Feb 25, 2025

OncoMark

DOI
https://dx.doi.org/10.17504/protocols.io.yxmvm98xol3p/v1

Shreyansh Priyadarshi[1]

[1]Ashoka University

Shreyansh Priyadarshi
Ashoka University

OPEN ACCESS

DOI: https://dx.doi.org/10.17504/protocols.io.yxmvm98xol3p/v1

Protocol Citation: Shreyansh Priyadarshi 2025. OncoMark. **protocols.io** https://dx.doi.org/10.17504/protocols.io.yxmvm98xol3p/v1

Protocol status: Working
We use this protocol and it's working.

Created: February 24, 2025

Last Modified: February 25, 2025

Protocol Integer ID: 123312

Keywords: cancer hallmark activity from transcriptomic data, cancer hallmark activity, cell transcriptomic, transcriptomic data, validating oncomark, oncomark, hallmark annotation, oncomark this protocol, single cell, cancer, learning framework, cell

Disclaimer

Abstract

This protocol details the methodology for developing and validating OncoMark, a neural multi-task learning framework designed to predict cancer hallmark activity from transcriptomic data. It covers synthetic data generation from single-cell transcriptomics, hallmark annotation, model architecture, training, validation, and statistical analyses.

1 **OncoMark: A Neural Multi-Task Learning Framework for Predicting Cancer Hallmark Activity from Biopsy Transcriptomic Data**

2 **Abstract**

This protocol details the methodology for developing and validating OncoMark, a neural multi-task learning framework designed to predict cancer hallmark activity from transcriptomic data. It covers synthetic data generation from single-cell transcriptomics, hallmark annotation, model architecture, training, validation, and statistical analyses.

3 **1. Materials & Resources**

4 **1.1. Software & Tools**

- **Python** (RRID: SCR_008394)
- **TensorFlow** (RRID: SCR_016345)
- **Scikit-learn** (RRID: SCR_002577)
- **ComBat** for batch effect correction (RRID: SCR_012835)
- **UCell** for Digital Scoring (**GitHub Repository**)

5 **1.2. Data Sources**

- **Single-cell transcriptomic data**: Weizmann 3CA repository
- **Bulk transcriptomic datasets**:
- The Cancer Genome Atlas (TCGA)
- MET500
- POG570
- Cancer Cell Line Encyclopedia (CCLE)
- Therapeutically Applicable Research to Generate Effective Treatments (TARGET)
- Pan-Cancer Analysis of Whole Genomes (PCAWG)
- **Normal datasets:**
- GTEx, ANTE

6 **1.3. Computational Environment**

- **Operating System:** Linux Ubuntu 20.04
- **RAM Requirements:** Minimum 64GB

7 **2. Step-by-Step Protocol**

8 **Step 1: Data Collection and Preprocessing**

1. **Download** single-cell transcriptomic data from the Weizmann 3CA repository.
2. **Perform quality control (QC):**

- Exclude cells with mitochondrial transcript content > 15%.

- Remove cells with < 200 or > 6000 detected transcripts.

1. **Download bulk transcriptomic datasets** from public repositories.
2. **Curate hallmark gene sets** from literature and databases.

9 **Step 2: Digital Scoring of Hallmarks**

1. **Compute digital scores** using UCell for each hallmark per single cell.
2. **Binarize scores** using Otsu's thresholding to classify hallmark presence or absence.
3. **Calculate tissue-specific thresholds** to refine hallmark classification.

10 **Step 3: Synthetic Biopsy Data Generation**

1. **Aggregate 200 hallmark-specific cells** from each patient sample to simulate biopsies.
2. **Ensure non-overlapping hallmark-positive and hallmark-negative samples** to avoid cross-contamination.

11 **Step 4: Model Architecture & Training**

1. **Define a multi-task learning model** with a shared base layer and hallmark-specific outputs.
2. **Split data:**

- Training: 85% (57,735 samples)
- Validation: 15% (10,195 samples)

1. **Use Adam optimizer** (learning rate: 0.0001) and binary cross-entropy loss.
2. **Train model for 50 epochs** with early stopping (patience = 6 epochs).

12 **Step 5: Model Validation**

1. **Use five-fold cross-validation repeated twice.**
2. **Validate on external datasets (n=95 patients).**
3. **Evaluate using F1-score, accuracy, precision-recall, and AUC-ROC.**

13 **Step 6: Statistical Analysis**

1. **Compare hallmark distributions** in cancer vs. normal samples using the Kolmogorov-Smirnov test.
2. **Perform logistic regression** to analyze hallmark-drug associations.
3. **Compute odds ratios (ORs) for hallmark co-occurrences.**

14 **Step 7: Deployment & Clinical Application**

1. **Deploy the trained model as a web tool** (OncoMark Web Server).
2. **Allow clinicians to upload tumor transcriptomic data** for real-time hallmark activity prediction.

15 **3. Expected Results**

- High accuracy in predicting hallmark presence.
- Clear separation of hallmark distributions in cancer vs. normal tissues.
- Identification of hallmark-drug interactions.

16 **4. Troubleshooting**

	Issue	Possible Cause	Solution
	Low model accuracy	Overfitting	Increase dropout, apply L2 regularization
	Batch effects	Dataset heterogeneity	Use ComBat for batch correction
	Imbalanced hallmark labels	Unequal representation	Perform synthetic data augmentation

17 **5. Data Sharing & Accessibility**

- **Synthetic Data:** https://doi.org/10.5061/dryad.zw3r228jc
- **Codebase:** OncoMark GitHub
- **Python Package:** OncoMark on PyPI
- **Web Server:** OncoMark Web Server

This protocol follows best practices in **reproducible AI-driven biomarker discovery** and can be modified for additional hallmark analyses.

For questions, refer to the full documentation at **OncoMark Docs**.