Hypoxia classifier for transcriptome datasets

Laura Puente-Santamaria[a,b,c,*], Lucia Sanchez-Gonzalez[a], Ricardo Ramos-Ruiz[b] and Luis del Peso[a,b,d,e,f,*]

[a]Departamento de Bioquímica, Universidad Autónoma de Madrid (UAM), Madrid, 28029, Spain
[b]Instituto de Investigaciones Biomédicas "Alberto Sols" (CSIC-UAM), Madrid, 28029, Spain
[c]Genomics Unit Cantoblanco, Parque Científico de Madrid, C/ Faraday 7, Madrid, 28049, Spain
[d]IdiPaz, Instituto de Investigación Sanitaria del Hospital Universitario La Paz, Madrid, 28029, Spain
[e]CIBER de Enfermedades Respiratorias (CIBERES), Instituto de Salud Carlos III, Madrid, 28029, Spain
[f]Unidad Asociada de Biomedicina CSIC-UCLM, Albacete, 02006, Spain

ARTICLE INFO

Keywords:
Transcriptome classification
Hypoxia
Gene expression
RNA-seq
Spatial transcriptomics

ABSTRACT

Molecular gene signatures are useful tools to characterize the physiological state of cell populations according to their gene expression profiles. However, most molecular gene signatures have been developed under a very limited set of conditions and cell types, and are often restricted to a set of gene identities linked to an event or biological process, therefore making necessary to develop and test additional procedures for its application to new data.

Focusing on the transcriptional response to hypoxia, we aimed to generate widely applicable classifiers capable of detecting hypoxic samples while maintaining transparency and ease of use and interpretation. Here we describe several tree-based classifiers sourced from the results of a meta-analysis of 69 differential expression datasets which included 425 individual RNA-seq experiments from 33 different human cell types exposed to different degrees of hypoxia (0.1-5%O_2) for a time spanning between 2 and 48h.

These decision trees include both the identities of genes key in the response to hypoxia and defined quantitative boundaries, allowing for the classification of individual samples without needing a control or normoxic reference. Despite their simplicity and ease of use, these classifiers achieve over 95% accuracy in cross validation and over 80% accuracy when applied to additional challenging datasets. Moreover, the explicit structure of the trees allowed for the identification of relevant biological features in cases where prediction was not accurate. Finally, we demonstrate that the classifiers can be applied to spatial gene expression data to identify hypoxic regions within histological sections. Although we have focused on the identification of hypoxia, this method can be applied to detect activation of other processes or cellular states.

1. Introduction

A gene expression signature is a single or combined group of genes whose expression is altered in predictable way in response to a specific signal or cellular status. Gene Signatures are often derived from the set of differentially expressed genes (DEGs) identified when comparing two groups of transcriptomes, such as disease versus healthy controls or treated versus untreated samples. In turn, a gene signature can be of aid in trying to determine whether a given biological sample was exposed to that particular stimulus or belongs to the status defined by the gene set. Thus, reliable gene signatures can be used as surrogate markers for the activation of pathways or cellular status.

Hypoxia can be defined as the situation were oxygen supply does not meet cellular demand [1]. In response to hypoxia cells activate a gene expression program, under the control of the Hypoxia Inducible Factors (HIFs) [2], that aims to increase oxygen supply while reducing its consumption. Thus, this transcriptional response restores oxygen balance and, as such, it is central in maintaining tissue homeostasis. Importantly, oxygen homeostasis is disrupted in a number of prevalent pathologies including neoplasms [3] and cardio-respiratory diseases [4]. For all this reasons, the development of a hypoxic gene signature could be of practical interest to identify cells or samples that had been exposed to hypoxia. Accordingly, a number of studies have published hypoxic signatures [5, 6, 7, 8, 9, 10, 11, 12]. However, in spite of their merit, in all these cases the gene signature was derived from a limited set of related tumoral samples, raising the question of their applicability in other contexts. On the other hand, in almost all the cases, the gene signature is just a set of genes without any additional information reflecting their relative importance or their expected expression levels under normoxic/hypoxic conditions. Thus, based solely in the identities of the genes in the signature it is nearly impossible to classify an individual isolated sample as normoxic or hypoxic.

Herein we describe a novel tree-based classifier that accurately identify hypoxic cells or samples based on their gene expression profile. The identification is absolute, meaning that it does not require a set of normoxic reference samples to sort out the hypoxic ones. Thus, it can be applied to interrogate a single isolated sample. Finally, although the classifier implicitly contains information about the relative importance of the genes in the signature and their expression levels in hypoxia, it is simple enough to be interpreted and

*Corresponding author
lpsantamaria@iib.uam.es, laura.puente@fpcm.es (L. Puente-Santamaria); luis.peso@uam.es (L.d. Peso)
ORCID(s): 0000-0001-9034-4576 (L. Puente-Santamaria); 0000-0002-6331-9786 (R. Ramos-Ruiz); 0000-0003-4014-5688 (L.d. Peso)