

TCGA-LUSC transcriptomic analysis identifies and externally validates a tumor gene signature

Agata Gabara

2026-02-21

Abstract

Lung squamous cell carcinoma (LUSC) represents a major subtype of non-small cell lung cancer and is characterized by profound transcriptional dysregulation and limited availability of robust biomarkers. Publicly available cancer genomics datasets provide an opportunity to systematically identify molecular signatures associated with tumor biology and clinical outcome.

In this study, we performed a fully reproducible transcriptomic analysis of The Cancer Genome Atlas (TCGA) LUSC RNA-seq cohort (511 tumor and 51 normal lung samples). Raw count data were processed and analyzed using the DESeq2 statistical framework to identify differentially expressed genes between tumor and normal tissue. Functional interpretation was performed using Gene Ontology enrichment, and clinical relevance was assessed using Kaplan–Meier survival analysis.

We identified a consistent tumor gene signature characterized by activation of cancer-testis antigens and epithelial differentiation programs typical for squamous carcinomas. Importantly, the signature was externally validated in an independent Gene Expression Omnibus cohort (GSE33479), where it separated tumor from normal samples and reproduced expression differences observed in TCGA.

These findings demonstrate that publicly available transcriptomic data can be used to derive biologically meaningful and reproducible cancer biomarkers and highlight candidate genes for future diagnostic and therapeutic investigation.

Introduction

Lung cancer remains the leading cause of cancer-related mortality worldwide. Among non-small cell lung cancers, lung squamous cell carcinoma (LUSC) constitutes a major histological subtype associated with smoking exposure and distinct molecular characteristics. Despite extensive genomic characterization efforts, clinically actionable biomarkers for LUSC remain limited compared with lung adenocarcinoma.

Large-scale cancer genomics initiatives such as The Cancer Genome Atlas (TCGA) have generated comprehensive transcriptomic datasets that enable systematic investigation of tumor biology. RNA sequencing allows unbiased measurement of gene expression and provides a powerful framework for identifying transcriptional programs underlying malignant transformation.

Previous studies have shown that squamous tumors are characterized by activation of keratinization pathways, epithelial differentiation programs, and immune-related signaling. However, many computational analyses remain difficult to reproduce due to incomplete reporting of methods or unavailable code.

Here, we present a fully reproducible bioinformatics workflow for the analysis of TCGA-LUSC RNA-seq data. The objectives of this study were to:

1. identify genes differentially expressed between tumor and normal lung tissue
2. characterize biological pathways associated with LUSC
3. evaluate clinical relevance using survival analysis
4. validate a tumor gene signature in an independent external dataset

By combining reproducible computational analysis with external validation, this work demonstrates how open cancer genomics resources can be used to derive candidate diagnostic and prognostic biomarkers.