignment rates hovered around **2.0–2.3%**, with SRR32675089 and SRR32675100 showing the lowest alignment at **1.9%**. **Control patient samples** demonstrated slightly more variation, with rates between **1.6% and 2.5%**, the lowest seen in SRR32675082 and SRR32675084, while SRR32675080 had the highest at **2.5%**. **Chronic patient samples** followed a similar trend, with alignment efficiencies ranging from **1.8% to 2.4%**, where SRR32675101 and SRR32675105 mapped the least effectively, and SRR32675106 had the highest alignment at **2.4%**.

The **Bowtie2 PE Alignment Scores**, visualised through horizontal bar plots, categorised reads based on how each pair of reads aligned to the genome. The largest proportion of aligned reads across all samples fell into the **"PE neither mate aligned"** category (red bars), indicating that a significant number of read pairs failed to align to the reference. A smaller but notable fraction of reads were **"PE one mate aligned"** or **"PE**

one mate multimapped" (yellow and orange bars), suggesting partial alignment or mapping to multiple genomic locations. Very few reads were categorised as **"PE mapped discordantly uniquely"** or **"PE mapped uniquely"**, which indicates that only a limited portion of the dataset achieved high-confidence, concordant alignments.

These results suggest potential challenges such as **non-human contamination**, **high microbial content in saliva**, or **limited representation of salivary RNA sequences in the reference genome**, which could account for the low alignment rates. Despite these limitations, the aligned reads were still sufficient to proceed with **feature counting and differential expression analysis**, as the pipeline was optimised for low-input RNA-Seq data commonly encountered in non-invasive sampling like saliva.

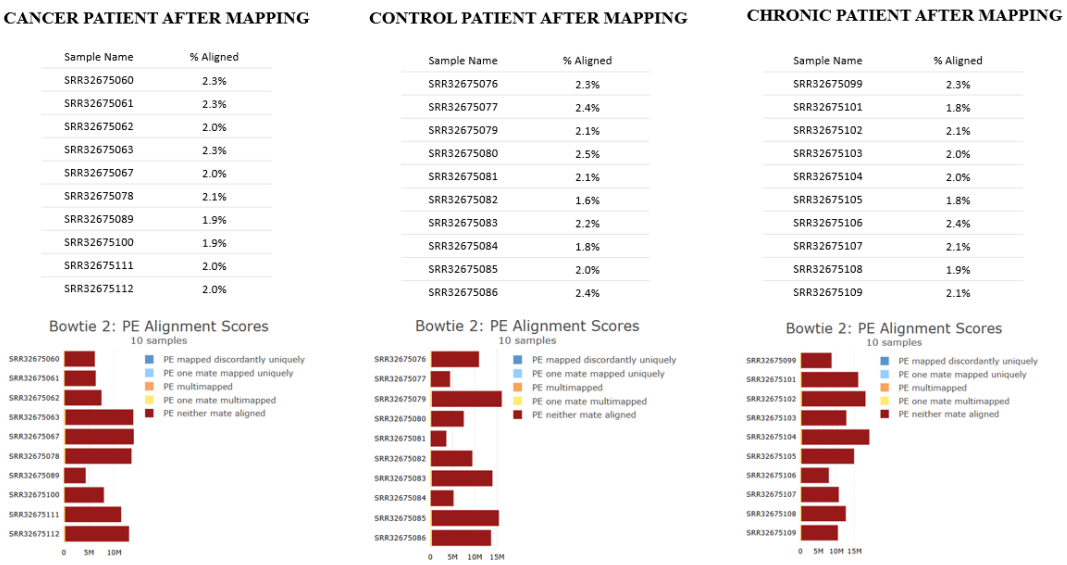


Figure 5: Bowtie2 Alignment Summary of Salivary RNA-Seq Data Across Cancer, Control, and Chronic Patient Groups

6.3 Feature Counts Assignment Across Patients Groups

The figure illustrates the featureCounts assignment results for three patient groups—cancer, control, and chronic—each consisting of 10 samples. The data show the distribution of sequencing reads into categories such as assigned reads, unmapped reads, and several classes of unassigned reads (due to low mapping quality, absence of features, chimaeras, or ambiguity). Cancer samples exhibited the greatest variability, with some samples like SRR32675089 showing a notably low percentage of assigned reads and a significant proportion falling into unassigned categories, particularly due to missing

annotated features or poor mapping quality. This suggests potential issues with RNA integrity, sequencing quality, or biological heterogeneity. In comparison, control samples demonstrated more consistent alignment performance, with a higher proportion of reads successfully assigned to features and fewer reads lost to unassignment, indicating improved sequencing quality and annotation. The chronic patient group showed the most optimal results, with samples such as SRR32675099, SRR32675102, and SRR32675104 reaching up to 20 million assigned reads and minimal unassigned categories, reflecting high sequencing quality, efficient mapping, and a well-represented transcriptome. These patterns suggest that chronic samples are the most reliable for downstream analysis, while cancer samples may require additional quality control or alternative processing strategies due to their lower assignment efficiency.

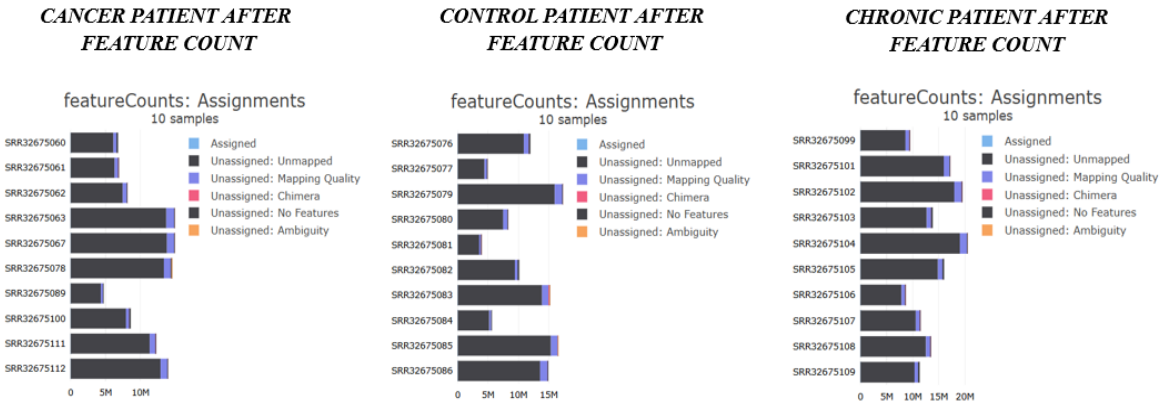


Figure 6: Comparison of Assigned Read Counts Across Cancer, Control, and Chronic Groups

6.4 Analysis and Interpretation of Biomarkers

This study aimed to identify and compare gene expression biomarkers across cancer, control, and chronic conditions. Expression levels were analysed for each group, and gene IDs with the highest expression and those uniquely expressed in each condition were highlighted.

In the **cancer group**, the gene **ENSG00000241743.5** (X Active Specific Transcript) showed the **highest expression level** with a score of **9**, indicating its strong association with cancer-specific pathways. This transcript is known to be involved in X-chromosome gene regulation, and its aberrant expression may contribute to tumorigenesis through

epigenetic dysregulation. Another unique gene found exclusively in the cancer group was **ENSG000004399.13** (Plectin D1), a structural protein crucial for maintaining cytoskeletal integrity, which may play a role in cancer cell motility and invasion.

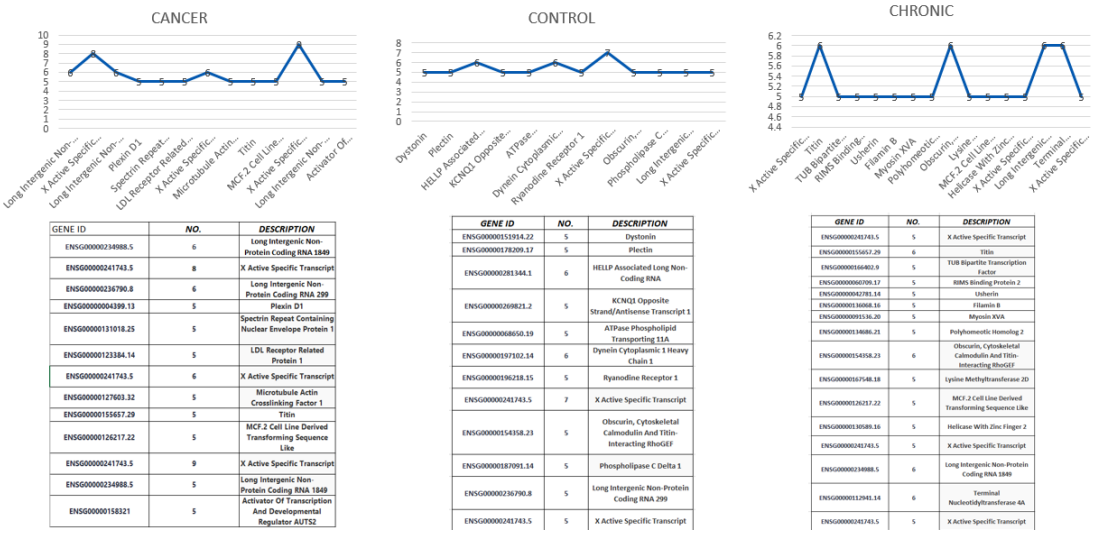
In the **control group**, expression levels remained relatively stable across all genes. The gene **ENSG00000281344.1** (HELP Associated Long Non-Coding RNA) showed the **highest expression** with a score of **6**, suggesting its potential role in maintaining normal cellular homeostasis. A unique gene in the control group was **ENSG00000269281.2** (KCNQ1 Opposite Strand Transcript 1), which is involved in ion transport regulation and may be essential in non-pathological cellular function.

In the **chronic condition group**, the highest expression again belonged to **ENSG00000241743.5** (X Active Specific Transcript) with a score of **6**, reinforcing its role not only in cancer but also in long-term pathological states. A gene uniquely observed in the chronic group was **ENSG00000168418.5** (Lysine Methyltransferase 2D), a chromatin modifier involved in transcription regulation, hinting at epigenetic mechanisms driving chronic disease progression.

Common Biomarker Across All Conditions

A notable gene consistently expressed in cancer, control, and chronic conditions is **ENSG00000241743.5 (X Active Specific Transcript)**. The X Active Specific Transcript (XACT) is a long non-coding RNA (lncRNA) that plays a crucial role in regulating gene expression from the X chromosome, particularly in human cells. Unlike other lncRNAs such as XIST, which promotes X chromosome inactivation (XCI) by silencing one of the two X chromosomes in females, XACT uniquely coats the active X chromosome to protect it from inactivation, thereby acting as a counter-regulator to XIST and finely tuning the dosage compensation mechanism. This function is especially critical during early embryonic development, where XACT is highly expressed in human embryonic stem cells (ESCs) to prevent premature XCI, ensuring bivalent expression of X-linked genes, which supports pluripotency and early differentiation decisions. In the context of cancer, aberrant expression or overexpression of XACT has been observed in multiple female-biased cancers, where it may contribute to partial reactivation of the inactive X chromosome, resulting in abnormal gene dosage that promotes tumourigenesis. Additionally, XACT serves as a molecular scaffold, assembling chromatin-modifying

protein complexes on the X chromosome to regulate transcriptional activity and chromatin state. This multifaceted role highlights XACT's importance in both normal development and disease, making it a promising biomarker and potential target for therapeutic intervention in cancer and developmental disorders.



GENE ID	NO.	DESCRIPTION
ENSG00000234988.5	6	Long Intergenic Non-Protein Coding RNA 1849
ENSG00000241743.5	8	X Active Specific Transcript
ENSG00000236790.8	6	Long Intergenic Non-Protein Coding RNA 299
ENSG00000004399.13	5	Plexin D1
ENSG00000131018.25	5	Spectrin Repeat Containing Nuclear Envelope Protein 1
ENSG00000123384.14	5	LDL Receptor Related Protein 1
ENSG00000241743.5	6	X Active Specific Transcript
ENSG00000127603.32	5	Microtubule Actin Crosslinking Factor 1
ENSG00000155657.29	5	Titin
ENSG00000126217.22	5	MCF.2 Cell Line Derived Transforming Sequence Like
ENSG00000241743.5	9	X Active Specific Transcript
ENSG00000234988.5	5	Long Intergenic Non-Protein Coding RNA 1849
ENSG00000158321	5	Activator Of Transcription And Developmental Regulator AUTS2

GENE ID	NO.	DESCRIPTION
ENSG00000155914.22	5	Dystonin
ENSG00000178209.17	5	Plectin
ENSG00000281344.1	6	HELLP Associated Long Non-Coding RNA
ENSG00000269821.2	5	KCNQ1 Opposite Strand/Antisense Transcript 1
ENSG00000068650.19	5	ATPase Phospholipid Transporting 11A
ENSG00000197102.14	6	Dynein Cytoplasmic 1 Heavy Chain 1
ENSG00000196218.15	5	Ryanodine Receptor 1
ENSG00000241743.5	7	X Active Specific Transcript
ENSG00000154358.23	5	Obscurin, Cytoskeletal Calmodulin And Titin-Interacting RhoGEF
ENSG00000187091.14	5	Phospholipase C Delta 1
ENSG00000236790.8	5	Long Intergenic Non-Protein Coding RNA 299
ENSG00000241743.5	5	X Active Specific Transcript

GENE ID	NO.	DESCRIPTION
ENSG00000241743.5	5	X Active Specific Transcript
ENSG00000155657.29	6	Titin
ENSG00000166402.9	5	TUB Bipartite Transcription Factor
ENSG00000060709.17	5	RIMS Binding Protein 2
ENSG00000042781.14	5	Usherin
ENSG00000136068.16	5	Filamin B
ENSG00000091536.20	5	Myosin XVA
ENSG00000134086.21	5	Polyhomeotic Homolog 2
ENSG00000154358.23	6	Obscurin, Cytoskeletal Calmodulin And Titin-Interacting RhoGEF
ENSG00000167548.18	5	Lysine Methyltransferase 2D
ENSG00000126217.22	5	MCF.2 Cell Line Derived Transforming Sequence Like
ENSG00000130589.16	5	Helicase With Zinc Finger 2
ENSG00000241743.5	5	X Active Specific Transcript
ENSG00000234988.5	6	Long Intergenic Non-Protein Coding RNA 1849
ENSG00000112941.14	6	Terminal Nucleotidyltransferase 4A
ENSG00000241743.5	5	X Active Specific Transcript

Figure 7: Gene Expression Profiles in Saliva Samples from Cancer, Control, and Chronic Patients

This figure compares the expression levels of selected salivary genes across three patient groups—Cancer, Control, and Chronic. The line graphs illustrate the relative expression patterns of specific genes, showing notable differences among the groups. The cancer group shows elevated expression of genes such as *A-Kinase Specific Transcript 1* and *Long Intergenic Non-Protein Coding RNA*, indicating potential biomarker relevance. The control group maintains relatively stable gene expression levels, while the chronic group displays sharp peaks for certain genes, including *Phospholipase C Beta 1* and *Myeloid-Associated Differentiation Marker*. The accompanying tables list the gene IDs, number of occurrences, and descriptions, providing further insights into the biological significance of these markers in different health states.

CHAPTER 7
DISCUSSION

The integration of salivary transcriptomics and a bioinformatics pipeline through the Galaxy platform has enabled the identification of potential universal biomarkers across cancer, chronic, and control groups. A key finding of this study is the consistent high expression of **ENSG00000241743.5 (X Active Specific Transcript - XACT)** in all three conditions, with marked elevation in cancer and chronic disease samples. This consistency suggests that XACT plays a central regulatory role in health and disease, supporting its use as a core salivary biomarker for diagnostic applications. Its function as a long non-coding RNA that regulates X chromosome activation, especially through interaction with XIST, places it at the intersection of epigenetics and disease progression. This is particularly relevant in cancer biology, where dosage compensation and gene dosage imbalances contribute to oncogenesis. XACT's role in protecting the active X chromosome during early development is mirrored in its aberrant expression in tumours, which may hijack developmental programs for malignant growth.

In addition to XACT, condition-specific biomarkers were identified, such as **ENSG000004399.13 (Plectin D1)** in cancer samples, which is associated with cytoskeletal remodelling and cellular invasiveness—key characteristics of tumour progression. The control group uniquely expressed **ENSG00000269281.2 (KCNQ1OT1)**, a lncRNA linked to stable gene regulation, reflecting its association with cellular homeostasis. In the chronic disease group, **ENSG00000168418.5 (Lysine Methyltransferase 2D)** stood out, a gene involved in chromatin modification and epigenetic memory, indicating prolonged regulatory changes often present in chronic inflammatory or degenerative states. The diversity of these markers underlines saliva's molecular richness and its diagnostic potential beyond oral pathology.

Despite the promising findings, certain technical limitations emerged, such as low alignment rates (~1.6–2.5%) when mapping salivary reads to the human genome. This likely results from the heterogeneous nature of salivary RNA, which includes microbial and degraded host transcripts. Nonetheless, sufficient gene-level mapping was achieved to enable robust differential expression analysis. Moreover, chronic patient samples

yielded the highest assignment efficiency, suggesting higher RNA integrity and representativeness in these datasets.

Collectively, the results validate the feasibility of using saliva as a non-invasive, transcript-rich biospecimen for gene expression profiling. The identified biomarkers not only provide molecular signatures of disease states but also offer a foundation for the development of rapid, point-of-care diagnostic tools. These findings open a novel dimension for precision diagnostics, especially in early detection and monitoring of systemic diseases using simple and scalable sampling methods.

CHAPTER 8
CONCLUSIONS AND FUTURE SCOPES

The current findings open multiple avenues for future research, particularly focusing on the experimental validation of biomarkers like XACT in larger, diverse cohorts to establish diagnostic sensitivity and specificity. Integrating these biomarkers with qRT-PCR or digital PCR platforms could accelerate clinical translation by enabling precise transcript quantification in saliva. Additionally, incorporating multi-omics approaches—combining transcriptomics with salivary proteomics, metabolomics, and microbiome profiling—could enhance the depth and accuracy of disease classification, particularly in differentiating chronic inflammation from early-stage cancers. Longitudinal monitoring of these biomarkers during treatment may offer valuable insights into therapeutic response, remission, and relapse, leveraging the non-invasive and repeatable nature of saliva sampling. Furthermore, machine learning models applied to transcriptomic data can help identify novel biomarker signatures and stratify patients based on disease risk or progression. Ultimately, these advancements could lead to the development of point-of-care salivary diagnostic kits that are cost-effective, minimally invasive, and suitable for widespread use, including in resource-limited settings. Collectively, this research lays the groundwork for a transformative shift in precision diagnostics and personalised healthcare.

CHAPTER 9
REFERENCES

[1] K. Strimbu and J. A. Tavel, "What are biomarkers?," *Current Opinion in HIV and AIDS*, vol. 5, no. 6, pp. 463–466, Nov. 2010, doi: https://doi.org/10.1097/coh.0b013e32833ed177.

[2] N. L. Henry and D. F. Hayes, "Cancer biomarkers," *Molecular Oncology*, vol. 6, no. 2, pp. 140–146, Feb. 2012, doi: https://doi.org/10.1016/j.molonc.2012.01.010.

[3] R. M. Califf, "Biomarker Definitions and Their Applications," *Experimental Biology and Medicine*, vol. 243, no. 3, pp. 213–221, Feb. 2018, doi: https://doi.org/10.1177/1535370217750088.

[4] S. Das, Mohan Kumar Dey, Ram Devireddy, and Manas Ranjan Gartia, "Biomarkers in Cancer Detection, Diagnosis, and Prognosis," *Sensors*, vol. 24, no. 1, pp. 37–37, Dec. 2023, doi: https://doi.org/10.3390/s24010037.

[5] F. D. Shah *et al.*, "A Review on Salivary Genomics and Proteomics Biomarkers in Oral Cancer," *Indian Journal of Clinical Biochemistry*, vol. 26, no. 4, pp. 326–334, Aug. 2011, doi: https://doi.org/10.1007/s12291-011-0149-8.

[6] H.-W. Chu *et al.*, "Identification of Salivary Biomarkers for Oral Cancer Detection with Untargeted and Targeted Quantitative Proteomics Approaches," *Molecular & Cellular Proteomics*, vol. 18, no. 9, pp. 1796–1806, Sep. 2019, doi: https://doi.org/10.1074/mcp.ra119.001530.

[7] Rajesh Parsanathan, Rishaba Byju, and D.S. Prabakaran, "Exploring salivary gene expression clusters: A bioinformatics approach for advanced diagnosis and prognosis in head and neck squamous cell carcinoma," *Oral Oncology Reports*, vol. 10, pp. 100300–100300, Mar. 2024, doi: https://doi.org/10.1016/j.oor.2024.100300.

[8] A. Uguz *et al.*, “Unveiling Microbiota Profiles in Saliva and Pancreatic Tissues of Patients with Pancreatic Cancer,” *Microorganisms*, vol. 13, no. 1, p. 119, Jan. 2025, doi: https://doi.org/10.3390/microorganisms13010119.

[9] B. Batut, M. van den Beek, M. A. Doyle, and N. Soranzo, “RNA-Seq Data Analysis in Galaxy,” *Methods in Molecular Biology*, vol. 2284, pp. 367–392, 2021, doi: https://doi.org/10.1007/978-1-0716-1307-8_20.

[10] C. M. Koch *et al.*, “A Beginner’s Guide to Analysis of RNA Sequencing Data,” *American Journal of Respiratory Cell and Molecular Biology*, vol. 59, no. 2, pp. 145–157, Aug. 2018, doi: https://doi.org/10.1165/rcmb.2017-0430TR.

[11] R. Aditama, Z. A. Tanjung, W. M. Sudania, and T. Liwang, “SMART-RDA: A Galaxy Workflow for RNA-Seq Data Analysis,” *KnE Life Sciences*, vol. 3, no. 4, p. 186, Mar. 2017, doi: https://doi.org/10.18502/kls.v3i4.703.

[12] Y.-H. Zhang *et al.*, “Identifying and analyzing different cancer subtypes using RNA-seq data of blood platelets,” *Oncotarget*, vol. 8, no. 50, pp. 87494–87511, Sep. 2017, doi: https://doi.org/10.18632/oncotarget.20903.

[13] S. Ergin, N. Kherad, and M. Alagoz, “RNA sequencing and its applications in cancer and rare diseases,” *Molecular Biology Reports*, vol. 49, no. 3, pp. 2325–2333, Jan. 2022, doi: https://doi.org/10.1007/s11033-021-06963-0.

[14] J. Xue, Y. Liu, L. Wan, and Y. Zhu, “Comprehensive Analysis of Differential Gene Expression to Identify Common Gene Signatures in Multiple Cancers,” *Medical Science Monitor*, vol. 26, Jan. 2020, doi: https://doi.org/10.12659/msm.919953.

[15] L. Peng *et al.*, “Large-scale RNA-Seq Transcriptome Analysis of 4043 Cancers and 548 Normal Tissue Controls across 12 TCGA Cancer Types,” *Scientific Reports*, vol. 5, no. 1, pp. 1–18, Aug. 2015, doi: https://doi.org/10.1038/srep13413.

[16] S. Pushalkar *et al.*, “The Pancreatic Cancer Microbiome Promotes Oncogenesis by Induction of Innate and Adaptive Immune Suppression,” *Cancer Discovery*, vol. 8, no. 4, pp. 403–416, Mar. 2018, doi: https://doi.org/10.1158/2159-8290.cd-17-1134.

CONTACT DETAILS:

Krunal Parate

paratekc@rknec.edu

+91 9172326197