Classification of regulatory sequences using machine learning techniques

Boris Shilov, Wim Thiels, Lucas Coppens, Zixuan Xie

Introduction

Melanoma is an aggressive cancer with a high level of therapy-resistance due to its high degree of heterogeneity and plasticity. Despite the great efforts put into the field of melanoma treatment, resulting in great advances, many challenges remain. The major challange in fighting melanoma relapse continues to be the tumor heterogeneity, as melanoma comprises a wide variety of phenotypically distinct subpopulations of cancer cells. To be able to adress this issue therapeutically, the underlying mechanisms of heterogeneity need to be characterized. In this paper, a distinction between two major types of melanoma cells is made: invasive and proliferative. (Verfaillie et al. 2015; Shannan et al. 2015)

Transcriptional reprogramming of melanoma cells in proliferative state into melanoma cells with invasive characteristics is a critical event at the origin of metastatic spreading of melanoma. Invasive cells have acquired the ability to migrate to other tissues, enter the bloodstream and therefore lie at the basis of the metastatic spreading of cancer in the body. While the transitional mechanisms from proliferative to invasive cancer cell are yet to be characterized more extensively, it is sure that one event lies at the basis of this transition: the transcriptional reprogramming of the cell. Studying the involved genes and regulatory elements using various bioinformatics approaches is therefore a hot topic in the area of melanoma research. Decoding the regulatory landscape could result into the ability to push melanoma cells towards a different cell state, which would be an interesting target from a therapeutic point of view. (Verfaillie et al. 2015)

Transcriptomic, open chromatin and histone modification maps of melanoma cultures were constructed, revealing thousands of active cis-regulatory regions, both for proliferative and invasive cells. (Verfaillie et al. 2015) It should be possible to discern which cis-regulatory regions are useful for the classification of cell states in melanoma samples if such states are truly distinct in terms of regulatory landscape. The aim of the project was the construction of classifiers predicting whether a regulatory region would be active in proliferative or invasive cell states. Another useful insight that would be gained from these classifiers, is which regulatory regions are the most significant for distinguishing between cell states, thus giving information about the underlying mechanisms and the critical genes and regulatory elements involved in cancer cell state transitions.

Two distinct machine learning techniques were used for the creation of such classifiers, namely the random forests ensemble method and deep learning, with the use of convolutional neural networks. Both models were trained on the same training set, which comprises the dataset of active cis-regulatory regions, mentioned hereabove. In this paper, both methods are described and their results are evaluated and compared.

Methods

Results and discussion

References

Shannan et al. 2015. "Heterogeneity in Melanoma." *Cancer Treatment and Research*.

Verfaillie et al. 2015. "Decoding the Regulatory Landscape of Melanoma Reveals Teads as Regulators of the Invasive Cell State." *Nature Communications*.