Brazilian Dental Science

unesp
UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Instituto de Ciência e Tecnologia
Campus de São José dos Campos

ORIGINAL ARTICLE

DOI: https://doi.org/10.4322/bds.2026.e4804

Stochastic gene expression modeling using the Gillespie algorithm: insights into oral cancer genomic variability

Modelagem estocástica da expressão gênica usando o algoritmo de gillespie: percepções sobre a variabilidade genômica no câncer bucal

Pradeep Kumar YADALAM[1], **Carlos Martin ARDILA**[1,2]

1 - Saveetha University, Saveetha Institute of Medical and Technical Sciences, Saveetha Dental College and Hospital, Department of Periodontics, Chennai, Tamil Nadu, India
2 - Universidad de Antioquia U de A, Faculty of Dentistry, Basic Sciences Department, Biomedical Stomatology Research Group, Medellín, Antioquia, Colombia

How to cite: Yadalam PK, Ardila CM. Stochastic gene expression modeling using the Gillespie algorithm: insights into oral cancer genomic variability. Braz. Dent. Sci. 2025;29:e4804. https://doi.org/10.4322/bds.2026.e4804

ABSTRACT

Objective: Gene expression is a complex and dynamic process influenced by various factors, particularly in diseases like oral cancer. This study applies the Gillespie algorithm to oral cancer genomic data, emphasizing its utility in exploring gene regulation and variability in tumorigenesis. **Material and Methods:** The study analyzed gene expression profiles from the NCBI GEO dataset (GSE30784), which includes data from 167 oral squamous cell carcinomas, 17 dysplasia cases, and 45 normal oral tissues. The Gillespie algorithm was employed to simulate stochastic processes governing gene expression, focusing on transcription and degradation reactions. The method involves initializing systems, calculating reaction propensities, and generating time-series data to model the time evolution of gene expression systems. **Results:** The analysis revealed key insights into transcriptional dynamics, highlighting variability in transcription rates and degradation rates. The study observed a theoretical mean expression level of 5.0 compared to an ensemble mean of 4.7515, indicating stochastic fluctuations. The ensemble Coefficient of Variation (CV) of 0.4125 quantified variability, while the high autocorrelation value (0.8339) indicated that gene expression is significantly influenced by preceding states. These findings provide a normalized measure of gene expression variability and underscore the influence of stochastic processes on cellular systems. **Conclusion:** The Gillespie algorithm effectively models the stochastic nature of gene expression, uncovering intrinsic noise and variability in oral cancer. By demonstrating the role of transcriptional stochasticity in cellular heterogeneity, this study provides a robust framework for investigating gene regulation in disease contexts, such as cancer progression and drug resistance.

KEYWORDS

Computational biology; Gene expression regulation; Mathematical models; Mouth neoplasms; Stochastic processes.

RESUMO

Objetivo: A expressão gênica é um processo complexo e dinâmico influenciado por múltiplos fatores, particularmente em doenças como o câncer bucal (CB). Este estudo aplica o algoritmo de Gillespie a dados genômicos de CB, enfatizando sua utilidade na exploração da regulação gênica e variabilidade na tumorigênese. **Material e Métodos:** Foram analisados perfis de expressão gênica do conjunto de dados NCBI GEO (GSE30784), que inclui dados de 167 carcinomas de células escamosas orais, 17 casos de displasia e 45 tecidos orais normais. O algoritmo de Gillespie foi empregado para simular processos estocásticos que governam a expressão gênica, focando em reações de transcrição e degradação. O método envolve a inicialização dos sistemas, cálculo das propensões das reações e geração de séries temporais para modelar a evolução temporal dos sistemas de expressão gênica. **Resultados:** A análise revelou percepções importantes sobre a dinâmica transcricional, destacando a

variabilidade nas taxas de transcrição e de degradação. O estudo observou um nível médio teórico de expressão de 5,0, comparado a uma média de conjunto de 4,7515, indicando flutuações estocásticas. O Coeficiente de Variação do conjunto, de 0,4125, quantificou a variabilidade, enquanto o alto valor de autocorrelação (0,8339) indicou que a expressão gênica é significativamente influenciada pelos estados precedentes. Esses achados fornecem uma medida normalizada da variabilidade da expressão gênica e ressaltam a influência de processos estocásticos nos sistemas celulares. **Conclusão:** O algoritmo de Gillespie modela efetivamente a natureza estocástica da expressão gênica, revelando o ruído intrínseco e a variabilidade no CB. Ao demonstrar o papel da estocasticidade transcricional na heterogeneidade celular, este estudo oferece uma estrutura robusta para investigar regulação gênica em contextos de doença, como a progressão do câncer e resistência a fármacos.

PALAVRAS-CHAVE

Biologia computacional; regulação da expressão gênica; Modelos teóricos; Neoplasias bucais; Processos estocásticos.

INTRODUCTION

Gene expression is a highly dynamic and stochastic process influenced by various internal and external factors [1,2]. Understanding the intricacies of gene expression is crucial, especially in the context of diseases such as oral cancer, where dysregulation can lead to malignancy and poor patient outcomes. Oral cancer continues to be a significant global health challenge, with rising incidence rates and complex molecular underpinnings. Recent advancements in high-throughput sequencing technologies have facilitated the generation of extensive genomic data, revealing intricate gene regulatory networks and signaling pathways involved in tumorigenesis. These insights underscore the urgent need for computational approaches that can capture the inherent variability in gene expression within cancer cells. By integrating stochastic modeling approaches, researchers can unravel the complexities of gene expression variability observed in oral cancer cells [3], shedding light on the mechanisms contributing to tumor heterogeneity and treatment resistance.

The integration of omics data has revolutionized our understanding of biological systems, and genome-scale metabolic modeling (GSMM) [4,5] has been instrumental in analyzing this data. Genome-scale metabolic models (GEMs) provide a robust framework for integrating multiple omics datasets, enabling a comprehensive understanding of cellular metabolism and its underlying mechanisms. GEMs serve as a valuable tool for predicting metabolic capabilities and identifying key regulatory nodes in biological systems. This approach has proven especially transformative in fields like oncology, where GEMs are now being leveraged to understand metabolic reprogramming in tumor cells and identify novel therapeutic targets. For example, GEMs have been successfully applied to catalog human gut microbes, marking a significant achievement in the field of systems biology.

Stochastic gene expression modeling is a method that captures the variability and noise in biological systems, which can have profound implications for cellular behavior, such as differentiation, adaptation to environmental changes, and response to therapies, particularly in diseases like cancer [6]. Factors contributing to the stochastic nature of gene expression include transcriptional noise, mRNA degradation, translation variability, and the cell microenvironment. The Gillespie algorithm is a widely used stochastic simulation method that models how chemical reactions occur in a system, including gene expression. It operates on discrete events, reaction rates, simulation steps, and an iterative process, allowing researchers to generate trajectories representing changes in mRNA and protein levels over time. Stochastic gene expression modeling applications include cancer heterogeneity, drug resistance, cellular differentiation, and synthetic biology [7]. By leveraging stochastic models, researchers can better characterize tumor microenvironments and explore cell fate decisions critical to cancer progression. Synthetic biology can also be applied to design synthetic gene circuits that exploit stochastic effects to create new functions. However, challenges such as data integration, parameter estimation, translation to clinical settings, and computational complexity must be addressed. By focusing on the stochastic dynamics underpinning gene expression, researchers can enhance our understanding of the intricate interplay between genetics and the environment, ultimately paving the way for more effective, personalized medical interventions [5].

Traditional deterministic models often fail to capture gene expression's inherent randomness and cellular noise, necessitating stochastic modeling approaches. Among these, the Gillespie algorithm is a powerful tool that simulates the probabilistic nature of biochemical reactions within cells, providing insights into gene expression dynamics at a single-cell level. Gene expression, a complex biological process, involves the transcription of genetic information from DNA to RNA, leading to protein synthesis. Dysregulation of gene expression is crucial in diseases like oral cancer, necessitating sophisticated modeling techniques to understand its mechanistic underpinnings. In contrast to deterministic models that assume a fixed set of parameters and outcomes, stochastic models account for randomness and variability in biological processes. Gene expression is influenced by genetic and epigenetic factors and random fluctuations in the cellular environment, leading to differences in gene expression levels among individual cells within the same tissue. This variability highlights the need for advanced simulation techniques to uncover critical regulatory mechanisms driving disease. The Gillespie algorithm allows for the simulation of discrete events in gene expression, enabling researchers to examine the probabilistic interactions between genes, regulatory elements, and environmental factors. This approach helps understand the distribution and variability of gene expression levels and offers a framework for predicting the behavior of cellular systems under different conditions. By applying the Gillespie algorithm [8,9] to genomic data from oral cancer patients, we aim to uncover novel insights into the stochastic processes driving tumor development and progression, ultimately aiding in developing targeted therapeutic strategies.

The Gillespie algorithm is a computational method used to simulate the time evolution of chemical reactions in a stochastic fashion. It tracks molecular species interactions and can be used to model gene expression, including transcription and translation, allowing researchers to examine variability and distribution patterns in experimental data. Its capacity to handle single-cell resolution data makes it especially valuable in studying heterogeneity within tumor populations. The Gillespie algorithm can study gene expression dynamics in oral cancer, enabling researchers to model tumor heterogeneity, investigate gene regulation networks, simulate treatment response, and guide personalized medicine [10,11]. This approach helps identify subpopulations with distinct gene expression patterns, predict resistance mechanisms, and identify biomarkers for treatment efficacy, contributing to improved patient outcomes in oral cancer. The study uses the Gillespie algorithm to analyze oral cancer genomic data, highlighting its role in understanding gene regulation.

MATERIAL AND METHODS

Gene expression data

Using the NCBI GEO dataset [12], specifically GSE30784, gene expression was studied to distinguish oral squamous cell carcinomas (OSCC) from controls. RNA from 167 OSCC, 17 dysplasia, and 45 normal oral tissues was analyzed for omics.

Overview of the Gillespie algorithm

The Gillespie algorithm is a stochastic simulation technique used to model the time evolution of chemical systems, such as gene expression [13,14], where reactions occur randomly. It is particularly suited for systems with small numbers of molecules, where stochastic effects dominate. The algorithm involves initializing a system with initial mRNA molecule numbers, transcription, and degradation rates. It calculates propensities for each reaction, calculates time steps, and selects the reaction based on cumulative propensities. The system is assumed to be well-mixed, and only transcription and degradation reactions are considered. The algorithm generates a time series of mRNA molecule counts, capturing the stochastic dynamics of gene expression. The output can be used to calculate statistical properties like mean, variance, and autocorrelation. Applications include modeling intrinsic noise in gene expression and simulating single-cell gene expression trajectories.

The Gillespie algorithm, also known as the Stochastic Simulation Algorithm (SSA), is a mathematical method for simulating the time evolution of systems described by discrete stochastic processes. It is commonly applied in fields such as biochemistry, genetics, and ecology to model the behavior of interacting particles, such as molecules reacting within a cell. The algorithm captures the random nature of molecular interactions and is particularly useful for systems with small numbers of particles, where significant statistical fluctuations occur.

The key steps involved in executing the Gillespie algorithm include defining the system, initializing variables, calculating propensities, computing total propensity, sampling two random numbers from a uniform distribution, determining the time until the next reaction, updating the current time, selecting the next reaction, updating molecule counts, and repeating the process until the desired simulation time is reached or a predefined number of reactions have been simulated.

By incorporating these steps iteratively, researchers can simulate stochastic processes that are otherwise inaccessible through deterministic modeling, enabling a deeper understanding of biological interactions. The Gillespie algorithm is a powerful tool for simulating stochastic processes in biochemical systems, as it accounts for intrinsic randomness, making it suitable for applications where deterministic models fall short.

Analysis framework

Temporal Statistics Analysis: The study aims to analyze the change in mRNA levels and noise in gene expression over time using running averages, coefficients of variation, and line plots. It also compares theoretical mean and noise levels to observed data.

Burst and Distribution Analysis: This analysis aims to characterize stochastic bursts in gene expression and validate the Poisson nature of mRNA distributions. Burst sizes, waiting times, and various statistical visualizations are used to explore their properties and variability.

State Transitions Analysis: The study examines temporal correlation and mRNA-level transitions using scatter and density plots and kernel density estimation to identify significant changes over time.

Stochastic Gene Expression Analysis: This analysis provides a detailed understanding of stochastic gene expression dynamics, using plots of single-cell and multi-cell trajectories, histograms, and box plots to analyze mRNA levels and noise distributions.

Gene expression analysis

Gene expression analysis involves several steps, including temporal statistics analysis, burst and distribution analysis, state transitions analysis, and stochastic gene expression analysis. Temporal statistics analysis helps track dynamic expression patterns by analyzing mRNA levels. This approach smooths out short-term fluctuations in data, quantifies variability relative to the mean, and uses line plots to identify trends. Burst and distribution analysis focuses on the discrete nature of gene expression, recognizing that mRNA production may occur in bursts. Understanding burst characteristics enhances insights into transcriptional control mechanisms.

State transition analysis is critical for elucidating gene expression dynamics and identifying regulatory switches or states in gene behavior. It uses temporal correlation and transitions, scatter plots and density plots, and kernel density estimation to identify underlying patterns in mRNA level changes. Stochastic gene expression analysis integrates multiple data to provide a cohesive view of gene expression variability. Single-cell and multi-cell trajectory plots are used to illustrate variability at the individual cell level and within a population. These analyses help uncover the complexities of stochastic processes, such as feedback, interactions, and regulatory networks that govern gene expression.

All analyses were performed using R (v4.3.1) and Python (v3.10.8). Differential expression analysis was conducted with DESeq2 (v1.42.0) and edgeR (v3.44.2), considering FDR-adjusted $p < 0.05$ as significant. Stochastic simulations were implemented in Python utilizing NumPy (v1.25.2) and SciPy (v1.11.3), following the Gillespie algorithm. Visualization was carried out with ggplot2 (v3.4.4), matplotlib (v3.7.2), and seaborn (v0.12.2).

RESULTS

The results of this study are presented under four main analyses: Temporal Statistics Analysis, Burst and Distribution Analysis, State Transitions Analysis, and Stochastic Gene Expression Analysis. Each analysis provides insight into the stochastic nature of gene expression and its variability, leveraging theoretical models and simulation data. These results collectively underline the importance of stochastic modeling in understanding biological processes and cellular behavior.

1. Temporal statistics analysis

The mRNA expression levels fluctuated around a theoretical mean of 5 molecules, with initial deviations attributed to stochastic bursts.

Noise levels decreased as the system approached equilibrium. The running averages revealed a gradual stabilization, emphasizing the transient nature of initial noise before reaching a steady state.

2. Burst and distribution analysis

The study found that bursts were small, indicating stochastic transcription. Waiting times between transcription events followed an exponential distribution, confirming the random nature of the process. Poisson validation further demonstrated that mRNA levels adhered to a Poisson distribution, though slight overdispersion was noted at higher quantiles. A fan factor value close to 1 for small time windows indicated behavior consistent with Poisson processes.

3. State transitions analysis

The scatter plot demonstrated a strong correlation between mRNA counts at different time points, with deviations attributable to noise. High-density regions in the plot confirmed temporal correlation, indicating that past expression levels influence current states. Kernel density estimates highlighted significant regions of state transitions, emphasizing regulatory dynamics.

4. Stochastic gene expression analysis

The study observed random fluctuations in mRNA levels, indicative of bursts in transcription and degradation. Significant variability was noted between individual cells, underscoring the inherently stochastic nature of gene expression. mRNA levels followed a Poisson distribution, with a peak near the mean. Moderate coefficients of variation (CV) values reflected a balance between noise and stability. These findings validate the stochastic models and highlight the dynamic equilibrium in cellular processes.

Key parameters

The detailed examination of key parameters provides a nuanced understanding of gene expression dynamics. Each parameter elucidates specific aspects of stochastic behavior, offering a quantitative perspective on transcription and degradation processes.

Transcription rate (k_transcription): 0.5

This indicates that, on average, 0.5 mRNA molecules are produced per minute. This rate is crucial for determining how quickly genes are expressed.

Degradation Rate (k_degradation = 0.1 min^{-1}). This value indicates that each mRNA molecule has a 10% chance of being degraded per minute. The balance between transcription and degradation rates influences overall mRNA levels in the cell.

Theoretical Mean: 5.0. This represents the expected average number of mRNA molecules predicted by the model under ideal conditions.

Ensemble Mean: 4.7515. The average number of mRNA molecules observed in the simulation, slightly lower than the theoretical mean, suggests some variation in expression.

Ensemble Standard Deviation: 0.5792 This statistic measures the variation or dispersion of mRNA levels around the ensemble mean. A higher standard deviation suggests greater variability in expression levels among different cells.

Ensemble CV (Coefficient of Variation): 0.1219. This value, calculated as the standard deviation divided by the mean, provides a normalized measure of variability. A CV of 0.12 indicates moderate noise in gene expression.

Mean CV Across Trajectories: 0.4125. This represents the average variability across multiple cell trajectories, indicating significant intrinsic noise in gene expression.

Mean Autocorrelation (Lag 1): 0.8339. This high autocorrelation value indicates that mRNA levels at one time point strongly correlate with those at the next time point, suggesting that past expression levels influence future levels.

Stochastic simulations used Python 3.10.8 and Gillespie's SSA for a two-state promoter: DNA_off ⇄ DNA_on with k_on = 0.05 min^{-1}, k_off = 0.10 min^{-1}; on, transcription at k_txn = 1.5 min^{-1} and mRNA degrades at k_deg = 0.10 min^{-1}. Values matched reported mean (~5 molecules), overdispersion, and an autocorrelation ≈ 0.83. Simulated for T = 200 min, discarding 40 min, sampling every 2 min for 100 points. Ran 2,000 trajectories, seed 42. Convergence when mean and Fano factor vary ≤2% over last 50%. Used NumPy/SciPy for stats; statsmodels for distribution checks.

The study reveals that gene expression is stochastic, with fluctuations around the mean expression level indicating a dynamic equilibrium. Multiple cell trajectories exhibited varying expression patterns, even under identical conditions. The observed mRNA levels followed

a Poisson distribution, accurately representing expected stochastic behavior. Additionally, strong autocorrelation at lag 1 suggests that gene expression levels exhibit memory effects, where past levels significantly influence future levels. The substantial intrinsic noise observed aligns with experimental data, confirming the inherently stochastic nature of gene expression processes.

The following figures provide visual support for the results described above. Each figure or group of figures represents specific aspects of the analysis, such as temporal dynamics, distribution characteristics, and variability across cells. These visualizations aim to clarify the inherent stochasticity and provide a comparative framework for understanding gene expression dynamics.

DISCUSSION

Oral cancer, particularly squamous cell carcinoma of the head and neck, is a growing global health concern. Genomic studies have identified various genetic alterations associated with oral cancer, including mutations, copy number variations (CNVs), and epigenetic changes. These genomic alterations contribute to the heterogeneity of tumors and influence their behavior, treatment responses, and patient outcomes [15]. Recent advancements in omics technologies have generated a wealth of biological data. Integrating these data within mathematical models is essential to fully leveraging their potential. Genome-scale metabolic models (GEMs) provide a robust framework for studying complex biological systems. GEMs have significantly contributed to our understanding of human metabolism, including the intrinsic relationship between the gut microbiome and the host metabolism. The integration of omics data with GEMs [9] can lead to new insights and advance our understanding of molecular mechanisms in human health and disease. Understanding the genomic landscape of oral cancer is crucial for developing stochastic models that can simulate tumor progression, treatment responses, and the emergence of resistance to therapies. BioModels can be applied to oral cancer research by simulating pathways, gene regulatory networks, and drug targets and integrating genomics data to create personalized models [16]. Retrieved genomic data GSE30784 [17] identified 131 differentially expressed probe sets and developed two models to distinguish oral squamous cell carcinoma (OSCC) from controls, demonstrating high sensitivity and specificity in validation sets. Hence, we used this dataset for stochastic analysis.

Key genomic alterations in OSCC [18,19] include mutations in the TP53 tumor suppressor gene, activating mutations in HRAS, and PIK3CA gene mutations [17]. Significant CNVs observed in OSCC include amplifications of oncogenes, deletions of tumor suppressor genes, epigenetic changes, and transcriptomic variability. Stochastic modeling of oral cancer involves simulating the dynamics of tumor evolution, treatment effects, and clonal diversity using the genomic alterations that have been identified. The Stochastic Model of Intra-tumor Heterogeneity (SMITH) [6] is an efficient cancer evolution model that combines branching and confinement mechanisms to limit clonal growth based on the size of individual clones and the overall tumor population, thereby enhancing the understanding of selection and mutation rates. One previous study [5] proposes a probabilistic model of gene expression in prokaryotes, assuming known network interactions. It introduces a method for estimating unknown parameters from sparse, irregularly sampled, and noisy time-course data, applicable to gene networks of arbitrary size. Additionally, another study showed an interactive workflow for model exploration of gene regulatory networks, utilizing semi-supervised learning and human-in-the-loop labeling of data. This approach reduces computational demands and enables the rapid discovery of interesting model behaviors.

Our stochastic simulations show transcription occurs in bursts with exponential waiting times, leading to Poisson-like but slightly over-dispersed mRNA distributions. The lag-1 autocorrelation ($\sim$0.83) indicates strong memory effects in gene expression. These results confirm that single-gene stochastic fluctuations can cause cell-to-cell variability, a key factor in oral cancer heterogeneity. This intrinsic noise may explain differences in drug response and resistance subclones. Stochastic models [7] are increasingly used in oral cancer research to understand tumor progression, therapeutic resistance, personalized medicine, and biomarker discovery. However, challenges like data quality, integration, parameter estimation, computational complexity, and experimental validation persist. By capturing tumor evolution variability, these models can inform therapeutic strategies,

improve personalized medicine outcomes, and enhance the management and treatment of cancer. Stochastic models inform therapy by quantifying heterogeneity and predicting subpopulation dynamics. Gene expression is a crucial aspect of cellular behavior and development, with stochastic fluctuations playing a significant role in cancer progression [20]. Oral cancer, a subset of head and neck cancers, presents unique challenges for effective treatment due to its heterogeneous nature driven by genetic and epigenetic alterations. Understanding the stochastic dynamics underlying gene expression in oral cancer is essential for unraveling tumor biology and improving therapeutic strategies [3,14,21]. The study analyzed mRNA expression levels, revealing a stochastic nature with fluctuations around the theoretical mean of 5 molecules per cell. The study found small bursts indicating stochastic transcription while waiting times followed an exponential distribution. The mRNA levels followed a Poisson distribution, peaking near the mean. The study also found a strong correlation between mRNA counts at different time points, confirming the temporal correlation. The study also found random fluctuations in mRNA levels, indicating bursts of transcription and degradation. The study highlighted the importance of temporal analysis and the validation of statistical models in understanding biological processes and cellular behavior. Key parameters included the transcription and degradation rates, the theoretical mean, the ensemble mean, the ensemble standard deviation, the ensemble CV, the mean CV across trajectories, and the mean autocorrelation. Incorporating additional analysis of temporal fluctuations and regulatory feedback mechanisms could further enhance the utility of these stochastic models. In oral squamous cell carcinoma, transcriptional noise relates to treatment resistance and clonal diversity evolution [22]. The study confirms the inherently stochastic nature of gene expression, as shown in Figures 1-4.

Understanding how stochastic processes influence treatment outcomes could inform more personalized therapeutic strategies. The cellular microenvironment, characterized by interactions between tumor cells, stromal cells, and extracellular matrix, can further modulate gene expression dynamics. Integrating high-throughput sequencing technologies with genomic data can refine stochastic models of gene expression, shedding light on pathways that underpin tumorigenesis and metastasis. However, challenges remain, such as the complexity of biological systems and incomplete data. Collaborative efforts among computational biologists, geneticists, oncologists, and therapeutic developers are crucial to bridge the gap between theoretical modeling and real-world applications.

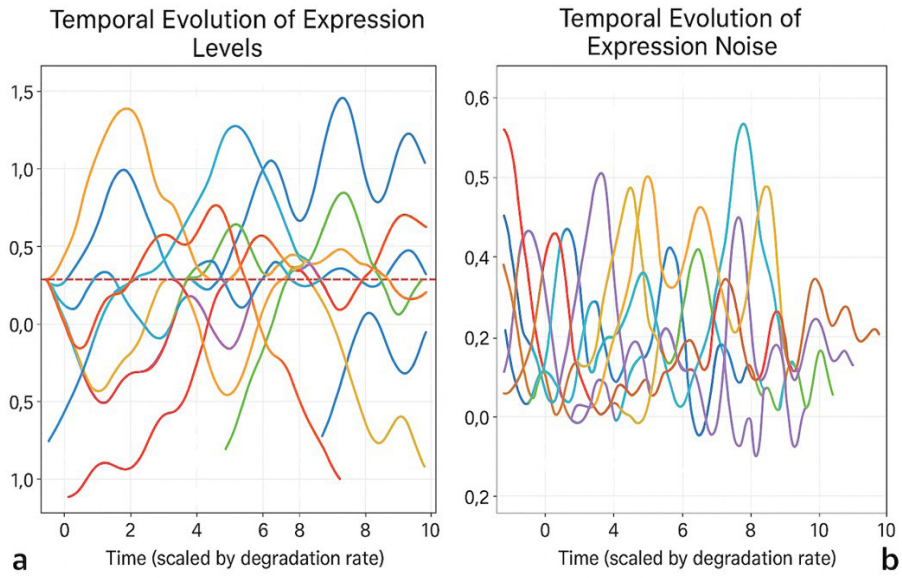

Figure 1 - a: Temporal evolution of mRNA levels in a single-cell system. The running average is plotted against time, with the red dashed line indicating the theoretical mean. Most trajectories converge near the mean as the system stabilizes; b: Temporal evolution of expression noise in a cell system. Variability decreases over time as equilibrium is approached.

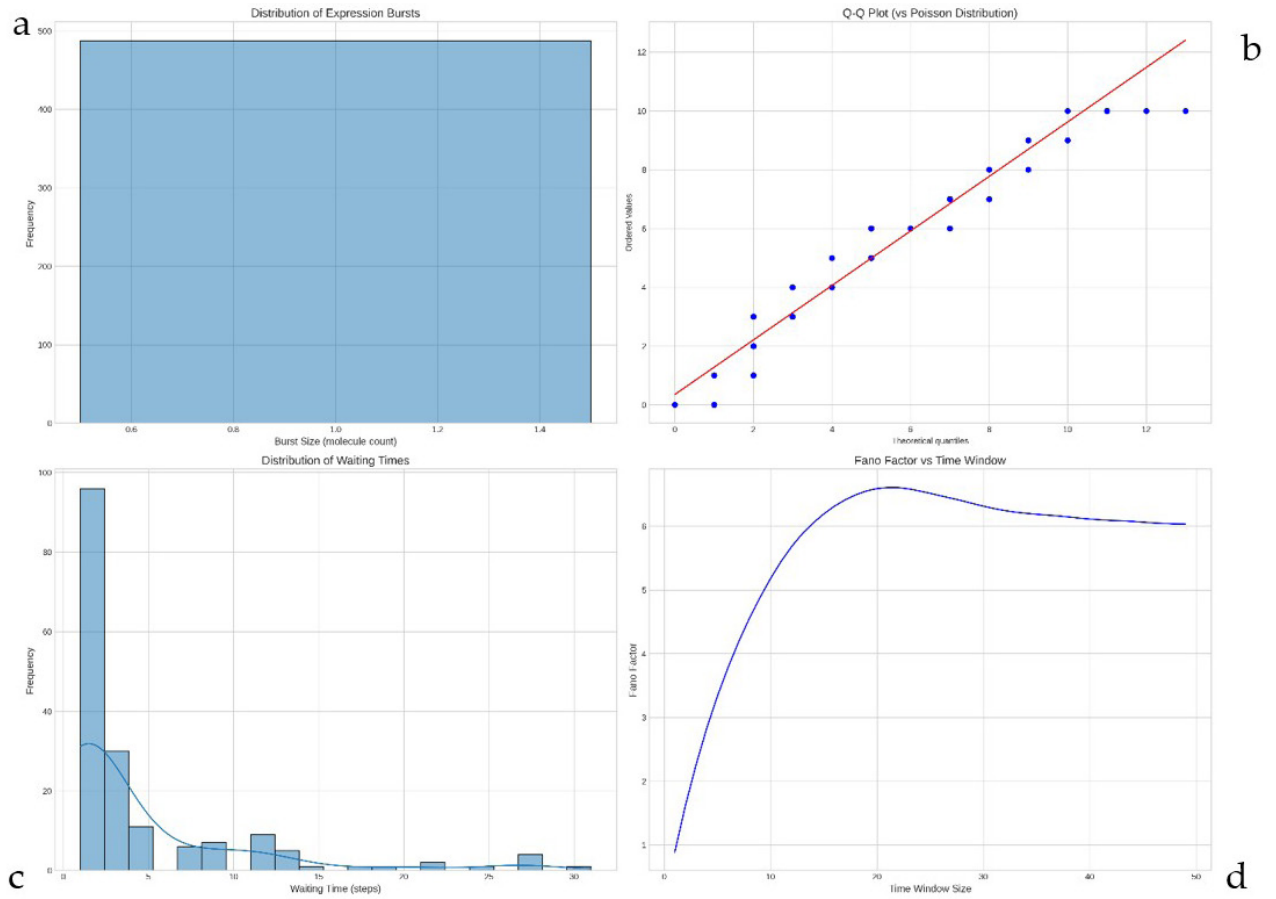

Figure 2 - Burst and Distribution Analysis. a: Histogram showing a uniform distribution of burst sizes. Small bursts predominate, reflecting stochastic transcription; b: Q-Q plot comparing observed data to the Poisson distribution. Theoretical quantiles align with observed values, with overdispersion noted at higher quantiles; c: Histogram of waiting times between transcriptional bursts. The exponential-like distribution suggests random transcription dynamics; d: Fano factor as a function of time window size. The Fano factor peaks and stabilizes over longer observation periods, confirming overdispersion in mRNA levels.

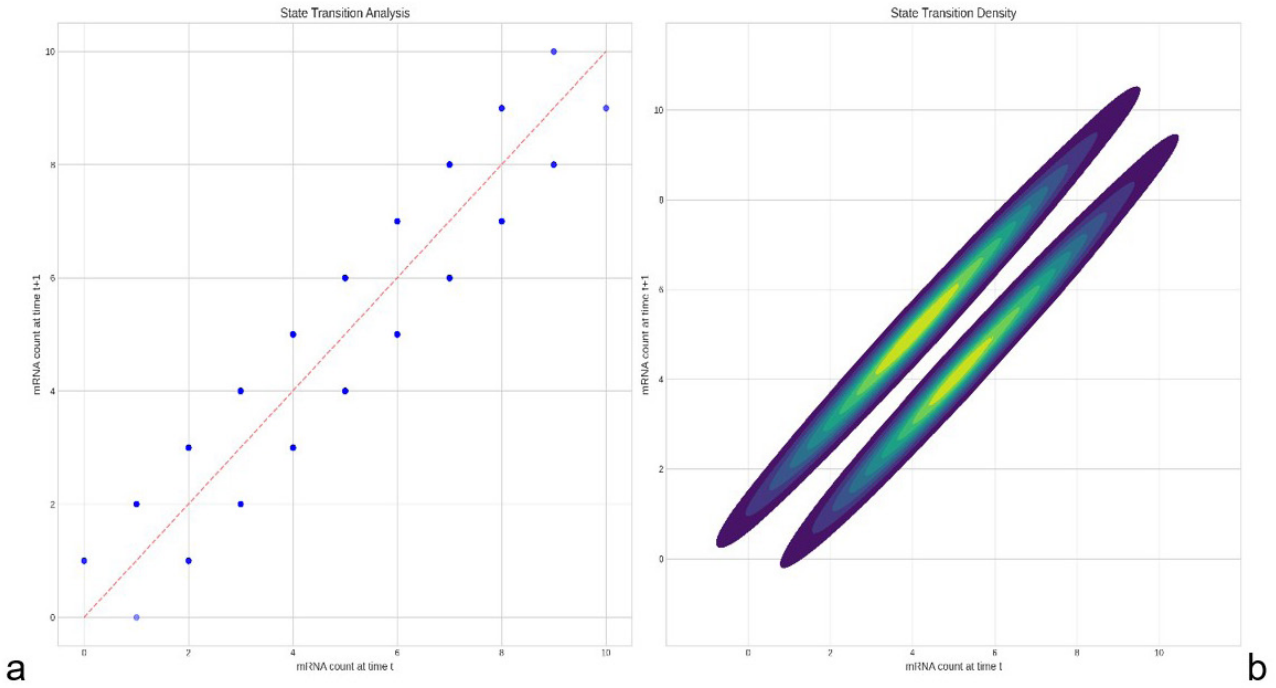

Figure 3 - State Transitions and Correlation Analysis. a: Scatter plot showing mRNA count correlation between consecutive time points. Variability increases with higher expression levels, and density is highest along the diagonal; b: Density plot of the same data, emphasizing regions of high concentration.

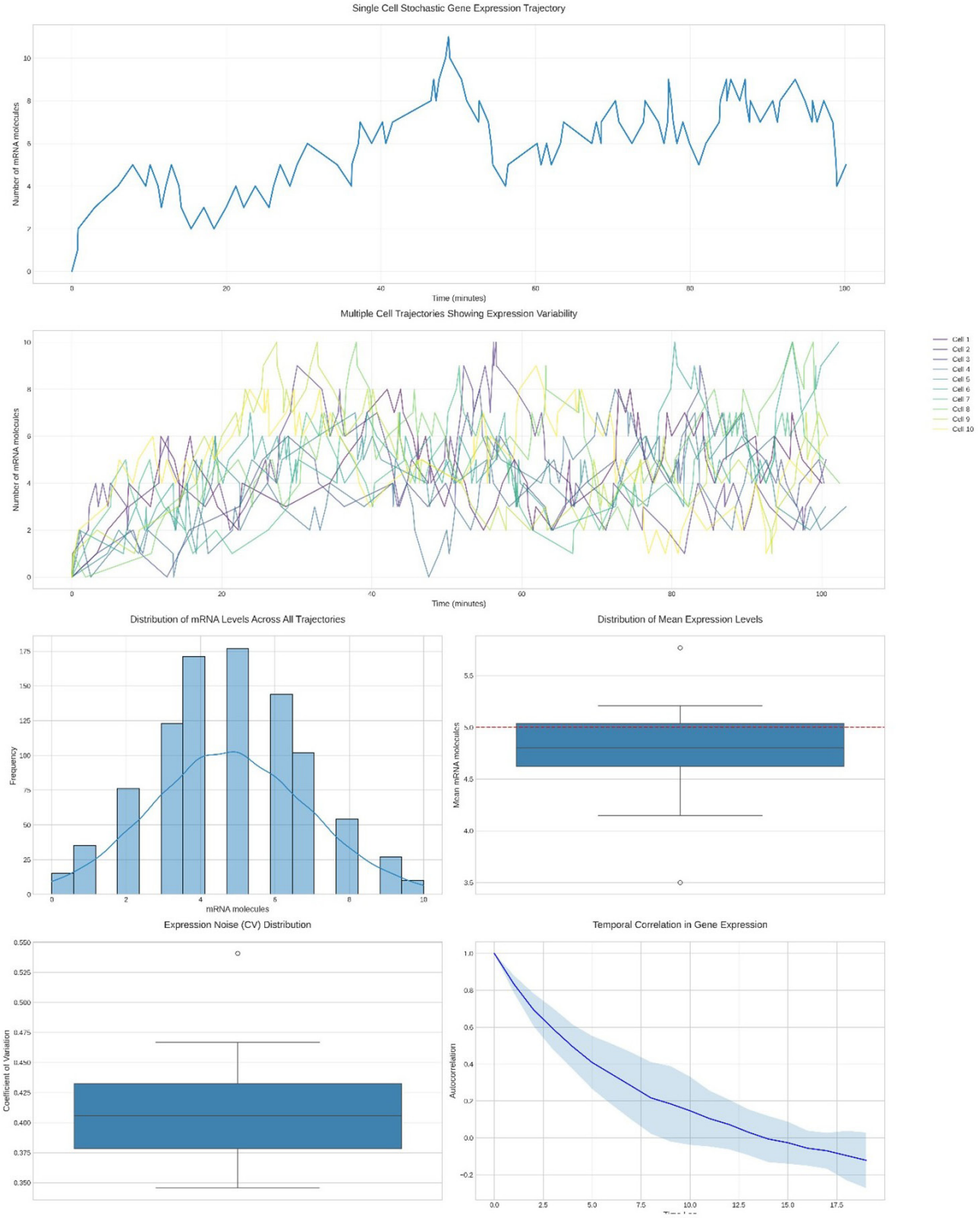

Figure 4 - Stochastic Gene Expression Variability. a: Single-cell stochastic gene expression trajectory. Irregular fluctuations are evident over approximately 100 time points, with mRNA molecule counts ranging from 0 to 10. b: Multiple cell trajectories showing expression variability. Each trajectory, represented by a distinct color, illustrates cell-to-cell differences. c: Frequency distribution of mRNA molecules across all trajectories. A fitted curve approximates a normal distribution, with an x-axis range of 0 to 10 mRNA molecules. d: Box plot of the coefficient of variation (CV) for gene expression, showing median, quartiles, and outliers. The accompanying plot displays the distribution of mean expression levels. e and f: Autocorrelation plots demonstrating a decrease in correlation over time. The confidence interval is shown as a shaded region, with time lag represented on the x-axis.

Additionally, the development of standardized protocols for integrating heterogeneous data types will enhance reproducibility and robustness in model predictions. This paper aims to explore the application of the Gillespie algorithm in modeling stochastic gene expression dynamics within oral cancer genomic data, highlighting the value of integrating computational modeling with genomic data analysis.

The Gillespie algorithm [13,14] was employed to model the stochastic nature of gene expression, with a focus on transcriptional bursts and degradation events. The study confirmed the model's accuracy by analyzing mRNA levels over time, which revealed random fluctuations around a theoretical mean of 5.0 molecules. The balance between stochasticity and regulatory mechanisms was also observed, with the ensemble mean of 4.7515 mRNA molecules closely matching the theoretical predictions. The analysis of transcriptional bursts and Poisson-like mRNA distributions confirmed the validity of the stochastic modeling approach. The study also found strong autocorrelation at short-time lags, indicating memory effects in gene expression. The significant variability in mRNA levels across cells underscored the role of intrinsic noise in generating phenotypic diversity. The balance between noise and regulation maintains cellular function while allowing phenotypic plasticity. Similar to this study, discrete stochastic models of gene regulatory networks are crucial for biological inquiry. An interactive workflow utilizing semi-supervised learning and human-in-the-loop labeling enables the rapid discovery of model behaviors, thereby reducing manual work and computational demands in model exploration. The study acknowledges simplified assumptions [15,21], neglects translation and post-transcriptional regulation, and focuses on a single gene. It also assumes constant rates, which may not fully capture the complexity of biological systems. Future directions include exploring transcriptional bursting variations across genes, cell types, and environmental conditions. The technical improvements include the development of more efficient simulation algorithms for larger-scale systems, the integration of spatial gene expression aspects, and the inclusion of multi-scale modeling approaches. Future research should investigate biological factors, such as regulatory feedback loops, transcription factor dynamics, and chromatin remodeling, to gain a deeper understanding of gene expression. Experimental validation and transcriptional bursting variations could enhance gene regulation, leading to applications in synthetic biology, biotechnology, and medicine [23,24]. While this study modeled a simple gene expression system; future work should explore multi-gene and pathway-level stochastic models. These could reveal how transcriptional noise spreads through oncogenic pathways, offering insights into tumorigenesis and metastasis.

CONCLUSION

The study employs the Gillespie algorithm to investigate the stochastic nature of gene expression, revealing intrinsic noise and variability resulting from transcriptional bursts and degradation events. Despite its limitations, the study provides a foundation for future research on the stochastic nature of gene regulation and its implications for cellular function and diversity, which are crucial for fields such as developmental biology and disease research. The study emphasizes the significance of stochastic processes in understanding cellular function and diversity. It has implications for developmental biology, synthetic biology, and disease research. The Gillespie algorithm provides a quantitative framework for studying gene expression variability, and future work will enhance our understanding of cellular regulation and phenotypic diversity. Expanding these insights to broader genomic and transcriptomic datasets could drive significant advances in personalized medicine and targeted therapies. The study highlights that stochasticity in gene expression is a fundamental feature of cellular systems, contributing to biological function and adaptation.

Acknowledgements

None

Author's Contributions

PKY, CMA: Conceptualization. PKY, CMA: Data Curation. PKY, CMA: Formal Analysis. PKY, CMA: Funding Acquisition. PKY, CMA: Investigation. PKY, CMA: Methodology. PKY, CMA: Project Administration. PKY, CMA: Resources. PKY, CMA: Software. PKY, CMA: Supervision. PKY, CMA: Validation. PKY, CMA: Visualization. PKY, CMA: Writing – Original Draft Preparation. PKY, CMA: Writing – Review & Editing.

Conflict of Interest

No conflicts of interest declared concerning the publication of this article.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Regulatory Statement

Not required

References

1. Whitmore SE, Lamont RJ. Oral bacteria and cancer. PLoS Pathog. 2014;10(3):e1003933. https://doi.org/10.1371/journal.ppat.1003933. PMid:24676390.
2. Balkwill F, Mantovani A. Inflammation and cancer: back to Virchow? Lancet. 2001;357(9255):539-45. https://doi.org/10.1016/S0140-6736(00)04046-0. PMid:11229684.
3. Pathoor N, Sankar Ganesh P. Harnessing natural quorum sensing inhibitors to prevent porphyromonas gingivalis-induced oral squamous cell carcinoma: an innovative approach. Galician Med J. 2024;31(4)
4. Wrede F, Hellander A. Smart computational exploration of stochastic gene regulatory network models using human-in-the-loop semi-supervised learning. Bioinformatics. 2019;35(24):5199-206. https://doi.org/10.1093/bioinformatics/btz420. PMid:31141124.
5. Sen P, Orešič M. Integrating omics data in genome-scale metabolic modeling: a methodological perspective for precision medicine. Metabolites. 2023;13(7):855. https://doi.org/10.3390/metabo13070855. PMid:37512562.
6. Streck A, Kaufmann TL, Schwarz RF. SMITH: spatially constrained stochastic model for simulation of intra-tumour heterogeneity. Bioinformatics. 2023;39(3):btad102. https://doi.org/10.1093/bioinformatics/btad102. PMid:36825830.
7. Wrede F, Hellander A. Smart computational exploration of stochastic gene regulatory network models using human-in-the-loop semi-supervised learning. Bioinformatics. 2019;35(24):5199-206. https://doi.org/10.1093/bioinformatics/btz420. PMid:31141124.
8. Lohfeld S, Barron V, McHugh PE. Biomodels of bone: a review. Ann Biomed Eng. 2005;33(10):1295-311. https://doi.org/10.1007/s10439-005-5873-x. PMid:16240079.
9. Cinquemani E, Milias-Argeitis A, Summers S, Lygeros J. Stochastic dynamics of genetic networks: modelling and parameter identification. Bioinformatics. 2008;24(23):2748-54. https://doi.org/10.1093/bioinformatics/btn527. PMid:18845579.
10. Li C, Donizelli M, Rodriguez N, Dharuri H, Endler L, Chelliah V, et al. BioModels Database: an enhanced, curated and annotated resource for published quantitative kinetic models. BMC Syst Biol. 2010;4(1):92. https://doi.org/10.1186/1752-0509-4-92. PMid:20587024.
11. Juty N, Ali R, Glont M, Keating S, Rodriguez N, Swat MJ, et al. BioModels: content, features, functionality, and use. CPT Pharmacometrics Syst Pharmacol. 2015;4(2):e3. https://doi.org/10.1002/psp4.3. PMid:26225232.
12. Barrett T, Troup DB, Wilhite SE, Ledoux P, Evangelista C, Kim IF, et al. NCBI GEO: archive for functional genomics data sets--10 years on. Nucleic Acids Res. 2011;39(Database issue):D1005-10. https://doi.org/10.1093/nar/gkq1184. PMid:21097893.
13. Dykeman EC. An implementation of the Gillespie algorithm for RNA kinetics with logarithmic time update. Nucleic Acids Res. 2015;43(12):5708-15. https://doi.org/10.1093/nar/gkv480. PMid:25990741.
14. Rijal K, Mehta P. A differentiable Gillespie algorithm for simulating chemical kinetics, parameter estimation, and designing synthetic biological circuits. eLife. 2025;14:RP103877. https://doi.org/10.7554/eLife.103877.3. PMid:40095799.
15. Tsai MH, Chen MY, Huang SG, Hung YC, Wang HC. A bio-inspired computing model for ovarian carcinoma classification and oncogene detection. Bioinformatics. 2014;31(7):1102-10. https://doi.org/10.1093/bioinformatics/btu782. PMid:25429060.
16. Malik-Sheriff RS, Glont M, Nguyen TVN, Tiwari K, Roberts MG, Xavier A, et al. BioModels-15 years of sharing computational models in life science. Nucleic Acids Res. 2020;48(D1):D407-15. PMid:31701150.
17. Chen C, Méndez E, Houck J, Fan W, Lohavanichbutr P, Doody D, et al. Gene expression profiling identifies genes predictive of oral squamous cell carcinoma. Cancer Epidemiol Biomarkers Prev. 2008;17(8):2152-62. https://doi.org/10.1158/1055-9965.EPI-07-2893. PMid:18669583.
18. Sankar Ganesh P, Naseef Pathoor N, Kanna Gopal R. Letter to the editor regarding, "Impact of post-operative transoral robotic surgery hemorrhage on adjuvant treatment delays in patients with oropharyngeal squamous cell carcinoma.". Oral Oncol. 2024;159:107091. https://doi.org/10.1016/j.oraloncology.2024.107091. PMid:39520948.
19. Pathoor NN, Sankar Ganesh P, Gopal RK. Letter to the editor regarding, "Exploring diagnostic frontiers in oral squamous cell carcinoma: a comprehensive review from immunohistochemistry to genomic profiling". Oral Oncol Rep. 2024;11:100607. https://doi.org/10.1016/j.oor.2024.100607.
20. Badri H, Leder K. Optimal treatment and stochastic modeling of heterogeneous tumors. Biol Direct. 2016;11(1):40. https://doi.org/10.1186/s13062-016-0142-5. PMid:27549860.
21. Marin-Riera M, Brun-Usan M, Zimm R, Välikangas T, Salazar-Ciudad I. Computational modeling of development by epithelia, mesenchyme and their interactions: a unified model. Bioinformatics. 2016;32(2):219-25. https://doi.org/10.1093/bioinformatics/btv527. PMid:26342230.
22. Cheng X, Shen S. Transcriptional reprogramming in oral squamous cell carcinoma. Sci Rep. 2025;15(1):18210. https://doi.org/10.1038/s41598-025-01364-w. PMid:40414942.
23. Almeida Rodrigues J, Dias Pereira dos Santo H. Is artificial intelligence really a future trend in health care? Braz Dent Sci. 2021;24(3):1-4. https://doi.org/10.14295/bds.2021.v24i3.3108.
24. Bandeira CM, de Almeida AA, Carta CFL, Almeida JD, Tango EK. How to improve the early diagnosis of oral cancer? Braz Dent Sci. 2017;12(15):25-31. https://doi.org/10.14295/bds.2017.v20i4.1439.

Carlos Martin Ardila
(Corresponding address)
Universidad de Antioquia U de A, Faculty of Dentistry, Basic Sciences Department, Biomedical Stomatology Research Group
Medellín, Antioquia, Colombia
Email: martin.ardila@udea.edu.co

Editor-in-chief: Sergio Eduardo de Paiva Gonçalves
Section Editor: Renata Falchete do Prado.

Date submitted: 2025 May 08
Accept submission: 2025 Oct 09