# A Computational Approach for Identification of Salivary Biomarkers

Krunal Parate[1*] and Arpita Parakh[2]

[1,2]Department of Biomedical Engineering, Shri Ramdeobaba College of Engineering and Management, Nagpur, India
paratekc@rknec.edu*, parakh@rknec.edu

**Abstract.** Atransparent,readilyaccessiblebodilyfluid,salivais essential for maintaining oral health. Numerous biological substances that are useful for illness detection are also present. Saliva has drawn interest recently as a non-invasive, reasonably priced method of diagnosing a range of illnesses. The kinds of chemicals found in saliva that can function as biomarkers are described in this overview, including proteins, Deoxyribonucleic acid (DNA),Ribonucleic acid(RNA), and metabolites. Diabetes, heart disease, and oral cancer can all be detected and tracked with the use of these biomarkers. Saliva collection is easy, safe, and painless, which makes it perfect for routinemedical examinations. The difficulties in using saliva for diagnosis are also covered in this research, including sample storage and guaranteeing reliable results. In spite of these obstacles, continuous. Saliva-based testing is getting better because to continued research, despite these obstacles. Saliva biomarkers have the potential to be a standard component of medical diagnostics in the future, particularly for early illness identification.

**Keywords:**Saliva, Biomarkers, Non-invasive diagnosis, Salivary diagnostics, Oral health, Disease detection, Point-of-care testing.

## 1 Introduction

Biomarkers are overview indicators that reflect biological conditions, disease progression, or response to therapy. In contemporary healthcare, they are essential instruments for prognosis, early diagnosis, therapeutic response monitoring, and treatment strategy customization [1][2][3][4]. Biomarkers in oncology help direct clinical judgement and better the effectiveness of clinical trials and medication expansion [2][4]. Each of the biomarkers in their classification are diagnostic, prognostic, predictive, and pharmacodynamics has a unique function in increase patient care [3].

Today's research work enables the use of saliva as a diagnostic fluid with great potential, despite the traditional accepted biomarker research requiring invasive fluid collection, including blood or tissue biopsies. This is because saliva offers a non-invasive,readily available, and reasonably cost-effective method to identify molecular expression indicative of diseases, particularly systemic and oral cancers [5][6][7]. The fluid contains proteins, metabolites,Deoxyribonucleic acid (DNA), and Ribonucleic acid (RNA), which provide information about systemic conditions and support real-time observation of diseases [5][6]. Moreover, shifts in the composition of the human


S. Bhalerao et al. (eds.), *Proceedings of the 2nd International Conference on Recent Advancement and Modernization in Sustainable Intelligent Technologies & Applications (RAMSITA-2026)*, Advances in Intelligent Systems Research 207,
https://doi.org/10.2991/978-94-6239-678-4_2

mouth's microbial flora have been associated with both oral cancers and systemic cancers such as pancreatic cancer, implying that it is feasible to use saliva to identify cancers that are not located in the mouth [8]. The capability to use saliva to transmit gene expression information indicative of both systemic and focal diseases farther accentuates the significance and utility of saliva. For instance, it has been shown to be related with the upregulation of key genes and specific expression patterns for both head and neck squamous cell carcinoma (HNSCC) [7], oral squamous cell carcinoma (OSCC) [6], and potential pancreatic carcinogenesis [8].

Galaxy and other bioinformatics platforms are becoming essential for analyzing such intricate transcriptome data. From preprocessing raw data to visualizing and analyzing differential gene expression, Galaxy enables researchers to carryoutend-to-endRibonucleic acid sequencing (RNA-Seq)workflows[9][10][11]. Because of its accessibilityandreproducibility, morepeople withoutextensiveprogramming-knowledgecanparticipatein genomics research [10]. In order to find identical salivary biomarkers—genes that are consistently differentially expressedacrossdiseasetypes—weprocessedRibonucleic acid sequencing (RNA-Seq)data from the saliva samples of cancer patients, patients with chronic diseases, and healthy controls using Galaxy.

Findingtheseuniformlyexpressed salivarybiomarkersisthe primary goal of this research in order to support the non- invasive identification, categorization, and surveillance of malignancies and chronic illnesses. Such biomarkers could transform screening procedures and facilitate individualized treatment planning, which is crucial given the rising prevalence of cancer and the need for early detection measures [12][13]. Additionally, pan-cancer investigations have shown that many malignancies, including lung, liver, breast, and cervical cancer, share gene expression profiles [14][15],highlighting the possibilityof finding biomarkers that are not specific to any one form of cancer.

Inconclusion,apotentandnon-invasivemethodforlocating universal biomarkersacross a varietyofmedical states isthe combination of salivary transcriptomic and bioinformatics. We have shown a repeatable and user-friendly method for analyzing Ribonucleic acid sequencing (RNA-Seq)data from saliva using the Galaxy platform,whichhasallowedustoidentifygenesthatareconsistently expressed differently in cancer, chronic illness, and healthy people [9][10][11]. Saliva's molecular diversity and capacity to represent both local and systemic health conditions, in addition to its affordabilityand ease of use, make it a promising diagnostic fluid [5][6][7][8]. The discovery of these salivary biomarkers presents a viable avenue for illness classification, early detection, and monitoring across a variety of disorders. This study emphasizestheimportanceofascalable,patient-friendly,and precision-focused diagnostic methodology that supports the objectives of personalized medicine and has a wider public health impact as transcriptome findings are progressively incorporated into clinical practice[4][12][13][14][15]. To fully understand the therapeutic application of these findings, more studyand validation in bigger cohorts will be essential.

## 2 Materials

A bioinformatics pipeline was established using the **Galaxy platform**toidentifypotentialsalivarybiomarkersassociated with disease states. Salivary Ribonucleic acid se-

quencing (RNA-Seq)datasets were retrieved from the **NCBI (National Centre of Biotechnology Information) Sequence Read Archive (SRA)** under accession number **SRP493350** as shown in table1whichincludessamplesfromcancerpatients,chronic disease patients, and healthy controls.

**Table1:**Ribonucleic acid sequencing (RNA-Seq) SampleAccessionNumbersGroupedby Patient Category (Cancer, Control, and Chronic)

| Patients | Cancer | Control | Chronic |
|---|---|---|---|
| 1 | | SRR32675112SRR32675086SRR32675109 | |
| 2 | | SRR32675111SRR32675085SRR32675108 | |
| 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 |

# 3 Methodology

The bioinformatics workflow used for saliva Ribonucleic acid sequencing (RNA-Seq) data analysis (Figure 1) includes data retrieval, quality control, adapter trimming, mapping, feature counting, and downstream analysis.

## 3.1 Data Retrieval and uploading

Ribonucleic acid sequencing (RNA-Seq)data for this study were obtained from the **NCBI Sequence Read Archive (SRA)**. FASTQ files corresponding to saliva samples from **cancer patients, individuals with chronic diseases, and healthy controls** were downloaded using unique SRA accession numbers. These raw sequencing files, containing nucleotide sequences and quality scores, were uploaded to the **Galaxy platform** using the Upload Data tool. Centralizing the data within Galaxy enabled standardized quality control, read processing, and downstream transcriptome analysis for biomarker discovery in a reproducible workflow.

## 3.2 Quality Control

FastQC was used to assess the quality of raw paired-end Ribonucleic acid sequencing(RNA-Seq) data obtained from NCBI SRA. Quality checks were performed separately for forward and reverse reads, evaluating key metrics such as per-base sequence quality, GC content, sequence length distribution, and adapter contamination. The reports helped identify low-quality regions and the need for adapter trimming, ensuring that only high-quality data were used. This quality control step improved the reliability and reproducibility of subsequent transcriptome analyses.

### 3.3 Adapter Trimming

Trimmomatic was used to preprocess paired-end reads by removing adapter/primer sequences and trimming low-quality bases while preserving read pairing integrity. The 16S rRNA primer sequences—**16S-forward (CCTACGGGNGGCWGCAG)** and **16S-reverse (GACTACHVGGGTATCTAATCC)** [16] were specifically trimmed. Reads shorter than **36 bp** and bases with quality scores below Q**20** were discarded. This cleaning step produced high-quality, uniform FASTQ files across control, cancer, and chronic samples, enabling accurate downstream alignment and transcriptome analysis.

### 3.4 Mapping with Reference Genome

As shown in figure 1, The quality-filtered paired-end FASTQ files were aligned to the human reference genome **hg38 (GRCh38)** using **Bowtie2** to determine the genomic origin of each read. Alignment was performed in paired-end mode to maintain read pairing and improve mapping accuracy. The use of the updated hg38 genome enabled comprehensive mapping of both coding and non-coding regions. The alignment process generated **BAM files**, which were used for downstream transcript quantification and differential gene expression analysis across cancer, chronic, and control saliva samples.

### 3.5 Feature Counts

After generating the gene count matrix using FeatureCounts, low-expression genes (counts < 5) were filtered out to reduce noise. Remaining genes were annotated using GeneCards, enabling identification of disease-specific salivary biomarkers across cancer, chronic, and control groups as shown in figure 1.

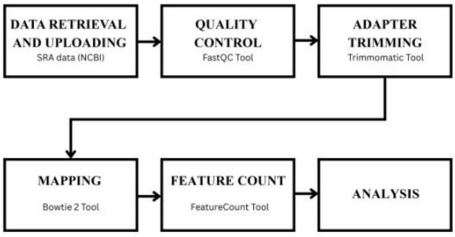


**Fig.1.Workflow of Salivary RNA-Seq Data Analysis for Biomarker Identification**

Showsthe RNA-Seq analysis workflow used in this study. Salivary RNA-Seq data were obtained from the NCBI SRA database, quality-checked using FastQC, trimmed with Trimmomatic, aligned to the human reference genome using Bowtie2, and quantified with FeatureCounts to identify differentially expressed salivary biomarkers.

# 4 Results

## 4.1 Post-Trimming Quality Assessment Using FastQC

AfterapplyingTrimmomaticforadapterremovalandquality trimming, all Ribonucleic acid sequencing (RNA-Seq) samples underwent quality assessment using **FastQC**to evaluate the efficiency of preprocessingandensuresuitabilityfordownstreamanalysis. The results demonstrated consistently high-quality reads across all three patient groups: cancer, control, and chronic.

### 4.1.1 CancerPatientSamples

There were 13.8 million reads in all, with a GC content of 43.0%andaduplication level of26.5%intheforwardreads. The reverse reads had 7.5 million sequences, a greater GC contentof52.0%,andalower duplicationrateof19.2%.The majority of bases received Phred values above 30 at every position, indicating great sequencing accuracy, according to themean qualityscore plot. At thehighest qualityrange,the per-sequence quality scores maximized, indicating few sequencing errors. After trimming, the read length was uniform, as evidenced by the sharp peak at 150 bp in the sequence length distribution graph (figure 2). The bimodal pattern in the GC content distribution was probably caused by a combination oftranscriptomicandgenomiccontent. Asseen in figure 2, the sequence duplication plot verified a modest proportionofduplicatedsequences,whichistypicalof Ribonucleic acid sequencing(RNA-Seq) data including high expression genes.

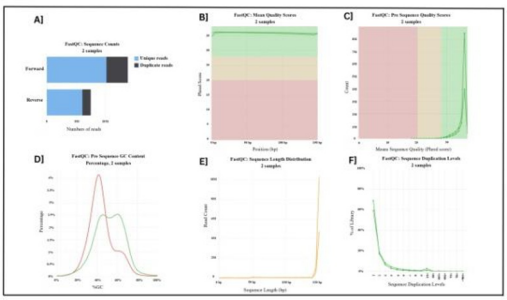


**Fig. 2. Quality Control Metrics of Cancer RNA-Seq Samples**
Quality control metrics of forward and reverse Ribonucleic acid sequencing (RNA-Seq) reads. The sequence count analysis (A) shows the total number of reads along with unique and duplicated reads for each sample. Mean Phred quality scores across read positions (B) remain consistently above 30, indicating high sequencing quality. The per-sequence quality score distribution (C) confirms that most reads exhibit high base-calling accuracy. GC content distribution (D) reveals differences in nucleotide composition between the forward and reverse samples. Sequence length distribution (E) demonstrates uniform read lengths of approximately 150 bp, characteristic of Illumina sequencing. Sequence duplication levels (F) indicate higher duplication in the forward sample compared to the reverse sample.

### 4.1.2 ControlPatientSamples

Both forward and reverse readings displayed comparable qualityattributesin control samples as shown in figure 3. Thereversereadshada significantly higher number of reads (16.0 million), 34.9% duplication,and 49.0%GCthan theforwardreads,whichhad 7.5 million sequences and 34.7% duplication. With Phred scores mostly above 30, the mean and per-sequence quality score graphs once more verified consistently high-quality data.Effectivetrimmingwasconfirmedbytheclearfocusof the sequence length distribution at 150 bp. Around 50% of the GC content was more central, suggesting balanced nucleotide representation. As seen in figure 3, the over- representation of housekeeping or stable genes in healthy controls may have contributed to the slightly greater duplication levels compared to cancer samples.

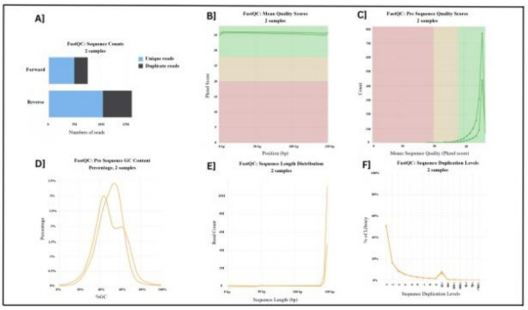


**Fig. 3. Quality Control Metrics of Control RNA-Seq Samples**
Quality control metrics of forward and reverse sequencing reads. Sequence counts (A) indicate total reads and duplication levels, with both samples showing relatively high duplication. Mean quality scores across read positions (B) and per-sequence quality scores (C) demonstrate consistently high sequencing accuracy. GC content distribution (D) is similar between samples, suggesting no strong GC bias. Sequence length distribution (E) shows uniform read lengths (~150 bp), and duplication analysis (F) highlights substantial duplication in both forward and reverse samples.

### 4.1.3 Chronic Patient Samples

Reversereadsshowed22.5%duplication,45.0%GCcontent, and18.0millionreads,whereasforwardreadsshowed20.6% duplication, 48.0% GC content, and 12.7 million sequences for chronic patient samples. Phred scores well within the green zone in every position, according to the quality score graphs, which stayed consistent with those of the other groups. Similar to previous groups, the length distribution displayed a clear peak at 150 bp, and the GC content plot exhibited expected fluctuation in transcriptome-derived samples.Outofthethreegroups,sequenceduplicationlevels were the lowest, indicating a more varied transcriptome profile (see figure 4).

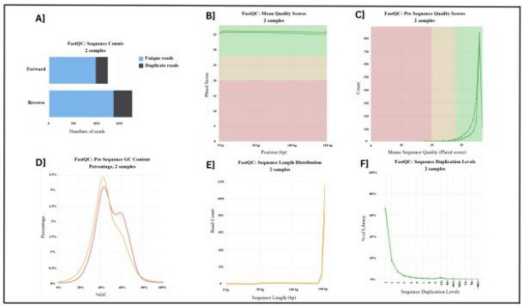

**Fig. 4.Quality Control Metrics of Chronic RNA-SeqSamples**
Quality control metrics for forward and reverse sequencing reads. Sequence counts (A) show total reads and moderate duplication levels. Mean quality scores across read positions (B) and per-sequence quality scores (C) indicate consistently high sequencing quality. GC content distribution (D) shows similar, unbiased nucleotide composition. Sequence length distribution (E) reveals uniform read lengths (~150 bp), and duplication analysis (F) confirms acceptable library complexity.

## 4.2 Mapping and Alignment Efficiency with Bowtie2

Following qualitytrimming, Bowtie2 was used tomatch the cleaned paired-end reads from the cancer, control, and chronic saliva samples to the human reference genome (hg38).Withalignmentpercentagestypicallyfallingbetween 1.6% and 2.5%, the total alignment efficiency was comparativelylowacrossall samplegroups, as shown in figure 5.Themajorityof alignmentrates for cancer patientsamples werebetween 2.0 and 2.3%; the lowest alignment rates, 1.9%, were found for SRR32675089 and SRR32675100. With rates ranging from 1.6% to 2.5%, control patient samples showed somewhat greater range; SRR32675080 had the highest rate at 2.5%, while SRR32675082 and SRR32675084 had the lowest. Similartrendswereobservedinchronicpatientsamples,with alignment efficiencies ranging from 1.8% to 2.4%. Figure 5 illustrates that SRR32675106 had the highest alignment at 2.4%, whereas SRR32675101 and SRR32675105 mapped the least successfully.

Reads were categorized using the Bowtie2 PE Alignment Scores, which were displayed as horizontal bar graphs in figure 5, according to how each pair of reads matched with the genome.Aconsiderablenumberofreadpairsdidnotalignto the reference, as evidenced by the fact that the biggest percentage of aligned reads across all samples fell into the "PE neither matealigned" category(red bars). "PE onemate aligned" or "PE one mate multimapped" (yellowand orange bars) represented a smaller but significant number of reads, indicating partial alignment or mapping to several genomic sites. Only a small percentage of the dataset obtained high- confidence, concordant alignments, as evidenced by the extremely small number of reads classified as "PE mapped discordantly uniquely" or "PE mapped uniquely."

Thesefindingspointtopossibleissuesthatcouldexplainthe low alignment rates, as discussed in figure 5, such as non-human contamination, a high microbial presence in saliva, or a lack ofrepresentation ofsalivaryRNAsequencesinthereferencegenome.Despite these drawbacks, the pipeline was designed to handle low- input Ribo-

nucleic acid sequencing (RNA-Seq) data, such as saliva, which is frequently found in non-invasive sampling, thus the alignedreads were stillenoughtomoveonwithfeature-countinganddifferential expression analysis, as shown in figure 5.

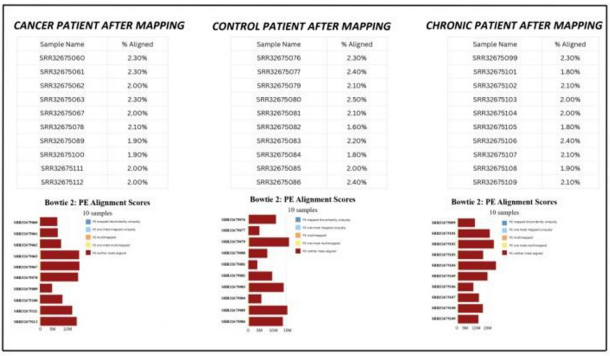


**Fig. 5.Bowtie2 Mapping Statistics Across Patient Groups**
Bowtie2 alignment summary of salivary Ribonucleic acid sequencing (RNA-Seq) data across cancer, control, and chronic patient groups. Alignment rates per sample ranged from 1.6% to 2.5%, reflecting the presence of microbial or non-human Ribonucleic acid (RNA). PE alignment score plots show most reads as "neither mate aligned," with smaller fractions uniquely, partially, or multimapped, providing an overview of mapping efficiency.

### 4.3 Feature Counts Assignment Across Patient Groups

The feature Counts assignment findings for three patient group-scancer,control,andchronic—each withtensamples are shown in the figure 6. The information displays how sequencing reads are distributed across many categories, includingassignedreads, unmappedreads,and other classes of unassigned reads (caused by ambiguity, chimaeras, poor mappingquality, or lackoffeatures).Thehighestvariability was seen in cancer samples, such as SRR32675089, which had a relatively low percentage of assigned reads and a sizablenumberthatfellintounassignedcategories,mostlyas a result of poor mapping quality or missing annotation features.

This raises the possibility of problems with biological heterogeneity, sequencingquality, or Ribonucleic acid (RNA) integrity. With a greater percentage of reads successfullyassigned to features and fewer reads lost to assignment, control samples showed more consistent alignment performance in contrast, suggesting better annotation and sequencing quality. With samples like SRR32675099, SRR32675102, and SRR32675104 reaching up to 20 million assigned reads and few unassigned categories, the chronic patient group demonstrated the best outcomes, demonstrating good sequencing quality, effective mapping, and a well- represented transcriptome. As illustrated in figure 6, these patternsimplythat chronic samples arethemost dependable for downstream analysis, but cancer samples might need more qualitycontrol or other processing techniques because of their lower assignment efficiency.

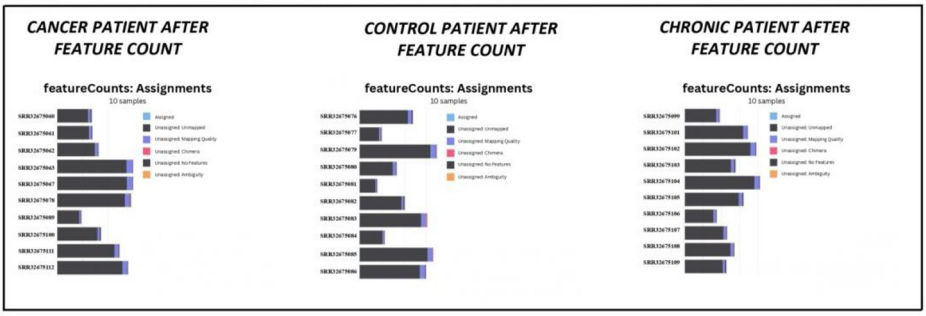


**Fig. 6.Comparison of Read Assignment in Cancer, Control, and Chronic Samples**
Comparison of assigned read counts across cancer, control, and chronic patient groups. Chronic samples show the highest proportion of assigned reads, indicating effective gene annotation. Control samples display intermediate assignment levels, while cancer samples exhibit greater variability and lower read assignment.

## 4.4 Analysis and interpretation of Biomarkers in Cancer, Control, and Chronic Groups

Finding and comparing gene expression biomarkers for cancer, control, and chronic illnesses was the goal of this work.GeneIDswiththegreatestexpressionandthosethat wereexpressedonlyineachconditionwerehighlightedafter expression levels for each group were examined.

With a score of9, the gene **ENSG00000241743.5(X Active Particular Transcript)** displayed the greatest level of expression in the cancer group, suggesting a strong correlationwithpathwaysparticulartocancer (see figure 7).Thistranscript is known to play a role in the control of X-chromosome genes, and through epigenetic dysregulation, its abnormal expression mayaidin the development of tumor's. Another distinct gene that was only discovered in the cancer group was **ENSG000004399.13 (Plectin D1)**,a structural protein thatisessentialfor preservingcytoskeletalintegrityandmay be involved in the invasion and motility of cancer cells.

Inthecontrolgroup,expressionlevelsremainedrathersteady across all genes. With a score of 6, the gene **ENSG00000281344.1(HELPAssociated Long Non-Coding RNA)**displayedthehighestexpression,indicatingapossible function in preserving regular cellular homeostasis. As shown in figure7, **ENSG00000269281.2(KCNQ1Opposite Strand Transcript 1)**, a unique gene in the control group, is engaged in ion transport regulation and may be crucial for non-pathological cellular function.

With a score of 6, **ENSG00000241743.5 (X Active Specific Transcript)** once again had the highest expression in the groupwithchronicconditions,confirmingitsinvolvementin both cancer and chronic pathological states. **ENSG00000168418.5 (Lysine Methyltransferase 2D),** a chromatin modifier involved in transcription control, was a gene that was only found in the chronic group. This finding suggests that epigenetic mechanisms are responsible for the advancement of chronic disease.

### 4.4.1 CommonBiomarkerAcrossAllConditions

As shown in Figure 7, **ENSG00000241743.5 (XACT)** is consistently expressed in cancer, control, and chronic disease samples. XACT is a long non-coding RNA that regulates X-chromosome gene expression by coating the **active X chromosome** and counteracting **XIST-mediated X-chromosome inactivation**. This role is crucial during early embryonic development, where XACT maintains proper gene dosage and supports pluripotency. Aberrant or overexpression of XACT has been reported in female-biased cancers, potentially causing abnormal X-linked gene dosage and promoting tumorigenesis. Due to its regulatory and scaffolding functions in chromatin modulation, XACT represents a potential **biomarker and therapeutic target** in cancer and developmental disorders.


**Fig. 7.Expression Analysis of Key Biomarker Genes in Saliva Samples**

Gene expression profiles in saliva samples from cancer, control, and chronic patients. Line graphs show relative expression patterns of selected genes, with elevated levels in cancer samples (e.g., A-Kinase Specific Transcript 1, long intergenic non-coding RNAs) and sharp peaks in chronic samples (e.g., Phospholipase C Beta 1, Myeloid-Associated Differentiation Marker). Tables provide gene IDs, counts, and descriptions for biological context.

## 5 Discussion

Potential universal salivary biomarkers across cancer, chronic, and control groups were identified using the Galaxy platform integrating transcriptomics and bioinformatics pipelines [9][10][11]. **ENSG00000241743.5 (X Active Specific Transcript, XACT)** showed consistently high expression across all conditions, with notable increases in cancer and chronic disease samples, supporting its role as a primary diagnostic biomarker [1][2][3][4]. As a long non-coding RNA regulating X chromosome

activity via interaction with XIST [13][14], XACT is central to epigenetic control and disease progression. Its dysregulation in tumors may alter developmental programs, contributing to malignant growth [2][14][15].

BesidesXACT,condition-specificbiomarkerswerefoundin cancer samples, such as **ENSG000004399.13 (Plectin D1)**, whichislinkedtocytoskeletalremodelingandcellular invasiveness, two important aspects of tumor growth [4][12][15]. A lncRNA associated with stable gene regulation, **ENSG00000269281.2 (KCNQ1OT1)**, was specifically expressed in the control group, indicating its connection to cellular homeostasis [1][5]. A gene involved in chromatin modification and epigenetic memory, **ENSG00000168418.5 (Lysine Methyltransferase 2D)** stood out in the chronic disease group. This gene indicates sustainedregulatoryalterationsthatare frequentlypresent in chronic inflammatory or degenerative conditions [2][14]. Saliva's molecular richness and diagnostic potential beyond oral pathology are highlighted by the variety of these indicators [5][6][7].

Despite the encouraging results, some technical issues surfaced, such as low alignment rates (~1.6–2.5%) when mappinghumangenomereadstosalivary-reads.Thefactthat salivary Ribonucleic acid (RNA) is diverse and contains both microbial and degraded host transcripts is probably the cause of this [6][9][11].However,enoughgene-levelmappingwasaccomplished to allow for reliable investigation of differential expression [9][10].Additionally,samplesfromchronicpatientshadthe highest assignment efficiency, indicating that these datasets had higher Ribonucleic acid (RNA) integrity and representativeness.

All of the findings support the viability of employing saliva as a transcript-rich, non-invasive bio specimen for gene expression profiling [5][6][7][8]. The discovered biomarkers serve as a basis for the creation of quick, point-of-care diagnostic instruments in addition to providing molecular signatures of disease states [4][7]. These results provide a new avenuefor precision diagnostics, particularlyin theuse of straightforward and scalable sample techniques for the early identification and tracking of systemic disorders [3][4][8][23].

## 6 Future Scopes

The recent findings pave the way for a number of future studies,withaparticularemphasisonestablishingdiagnostic sensitivityandspecificitythroughexperimentalvalidationof biomarkers such as XACT in larger, diverse cohorts. By providing accurate transcript quantification in saliva, integrating these biomarkers with qRT-PCR or digital PCR systems may hasten clinical translation. Furthermore, integratingmulti-omicstechniques—combining transcriptomicwithmetabolomics,salivaryproteomics,and microbiome profiling—could improve the precision and depth of disease categorization, especially when it comes to distinguishingbetweenearly-stagemalignanciesandchronic inflammation. Using the non-invasive and repeatable nature ofsalivasample,longitudinalmonitoringofthesebiomarkers throughouttherapymayprovideinsightfulinformationabout therapeutic response, remission, and recurrence. Additionally, by applying machine learning models to transcriptome-data,itispossibletostratifypatientsaccording to the risk or development of their disease

and find new biomarker signatures. The creation of point-of-care salivary diagnostic kits that are affordable, minimally intrusive, and appropriatefor generaluse—includingin environmentswith low resources—may ultimately result from these developments. All of this research points to a revolutionary change in customized healthcare and precision diagnostics.

## References


1. Strimbu K, Tavel JA (2010) What are biomarkers? *Current Opinion in HIV and AIDS* 5(6):463–466. https://doi.org/10.1097/COH.0b013e32833ed177
2. Henry NL, Hayes DF (2012) Cancer biomarkers. *Mol Oncol* 6(2):140–146. https://doi.org/10.1016/j.molonc.2012.01.010
3. Califf RM (2018) Biomarker definitions and their applications. Exp Biol Med (Maywood) 243(3):213–221. https://doi.org/10.1177/1535370217750088
4. Das S, Dey MK, Devireddy R, Gartia MR (2023) Biomarkers in cancer detection, diagnosis, and prognosis. Sensors (Basel) 24(1):37. https://doi.org/10.3390/s24010037
5. Shah FD, Begum R, Vajaria BN, Patel KR, Patel JB, Shukla SN, Patel PS (2011) A review on salivary genomics and proteomics biomarkers in oral cancer. Indian J Clin Biochem 26(4):326–334. https://doi.org/10.1007/s12291-011-0149-8
6. Chu HW, Chang KP, Hsu CW, Chang IY, Liu HP, Chen YT, Wu CC (2019) Identification of salivary biomarkers for oral cancer detection with untargeted and targeted quantitative proteomics approaches. Mol Cell Proteomics 18(9):1796–1806. https://doi.org/10.1074/mcp.RA119.001530
7. Parsanathan R, Byju R, Prabakaran DS (2024) Exploring salivary gene expression clusters: A bioinformatics approach for advanced diagnosis and prognosis in head and neck squamous cell carcinoma. Oral Oncol Rep 10:100300. https://doi.org/10.1016/j.oor.2024.100300
8. Uguz A, Muftuoglu C, Mert U, Gumus T, Ece D, Asadi M, Ulusan Bagci O, Caner A (2025) Unveiling microbiota profiles in saliva and pancreatic tissues of patients with pancreatic cancer. Microorganisms 13(1):119. https://doi.org/10.3390/microorganisms13010119
9. Batut B, van den Beek M, Doyle MA, Soranzo N (2021) RNA-Seq Data Analysis in Galaxy. Methods Mol Biol 2284:367–392. https://doi.org/10.1007/978-1-0716-1307-8_20
10. Koch CM, Chiu SF, Akbarpour M, Bharat A, Ridge KM, Bartom ET, Winter DR (2018) A beginner's guide to analysis of RNA sequencing data. Am J Respir Cell Mol Biol 59(2):145–157. https://doi.org/10.1165/rcmb.2017-0430TR
11. Aditama R, Achmad Tanjung Z, Made Sudania W, Liwang T (2017) SMART-RDA: A Galaxy Workflow for RNA-Seq Data Analysis. KnE Life Sciences 3(4):186–193. https://doi.org/10.18502/kls.v3i4.703
12. Zhang YH, Huang T, Chen L, Xu Y, Hu Y, Hu LD, Cai Y, Kong X. Identifying and analyzing different cancer subtypes using RNA-seq data of blood platelets. Oncotarget. 2017 Sep 15;8(50):87494-87511. doi: 10.18632/oncotarget.20903. PMID: 29152097; PMCID: PMC5675649.
13. Ergin S, Kherad N, Alagoz M (2022) RNA sequencing and its applications in cancer and rare diseases. *Molecular Biology Reports* 49(3):2325–2333. https://doi.org/10.1007/s11033-021-06963-0
14. Xue J-M, Liu Y, Wan L-H, Zhu Y-X (2020) Comprehensive analysis of differential gene expression to identify common gene signatures in multiple cancers. Med Sci Monit 26:e919953. https://doi.org/10.12659/MSM.919953

15. Peng L, Bian X-W, Li D-K, Xu C, Wang G-M, Xia Q-Y, Dai Z, Li J, Liu F, Zhao C-H, Li H-Y, Zhang Y, Jiang W, Zhang H-Y, Chen Z-H, Li N, Lai M-D (2015) Large-scale RNA-Seq transcriptome analysis of 4043 cancers and 548 normal tissue controls across 12 TCGA cancer types. Sci Rep 5:13413. https://doi.org/10.1038/srep13413
16. Pushalkar S, Hundeyin M, Daley D, Zambirinis CP, Kurz E, Mishra A, Mohan N, Aykut B, Usyk M, Torres LE, Werba G, Zhang K, Guo Y, Li Q, Akkad N, Lall S, Wadowski B, Gutierrez J, Kochen Rossi JA, Herzog JW, Diskin B, Torres-Hernandez A, Leinwand J, Wang W, Taunk PS, Savadkar S, Janal M, Saxena A, Li X, Cohen D, Sartor RB, Saxena D, Miller G (2018) The pancreatic cancer microbiome promotes oncogenesis by induction of innate and adaptive immune suppression. *Cancer Discov* 8(4):403–416. https://doi.org/10.1158/2159-8290.CD-17-1134
17. Nalluri H, Jensen E, Staley C (2021) Role of biliary stent and neoadjuvant chemotherapy in the pancreatic tumor microbiome. *BMC Microbiology* 21:280. https://doi.org/10.1186/s12866-021-02339-3
18. Guan SW, Lin Q, Yu HB (2023) Intratumour microbiome of pancreatic cancer. *World Journal of Gastrointestinal Oncology* 15(5):713–730. https://doi.org/10.4251/wjgo.v15.i5.713
19. Fu X, Shama A, Norbäck D, Chen Q, Xia Y, Zhang X, Sun Y (2024) Exploring the role of indoor microbiome and environmental characteristics in rhinitis symptoms among university students. *Front Microbiomes* 3:1277177. https://doi.org/10.3389/frmbi.2024.1277177
20. Olson SH, Satagopan J, Xu Y, Ling L, Leong S, Orlow I, Saldia A, Li P, Nunes P, Madonia V, Allen PJ, O'Reilly E, Pamer E, Kurtz RC (2017) The oral microbiota in patients with pancreatic cancer, patients with IPMNs, and controls: a pilot study. *Cancer Causes Control* 28(9):959–969. https://doi.org/10.1007/s10552-017-0933-8
21. Okuda S, Hirose Y, Takihara H, Okuda A, Ling Y, Tajima Y, Shimada Y, Ichikawa H, Takizawa K, Sakata J, Wakai T (2022) Unveiling microbiome profiles in human inner body fluids and tumor tissues with pancreatic or biliary tract cancer. *Sci Rep* 12:8766. https://doi.org/10.1038/s41598-022-12658-8
22. Vogtmann E, Han Y, Caporaso JG, Bokulich N, Mohamadkhani A, Moayyedkazemi A, Hua X, Kamangar F, Wan Y, Suman S, Zhu B, Hutchinson A, Dagnall C, Jones K, Hicks B, Shi J, Malekzadeh R, Abnet CC, Pourshams A (2020) Oral microbial community composition is associated with pancreatic cancer: a case–control study in Iran. *Cancer Med* 9(2):797–806. https://doi.org/10.1002/cam4.2660
23. Herremans KM, Riner AN, Cameron ME, McKinley KL, Triplett EW, Hughes SJ, Trevino JG (2022) The oral microbiome, pancreatic cancer and human diversity in the age of precision medicine. *Microbiome* 10:93. https://doi.org/10.1186/s40168-022-01262-7