Bioinformatics, 36(7), 2020, 2293–2294
doi: 10.1093/bioinformatics/btz914
Advance Access Publication Date: 6 December 2019
Applications Note

Gene expression

SpaCell: integrating tissue morphology and spatial gene expression to predict disease cells

Xiao Tan†, Andrew Su†, Minh Tran and Quan Nguyen *

Division of Genetics and Genomics, Institute for Molecular Bioscience, The University of Queensland, Brisbane 4072, QLD, Australia

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
Associate Editor: Anthony Mathelier

Received on August 12, 2019; revised on November 22, 2019; editorial decision on November 28, 2019; accepted on December 4, 2019

Abstract

Motivation: Spatial transcriptomics (ST) technology is increasingly being applied because it enables the measurement of spatial gene expression in an intact tissue along with imaging morphology of the same tissue. However, current analysis methods for ST data do not use image pixel information, thus missing the quantitative links between gene expression and tissue morphology.
Results: We developed a user-friendly deep learning software, SpaCell, to integrate millions of pixel intensity values with thousands of gene expression measurements from spatially barcoded spots in a tissue. We show the integration approach outperforms the use of gene-count data alone or imaging data alone to build deep learning models to identify cell types or predict labels of tissue images with high resolution and accuracy.
Availability and implementation: The SpaCell package is open source under an MIT licence and it is available at https://github.com/BiomedicalMachineLearning/SpaCell.
Contact: quan.nguyen@uq.edu.au
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Spatial transcriptomics (ST) technology is emerging as an important platform for measuring molecular biological processes at the tissue level (Burgess, 2019). Different from other genomics technologies, ST does not require dissociating cells from the original tissue. Molecular measurements can be mapped back to the spatial location of the cells in tissue via spatial barcodes, adding a novel spatial data dimension to gene expression data. Moreover, platforms such as Slide-seq generate a tissue image and a gene expression profile of the same tissue, allowing the integration of tissue morphology and spatial gene expression (Salmén *et al.*, 2018).

However, incorporating imaging data to gene expression data is a new analysis area, while current analysis pipelines mainly focus on using expression values but not image pixel values. Image pixel intensity data contain informative features that can be used for diagnosing diseases such as for cancer staging (Coudray *et al.*, 2018). Although machine learning methods exist for analyzing imaging data (Komura and Ishikawa, 2018), these methods do not utilize molecular data. Advances in genomics technologies create new data types for novel machine learning applications to combine molecular measurements with image pixel data to characterize tissue morphological images beyond current pathological annotation (Hekler *et al.*, 2019). Existing methods for spatial data analysis, however, use gene expression, but not image pixel information (Dries *et al.*, 2019; Navarro *et al.*, 2017). We developed SpaCell with a comprehensive workflow to utilize both pixel and gene expression data to train neural network (NN) models for cell-type and disease-stage classification.

2 Main workflow

SpaCell's workflow (Fig. 1) starts with two-stream data preprocessing. For image preprocessing, SpaCell first removes any colour cast, which is the background difference from the white background in the Hematoxylin and Eosin staining image, then performs stain normalization to overcome inconsistencies in the staining process (Macenko *et al.*, 2009) (Supplementary Methods). Then, high images are tiled into small tiles and the tiles are resized to 299× 299 pixels, where each tile contains one spot. To increase model performance and generalizability, SpaCell performs random rotation and *Z*-transform of the tiled images for each training step. For count matrix preprocessing, gene counts are mapped read counts to each ST spot, recovered by spatial barcodes. A large range of programs developed for single-cell data analysis are available for users to process and normalize count data. SpaCell has built-in and fast options to remove unreliably detected spots and genes, followed by library-size normalization.

In the cell-type classification model (Supplementary Methods), SpaCell analyses one high-resolution image and its spatial count matrix. To extract a latent feature vector for each image tile, we fit pretrained convolutional neural network (CNN) weights from the

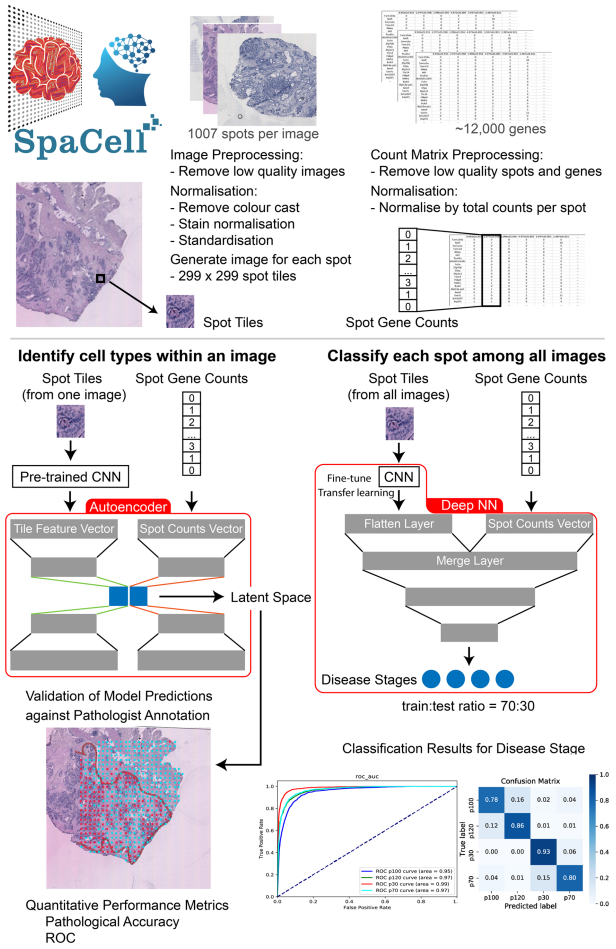

Fig. 1. Two main analysis workflows of SpaCell: cell-type identification and disease-stage classification

ResNet50 model to utilize big data in the ImageNet database. For each tile feature vector and its corresponding spot gene counts, we trained two autoencoders to find two latent spaces of equal dimension, which are then concatenated into one latent vector. For all spots, the latent vectors are then combined to form a latent matrix representative for both image and gene-count data, which are then used to perform clustering (default as *K*-means clustering) to identify cell types. After spot clustering, SpaCell provides visualization functions to evaluate the model performance. Importantly, to enable quantitative comparison to pathological annotation information, we devised an approach to automatically and accurately detect and map annotation contours from a low-resolution image to a whole slide image. After mapping, we use spot coordinates and the contour-masked regions to compare computational prediction with pathological annotation.

In the disease-stage prediction model (Supplementary Methods), SpaCell uses hundreds of images and corresponding count matrices. Both tiles and count data are input into a fully connected model, initially as two streams (Fig. 1). Each image tile is initiated by a CNN with weights pre-trained on the ImageNet dataset and these weights are trainable together with parameters from the count stream. To increase model generalization and reduce over-fitting, the following strategies are applied: random sampling of images stratified by labels ensuring unseen test images, 5-fold cross-validation, drop-out and L2 penalization. Following model training, users can apply evaluation functions in SpaCell for quantitative analysis of model performance such as test accuracy, ROC curves and confusion matrix.

SpaCell models were tested on a prostate cancer (Berglund *et al.*, 2018) and amyotrophic lateral sclerosis (Maniatis *et al.*, 2019) datasets (Supplementary Methods), which represent a dataset with few images and high resolution compared to a dataset with more images and lower resolution. By testing >40 models, we consistently found that the combination of pixel and gene expression data improved model performance by 8–14% in accuracy, precision, *F*-score and area under the curve in cell-type models (Supplementary Figs S1 and S2) and 4% in disease-stage classification models (Supplementary Fig. S3).

3 Implementation

SpaCell has been developed with Python 3.7 as a user-friendly software. Installation and tutorials are described in the SpaCell GitHub page and PyPI repository. Changes in the parameter settings are kept in the config file for reproducibility. SpaCell uses Keras and TensorFlow backend which are portable between platforms and supports graphics processing units distribution to accelerate the training step.

4 Conclusion

SpaCell is a pioneering software program implementing deep NNs for integrating image pixel data and spatial gene expression data for biomedical research. We show that SpaCell can automatically and quantitatively identify cell types and disease stages. We tested over 40 models and consistently found that the integration of both data types increased model performance compared to using one type of data input. Moreover, SpaCell prediction results have higher resolution, specific to thousands of spatial spots, compared to typical pathological annotation with several large regions. We expect that our model can be applied to any type of spatial omics data that have both images and expression values.

Acknowledgements

We thank Prof Joakim Lundeberg and Dr Emelie Berglund for sharing the spatial data and members in Nguyen's Biomedical Machine Learning Lab for helpful discussion.

Funding

This work was supported by the Australian Research Council Discovery Early Career Researcher Award (DE190100116), the University of Queensland and the Genome Innovation Hub.

Conflict of Interest: none declared.

References

Berglund,E. *et al.* (2018) Spatial maps of prostate cancer transcriptomes reveal an unexplored landscape of heterogeneity. *Nat. Commun.*, **9**, 2419.

Burgess,D.J. (2019) Spatial transcriptomics coming of age. *Nat. Rev. Genet.*, **20**, 317.

Coudray,N. *et al.* (2018) Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nat. Med.*, **24**, 1559–1567.

Dries,R. *et al.* (2019) Giotto, a pipeline for integrative analysis and visualization of single-cell spatial transcriptomic data. *bioRxiv*. doi: 10.1101/701680.

Hekler,A. *et al.* (2019) Deep learning outperformed 11 pathologists in the classification of histopathological melanoma images. *Eur. J. Cancer*, **118**, 91–96.

Komura,D. and Ishikawa,S. (2018) Machine learning methods for histopathological image analysis. *Comput. Struct. Biotechnol. J.*, **16**, 34–42.

Macenko,M. *et al.* (2009). A method for normalizing histology slides for quantitative analysis. In: 2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro, pp. 1107–1110.

Maniatis,S. *et al.* (2019) Spatiotemporal dynamics of molecular pathology in amyotrophic lateral sclerosis. *Science*, **364**, 89–93.

Navarro,J.F. *et al.* (2017) ST pipeline: an automated pipeline for spatial mapping of unique transcripts. *Bioinformatics*, **33**, 2591–2593.

Salmén,F. *et al.* (2018) Barcoded solid-phase RNA capture for spatial transcriptomics profiling in mammalian tissue sections. *Nat. Protoc.*, **13**, 2501–2534.