Layer-wise relevance propagation reveals novel insight in cancer patient segmentation

Shoko Iwai*, Hongye Yang, Jason K Ellis, Ryan Monsurate, Rajeev Dutt. AI Dynamics, Bellevue, WA *shoko@aidynamics.com

AI DYNAMICS

ABSTRACT To provide an appropriate treatment to cancer patients, understanding each patient's complex pathologies is important. The variation in gene expression of the cancer tissue has been used to study such complex conditions. Machine learning methods are among many tools to make biological discoveries from the gene expression data. Deep learning algorithms are promising machine learning methods, however, one of the biggest concerns has been its luck of explainability. Here we used multilayer perceptron network to build a model for previously published cancer clusters based on gene expression data, then applied layer-wise relevance propagation to determine which set of genes "explain" the model's prediction. A set of model explaining genes were different from those identified by conventional differentially expressed genes between clusters. Our approach revealed novel biological hypotheses to be investigated.

METHODS Processed and normalized mRNA gene expression data used in Campbell et al. (2018) was obtained from Genomic Data Commons website (https://gdc.cancer.gov/about-data/publications/PanCan-Squamous-2018). Cancer-related genes were selected as described in Campbell et al. (2018). Gene abundance was scaled using z-score for machine learning. Six clusters defined by k-means in Campbell et al. (2018) were used as labels for the model. Layer-wise relevance propagation (LRP) was performed using iNNvestigate (Alber et al., 2019) with multiple decomposition methods. LRP scores and normalized gene expression levels were compared across clusters by Welch test and p-values were adjusted by Benjamini-Hochberg method. Gene set enrichment analysis was performed using R package, ReactomePA (Yu and He, 2016).

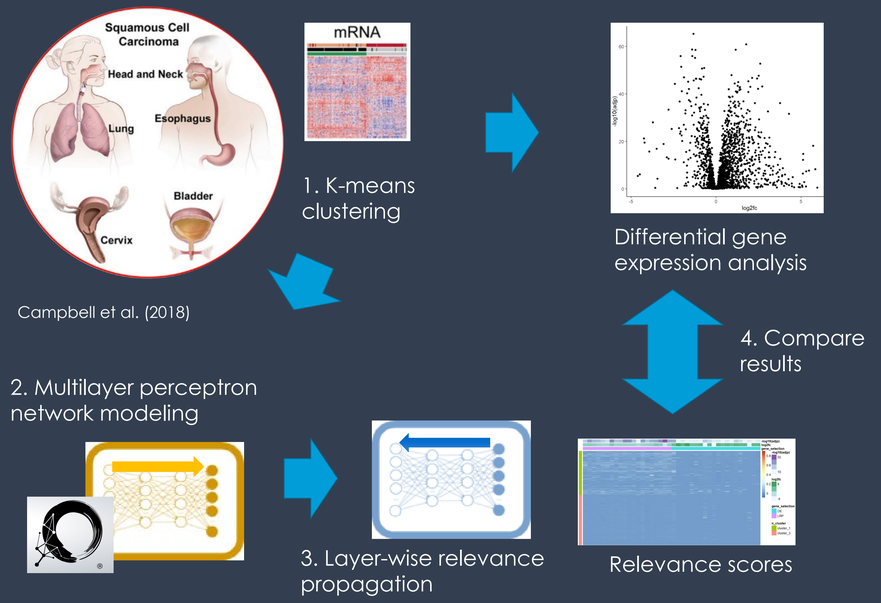

Fig. 1. Analysis overview. (1) Six clusters were defined by k-means in Campbell et al. (2018). (2) A multilayer perceptron network model was constructed. (3) Layer-wise relevance propagation (LRP) was performed. (4) LRP scores and differentially expressed genes (conventional method) were compared.

RESULTS & DISUCSSION

Multilayer perceptron network model explains carcinoma clusters. A multilayer perceptron (MLP) network model was constructed with hidden layers of 256, 128, 64 and 2 neurons. We split the whole dataset into a training set and a validation set. 80% of the dataset went into the training set and 20% of the dataset goes into the validation set. We obtained a validation accuracy of 98% and area under curve of 0.98.

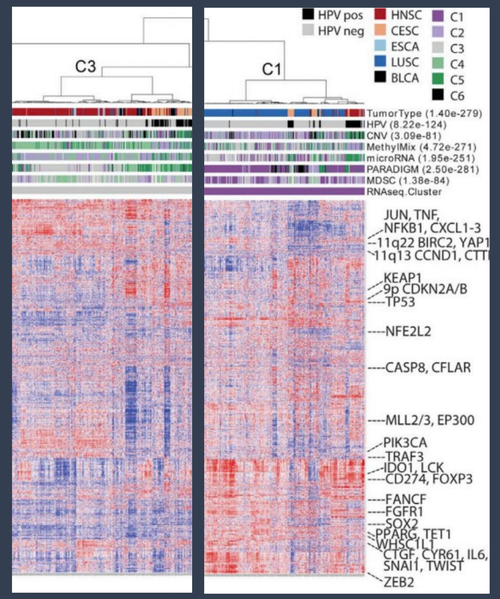

Fig. 2. K-means clustering results of cancer gene expression in squamous cell carcinomas. Only clusters 1 and 3 are shown. (Campbell et al., 2018)

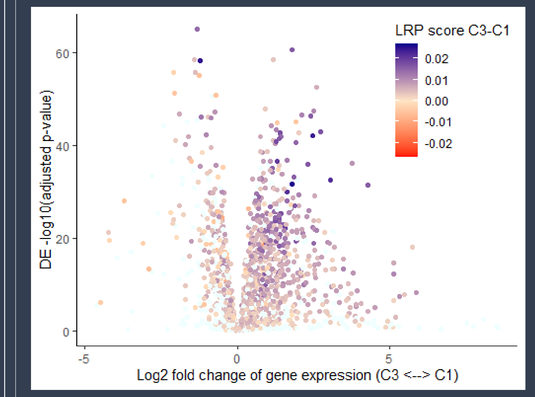

Fig. 3. Volcano plot of differentially expressed genes colored by LRP score differences between two clusters. Large LRP score differences do not always match to highly differentially expressed genes.

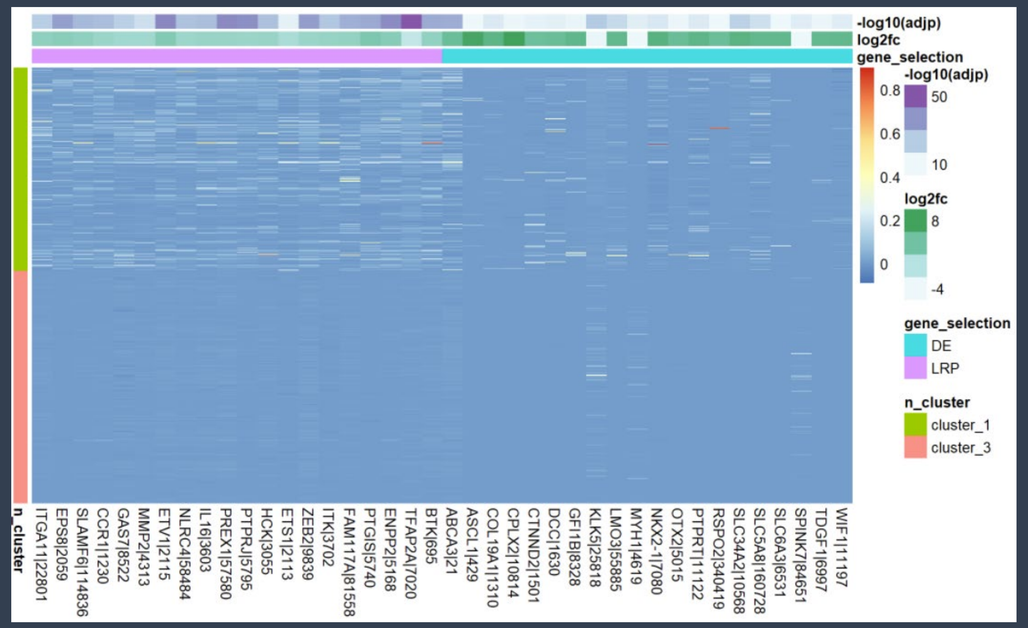

Fig. 4. Genes with high LRP scores (Alpha2Beta1) do not always have large effect size. DE, significantly differentially expressed genes with largest effect size; LRP, significantly differential LRP scores with largest abundance differences. Significance was determined by adjusted p-value <0.05. Adjusted p-value and log2 fold change of the gene expression abundance are shown at the top of the heatmap.

LRP identified genes and pathways are different from those identified by conventional differential gene expression analysis. Gene set enrichment analysis identified biological pathways associated with the highly differential LRP scores (LRP-GSE) and highly differentially expressed genes (DE-GSE). DE-GSE includes GPCR related pathways, whereas LRP-GSE includes GTPase cycle related pathways. LRP identified pathways may better represent a composite view of the cell response.

CONCLUSION We employed LRP to explain underlying biology associated with a deep learning model developed on squamous cancer gene expression data. The results clearly showed neural network uses distinct genes and pathways from those identified differentially expressed by conventional univariate tests. Although additional biological analysis is necessary to validate the importance of identified genes and pathways, the methods added new insights to generating biological hypothesis to test, eventually helping to develop tailored treatments for patient groups.

REFERENCES

Alber M et al. (2019) iNNvestigate neural networks! J Machine Learning Res. 20.
Campbell JD et al., (2018) Genomic, pathway network, and immunologic features distinguishing squamous carcinomas. Cell Rep. 23(1):194-212.e6.
Yu G and He QY (2016) ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization. Mol Biosyst 12(2):477-479.