Predicting Tumor Mutational Burden and Survival in Head and Neck Squamous Cancer Patients Using Machine Learning and Bioinformatics Approaches

Nguyen-Kieu Viet-Nhi
Taipei Medical University
Taipei, Taiwan
d142110002@tmu.edu.tw

Le Nguyen Quoc Khanh
Taipei Medical University
Taipei, Taiwan
khanhlee@tmu.edu.tw

Vu Cong Truc
Uni of Medicine and Pharmacy at Ho Chi Minh City, Vietnam
trucvucong@gmailcom

Thi Thuy Nguyen
Hue University of Medicine and Pharmacy, Hue City, Vietnam
ntthuy.ub@hueuni.edu.vn

Tran Nguyen Anh Duy
Can Tho University of Medicine and Pharmacy, Vietnam
d142111006@tmu.edu.tw

Nguyen Tu Thai Bao
Can Tho University of Medicine and Pharmacy, Vietnam
d142111005@tmu.edu.tw

How Tseng
Taipei Medical University
Taipei, Taiwan
tsenghow@tmu.edu.tw

Shih-Han Hung
Taipei Medical University
Taipei, Taiwan
seedturtle@tmu.edu.tw

CCS CONCEPTS

• Applied Computing → Machine Learning → Computational Genomics → Bioinformatics.

KEYWORDS

Tumor mutational burden; survival; head and neck squamous cancer; machine learning; bioinformatics

ACM Reference format:

Nguyen-Kieu Viet-Nhi, Le Nguyen Quoc Khanh, Vu Cong Truc, Nguyen Thi Thuy, Tran Nguyen Anh Duy, Nguyen Tu Thai Bao, How Tseng and Shih-Han Hung. 2023. Predicting Tumor Mutational Burden and Survival in Head and Neck Squamous Cancer Patients Using Machine Learning and Bioinformatics Approaches. In Houston '23: *The 14th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, September 03-06, 2023, Houston, TX. ACM, New York, NY, USA, 1 page. https://doi.org/10.1145/3584371.3613038

ABSTRACT

Head and neck squamous cancer (HNSC) is a prevalent malignancy with a complex genetic profile. Tumor Mutational Burden (TMB) is an emerging biomarker associated with prognostic and therapeutic implications. In this study, we aimed to develop machine-learning models for predicting TMB and patient survival using RNA sequencing data and clinical features.

Methods: We collected RNA sequencing data and clinical information of HNSC patients from Gene Expression Omnibus (GEO) (GSE142083) and The Cancer Genome Atlas (TCGA) databases. Machine learning models, including Random Forest, Support Vector Machine (SVM), k-Nearest Neighbors (kNN), and Logistic Regression, were built to predict TMB levels based on patient gene expression profiles. The top 100 important features (genes) were selected from these models to create a survival prediction model.

Results: Among the tested models, Random Forest showed the highest accuracy (0.8011), followed by SVM (0.796), kNN (0.777), and logistic regression (0.704). Using the top 100 important genes, we developed a model to predict HNSC patient survival (under 3 years, 3-5 years, and over 5 years). Random forest achieved an accuracy of 0.70, while SVM and kNN reached 0.65. We identified five genes (KRT14, KRT6B, COL1A1, FN1, KRT6C) most closely related to TMB and patient survival. Through KEGG pathway analysis and neural network approaches, we discovered that these genes play a significant role in three pathways: PI3K-Akt signaling pathway, Human papillomavirus infection, and Bacterial invasion of epithelial cells.

In Conclusion, our study highlights the potential of machine learning in integrating bioinformatics for predicting TMB and patient survival in HNSC. The identified genes (KRT14, KRT6B, COL1A1, FN1, KRT6C) and related pathways may serve as potential biomarkers and therapeutic targets in HNSC treatment and prognosis.

BCB'23, September 3-6, 2023, Houston, Texas USA

ACM ISBN 979-8-4007-0126-9/23/09
https://doi.org/10.1145/3584371.3613038