Highlights

Hypoxia classifier for transcriptome datasets

Laura Puente-Santamaria,Lucia Sanchez-Gonzalez,Ricardo Ramos-Ruiz,Luis del Peso

- Description of a method to generate simple tree-based gene signatures that combine gene identity and expression.
- Tree-based gene signatures can be easily applied for accurate absolute classification of samples and regions in spatial gene expression data.
- The explicit structure of the classifiers permits the intuitive interpretation of classification allowing for novel biological insights.

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)

applied without the need for sophisticated computational tools.

2. Materials and Methods

2.1. RNA-seq data download and processing

Raw reads of the RNA-seq experiments were downloaded from Sequence Read Archive [13]. Pseudocounts for each gene were obtained with salmon [14] using RefSeq [15] mRNA sequences for human genome assembly GRCh38/hg38 and mouse genome assembly mm10 as references.

Reads counts of tumoral and healthy samples were downloaded from the TCGA data portal and transformed to counts per million.

Spatial gene expression datasets were downloaded from 10X Genomics website [16, 17, 18, 19]. Raw read counts are normalized with sctransform[20] following Seurat v4.0.4[21] standard pipeline for analysis, visualization, and integration of spatial datasets.

2.2. Generation of a classifier

To generate the classifier we made use of 425 transcriptomic profiles of hypoxia-exposed cells and their normoxic counterparts described in a recent study [22]. From the gene pseudocounts in each sample, we calculated each gene's ranking percentile and used this information in downstream analyses. We used the R package randomForest [23] to perform feature selection and the R package rpart [24] to generate decision trees. Each decision tree was evaluated by cross-validation using 70% of the available RNA-seq experiments as a training set and the remaining 30% as a validation set. By default, a sample is classified as hypoxic when the tree assigns it a probability over 50% of being hypoxic, even though this threshold can be made stricter or laxer.

Results

Generation of a hypoxic classifier

Results from our previous work [22] on differential expression triggered by hypoxia indicate that, even for those genes showing a significant regulation in the ensemble of datasets, the response to hypoxia could be in large part cell-specific. Thus, we sought to identify a minimal set of genes that could be used as a reliable readout of exposure to hypoxia and to develop a simple, easy to use, classifier that could identify whether an individual sample is hypoxic based on its gene expression.

With the goal of making the model as widely applicable as possible we chose to use as input data the percentile of each gene on a gene expression ranking, thus minimizing the effects of read depth, different normalization methods, possible rRNA contamination, and other factors that influence RNA quantification. Thus, we first constructed a gene ranking matrix for a set of 425 individual RNA-seq samples derived from published transcriptomic analysis of hypoxic cells and controls. For subsequent analyses we kept the subset of 178 genes both significantly up-regulated by hypoxia ($LFC > 0.7$, $FDR < 0.01$) and widely expressed (detectable in $\geq 90\%$ of the analyzed subsets) according to a meta-analysis performed on this data set [22].

In order to select the most informative genes in this subset, we used 1000 iterations of a random forest classifier sampling 70% of the RNA-seqs at each iteration. As a measure of each gene's importance we use the mean decrease in accuracy (MDA) across all iterations. The 20 genes with average MDA over 4 (1A) were selected to train 10000 decision trees randomly sampling 70% of the individual RNA-seq experiments and using the remaining 30% as a validation set. The 276 trees with an accuracy over 0.95 on the validation set were selected to further test their performance (Supplementary table S1-3, "Cross validation").

Only 16 genes are used in all of the 276 decision trees, with half of them (EGLN1, MIR210HG, NDRG1, ANKRD37, TCAF2, PFKFB3, BHLHE40, and MAFF) being included overwhelmingly more frequently (Fig. 1B). Most of these genes have already been linked to the transcriptional response to hypoxia [25, 26, 27, 28, 29, 30, 31, 32], even though in some cases their particular role in it has not yet been defined.

In these classification trees the rank percentile of the gene in each decision point is used to determine the branch followed for the classification of the sample, hence final sample label is assigned based on the relative (percentile rank) expression values of the genes in the tree. As seen in figure 1C, the split point for most genes is limited to a narrow range of rank percentiles, with ANKRD37 and BHLHE40 being the exception. These genes show two differentiated split points that depend on the identity of remaining genes in the tree. ANKDR37's split point depends on whether its combined together with NDRG1 or TCAF2, while BHLHE40's depends on whether the tree includes MIR210HG. Figure 1D represents gene identity and split points for the 10 best trees according to cross-validation accuracy. Both in this top 10, as well as in the whole set of 276 trees we find a limited number of topologies present, with MIR210HG, EGLN1, MAFF, NDRG1, and PFKFB3 forming the most common combination, closely followed by ANKRD37, NDRG1, and BHLHE40.

Evaluation and validation of the resulting decision trees

In order to evaluate the performance of each of the decision trees and test particular strengths and weaknesses of each model, we tested them on a series of datasets that were not part of the training nor cross-validation sets and had some differential feature that posed a challenge to the classification (Supplementary table S1-1 "RNA-seq metadata"). Based on the results described in this section, we will choose individual trees better suited to be used as general and robust hypoxia classifiers.

First we chose the time series experiments available on PRJNA561635 [33], consisting on RNA-seqs of HUVEC cells exposed to different oxygen concentrations at nine

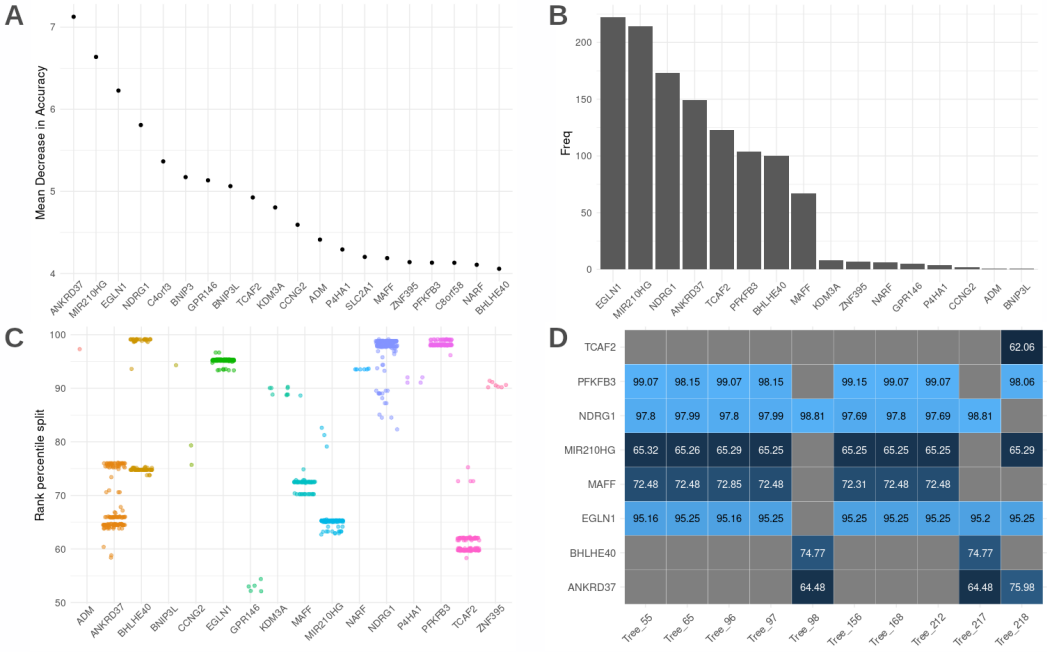

Figure 1: Generating expression based tree classifiers to identify hypoxic samples. A: median decrease in accuracy index of the 20 most important genes according to 1000 random forest iterations. **B**: Frequency of each gene being used as a predictor variable in the classification trees. **C**: Split points for the rank percentile (100 being the most expressed gene, 0, the least) of the genes used in all the models with accuracy > 0.95. **D**: Split points for the rank percentile of the genes used in the 10 best performing models according to cross-validation accuracy.

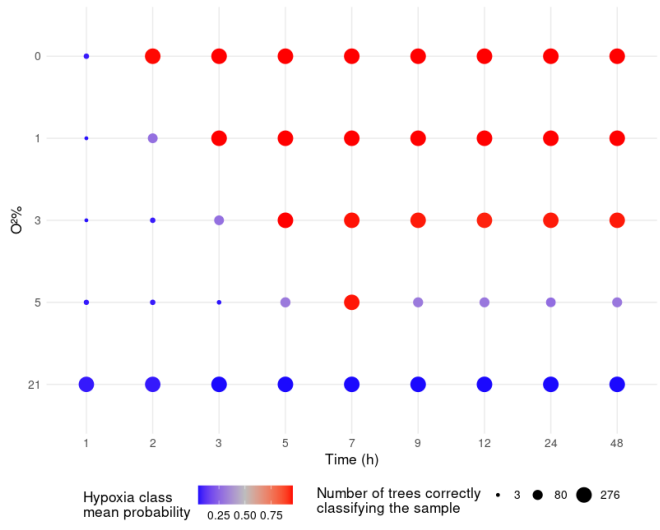

Figure 2: Validation set, PRJNA561635 time series.. Each of the dots represents one of the samples in the five time series in PRJNA561635, ordered by oxygen tension and time. The color of each dot represents the mean probability of each sample to be classified as hypoxic, while the size of the dot is proportional to the number of trees correctly classifying each sample.

time points. The main challenge with this validation set is detecting early stages of hypoxia (1-3 hours), where most hypoxia-target genes have just barely began to accumulate, and moderately low oxygen concentrations, since 5% O_2 is within the range of physiological oxygen concentration for many tissues [34] and in vitro triggers a weaker transcriptional response for many genes [35]. As shown in figure 2 and Supplementary table S1-4 "PRJNA561635", all decision trees correctly classified normoxic samples and samples exposed to oxygen levels below 3%O_2 for at least 5h. It is worth highlighting that around a third of the threes were also able to detect earlier stages of hypoxia (2h 1%O_2, 3h 3%O_2) as well as most samples grown at 5%O_2. In addition, these results clearly show that the lower the oxygen tension, the strongest the signal detected at early times of exposure.

The next validation set consisted of four studies on specific fractions of RNA: newly transcribed RNA(4sU labeling RNA-seq and GRO-seq) and actively translated RNA (polysomal RNA-seq) (Supplementary table S1-5, "RNA fractions"). In this case the challenge stems from the different RNA fractions used in the derivation of the trees (total mRNA) and test datasets. In spite of the different source of RNA, all decision trees had an accuracy over 0.75, with 36% showing an accuracy over 0.9 in the classification of 4sU labeling, GRO-seq, and polysomal samples (3A). As shown in Figures 3B and 3C, with this dataset the main influence in tree accuracy is the use of MAFF or NDRG1 as predictor variables, since MAFF's raking percentile is not altered between normoxic and hypoxic conditions (3D).

Given the importance of oxygenation on tumor biology, we next decided to test the predictive potential of our classifier on tumor-derived gene profiles. As a first approximation, we tested whether our models could differentiate between clear cell renal carcinoma samples (ccRCC) and healthy

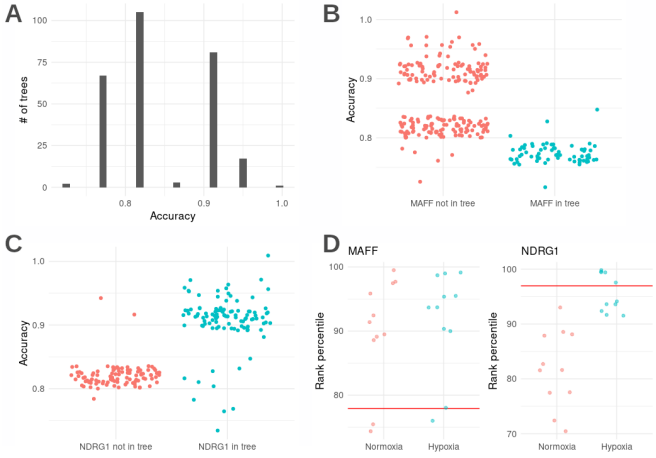

Figure 3: Validation set, RNA fractions. A: Histogram of the accuracies of the 276 classification trees in 22 datasets of specific RNA fractions. **B**: Accuracies of the classification trees according to the use of MAFF as a predictor variable. **C**: Accuracies of the subset of trees that do not use MAFF as a predictor variable, according to the use of NDRG1. **D**: Ranking percentiles of MAFF and NDRG1 in the samples forming the validation set. The red line represents the mean split point for each gene in the trees using either of them as a predictor variable.

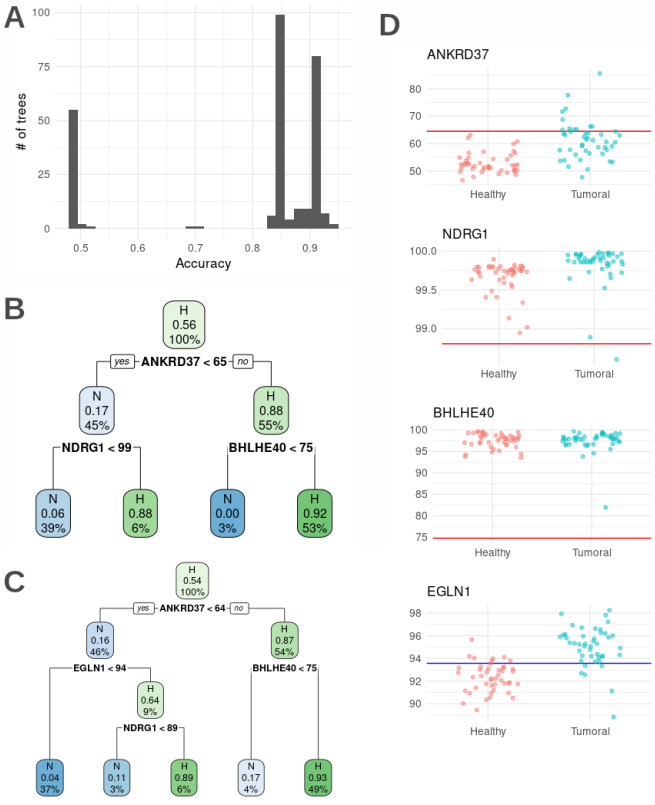

Figure 4: Validation set, clear cell renal carcinoma with VHL mutations. A: Distribution of classification accuracies of the 276 trees evaluated. **B**: Representative example of a poorly performing tree. **C**: Example of a tree with a similar structure, but better performance with ccRCC data (tree #241). **D**: Ranking percentiles of ANKRD37, NDRG1, BHLHE40, and EGLN1 in the samples making up the validation set. The red line represents the mean split point for each gene in the poorly performing trees. The blue line represents the split point for EGLN1 in tree #241.

adjacent tissue (Supplementary table S1-6, "ccRCC"). Over 80% of ccRCC show mutations in the von Hippel-Lindau (VHL) gene and the product of the VHL tumor suppressor gene is a key molecule controlling HIF stability, thus its mutation leads to chronic HIF activation, even in the presence of oxygen, leading to a hypoxia-like transcriptional pattern [36, 37]. Despite most of the trees reaching high accuracies in classifying samples in this dataset, as seen in Figure 4A, a subset of them miscalls half of the elements by overestimating the number of hypoxic samples. Out of the 58 trees with this issue, 42 share a common structure: the tree are composed of only three genes organized in two levels, with ANKRD37 being the root nodes and two branches, one evaluating NDRG1 and the other BHLHE40 (Fig. 4B). As both NDRG1 and BHLHE40 are highly expressed in healthy tissue samples (4D), those trees with the structure shown in Fig. 4B are not capable to differentiate between conditions. Interestingly, those trees with one extra node evaluating EGLN1 (Fig. 4C) increase their classification accuracy up to 0.88.

In order to investigate a wider range of tumors without restriction to those including VHL mutations, we have compared our classifiers against available hypoxia signatures applied to TCGA data [38]. We counted the proportion of samples classified as hypoxic by primary site (Fig. 5, Supplementary table S1-7, "TCGA hypoxic sample proportion"), and with the exception of renal carcinomas (considered more "hypoxic" on average by our model) the proportion of samples classified as hypoxic by our decision trees follows the median hypoxia scores calculated in [38] for each cancer type.

Finally, since the tree-classifier was derived from 425 human samples, we decided to test its performance on transcriptomes from other organisms. To that end, we gathered five studies in murine cells, totalling 34 individual RNA-seq experiments. Similar to the case of the ccRCC samples, we found that, albeit the majority of trees (160 out 276) were able to classify samples with an accuracy of about 79% or higher, a subset of classifiers performed poorly (Fig. 6A). Examining the possible causes of the stark differences in accuracy between trees, we found that most of the best classifiers did not included MIR210HG in their structure (Fig. 6B), which makes sense as this gene does not have any orthologs annotated in mouse (Fig. 6B). Since genes can be connected in complex regulatory networks, a missing variable can affect the efficiency of subsequent nodes in the tree, as we see with other genes in the subset of trees using MIR210HG in Figures 6C and 6D. In spite of this, the combination with other genes allow some MIR210HG-containing trees to achieve reasonable accuracies.

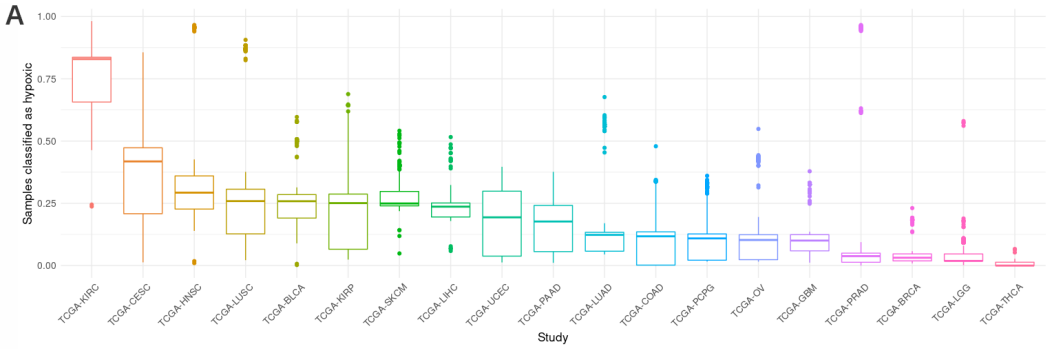

Figure 5: Validation set, RNA-seqs of tumoral samples in TCGA studies. Proportion of TCGA tumor samples classified as hypoxic by the decision trees by primary site.

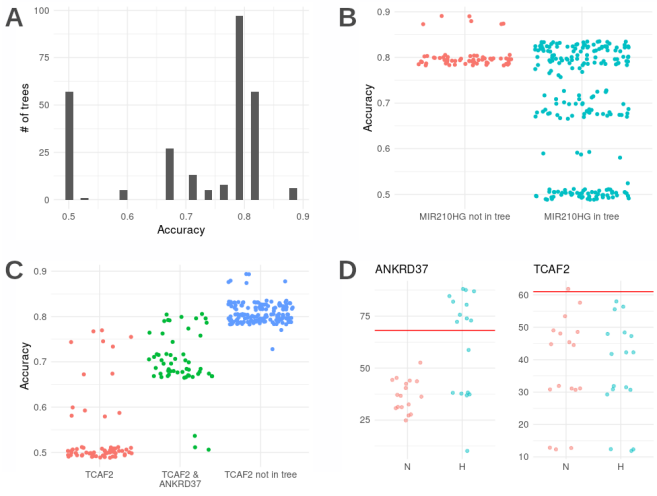

Figure 6: Validation set, murine cells RNA-seqs. A: Histogram of the accuracies of the 276 classification trees in 34 mouse RNA-seq datasets. **B**: Accuracies of the classification trees according to the use of MIR210HG as a predictor variable. **C**: Accuracies of the subset of MIR210HG containing classification trees according to the combined use of TCAF2 and ANKRD37. **D**: Ranking percentiles of ANKRD37 and TCAF2 in the samples forming the mouse dataset. The red line represents the mean split point for each gene in the trees using either of them as a predictor variable.

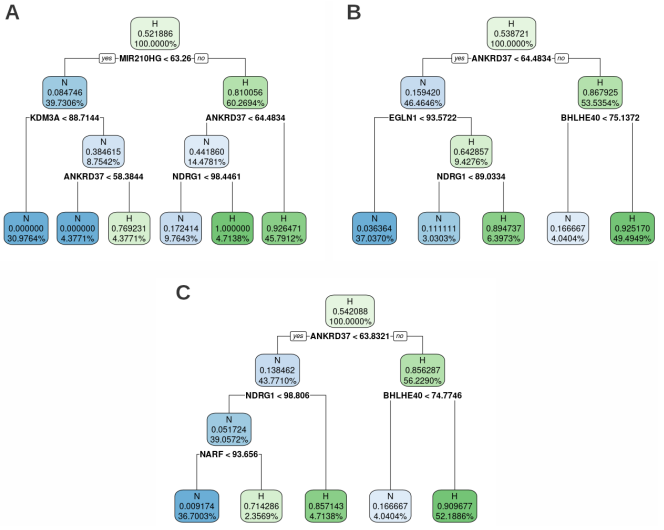

Figure 7: Selected decision trees. The labels in each node indicate the the node's class(N, normoxic, and H, hypoxic), the probability of samples in the node to be classified as hypoxic, and the percentage of samples of the training set in each node. **A**: Overall best performing tree (tree #125). **B**: Best performing tree among those not using MIR210HG (tree #241) **C**: Best performing tree on mouse data (tree #42)

Altogether the results presented above allowed us to identify a subset of three trees that accurately classify samples even from challenging datasets, with Fig. 7A being the best performing tree overall, especially detecting short exposure to hypoxia and mildly low oxygen levels without overestimating the number of hypoxic samples in other validation sets. Since many datasets may not include data on MIR210HG expression, we have also chosen the tree in Fig. 7B, the best performing among those that don't include said gene. Even though both trees perform reasonably well on mouse data, we have selected the additional tree in Fig. 7C for being the best performing in classifying murine samples specifically.

Finally, since it is unlikely that a single tree could accurately classify samples from all potential datasets, we tested whether classification would improve using several trees and generating a consensus. To this end, we compared the three trees selected (Fig. S2A-C) against three ensembles: the whole 276 collection (Fig. S2D) , the 20 trees with higher mean F_1-score (Fig. S2E) and the aforementioned three selected trees (Fig. S2F). The result of this analysis shows that, in general, the ensembles outperform individual trees. However, the difference is small and in some datasets individual datasets performed as well as the ensembles.

Application of the hypoxic classifier to spatial gene expression datasets

Finally, we have applied the classification trees to spatial gene expression datasets. To this end we made use of the transcriptomes derived from adult mouse kidney and different regions of a glioblastoma, prostate and colorectal

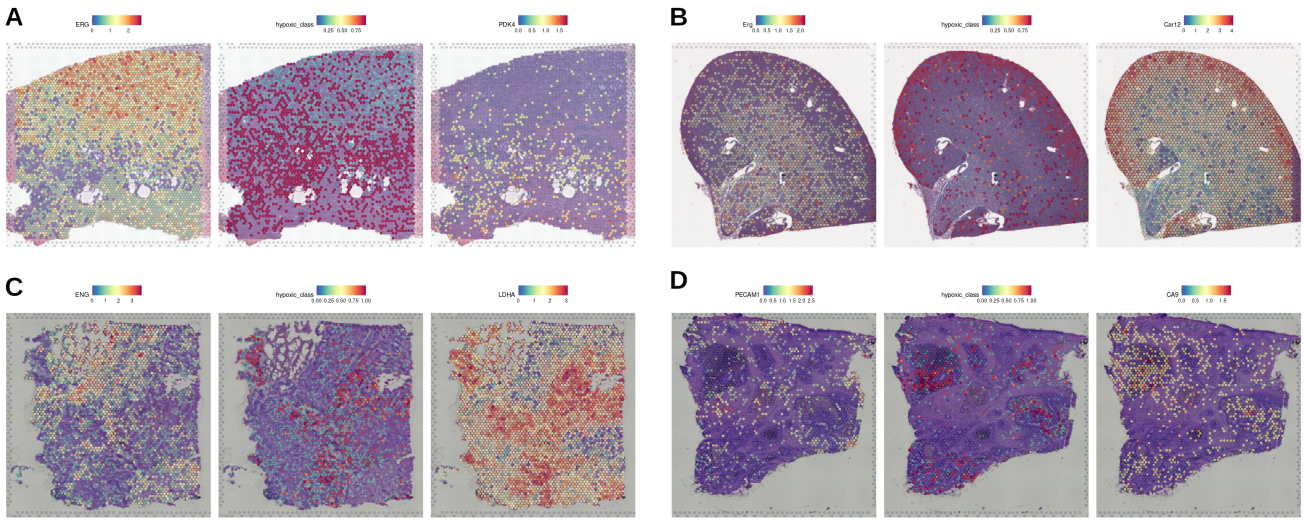

Figure 8: Hypoxia Classifier applied to 10X Genomics Spatial Gene Expression datasets. On the right column, ERG, ENG, and PECAM1 are highlighted as endothelial markers, while on the left PDK4, Car12, LDHA, and CA9 mark regions of active anaerobic glycolysis. The central column represents the probability of each spot to be classified as hypoxic. **A**: Human Prostate Cancer, Adenocarcinoma with Invasive Carcinoma (FFPE). **B**: Adult Mouse Kidney (FFPE). **C**: Human Glioblastoma. **D**: Human Colorectal Cancer.

human cancer specimens publicly available at 10X Genomics web site. Each spot in the samples was classified as normoxic/hypoxic by application of the tree classifier to the local transcriptome. Since the prostate cancer dataset did not include MIR210HG expression, we used tree #241 (Fig. 7B) for spot classification. With the mouse kidney dataset, we used decision tree #42 (Fig. 7C). As for the glioblastoma and colorectal cancer datasets, tree #125 (Fig. 7A) was chosen. As shown in figure 8, in all tumor cases spots classified as hypoxic are excluded from vascularized areas and show markers of active anaerobic glycolysis. In the case of normal kidney, hypoxic spots localize to the cortex known to be relatively hypoxic as compared to other regions [39]. Thus, altogether these results suggest that the tree classifier accurately identify hypoxic regions in spatial gene expression datasets. As further confirmation we proceeded to cluster the spots in each dataset (Supplementary figure S3) and examined differential expression between clusters containing high or low proportion of spots classified as hypoxic (Supplementary table S3-9 and Supplementary figure S4).

Discussion

In this work we aim to generate a gene signature that, in addition to define genes that characterize the response to hypoxia, could be used to of assess if an individual gene expression dataset corresponds to sample that has been exposed to low oxygen tension. This work is based on a meta-analysis of the transcriptomic response to hypoxia, generated through the integration of a corpus of 69 differential expression datasets which included 425 individual RNA-seq experiments from 33 different cell types exposed to different degrees of hypoxia (0.1-5%O_2) for a period of time spanning between 2 and 48h. As a first filter of the variables (genes) to be included in the signature, we selected those widely expressed and significantly up-regulated by hypoxia according to the meta-analysis. This step ensured that the resulting models can be applied to a large variety of tissues as well as minimizing the risks of a biased corpus of publicly available experiments. Then we applied data mining methods to identify sets of genes that best separated normoxic and hypoxic samples using a tree-like decision structure. Although the total number trees that achieved high accuracy was relatively large, only 16 out of the 20 pre-selected genes were required among all the trees, with many having the same structure and differing only slightly in the gene expression threshold. Moreover, the vast majority of trees included different combination of 3-5 genes from the set EGLN1, MIR210HG, NDRG1, ANKRD37, TCAF2, PFKFB3, BHLHE40, and MAFF (Supplementary table S1). In contrast with classical molecular signatures, the trees described herein provide not just a list of genes relevant to the process, but also a set of matching quantitative expression boundaries, which allows it to classify individual samples both from a binary perspective (hypoxic or normoxic sample) as well as a continuous one (probability of a sample to be classified as hypoxic, shown in figure 8). This feature allows the application of the classification trees to a wide range of gene expression datasets, from the conventional bulk RNA-seq by polyA capture and techniques to characterize newly transcribed RNA [40, 41] to spatially resolved transcriptomics and single cell RNA-seq. Importantly, gene's expression boundaries are represented as the percentile occupied by the gene in a ranked list of expression values from a given sample. Thus, this method can be applied to a diverse set of input formats: raw reads, counts per million, FPKM, variance stabilizing transformations, etc.

Should be noted that these classifiers are robust enough to predict the condition of samples from murine cells despite having being trained only with human datasets, as well as identify samples in which response to hypoxia is activated by mutations in specific signalling pathways (ccRCC dataset) or due to the pattern of vascularization and/or oxygen consumption (TCGA datasets, tumor specimens and normal kidney).

In regard to the classification of tumor samples according to their degree of hypoxia, we tested how each classification tree managed to separate tumor samples according to the 9 hypoxic scores calculated in [38] using different hypoxic signatures (Supplementary table S1-8, "TCGA hypoxia signatures"). To measure the performance of the classifier trees we use the location shift of the distribution of hypoxia scores between tumor samples in each class. In spite of its apparent usefulness in evaluating tree performance, there is not a consistent relationship between being able to separate tumor samples with high or low hypoxia scores and properly classifying validation datasets. As shown in figure S1 and table S2, the difference between distributions of hypoxia scores of tumor samples in each class ("location shift") often does not correlate positively with F_1-scores in the different validation sets. Considering all but one of this signatures have been defined in the context of tumor hypoxia, a poor correlation could be expected when comparing to F_1-scores of non-tumoral validation sets, but in the case of clear cell renal carcinoma dataset, 5 out of 9 signatures have is a significant negative correlation between location shift and validation set F_1-score. Since TCGA data does not include tumor oxygenation measures that could be used to verify our classifier's prediction, when selecting the best performing trees we have given more weight to the classification of validation sets than to TCGA hypoxia scores. Nevertheless, as shown in figure 5, with the exception of renal carcinomas, the proportion of tumor samples classified as hypoxic in each group resembles more closely the results obtained by Bhandari et al.

As further confirmation of the trees in comparison to classic gene signatures, we decided to test our trees in spatial gene expression datasets, using as negative control the expression levels of endothelial markers ERG, ENG and PECAM1/CD31 and as positive control, the expression levels of genes related to anaerobic glycolysis. As shown in figure 8, regions classified as hypoxic overlap those of active anaerobic glycolysis, meanwhile regions rich in endothelial markers tend to be classified as normoxic. After unsupervised clustering of the same datasets (Figure S3), differential expression between clusters overlapping normoxic and hypoxic areas highlighted genes linked to hypoxia and not included in our models, such as VEGFA or ENO1 (Supplementary table S3). Furthermore, when comparing the adjusted p-values of genes up-regulated between clusters that are also significantly up-regulated in the cited hypoxia meta-analysis [22] (random effect>0.7 and FDR<0.01) this group has significantly lower p-values than genes not linked to hypoxia, confirming an enrichment on hypoxia-related genes among those differentially expressed between hypoxic and normoxic clusters.

In addition to their remarkable performance, the structure of the decision trees allows for biological interpretation of the prediction's results. In this regard, the application of the decision trees to the challenging datasets provided relevant and novel insights into the underlying biological processes. For example, the analysis of the performance of different trees on the 4sU labeling, GRO-seq, and polysomal datasets suggest that the regulation of MAFF expression by hypoxia could imply biological processes other than transcription, such as mRNA translation or decay. Whereas its application to the ccRCC dataset revealed that BHLHE40 and NDRG1 are expressed at high levels in renal tissue which can hint to specific functions of these genes in kidney physiology. As seen with the mouse dataset in figure 6, missing data in one of the classifying variables (MIR210HG) could directly or indirectly hinder the performance of the trees. Thus, we tested if performance of the tree classifiers can be improved by generating a consensus. As we show in figure S2, an ensemble of the 20 trees with higher mean F_1-score (Fig. S2D) can overperform all individual trees and other ensembles in most cases (with the exception of tree #42 in the mouse dataset). A classification based just in the consensus of the three trees selected in this paper (Fig. S2F) can compensate for the shortcomings of each individual model while maintaining the ease of use intended for this project. Tree ensembles could be a better suited alternative for samples that are harder to classify or derived from a dataset very distantly related to the ones used to derive our tree classifiers.

In summary, herein we describe a ensemble of tree gene signatures that can be easily implemented to identify hypoxic samples based on their transcriptomic profile without the need for a reference. Given the importance of oxygen homeostasis in physiology and disease, this tool could be of help in a wide variety of research and clinical settings. Finally, in the view of its merits, we proposed the use of this method to define gene signatures, in contrast to the simple gene lists, to define other cellular processes.

The full collection of 276 decision trees, the validation datasets used, as well as a step-by-step guide to apply the classifiers to new data are available at the following GitHub repository: `www.github.com/LauraPS1/Hypoxia_Classifier`.

Acknowledgments

The results shown here are in part based upon data generated by the TCGA Research Network: `www.cancer.gov/tcga`.

This work was supported by Grants SAF2017-88771-R and PID2020-118821RB-I00 funded by MCIN/AEI/10.13039/501100011033 and by "ERDF A way of making Europe" and by grant IND2019/BMD-17134 funded by Autonomous Autonomous Community of Madrid.

References

[1] G. L. Semenza, Hypoxia-inducible factors in physiology and medicine, Cell 148 (2012) 399–408.

[2] G. P. Elvidge, L. Glenny, R. J. Appelhoff, P. J. Ratcliffe, J. Ragoussis, J. M. Gleadle, Concordant regulation of gene expression by hypoxia and 2-oxoglutarate-dependent dioxygenase inhibition: the role of HIF-1alpha, HIF-2alpha, and other pathways, J Biol Chem 281 (2006) 15215–15226.

[3] G. Yang, R. Shi, Q. Zhang, Hypoxia and oxygen-sensing signaling in gene regulation and cancer progression, International Journal of Molecular Sciences 21 (2020).

[4] O. A. Mesarwi, R. Loomba, A. Malhotra, Obstructive sleep apnea, hypoxia, and nonalcoholic fatty liver disease, American Journal of Respiratory and Critical Care Medicine 199 (2019) 830–841.

[5] F. M. Buffa, A. L. Harris, C. M. West, C. J. Miller, Large meta-analysis of multiple cancers reveals a common, compact and highly prognostic hypoxia metagene, British Journal of Cancer 102 (2010) 428–435.

[6] S. C. Winter, F. M. Buffa, P. Silva, C. Miller, H. R. Valentine, H. Turley, K. A. Shah, G. J. Cox, R. J. Corbridge, J. J. Homer, B. Musgrove, N. Slevin, P. Sloan, P. Price, C. M. West, A. L. Harris, Relation of a hypoxia metagene derived from head and neck cancer to prognosis of multiple cancers, Cancer Research 67 (2007) 3441–3449.

[7] H. B. Ragnum, L. Vlatkovic, A. K. Lie, K. Axcrona, C. H. Julin, K. M. Frikstad, K. H. Hole, T. Seierstad, H. Lyng, The tumour hypoxia marker pimonidazole reflects a transcriptional programme associated with aggressive prostate cancer, British Journal of Cancer 112 (2015) 382–390.

[8] A. Eustace, N. Mani, P. N. Span, J. J. Irlam, J. Taylor, G. N. Betts, H. Denley, C. J. Miller, J. J. Homer, A. M. Rojas, P. J. Hoskin, F. M. Buffa, A. L. Harris, J. H. Kaanders, C. M. West, A 26-gene hypoxia signature predicts benefit from hypoxia-modifying therapy in laryngeal cancer but not bladder cancer, Clinical Cancer Research 19 (2013) 4879–4888.

[9] B. S. Sørensen, K. Toustrup, M. R. Horsman, J. Overgaard, J. Alsner, Identifying ph independent hypoxia induced genes in human squamous cell carcinomas in vitro, Acta Oncologica 49 (2010) 895–905.

[10] G. P. Elvidge, L. Glenny, R. J. Appelhoff, P. J. Ratcliffe, J. Ragoussis, J. M. Gleadle, Concordant regulation of gene expression by hypoxia and 2-oxoglutarate-dependent dioxygenase inhibition: The role of hif-1α, hif-2α, and other pathways*, Journal of Biological Chemistry 281 (2006) 15215–15226.

[11] Z. Hu, C. Fan, C. Livasy, X. He, D. S. Oh, M. G. Ewend, L. A. Carey, S. Subramanian, R. West, F. Ikpatt, O. I. Olopade, M. van de Rijn, C. M. Perou, A compact VEGF signature associated with distant metastases and poor outcomes, BMC Medicine 7 (2009).

[12] R. Seigneuric, M. H. Starmans, G. Fung, B. Krishnapuram, D. S. Nuyten, A. van Erk, M. G. Magagnin, K. M. Rouschop, S. Krishnan, R. B. Rao, C. T. Evelo, A. C. Begg, B. G. Wouters, P. Lambin, Impact of supervised gene signatures of early hypoxia on patient survival, Radiotherapy and Oncology 83 (2007) 374–382.

[13] R. Leinonen, H. Sugawara, M. Shumway, The sequence read archive, Nucleic Acids Research 39 (2011) 148–162.

[14] R. Patro, G. Duggal, M. I. Love, R. A. Irizarry, C. Kingsford, Salmon provides fast and bias-aware quantification of transcript expression, Nature Methods 14 (2017) 417–419.

[15] N. A. O'Leary, M. W. Wright, J. R. Brister, S. Ciufo, D. Haddad, R. McVeigh, B. Rajput, B. Robbertse, B. Smith-White, D. Ako-Adjei, A. Astashyn, A. Badretdin, Y. Bao, O. Blinkova, V. Brover, V. Chetvernin, J. Choi, E. Cox, O. Ermolaeva, C. M. Farrell, T. Goldfarb, T. Gupta, D. Haft, E. Hatcher, W. Hlavina, V. S. Joardar, V. K. Kodali, W. Li, D. Maglott, P. Masterson, K. M. McGarvey, M. R. Murphy, K. O'Neill, S. Pujar, S. H. Rangwala, D. Rausch, L. D. Riddick, C. Schoch, A. Shkeda, S. S. Storz, H. Sun, F. Thibaud-Nissen, I. Tolstoy, R. E. Tully, A. R. Vatsan, C. Wallin, D. Webb, W. Wu, M. J. Landrum, A. Kimchi, T. Tatusova, M. DiCuccio, P. Kitts, T. D. Murphy, K. D. Pruitt, Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation, Nucleic Acids Research 44 (2016) D733–D745.

[16] X. Genomics, Human prostate cancer, adenocarcinoma with invasive carcinoma (ffpe), spatial gene expression dataset by space ranger 1.3.0, 2021.

[17] X. Genomics, Adult mouse kidney (ffpe), spatial gene expression dataset by space ranger 1.3.0, 2021.

[18] X. Genomics, Human colorectal cancer: Whole transcriptome analysis, spatial gene expression dataset by space ranger 1.2.0, 2020.

[19] X. Genomics, Human glioblastoma: Whole transcriptome analysis, spatial gene expression dataset by space ranger 1.2.0, 2020.

[20] C. Hafemeister, R. Satija, Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression, Genome Biology 20 (2019).

[21] Y. Hao, S. Hao, E. Andersen-Nissen, W. M. M. III, S. Zheng, A. Butler, M. J. Lee, A. J. Wilk, C. Darby, M. Zagar, P. Hoffman, M. Stoeckius, E. Papalexi, E. P. Mimitou, J. Jain, A. Srivastava, T. Stuart, L. B. Fleming, B. Yeung, A. J. Rogers, J. M. McElrath, C. A. Blish, R. Gottardo, P. Smibert, R. Satija, Integrated analysis of multimodal single-cell data, Cell (2021).

[22] L. Puente-Santamaria, L. Sanchez-Gonzalez, B. P. Gonzalez-Serrano, N. Pescador, O. H. Martinez-Costa, R. Ramos-Ruiz, L. del Peso, Formal meta-analysis of hypoxic gene expression profiles reveals a universal gene signature and cell type-specific effects, bioRxiv (2021).

[23] A. Liaw, M. Wiener, Classification and regression by randomforest, R News 2 (2002) 18–22.

[24] T. Therneau, B. Atkinson, rpart: Recursive Partitioning and Regression Trees, 2019. URL: https://CRAN.R-project.org/package=rpart, r package version 4.1-15.

[25] W. G. Kaelin, P. J. Ratcliffe, Oxygen sensing by metazoans: The central role of the hif hydroxylase pathway, Molecular Cell 30 (2008) 393–402.

[26] X. Huang, Q. T. Le, A. J. Giaccia, MiR-210 - micromanager of the hypoxia pathway, Trends in Molecular Medicine 16 (2010) 230–237.

[27] H. Cangul, Hypoxia upregulates the expression of the NDRG1 gene leading to its overexpression in various human cancers, BMC Genetics 5 (2004).

[28] Y. Benita, H. Kikuchi, A. D. Smith, M. Q. Zhang, D. C. Chung, R. J. Xavier, An integrative genomics approach identifies Hypoxia Inducible Factor-1 (HIF-1)-target genes that form the core response to hypoxia, Nucleic Acids Research 37 (2009) 4587–4602.

[29] K. Saeki, H. Onishi, S. Koga, S. Ichimiya, K. Nakayama, Y. Oyama, M. Kawamoto, K. Sakihama, T. Yamamoto, R. Matsuda, Y. Miyasaka, M. Nakamura, Y. Oda, Fam115c could be a novel tumor suppressor associated with prolonged survival in pancreatic cancer patients, J Cancer 11 (2020) 2289–2302.

[30] M. Obach, Àurea Navarro-Sabaté, J. Caro, X. Kong, J. Duran, M. Gómez, J. C. Perales, F. Ventura, J. L. Rosa, R. Bartrons, 6-phosphofructo-2-kinase (pfkfb3) gene promoter contains hypoxia-inducible factor-1 binding sites necessary for transactivation in response to hypoxia*, Journal of Biological Chemistry 279 (2004) 53562–53570.

[31] A. V. Ivanova, S. V. Ivanov, A. Danilkovitch-Miagkova, M. I. Lerman, Regulation of stra13 by the von hippel-lindau tumor suppressor protein, hypoxia, and the ubc9/ubiquitin proteasome degradation pathway*, Journal of Biological Chemistry 276 (2001) 15306–15315.

[32] L. Chen, T. Fink, P. Ebbesen, V. Zachar, Temporal transcriptome of mouse atdc5 chondroprogenitors differentiating under hypoxic conditions, Experimental Cell Research 312 (2006) 1727–1744.

[33] J. Klomp, J. Hyun, J. E. Klomp, K. Pajcini, J. Rehman, A. B. Malik, Comprehensive transcriptomic profiling reveals sox7 as an early regulator of angiogenesis in hypoxic human endothelial cells, Journal of Biological Chemistry 295 (2020) 4796–4808.

[34] S. R. McKeown, Defining normoxia, physoxia and hypoxia in tumours - Implications for treatment response, British Journal of Radiology 87 (2014).

[35] T. Löfstedt, E. Fredlund, L. Holmquist-Mengelbier, A. Pietras, M. Ovenberger, L. Poellinger, S. Påhlman, Hypoxia inducible factor-2α in cancer, Cell Cycle 6 (2007) 919–926. PMID: 17404509.

[36] G. Guo, Y. Gui, S. Gao, A. Tang, X. Hu, Y. Huang, W. Jia, Z. Li, M. He, L. Sun, P. Song, X. Sun, X. Zhao, S. Yang, C. Liang, S. Wan, F. Zhou, C. Chen, J. Zhu, X. Li, M. Jian, L. Zhou, R. Ye, P. Huang, J. Chen, T. Jiang, X. Liu, Y. Wang, J. Zou, Z. Jiang, R. Wu, S. Wu, F. Fan, Z. Zhang, L. Liu, R. Yang, X. Liu, H. Wu, W. Yin, X. Zhao, Y. Liu, H. Peng, B. Jiang, Q. Feng, C. Li, J. Xie, J. Lu, K. Kristiansen, Y. Li, X. Zhang, S. Li, J. Wang, H. Yang, Z. Cai, J. Wang, Frequent mutations of genes encoding ubiquitin-mediated proteolysis pathway components in clear cell renal cell carcinoma, Nature Genetics 44 (2012) 17–19.

[37] P. Chittiboina, R. R. Lonser, Chapter 10 - von hippel–lindau disease, in: M. P. Islam, E. S. Roach (Eds.), Neurocutaneous Syndromes, volume 132 of *Handbook of Clinical Neurology*, Elsevier, 2015, pp. 139–156. URL: https://www.sciencedirect.com/science/article/pii/B978044462702500010X. doi:https://doi.org/10.1016/B978-0-444-62702-5.00010-X.

[38] V. Bhandari, C. Hoey, L. Y. Liu, E. Lalonde, J. Ray, J. Livingstone, R. Lesurf, Y. J. Shiah, T. Vujcic, X. Huang, S. M. Espiritu, L. E. Heisler, F. Yousif, V. Huang, T. N. Yamaguchi, C. Q. Yao, V. Y. Sabelnykova, M. Fraser, M. L. Chua, T. van der Kwast, S. K. Liu, P. C. Boutros, R. G. Bristow, Molecular landmarks of tumor hypoxia across cancer types, Nature Genetics 51 (2019) 308–318.

[39] R. G. Evans, D. W. Smith, C.-J. Lee, J. P. Ngo, B. S. Gardiner, What makes the kidney susceptible to hypoxia?, The Anatomical Record 303 (2020) 2544–2552.

[40] A. Garibaldi, F. Carranza, K. J. Hertel, Isolation of newly transcribed rna using the metabolic label 4-thiouridine, Methods in Molecular Biology 1648 (2017) 169–176.

[41] A. Gardini, Global run-on sequencing (GRO-Seq), Methods in Molecular Biology 1468 (2017) 111–120.

Supplementary Information

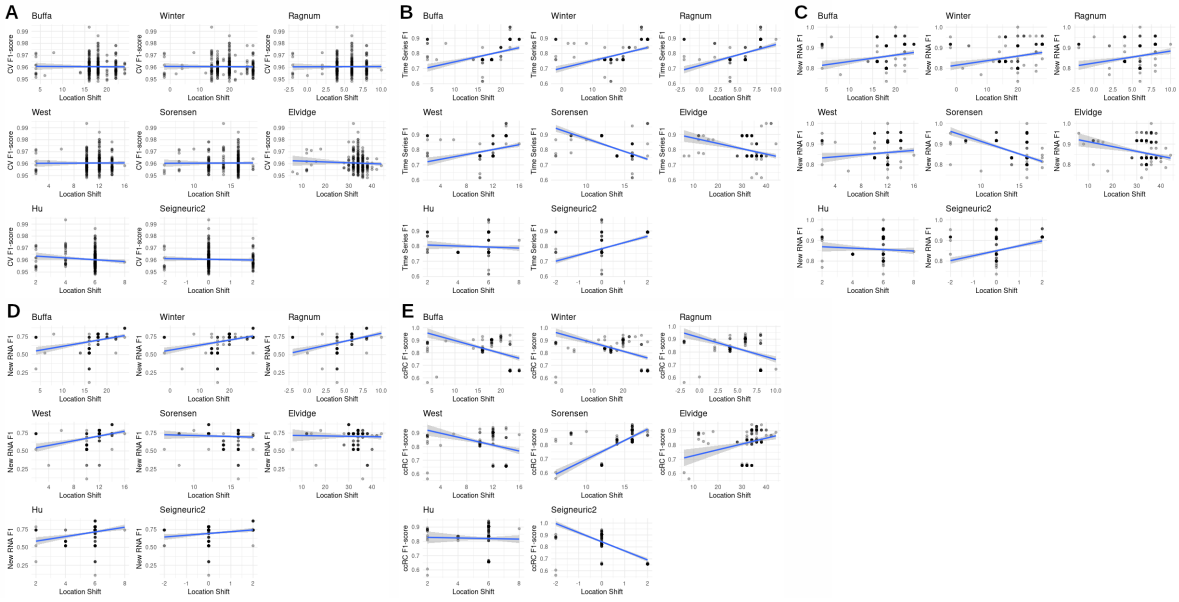

Figure S1: Correlation between two validation measurements. Validation set F1 score and location shift in TCGA's hypoxia signatures score. **A**:Cross-validation. **B**:PRJNA561635 time series. **C**: Newly transcribed RNA. **D**: Mouse RNA. **E**: ccRCC with VHL mutations.

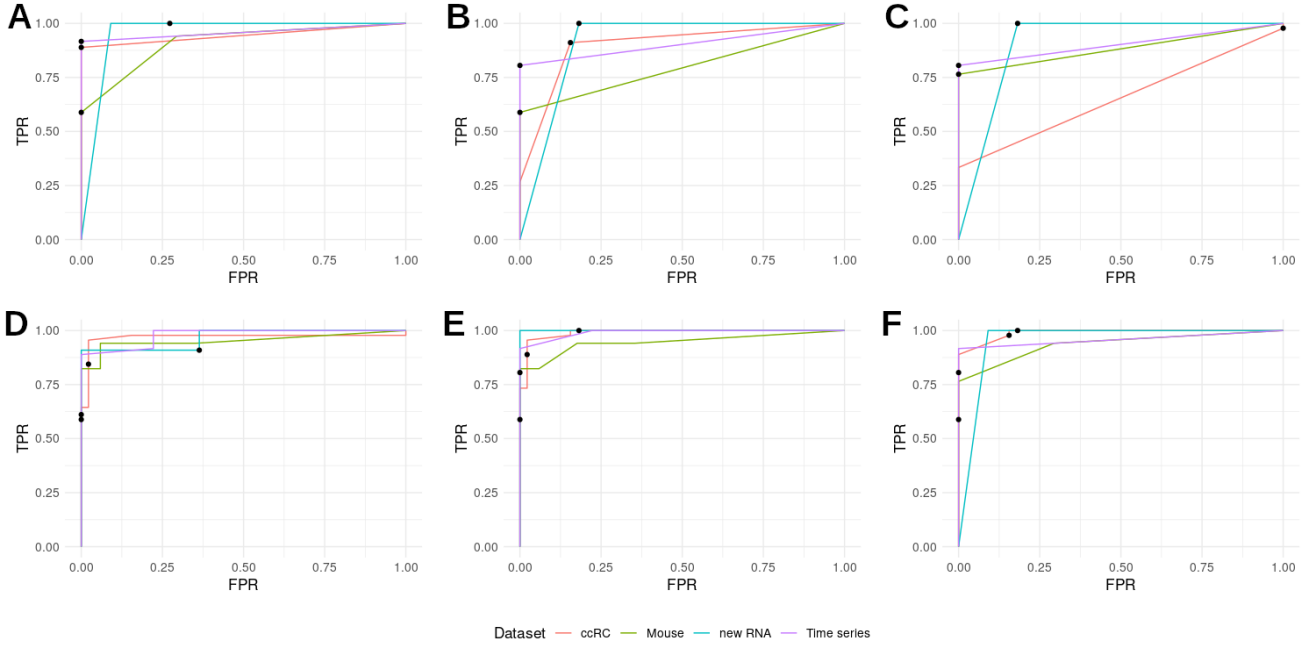

Figure S2: Tree ensembles performance. ROC curves for individual trees and three ensembles using four validation datasets. A sample was classified as hypoxic when the mean probability given by the trees surpassed a given threshold between 0 and 1. Black dots represent TPR/FPR values for a probability threshold of 0.5 to classify a sample as hypoxic. **A-C**: ROC curves for the individual trees selected for their performance. **A**: Tree #125 (Fig. 7A). **B**: Tree #241 (Fig. 7B). **C**: Tree #42 (Fig. 7C). **D-E**: ROC curves for three tree ensembles. **D**: all 276 trees. **E**: top 20 trees by medium F_1-score. **F**: three selected trees (#125, #241, #42).

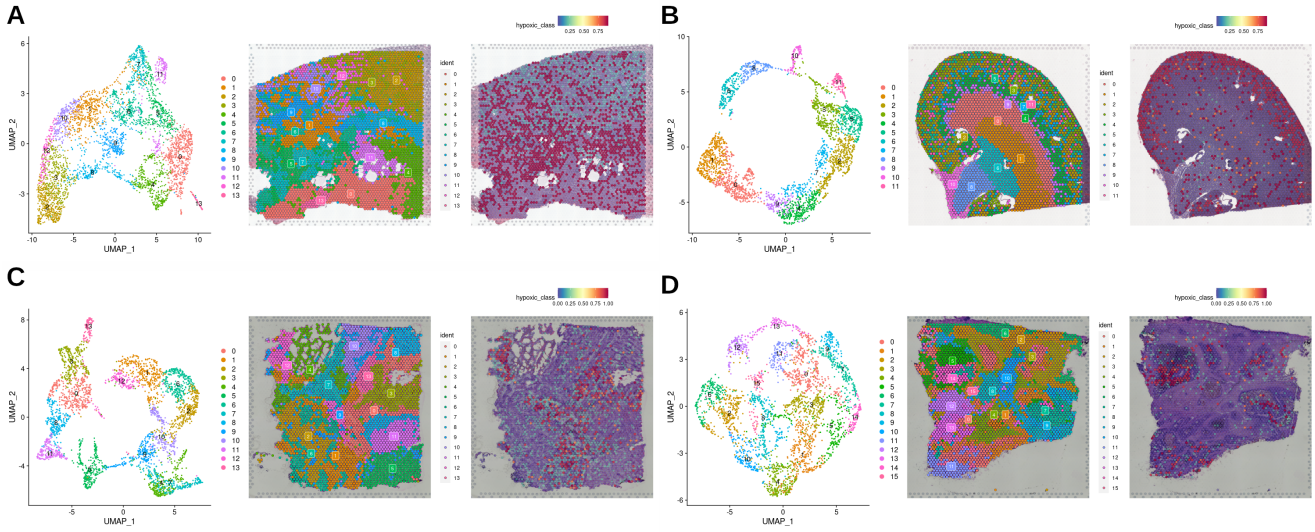

Figure S3: Unsupervised clustering matches areas classified as hypoxic. UMAP representations (first column) and microscopy overlay (second column) of spot clustering using a shared nearest neighbor (SNN) based algorithm. Spots marked as hypoxic by the decision trees group together and "hypoxic" clusters are positioned close by in the UMAP space. **A**: Human Prostate Cancer, Adenocarcinoma with Invasive Carcinoma (FFPE). **B**: Adult Mouse Kidney (FFPE). **C**: Human Glioblastoma. **D**: Human Colorectal Cancer.

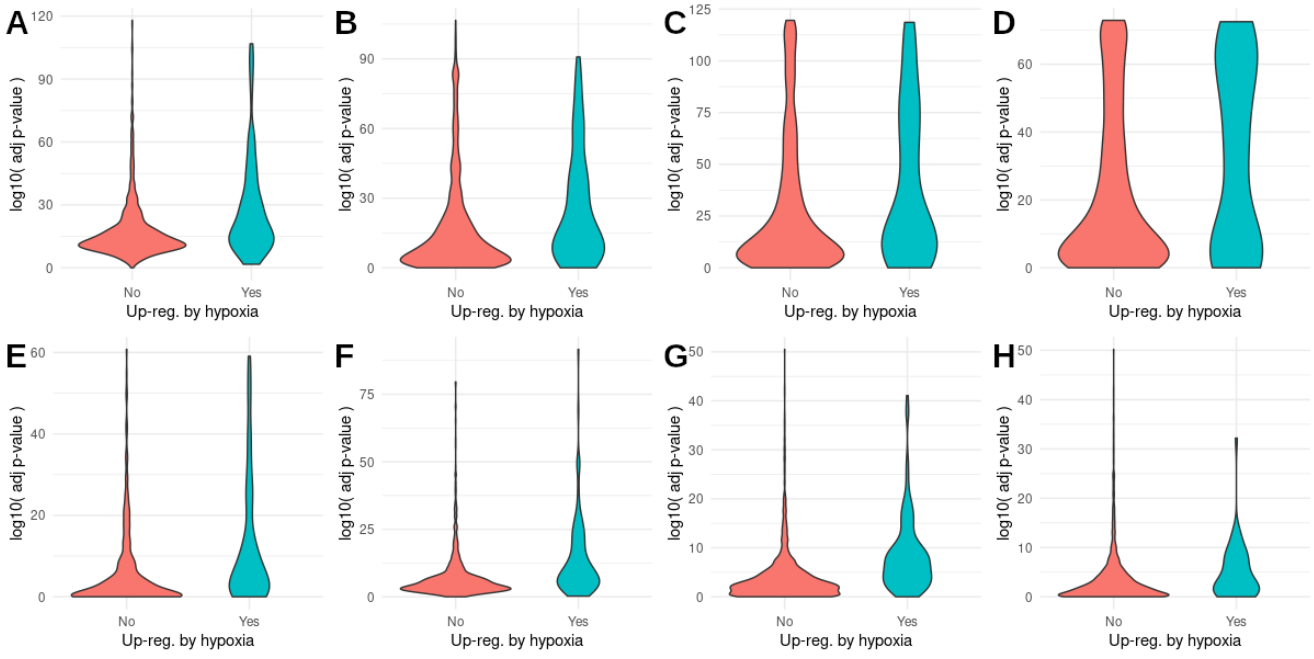

Figure S4: Hypoxic genes DE between Visium dataset clusters. -Log_{10}(FDR adjusted p-values) of genes significantly up-regulated between clusters, according to their up-regulation status in the previously cited hypoxia meta-analysis. **A**: Prostate cancer, C7 vs C0. **B**: Prostate cancer, C11 vs C2. **C**: Mouse kidney, C3 vs C1. **D**: Mouse kidney, C5 vs C6.**E**: Glioblastoma, C12 vs C8.**F**: Glioblastoma, C1 vs C5 . **G**: Colorectal cancer, C14 vs C2. **H**: Colorectal cancer, C7 vs C8.

	Cross-validation		PRJNA561635		RNA fractions		Mouse RNA		VHL mut ccRC	
Signature	r	p-value	r	p-value	r	p-value	r	p-value	r	p-value
Buffa	-0.01	8.64E-01	0.4	5.81E-12	0.23	1.49E-04	0.23	1.49E-04	-0.41	1.27E-12
Winter	-0.01	8.79E-01	0.49	4.17E-18	0.27	5.70E-06	0.27	5.70E-06	-0.47	3.22E-16
Ragnum	0	9.37E-01	0.46	7.20E-16	0.23	9.49E-05	0.23	9.49E-05	-0.38	4.18E-11
West	0.01	8.85E-01	0.31	9.44E-08	0.12	4.11E-02	0.12	4.11E-02	-0.29	9.69E-07
Sorensen	0.01	8.99E-01	-0.62	1.04E-30	-0.54	6.16E-22	-0.54	6.16E-22	0.65	1.15E-34
Elvidge	-0.05	3.76E-01	-0.29	1.01E-06	-0.23	9.93E-05	-0.23	9.93E-05	0.23	1.51E-04
Hu	-0.11	6.45E-02	-0.05	3.74E-01	-0.07	2.79E-01	-0.07	2.79E-01	-0.02	7.21E-01
Seigneuric2	-0.04	5.23E-01	0.58	3.41E-26	0.41	1.87E-12	0.41	1.87E-12	-0.72	2.96E-46
Buffa intra-tumour	0.09	1.53E-01	-0.22	1.87E-04	-0.29	7.49E-07	-0.29	7.49E-07	0.45	5.47E-15

Table S2

TCGA hypoxia signature scores and validation sets F1-scores. Pearson's correlation coefficients and FDR adjusted p-values between validation sets F1-scores and location shift in hypoxia scores between TCGA samples classified as hypoxic or normoxic

Table S1

Classification tree validation. **Sheet 1**: metadata for each of the experiments used in the validation process. **Sheet 2**: summary of tree variables and validation accuracy. **Sheets 3-6**: performance measurements for each human validation dataset. **Sheet 7**: Proportion of TCGA tumor samples classified as hypoxic by primary site. **Sheet 8**: difference in previously calculated hypoxia scores between samples classified as hypoxic and normoxic by each tree. **Sheet 9**: performance measurements for the mouse validation dataset.

Table S3

Differential expression in unsupervised clustering of spatial gene expression datasets: Differentially expressed genes between clusters containing a high proportion of spots classified as hypoxic and clusters with low or no hypoxic spots. **Sheet 1**: Prostate cancer, C7 vs C0. **Sheet 2**: Prostate cancer, C11 vs C2. **Sheet 3**: Mouse kidney, C3 vs C1. **Sheet 4**: Mouse kidney, C5 vs C6.**Sheet 5**: Glioblastoma, C12 vs C8.**Sheet 6**: Glioblastoma, C1 vs C5 . **Sheet 7**: Colorectal cancer, C14 vs C2. **Sheet 8**: Colorectal cancer, C7 vs C8. **Sheet 9**: Mann–Whitney test results for differences in p-values among genes up-regulated between clusters according to their up-regulation status in our hypoxia meta-analysis (random effect>0.7, FDR<0.01).