A workflow combining single-cell CRISPRi screening and a supervised autoencoder neural network to detect subtle transcriptomic perturbations induced by lncRNA Knock-Down

Marin Truchi[1‡], Caroline Lacoux[1‡], Cyprien Gille[2], Julien Fassy[1], Virginie Magnone[1], Rafael Lopez-Goncalvez[1], Cédric Girard-Riboulleau[1], Iris Manosalva-Pena[3], Marine Gautier-Isola[1], Salvatore Spicuglia[3], Georges Vassaux[1], Roger Rezzonico[1], Michel Barlaud[2‡], Bernard Mari*[1‡]

1 Université Côte d'Azur, IPMC, CNRS UMR7275, Valbonne, France
2 Université Côte d'Azur, I3S, CNRS UMR7271, Sophia Antipolis, France
3 Université Aix-Marseille, Inserm, TAGC, UMR1090, Marseille, France

‡These authors contributed equally to this work.
* Bernard.Mari@unice.fr

Abstract

Recent advances in cancer genomics have highlighted aberrant expression of various families of non-coding RNAs in all cancer types, including lung adenocarcinomas (LUAD). Here we aim to better understand the functions of long non coding RNAs (lncRNAs) regulated by the hypoxic response in LUAD cells, conditions that promote tumor aggressiveness and drug resistance. We performed a single-cell CRISPR-interference-based (CRISPRi) transcriptome screening (CROP-Seq) for HIF1A, HIF2, and a subset of lncRNA candidates regulated by hypoxia and/or potentially associated with LUAD prognosis. The mini-CROP-seq library of validated guides RNA (gRNA) was amplified and transduced in A549 LUAD cells cultured in normoxia or exposed to hypoxic conditions during 3, 6 or 24 hours. To overcome the challenge of detecting subtle gRNA-induced transcriptomic perturbation and classifying the most responsive cells, we used a new supervised autoencoding neural networks method (SAE), leveraging on both transcriptomic data and cell labels corresponding to known received gRNA. We first validated the SAE approach on HIF1A and HIF2 by confirming the specific effect of their knock-down during the temporal switch of the hypoxic response. Next, the SAE method was able to detect stable short hypoxia-dependent transcriptomic signatures induced by the knock-down of some lncRNA candidates, outperforming previously published machine learning approaches. This proof of concept demonstrates the relevance of the SAE approach for deciphering weak perturbations in single-cell transcriptomic data readout as part of CRISPR-based screening.

Introduction

Cancer cells in solid tumors, often suffer from hypoxic stress and adapt to this micro-environment via the activation of Hypoxia inducible factor (HIF), a heterodimeric transcription factor composed of either HIF-1α or HIF-2α (initially identified as endothelial PAS domain protein (EPAS1)) and HIF-1β /ARNT subunits [1–3]. In normoxia, HIFα is continuously degraded by an ubiquitin–dependent mechanism

mediated by interaction with to the von Hippel–Lindau (VHL) protein. Hydroxylation of proline residues in HIFα is necessary for VHL binding and is catalyzed by the α-ketoglutarate-dependent dioxygenases prolyl hydroxylases (PHD). During hypoxia, PHDs are inactive, leading to HIF-α stabilization, dimerization with HIF-1β and finally translocation into the nucleus to bind to E-box-like hypoxia response elements (HREs) within the promoter region of a wide range of genes that control cellular oxygen homeostasis, erythrocyte production, angiogenesis and mitochondrial metabolism [4]. These molecular changes are notably crucial for cells to adapt to stress by lowering oxygen consumption by shifting from oxidative metabolism to glycolysis. While HIF-1 and HIF-2 bind to the same HRE consensus sequence, they are non-redundant and have distinct target genes and mechanisms of regulation. It is generally accepted that the individual HIFs have specific temporal and functional roles during hypoxia, known as the HIF switch, with HIF-1 driving the initial response and HIF-2 directing the chronic response [5]. In most solid tumors, including lung adenocarcinoma (LUAD), the degree of hypoxia is associated with poor clinical outcome. Induction of HIF activity upregulates genes involved in many hallmarks of cancer, including metabolic reprogramming, epithelial-mesenchymal transition (EMT), invasion and metastasis, apoptosis, genetic instability and resistance to therapies.

Emerging evidence have highlighted that hypoxia regulates expression of a wide number of non-coding RNAs classes including microRNAs (miRNAs) and long non-coding RNAs (lncRNAs) that in turn are able to influence the HIF-mediated response to hypoxia [6–8]. LncRNAs constitute a heterogeneous class of transcripts which are more than 200 nt long with low or no protein coding potential, such as intergenic and antisense RNAs, transcribed ultraconserved regions (T-UCR) as well as pseudogenes. Recent advances in cancer genomics have highlighted aberrant expression of a wide set of lncRNAs [9], revealing their roles in regulating the genome at several levels, including genomic imprinting, chromatin state, transcription activation or repression, splicing and translation control [10]. LncRNAs can regulate gene expression through different mechanisms, as guide, decoy, scaffold, miRNA sponges or micropeptides. Of note, recent studies demonstrated the role of several lncRNAs in the direct and indirect regulation of HIF expression and pathway through diverse mechanisms [7]. Moreover, hypoxia-responsive lncRNAs have been shown to play regulatory functions in pathways associated with the hallmarks of cancer. For instance, the hypoxia-induced Nuclear-Enriched Abundant Transcript 1 (NEAT1) lncRNA has been associated with the formation of nuclear structures called paraspeckles during hypoxia as well as an increased clonogenic survival of breast cancer cells. Another highly studied lncRNA, Metastasis-Associated Lung Adenocarcinoma Transcript 1 (MALAT1, also known as NEAT2) has been found upregulated by hypoxia in LUAD A549 cells and associated with various cellular functions depending on tumor cell types including cell death, proliferation, migration and invasion [11]. Starting from an expression screening in LUAD patients samples and cell lines subjected to hypoxia, we have characterized a new nuclear hypoxia-regulated transcript from the Lung Cancer Associated Transcript (LUCAT1) locus associated with patient prognosis and involved in redox signaling with implication for drug resistance [12]. Additional promising lncRNA candidates regulated by hypoxia and/or associated with bad prognosis have been identified but deciphering the regulatory functions of these poorly annotated transcripts remains a major challenge. Pooled screening approaches using CRISPR-based technology have offered the possibility to evaluate mammalian gene function, including lncRNAs at genome scale levels [13]. More recently, they have been applied to cancer cell lines and have confirmed the oncogenic or tumor suppressor roles of some lncRNA [14]. This strategy is able to test a large number of candidates simultaneously but require well identified phenotypes such as cell proliferation, cell

viability, or cell migration. More subtle screens require techniques based on transcriptomic signatures [15] and approaches have been developed to combine CRISPR gene manipulation, including CRISPR interference and single-cell RNA-seq (scRNA-seq) based on droplet isolation, such as Perturb-seq [16], CROP-seq [17] and ECCITE-seq [18]. These methods combine the advantages of screening a large number of genes simultaneously and linking the modifications to the transcriptomic phenotype, all by breaking down the perturbation signal cell by cell [10, 16].

Deep neural networks have proven their efficiency for classification and feature selection in many domains, but have also been applied to omics data analyses [19, 20]. Among the proposed neural networks architectures, autoencoders are able to learn a representation of the data, typically in a latent space of lower dimension than the input space. As such, they are often used for dimensionality reduction [21] and have applications in the medical field as data denoisers or relevant feature selectors [22–24]. A widely used type of autoencoders is the Variational Autoencoder (VAE) [25]. This VAE adds the assumption that the encoded data follows a prior gaussian distribution, and thus combines the reconstruction loss with a distance function (between the gaussian prior and the actual learned distribution). For example, VAE has been applied to scRNA-seq to predict cell response to biological perturbations [26].

In the present work, we have developed a single-cell CRISPR-interference-based (CRISPRi) transcriptome screening based on the CROP-Seq approach to gain insight on the regulatory functions of hypoxia-regulated lncRNAs. As a proof-of-concept, a mini-CROP-seq library, including validated guide RNAs (gRNAs) targeting six previously identified lncRNA regulated by hypoxia and/or associated with bad prognosis [12] a well as the two master transcription factors of the hypoxic response (HIF1A and HIF2/EPAS1) and negative control guides, was used. To optimize analysis of fine-tuned regulations in this dataset, we have developed a supervised autoencoder (SAE) neural network [27], where we relax the parametric distribution assumption of classical VAE. It leverages on the known cell labels, corresponding to the received gRNA, and a classification loss to incite the latent space to fit the true data distribution. We first validated the approach on HIF1 and HIF2/EPAS1 knock-down, showing a good sensitivity to detect the known temporal switch between both regulators. We then applied the SAE to the cells treated with the different hypoxia-regulated lncRNA gRNAs to identify subtle signatures linked to the knock-down of the lncRNAs.

Materials and methods

Lentivirus production

Lentiviruses were produced using a standard Lipofectamine 2000TM transfection protocol, using one million HEK293 cells seeded in a 25 cm2 flask in DMEM medium supplemented with 10% bovine serum. A mixture of four plasmids (3 µg pMDLg/pRRE (addgene "12251"), 1.4 µg pRSV-Rev (addgene "12253"), 2 µg pVSV-G (addgene "12259") and 2.5 µg of the plasmid containing the expression cassette to package or the pooled CROP-seq guides) was transfected. Forty-eight hours later, the medium was collected, centrifuged for 5 minutes at 3000 rpm, and 2.5 mL supernatant containing the viral particles was collected and used to infect cells or aliquoted and stored at -80°C. Large scale preparations of lentivirus were produced at the Vectorology facility, PVM, Biocampus (CNRS UMS3426), Montpellier, France.

Generation of dCas9-expressing A549 cell line

The lung adenocarcinoma cell line A549 was infected with a lentivirus produced from the plasmid lenti- dCas9-KRAB-MeCP2 (a gift from Andrea Califano, addgene 122205) allowing the expression of a fusion protein MeCP2-KRAB-dCas9 and a gene conferring resistance to blasticidin. Infected cells were then grown in the presence of 10 µg/mL of blasticidin (Sigma). Selection of A549-KRAB-MeCP2 cells was complete within 3 to 5 days. Bulk blasticidin positive cells were amplified and cloned for the CRISPRi scRNA seq experiments. The best clone was selected according to the expression level of MeCP2-KRAB-dCas9 mRNA and to the most effective inhibition of NLUCAT1 using the NLUCAT1 sg3 RNA.

Cloning of individual guides in the CROPseq-Guide-Puro plasmid

The plasmid CROPseq-Guide-Puro (Datlinger et al. Nat Methods 2017) (a gift from C Bock, Addgene plasmid 86708) was digested using the restriction enzyme BsmBI (NEB R0580) for 2h at 50°C. The relevant fragments (around 8 kB) were gel-purified using the Qiagen Gel purification kit and stored at –20°C in 20-fmol aliquots. Guides against the targeted genes were cloned using the Gibson assembly method (NEBuilder HiFi DNA Assembly Master Mix, NEB E2621). Aliquoted, BsmBI-digested plasmid was mixed with 0.55 µL guide oligonucleotide (200nM, see Supplemental Table) in 10µl total volume, combined with 10µl 2X NEBuilder HiFi Assembling Master mix and the mixture was incubated at 50°C for 20 minutes. 8µL of NEBuilder Assembling mixture were incubated with 100 µL of Stabl2 competent E coli. The mixture was heat-shocked at 42°C for 45 seconds and transferred to ice for 2 minutes. SOC medium (900 µl) was added to the Stabl2-NEBuilder mixture and the mix was incubated at 37°C for 1 hour. Transformed bacterial cells (350µl) were plated onto LB agarose plates containing ampicillin (100µg/mL) and incubated overnight at 37°C. Individual colonies were picked and grown overnight in 5 mL of Terrific Broth medium containing 150µg/mL ampicillin and low-endotoxin, small scale preparation of plasmid DNA were performed using the ToxOut EndoFree Plasmid Mini Kit from BioVision (K1326-250). All plasmids were verified by Sanger sequencing with the primer 5'-TTGGGCACTGACAATTCCGT-3'.

Selection of the guides

A549-KRAB-MeCP2 cells were infected with lentivirus obtained from individual CROPseq-Guide-Puro plasmids, encoding individual guides. Infected cells were then grown in the presence of 1 µg/mL of puromycin (Sigma). A week later, total RNAs were purified from A549-KRAB-MeCP2 cells infected with guide encoding lentiviruses and RT-qPCR (primers sequences presented in Supplemental Table YYY) were performed to measure expression of the targeted genes. A validated guide was defined as a guide providing at least 75% inhibition of targeted gene expression compared to a control guide.

Lentiviral transduction with gRNA libraries and cell preparation for chromium scRNA-seq

A549-KRAB-MeCP2 cells were transduced with different amounts of the viral stock containing the library of pooled, selected sgRNA. After six hours, the virus-containing medium was replaced by fresh complete culture medium. Puromycin selection (1µg/ml) was started at 48 h post-transduction, and two days later, the plate with about 30% surviving cells was selected, corresponding roughly to a MOI=3. The cells were then

amplified under puromycin selection for 5 days. The cells were then plated and further cultured in normoxia or in hypoxic condition (1% O2) for 3h, 6 h or 24h. Cells were trypsinized counted and assessed for cell viability using the Countess 3 FL (Fisher Scientific). Samples were then stained for multiplexing using cell hashing [28], using the Cell Hashing Total-Seq-ATM protocol (Biolegend) following the protocol provided by the supplier, using 4 distinct Hash Tag Oligonucleotides-conjugated mAbs (TotalSeq™-B0255, B0256, B0257 and B0258). Briefly, for each condition, 1.106 cells were resuspended in 100µL of PBS, 2% BSA, 0.01% Tween and incubated with 10µL Fc Blocking reagent for 10 minutes at 4°C then stained with 0.5µg of cell hashing antibody for 20 minutes at 4°C. After washing with PBS, 2% BSA, 0.01% Tween, samples were counted and merged at the same proportion, spun 5 minutes 350 x g at 4°C and resuspended in PBS supplemented with 0.04% of bovine serum albumin at final concentration of 500 cells/µL. Samples were then adjusted to the same concentration, mixed in PBS supplemented with 0.04% of bovine serum albumin at a final concentration of 100 cells/µl and pooled sample were immediately loaded onto10X Genomics Chromium device to perform the single cell capture.

Generation of CROP-seq librairies and single-cell RNA-seq data processing

After single-cell capture on the 10X Genomics Chromium device (3' V3), libraries were prepared as recommended, following the Chromium Next GEM Single Cell 3' Reagent Feature Barcoding V3.1 kit (10X Genomics) and a targeted sgRNA amplification [29] with respectively 6, 8 and 10 PCR cycles. Libraries were then quantified, pooled (80% RNA libraries, 10% sgRNA libraries and 10% hashing libraries) and sequenced on an Illumina NextSeq 2000. Alignment of reads from the single cell RNA-seq library and unique molecular identifiers (UMI) counting, as well as oligonucleotides tags (HTOs) counting, were performed with 10X Genomics Cell Ranger tool (v3.0.2). Reads of the gRNA library were counted with CITE-seq-Count (v1.4.2). Counts matrices of total UMI, HTOs, and gRNA were thus integrated on a single object using Seurat R package (v4.1.0), from which the data were processed for analysis. HTOs and gRNA were demultiplexed with HTODemux() and MULTIseqDemux(autoThresh = TRUE) functions respectively, in order to assign treatment and received gRNA for each cell. Only cells identified as "Singlet" after both demultiplexing and passing quality control thresholds of UMI and mitochondrial content were kept. Inhibitions of target genes expression in presence of specific gRNA were validated in all 4 conditions, as well as their progressive upregulation (CYTOR, LUCAT1, NEAT1, SNHG12) or downregulation (HIF1A and SNHG21) during exposition to hypoxia.

Supervised Autoencoder Neural network framework

In this section, we provide the background of the supervised autoencoder (SAE) neural network [27], and the structured sparsity projection method for selecting features.

Figure 1 depicts the main constituent blocks of our proposed approach. Note that we added a "soft max" block to our SAE to compute the classification loss.

Let X be the dataset in , and Y the labels in $\{0, \ldots, k\}$, with k the number of classes. Let Z $\in$ be the encoded latent vectors, $\widehat{X} \in$ the reconstructed data and W the weights of the neural network. Note that the dimension of the latent space k corresponds to the number of classes.

The goal is to compute the network weights W minimizing the total loss which includes both the classification loss and the reconstruction loss. To perform feature selection, as biomedical datasets often present a relatively small number of informative

features, we also want to sparsify the network, following the work proposed in [30], [31], [32] and [33]. Thus, instead of the classical computationally expensive lagrangian regularization approach [34], we propose to minimize the following constrained approach :

$$Loss(W) = \mathcal{H}(Z, Y) + \lambda\psi(\widehat{X} - X) \text{ s.t. } \|W\|_1^1 \leq \eta. \quad (1)$$

We use the Cross Entropy Loss for the classification loss $\mathcal{H}$. We use the robust Smooth ℓ_1 (Huber) Loss [35] as the reconstruction loss ψ. The main difference with the criterion proposed in [25] is the introduction of the constraint on the weights W to sparsify the neural network. A classical approach is the Group LASSO method [36] which consists of using the $\ell_{2,1}$ norm for the constraint on W. However, the $\ell_{2,1}$ norm does not induce a structured sparsity of the network [37], which leads to negative effects on performance.

The main difference with the criterion proposed in [25] is the introduction of the constraint on the weights W to sparsify the neural network. Note that low values of η imply high sparsity of the network. To achieve structured sparsity (feature selection), we use the $\ell_{1,1}$ projection [30]. The basic idea of the $\ell_{1,1}$ projection is first to compute the radius t_i and then project the rows using the ℓ_1 adaptive constraint t_i [38, 39].

Fig 1. Supervised autoencoder framework

Following the work by Frankle and Carbin [31] further developed by [33], we follow a double descent algorithm, originally proposed as follows: after training a network, set all weights smaller than a given threshold to zero, rewind the rest of the weights to their initial configuration, and then retrain the network from this starting configuration while keeping the zero weights frozen. We replace the thresholding by our $\ell_{1,1}$ projection.

We implemented our SAE method using the PyTorch framework for the model, optimizer, schedulers and loss functions. We train the network using the classical Adam optimizer [40]. We used a symmetric linear fully connected network [27], with the encoder comprised of an input layer of d neurons, one hidden layer followed by a ReLU activation function and a latent layer of dimension $k = 2$ since we have two classes.

We compute features significance for the supervised autoencoder using the SHAP method, implemented in the captum python package [41]. The accuracy of the model was systematically computed for each SAE run using 4 folds cross-validation and a mean over 3 seeds.

Results

Single-cell CRISPRi screening of hypoxia-regulated lncRNA

In order to gain new insights into the molecular functions of 6 hypoxia-regulated lncRNA in LUAD cells we performed a single-cell CRISPRi transcriptome screening based on the CROP-Seq approach. We transduced A549 cells expressing double repressor Krab-MeCP2-dCas9 with a mini-library containing 12 validated gRNA targeting CYTOR (also known as LINC00152), LUCAT1, MALAT1, NEAT1, SNHG12 and SNHG21 as well as the two key regulators of the hypoxic response, HIF1A and HIF2 (**Table 1**). Two additional guides, with no effect on the genome, were used as negative controls. In order to mimic the hypoxic environment in which tumors develop in vivo, we equally divided the transduced dCas9-Krab-MeCP2 A549 cells in 4 samples that we then cultured in normoxia or in hypoxia during 3, 6 or 24 hours **Figure 2A**. Cells from each sample were labeled with a specific barcoded antibody (HTOs), pooled, and simultaneously sequenced using droplet based scRNA-seq (10X Genomics

Chromium). The received gRNA and the culture condition were subsequently assigned for each cell by demultiplexing both gRNA and HTOs counts respectively.

Table 1. gRNA library

gRNA	Target	Type	% Inhibition
HIF1A-sg1	HIF1A	Hypoxic response regulator	>95%
HIF1A-sg2			>95%
HIF2-sg5	HIF2/EPAS1		>95%
LINC00152-sg3	CYTOR	Hypoxia-regulated lncRNA	>75%
LUCAT-sg3	LUCAT1		>97%
LUCAT-sg5			>90%
MALAT-sg1	MALAT1		>95%
NEAT1-sg2	NEAT1		>85%
NEAT1-sg6			>95%
SNHG12-sg1	SNHG12		>75%
SNHG12-sg3			>90%
SNHG21-sg5	SNHG21		>85%
Neg-sg1	None	Negative control	None
Neg-sg2			

Fig 2. Single-cell CRISPRi screening A: Design of CROP-seq experiment. B: Heatmaps of gRNA counts or target gene RNA in each cell, labelled according to assigned gRNA and condition after demultiplexing. C: Heatmap of target gene RNA in each cell, labelled according to assigned gRNA and condition after demultiplexing. D: Supervised autoencoder classification workflow.

Overall, we found a balanced representation for each treatment and for each gRNA among the sequenced cells, except for the cells targeted by "SNHG12-sg3" which were depleted in all conditions (**Figure 2B**, **Table 2**). Moreover, the expression of this particular gRNA was lowly detected in those cells, confirming previous observations that this gRNA induced cell death and that only cells with low expression survive. Inhibition of target gene expression in the presence of their corresponding gRNAs were validated in all 4 conditions, as well as their progressive increase (CYTOR, LUCAT1, NEAT1, SNHG12) or decrease (HIF1A and SNHG21) during hypoxia exposure (**Figure 2C**).

Data analysis using the supervised autoencoder classification workflow

For each target gene in each condition, a matrix concatenating raw count from non-targeted control and gRNA-targeted cells was prepared, with cell labels as first row. Cells were pooled according to their targeted gene, and only the top 10 000 most expressed genes were kept. This matrix is the single input of the SAE classification workflow, which is carried out as follows **Figure 2D**.The first SAE run gives a classification score for both non-targeted control cells and for cells targeted for a particular gene. According to the classification, this specific score, called perturbation score, separates cells into 2 subsets : targeted cells with a score ¿ 0.5 are classified as "perturbed" cells, whereas targeted cells with a score ¡ 0.5 are classified as "non-perturbed" cells. A new matrix is generated, containing only the raw counts and labels of the selected perturbed cells and an equivalent number of randomly sampled non-targeted control cells in order to balance both classes size. The second SAE run gives a list of the most discriminant features between both classes, ranked by their

Table 2. Repartition of Doublet, Singlet, and Negative cells in all conditions after demulplexing

	Normoxia (%)	Hypoxia-3h (%)	Hypoxia-6h (%)	Hypoxia-24h (%)
Doublet	8,378	9,908	7,717	8,096
HIF1A-sg1	6,752	5,766	5,681	5,181
HIF1A-sg2	8,346	8,236	8,110	8,265
HIF2-sg5	4,720	5,063	3,859	5,157
LINC00152-sg3	4,720	4,918	5,645	5,108
LUCAT1-sg3	5,345	5,911	5,288	6,048
LUCAT1-sg5	8,690	8,842	9,218	10,120
MALAT1-sg1	6,471	6,686	7,181	5,976
NEAT1-sg2	5,220	5,354	4,823	5,373
NEAT1-sg6	3,501	4,288	3,930	3,952
SNHG12-sg1	4,189	3,125	3,823	3,759
SNHG12-sg3	2,094	2,253	1,751	1,639
SNHG21-sg5	6,690	6,492	6,967	7,398
Neg-sg1	6,346	5,838	6,717	6,554
Neg-sg2	7,221	6,783	7,503	7,373
Negative	11,316	10,538	11,790	10,000
Total cell number	3199	4128	2799	4150

weight in the learned latent space. The complete procedure is run multiple times with different initialization seeds in order to compute a mean and a standard deviation of the obtained ranks, which are used to evaluate the robustness of the perturbation signature. A cell is definitively considered as perturbed if it is classified as such in each run.

Knock-down of HIF1A and HIF2 differentially modulate the hypoxic response

In order to validate the approach, we first evaluated the transcriptomic perturbations induced by the knock- down of the two main regulators of the hypoxic response, HIF1A and HIF2. Globally, the inhibition of HIF1A induced a strong transcriptomic perturbation which affected more than 85% of targeted cells in all conditions (**Table 3**). Even in normoxic condition, the signature breadth was sufficient to allow a classification accuracy above 93%. Among the genes modulated independently from the hypoxic status, we found SNAPC1, IGFL2-AS1, BNIP3L and LDHA, whereas PGK1, PDK1, or BNIP3 modulations were specific to hypoxic conditions (**Figure 3A**). We also found gene modulations specific to early (KDM3A, HIPLDA, ZNF292, EGLN3) or late (SLC16A3, GPI, PGAM1, TPI1) hypoxic response, which correspond to the progressive establishment of the HIF1A-mediated metabolic switch [42]. In normoxia, the knock-down of HIF2 did not produce stable perturbations, except for its own target gene EPAS1 (**Figure 3B**). The 2 early time points of hypoxia exposure showed an improvement of the associated classification accuracy, which reflected a slight increase of the transcriptomic perturbation induced by HIF2 knock-down in these experimental settings. This early signature was mainly driven by genes involved in lipid metabolism ANGPTL4, IGFBP3 and HILPGA. Discrepancies between the results at 3h or 6h were mainly due to the lower number or targeted cells at 6h (104 instead of 202), which impacted the classification. At 24h of hypoxic exposure, the effect of HIF2 inhibition reached its maximum, with 84% perturbed cells and an accuracy of 97%. However, this signature was quite different from that of HIF1A-targeted cells under the same condition. Indeed, some upregulated (ALDH3A1,CPLX2,FTL,PAPPA) or downregulated (ATP1B1,FXYD2,ANXA4,LOXL2) genes in HIF2-targeted cells were

not modulated in HIF1A-targeted cells ((**Figure 3C**)). Moreover, several genes showed an opposite perturbation between the two groups of cells. This was the case for BNIP3, PGK1, GPI, FAM162A, SLC16A3, TPI1, or PGAM1 which were downregulated upon HIF1A inhibition but were found upregulated upon HIF2 inhibition after 24h of culture in hypoxia. These results were consistent with the known role of HIF2, which is activated upon prolonged exposure to hypoxia and is involved in the regulation of the chronic hypoxic response [5]. They also confirm that in LUAD cells, HIF1A and HIF2-regulated functions are specific, or even antagonistic for certain genes, which has been previously demonstrated in other cancers [43].

Table 3. Supervised autoencoder classification and accuracy for HIF1A and HIF2 gRNA targeted cells

	Treatment	Targeted cells	Perturbed cells (%)	Accuracy (%)
HIF1A	Normoxia	475	86,7	95,33
HIF1A	Hypoxia 3h	554	87,5	94,00
HIF1A	Hypoxia 6h	372	85,8	93,67
HIF1A	Hypoxia 24h	554	85,7	94,67
HIF2	Normoxia	147	11,6	66,67
HIF2	Hypoxia 3h	202	51	86,33
HIF2	Hypoxia 6h	104	38,5	82,33
HIF2	Hypoxia 24h	213	84	97,67

Fig 3. Knock-down of HIF1A and HIF2 differentially modulate the hypoxic response A: Top 20 discriminant features between perturbed and control cells for HIF1A for each treatment. Upregulated or downregulated genes are written in red or blue respectively. B: Top 20 discriminant features between perturbed and control cells for HIF2/EPAS1 for each treatment. C: Differentially expressed genes between perturbed and control cells for HIF1A and HIF2 for each treatment.

Knock-down of hypoxia-regulated lncRNAs results in weak and heterogeneous condition-dependent transcriptomic modulations

We then applied the SAE workflow to classify cells treated with the 6 gRNA targeting hypoxia-regulated lncRNAs and cultured in the 4 conditions. Globally, the SAE was able to classify perturbed and control cells with a good overall accuracy around 80%, except for SNHG12 and SNHG21 (**Table 4**). For the SNHG12 and SNHG21 datasets, the first round of SAE selected only around 10% of perturbed cells. Thus we could not run the SAE for the second round because of a too low number of cells for the 4 fold cross validation. For those 2 genes, we just reported the average accuracies obtained after the first round of the SAE. These initial good performances for LINC00152, MALAT1 and NEAT1 were not found to be associated with a specific transcriptomic perturbation signature, but were exclusively due to the strong inhibition of the target gene, as indicated by the high means and standard deviations of the ranks obtained for the other genes (**Figure 4A-C**). For example, the combined inhibition of CYTOR/LINC00152 with MIR4435-2HG, whose sequences are highly homologous (99% in the 220 bp region including the most efficient sgRNA), was sufficient to select half of the targeted cells with an accuracy above 85% regardless of the condition. Surprisingly, we did not detect any other stable perturbations than the target gene for both MALAT1 and NEAT1 targeted cells, as indicated by the obtained high means and standard

deviations of the computed ranks, while those two lncRNAs were previously associated with various gene regulation functions [44, 45]. As SNHG21 expression is relatively low in LUAD cells and is decreased by hypoxic stress, the extent of its inhibition was therefore weaker and not sufficient to distinguish targeted from control cells. Combined with the lack of transcriptomic effect induced by its knock-down, it explains the poor classification results and the randomness of features selected for cells targeted by this particular gene under all conditions (**Figure 4D**).

Table 4. Supervised autoencoder classification and accuracy for hypoxia regulated lncRNA gRNA targeted cells (*obtained without cell selection)

	Treatment	Targeted cells	Perturbed cells (%)	Accuracy (%)
LINC00152	Normoxia	147	55,1	87,67
LINC00152	Hypoxia 3h	200	59	91,67
LINC00152	Hypoxia 6h	150	58	86,33
LINC00152	Hypoxia 24h	209	42,6	85,00
LUCAT1	Normoxia	438	51,8	73,33
LUCAT1	Hypoxia 3h	583	77,7	82,33
LUCAT1	Hypoxia 6h	391	63,9	79,00
LUCAT1	Hypoxia 24h	666	90,4	85,67
MALAT1	Normoxia	205	37,1	82,67
MALAT1	Hypoxia 3h	269	45	82,33
MALAT1	Hypoxia 6h	194	37,1	78,00
MALAT1	Hypoxia 24h	241	26,1	78,67
NEAT1	Normoxia	274	59,1	86,33
NEAT1	Hypoxia 3h	390	73,3	89,00
NEAT1	Hypoxia 6h	237	52,7	85,33
NEAT1	Hypoxia 24h	566	58,7	87,33
SNHG12	Normoxia	196	14,8	65*
SNHG12	Hypoxia 3h	217	14,7	69*
SNHG12	Hypoxia 6h	154	3,9	67,4*
SNHG12	Hypoxia 24h	213	8,5	71*
SNHG21	Normoxia	211	5,2	60.7*
SNHG21	Hypoxia 3h	256	3,5	61.1*
SNHG21	Hypoxia 6h	192	6,8	59,1*
SNHG21	Hypoxia 24h	299	2	63,5*

Fig 4. Top 20 discriminant features between perturbed and control cells for LINC00152 (A), MALAT1 (B), NEAT1 (C), and SNHG21 (D) for each treatment. Upregulated or downregulated genes are written in red or blue respectively.

Knock-down of hypoxia-regulated lncRNA LUCAT1 leads to hypoxic condition-dependent transcriptomic modulations

The SAE outcomes were different for the classification of LUCAT1-targeted cells. Indeed, the transcriptomic inhibition of LUCAT1 resulted in a stable upregulation for

PCOLCE2 and ISCA1 in normoxia, HDHD2 after 3h, and 6h of hypoxia (**Figure 5A**). ISCA1 and HDHD2 encode for metal ion binding proteins, whereas TFCP2 is a known oncogene. After 24h of hypoxia, a completely different perturbation signature was found, with at least 6 stably modulated genes, including the upregulation of KDM5C, TMEM175 and NIT1, as well as the downregulation of ATP6AP1, PEX1 and PHF20. ATP6AP1 and PEX1 are respectively components of the V-ATPase and the peroxisomal ATPase complexes, while TMEM175 is a proton channel also involved in pH regulation. KDMC5 and PHF20 are both involved in chromatin remodeling and transcriptomic regulation, while NIT1 is associated to tumor suppressor functions. This particular signature allowed the classification of 90,4% of targeted cells with an accuracy of 85,67%. These results indicate that LUCAT1 inhibition induces hypoxic condition-dependent transcriptomic modulations that potentially impact tumor survival and gene regulatory processes during prolonged exposure to hypoxic conditions, completing our previous observations [12].

Fig 5. Top 20 discriminant features between perturbed and control cells for LUCAT1 (A) and SNHG12 (B) for each treatment. Upregulated or downregulated genes are written in red or blue respectively.

Supervised autoencoder classification revealed an anti-apoptotic signature expressed by a subset of SNHG12-targeted cells in response to the cytotoxic effect of one of its gRNA

Looking at the SAE classification outcomes for SNHG12-targeted cells, only about 15% of them were classified as perturbed in normoxia and after 3h of hypoxia, with a poor accuracy. The number of selected cells was even worse for a longer exposure to hypoxia (**Table 2**). Nevertheless, the ranked list of top discriminant features between the few perturbed cells and control cells obtained for the first two time points showed a notable perturbation signature. In normoxia, it was only composed of BAG1 upregulation, whereas after 3h of hypoxia exposure, this signature was completed by GAS5 (snoRNAs containing lncRNA gene), ARRB2, ATF5, and ETHE1 upregulations (**Figure 5B**). These 5 genes are all known anti-apoptotic factors. We hypothesized that this anti-apoptotic signature was expressed by a subset of LUAD cells that were actively escaping the cytotoxic effect we systematically observed for the most efficient of the two gRNAs selected for targeting SNHG12, (SNHG12-sg3) (**Table 1**). Indeed, most of the cells classified as perturbed were specific to this particular gRNA (**Figure 6**). As this signature progressively attenuated over time under hypoxic conditions, we speculate that the activation of this anti-apoptotic response may be inhibited by hypoxic stress, or that hypoxia may protect against the cytotoxic effect of this guide. These results demonstrate the precision of the SAE-based approach to detect a short signature, even restricted to a small subset of cells..

Fig 6. Percentages of targeted cells classified as perturbed or non-perturbed for each gRNA in each condition.

Comparison with other Machine learning methods for features selection

We compared the performance of our SAE, with or without selection of the most responsive cells, with classical machine learning methods such as the popular Partial

Least Squares regression (PLS) [32] and random forests (RF) [33] using 400 estimators (using the Gini importance (GI) for feature ranking). Note that the authors of RF proposes two measures for feature ranking, the variable importance (VI) and Gini importance (GI) : a recent study showed that if predictors are real with multimodal Gaussian distributions, both measures are biased [34]. We also used the Log Fold Change (LFC) results, which is a common analysis in scRNA-seq, even if this particular metric does not produce a value of accuracy. We ranked the top differentially expressed genes according to p-values obtained with Wilcoxon Rank Sum test and ajusted with Bonferroni correction. We performed this comparison for 3 representative datasets, namely HIF2-targeted cells versus 24h hypoxia-related control cells, LUCAT1-targeted cells versus 24h hypoxia -related control cells, and SNHG12-targeted cells versus 3h hypoxia-related control cells. For the first dataset, HIF2 inhibition induced a large perturbation signature. With an accuracy of 97.67%, the SAE outperformed both PLS and RF, for which we obtained an accuracy of 92.5% and 72.42% respectively (**Table 5**). Note that selecting the cells thanks to the confidence score improved significantly the accuracy of the SAE by 4.67%. However, the top 15 selected features between all methods were highly similar, because of the strong effect of HIF2 inhibition (**Figure 7A**).

Table 5. Comparison of the 15 first selected features between SAE (with or without selection), PLS, Random Forest and Log Fold Change, for HIF2 (EPAS1) targeted cells in hypoxia 24h

SAE (selection)	SAE (no selection)	PLS	Random Forest	Log Fold Change
97.67%	93%	92.5 %	72.42 %	
TMEM141	TMEM141	TMEM141	IGFBP3	TMEM141
IGFBP3	PGK1	IGFBP3	FTL	PGK1
BNIP3	IGFBP3	FAM162A	PGK1	IGFBP3
FAM162A	BNIP3	BNIP3	TMEM141	BNIP3
EPAS1	EPAS1	EPAS1	BNIP3	FAM162A
PGK1	FAM162A	PGK1	FAM162A	ENO1
FXYD2	FXYD2	TESC	EPAS1	EPAS1
ATP1B1	GPI	ATP1B1	FXYD2	FTL
TESC	SLC16A3	FXYD2	ATP1B1	FXYD2
SLC16A3	ANXA4	SLC16A3	ENO1	ATP1B1
GPI	TESC	LOXL2	GPI	TESC
ALDH3A1	FTL	GPI	TESC	GPI
ANXA4	PGAM1	ANXA4	PGD	LOXL2
GAL3ST1	ATP1B1	CYP1B1	ALDH1A1	PGAM1
FTL	LOXL2	DSP	ALDH3A1	VIM

We observed more extreme results for the LUCAT1 dataset. Indeed, with an accuracy of 85.67%,the SAE outperformed both PLS and RF, for which we obtained an accuracy of only 62% and 54.6% respectively (**Table 6**). Note that selecting the cells thanks to the confidence score improved the accuracy of the SAE by 1.67%. Furthermore, most of the obtained signatures were specific to a particular Machine Learning method, with some overlaps (**Figure 7B**) such as ATP6AP1 that is also detected by PLS. NIT1, PEX1, KDMC5, PHF20 and TMEM175 were all found specific to the SAE.

For the SNHG12 dataset, we stopped the workflow of SAE tests before the selection step since only 14% of perturbed cells were found perturbed. All methods performed similarly, with a poor average accuracy of 70% (**Table 7**). They all detected for instance the strong modulation of BAG1, but the other selected genes were mostly

Table 6. Comparison of the 10 first selected features between SAE (with or without selection), PLS, Random Forest and Log Fold Change (LFC), for LUCAT1 targeted cells in hypoxia 24h

SAE (selection)	SAE (no selection)	PLS	Random Forest	LFC
85.67%	84%	62%	54.6 %	
LUCAT1	LUCAT1	LUCAT1	LUCAT1	LUCAT1
ATP6AP1	ATP6AP1	PAIP2	BPTF	ATP6AP1
KDM5C	CCDC142	CFLAR	G3BP1	AHNAK2
TMEM175	NIT1	EIF3J	NKAPD1	AMIGO2
PEX1	PEX1	RAB13	PDCL	H1-3
NIT1	KDM5C	DKC1	HSF1	KLHL5
PHF20	PHF20	ATP6AP1	MFSD1	SLBP
ZNF181	GFRA1	APP	BCLAF1	H3C2
CCDC142	CMBL	H2AJ	UBE2R2	PAIP2
AGO2	ELF1	RPAIN	RPL14	MKI67

specific to each method (**Figure 7C**). Of note, the SAE was the only method that detected SNHG12 target gene inhibition. Taken together, our results indicate that the SAE approach is relevant to detect subtle perturbation signals found in CRISPRi screening with a scRNA-seq readout.

Table 7. Comparison of the 10 first selected features between SAE (without selection), PLS, Random Forest and Log Fold Change (LFC), for SNHG12 targeted cells in hypoxia 24h

SAE (no selection)	PLS	Random Forest	LFC
70%	70%	69%	
BAG1	BAG1	BAG1	BAG1
GAS5	RPL22L1	DPP3	RPL22L1
SNHG5	ETHE1	SNHG5	GAS5
ZMAT5	PPDPF	C12orf75	H4C3
SNHG12	NFKBIA	RPS27L	H1-3
RPL22L1	GAS5	FXYD2	MKI67
ATF5	TLE1	DCDC2	SNHG5
DCDC2	REXO2	FAM136A	KPNA2
FAM98C	SNHG5	MRPL14	NOP56
MAD1L1	PARD3	CXCL8	NOLC1

Fig 7. Intersection of the 15 first selected features lists between SAE with selection, SAE without selection, PLS, Random Forest and LFC for indicated datasets.

Discussion

Single-cell CRISPR(i)-based transcriptome screenings are powerful tools for simultaneously accessing the expression profiles of cells targeted by different gRNA, in order to infer target genes functions from the observed perturbations. However, these approaches are limited by the low molecule capture rate and sequencing depth provided by droplet-based scRNA-seq, which produce sparse and noisy data. Furthermore, the outcome of CRISPR-induced modification in each cell is a stochastic event, depending

among other things, on the expression levels of the transcribed gRNA and dCas9, as well as the accessibility of the target gene locus, that may be heterogeneously regulated at the epigenomic levels in the different cells. For these reasons, the induced perturbation signature and its detection are likely heterogeneous between cells, even when dCas9-expressing cells receiving the same gRNA have been cloned. Deciphering this heterogeneity in sparse data is even more complex when the targeted genes are not master geneq involved in signaling or regulatory pathways, such as transcription factors and receptors. In this respect, a previous study [46] has shown that this particular challenge cannot be met using conventional scRNA-seq analysis tools such as differential expression, which is clearly limited to the detection of weak and heterogeneous perturbation signals. This challenge seems even more complex for the study of perturbations mediated by knockdown of non-coding RNAs, which have been largely involved in the fine-tuning of gene expression regulation. To increase the sensitivity of single-cell CRISPR(i)-based transcriptome screenings, we propose here a powerful feature selection and classification approach based on a supervised autoencoder (SAE). It leverages in particular on the known cell labels initially given by gRNA counts demultiplexing to constrain the latent space to fit the original data distribution. Beyond high statistical accuracy, our SAE offered relevant properties that distinguishes it from classical classification methods : i) a stringent feature selection producing an interpretable readout of ranked top discriminant genes associated to their weights; ii) a classification score which allow the selection of the most perturbed cells and the eventual signal to obtain a more robust perturbation signature.

We first validated this approach by analyzing the perturbations associated with the knock- down of the two master regulators of the hypoxic response, HIF1A and HIF2. Using this classification workflow, we showed that the SAE was able to learn a latent space and a perturbation signature which can for exemple almost perfectly discriminate HIF2-targeted cells from their control in condition of prolonged hypoxia. The SAE classification accuracy provided a global perturbation score associated with HIF1A and HIF2 at each time point, reflecting the biological activity of each factor during the hypoxic response. We were able to recapitulate the known distinct influence and target specificity of HIF1 and HIF2 during the hypoxia time course [5], with notably i) a strong perturbation driven by HIF1 at early time points; ii) a progressive influence of HIF2 with a maximum effect observed at 24h of hypoxia; iii) a specificity regarding their targets, with sometimes an opposite regulation for some genes. Finally, this unique dataset provides a global and dynamic description of the transcriptomic modulations mediated by the two main regulators of the hypoxic response in LUAD A549 cells.

Surprisingly, we did not detect any relevant and stable perturbation in cells targeted for LINC00152, MALAT1, NEAT1 and SNHG21, in the four culture conditions. This result appears quite unexpected for MALAT1 and NEAT1, two of the most studied lncRNAs that are associated with various functions in cancer, including proliferation, migration, and invasion [44, 47]. In particular, it has been shown that MALAT1 knockout in the same cellular model (A549) modulated a set of metastasis-associated genes [48]. Although CRISPRi-mediated knock-down achieved an efficient knock-down (¿95%), it is however possible that based on the very high level of MALAT1, the remaining transcripts are sufficient to mediate the cellular function. Another possibility could be due to differences in methodology, notably the need to isolate single clones for the knockout protocol, a long procedure that can profoundly affect the transcriptome, compared with the CROP-seq approach performed on a bulk population prior to immediate single-cell isolation. A similar situation may occur for NEAT1, a highly abundant lncRNA acting as a structural scaffold of membraneless paraspeckle nuclear bodies. Moreover, NEAT1 can produces two isoforms, with a differential regulation upon stress and distinct functions [49]. Additional work will be thus necessary to

further analyze the relative proportion of the two isoforms in A549 cells and their potential function during hypoxia.

However, for LUCAT1-targeted cells after 24h of hypoxia exposure, we found a stable signature of 6 modulated genes, which are associated with pH or gene regulation. It suggested a potential capacity of LUCAT1 to promote tumor cell survival during prolonged hypoxia and to contribute to an aggressive phenotype in LUAD cells, as we previously demonstrated [12]. Finally, we also found a relevant signature in SNHG12-targeted cells, characterized by the upregulation of anti-apoptotic genes. As this signature is almost exclusive to cells targeted by the most effective gRNA against SNHG-12, which appeared to systematically induce cell death, we hypothesized that it is expressed by surviving cells. The potential pro-oncogenic role of the complex SNHG-12 locus, producing a lncRNA and 3 snoRNAs, should be pursued to decipher the molecular components associated with this phenotype, as also suggested by previous studies [50].

In this paper, we demonstrate that the SAE is highly relevant in situations in which low signals in a restricted number of cells need to be detected. However, SAE has some limitations. Like all statistical method, it is highly dependent on the number of samples/cells compared. Low cell number impact classification performance and can produce inconsistent results, such as better accuracy and robustness of selected features for HIF2-targeted cells after 3h of hypoxia compared with 6h exposure. In this context, the relevance of the top selected genes list and their superiority over other compared methods can be asserted by evaluating the robustness of the ranks and the classification accuracies. The size of the perturbation signatures obtained for LUCAT1 and SNHG12 datasets prevented the utilization of functional enrichment analysis to characterize their modulated functions. Furthermore, as these small signatures were found in specific subsets of targeted cells, it seems complicated to validate them using a bulk experimental approach that will average the signal across all cells. Despite these limitations, we believe that our approach is well suited to the particular deciphering of single cell CRISPR-based screen with omics readout, or for other similar assays to assess the effect of perturbation at the single cell level.

Availability of data and materials

We implemented the code with python. Functions and scripts are freely available at https ://github.com/MichelBarlaud/SAE Supervised-Autoencoder-Omics

Competing interests

The authors declare that they have no competing interests.

Fundings

We acknowledge the support from the Centre National de la Recherche Scientifique (CNRS), Université Côte d'Azur, Canceropôle PACA (Action Structurante CRISPR SCREEN), the French Government (National Research Agency, ANR) program "Investissements d'Avenir" UCAJEDI n° ANR-15-IDEX-01 (PERTURB- ENCODER) and ANR-22-CE17-0046-01 MIR-ASO, Plan Cancer 2018 "ARN non-codants en cancérologie : du fondamental au translationnel" (number 18CN045) and Fondation ARC.

Acknowledgments

The authors thank the technical support of the UCA GenomiX of the University Côte d'Azur. We also thank A. Monteil and C. Lemmers from the Vectorology facility, PVM, Biocampus Montpellier, CNRS UMS3426

References

1. Semenza GL. Hypoxia-inducible factors in physiology and medicine. Cell. 2012;148(3):399–408. doi:10.1016/j.cell.2012.01.021.
2. Rankin EB, Nam JM, Giaccia AJ. Hypoxia: Signaling the Metastatic Cascade. Trends in Cancer. 2016;2(6):295–304. doi:10.1016/j.trecan.2016.05.006.
3. Rankin EB, Giaccia AJ. Hypoxic control of metastasis. Science (New York, NY). 2016;352(6282):175–180. doi:10.1126/science.aaf4405.
4. Kaelin WG, Ratcliffe PJ. Oxygen sensing by metazoans: the central role of the HIF hydroxylase pathway. Molecular Cell. 2008;30(4):393–402. doi:10.1016/j.molcel.2008.04.009.
5. Koh MY, Powis G. Passing the baton: the HIF switch. Trends in Biochemical Sciences. 2012;37(9):364–372. doi:10.1016/j.tibs.2012.06.004.
6. Bertero T, Rezzonico R, Pottier N, Mari B. Chapter Three - Impact of MicroRNAs in the Cellular Response to Hypoxia. In: Galluzzi L, Vitale I, editors. International Review of Cell and Molecular Biology. vol. 333 of MiRNAs in Differentiation and Development. Academic Press; 2017. p. 91–158.
7. Choudhry H, Harris AL. Advances in Hypoxia-Inducible Factor Biology. Cell Metabolism. 2018;27(2):281–298. doi:10.1016/j.cmet.2017.10.005.
8. Barth DA, Prinz F, Teppan J, Jonas K, Klec C, Pichler M. Long-Noncoding RNA (lncRNA) in the Regulation of Hypoxia-Inducible Factor (HIF) in Cancer. Non-coding RNA. 2020;6(3):27. doi:10.3390/ncrna6030027.
9. Carlevaro-Fita J, Lanzós A, Feuerbach L, Hong C, Mas-Ponte D, Pedersen JS, et al. Cancer LncRNA Census reveals evidence for deep functional conservation of long noncoding RNAs in tumorigenesis. Communications Biology. 2020;3(1):1–16. doi:10.1038/s42003-019-0741-7.
10. Slack FJ, Chinnaiyan AM. The Role of Non-coding RNAs in Oncology. Cell. 2019;179(5):1033–1055. doi:10.1016/j.cell.2019.10.017.
11. Hu L, Tang J, Huang X, Zhang T, Feng X. Hypoxia exposure upregulates MALAT-1 and regulates the transcriptional activity of PTB-associated splicing factor in A549 lung adenocarcinoma cells. Oncology Letters. 2018;16(1):294–300. doi:10.3892/ol.2018.8637.
12. Moreno Leon L, Gautier M, Allan R, Ilié M, Nottet N, Pons N, et al. The nuclear hypoxia-regulated NLUCAT1 long non-coding RNA contributes to an aggressive phenotype in lung adenocarcinoma through regulation of oxidative stress. Oncogene. 2019;38(46):7146–7165. doi:10.1038/s41388-019-0935-y.

13. Liu SJ, Horlbeck MA, Cho SW, Birk HS, Malatesta M, He D, et al. CRISPRi-based genome-scale identification of functional long noncoding RNA loci in human cells. Science (New York, NY). 2017;355(6320):aah7111. doi:10.1126/science.aah7111.

14. Esposito R, Polidori T, Meise DF, Pulido-Quetglas C, Chouvardas P, Forster S, et al. Multi-hallmark long noncoding RNA maps reveal non-small cell lung cancer vulnerabilities. Cell Genomics. 2022;2(9):100171. doi:10.1016/j.xgen.2022.100171.

15. Gapp BV, Konopka T, Penz T, Dalal V, Bürckstümmer T, Bock C, et al. Parallel reverse genetic screening in mutant human cells using transcriptomics. Molecular Systems Biology. 2016;12(8):879. doi:10.15252/msb.20166890.

16. Replogle JM, Norman TM, Xu A, Hussmann JA, Chen J, Cogan JZ, et al. Combinatorial single-cell CRISPR screens by direct guide RNA capture and targeted sequencing. Nature Biotechnology. 2020;38(8):954–961. doi:10.1038/s41587-020-0470-y.

17. Datlinger P, Rendeiro AF, Schmidl C, Krausgruber T, Traxler P, Klughammer J, et al. Pooled CRISPR screening with single-cell transcriptome readout. Nature Methods. 2017;14(3):297–301. doi:10.1038/nmeth.4177.

18. Mimitou EP, Cheng A, Montalbano A, Hao S, Stoeckius M, Legut M, et al. Multiplexed detection of proteins, transcriptomes, clonotypes and CRISPR perturbations in single cells. Nature Methods. 2019;16(5):409–412. doi:10.1038/s41592-019-0392-0.

19. Lopez R, Regier J, Cole MB, Jordan MI, Yosef N. Deep generative modeling for single-cell transcriptomics. Nature Methods. 2018;15(12):1053–1058. doi:10.1038/s41592-018-0229-2.

20. Leclercq M, Vittrant B, Martin-Magniette ML, Scott Boyer MP, Perin O, Bergeron A, et al. Large-Scale Automatic Feature Selection for Biomarker Discovery in High-Dimensional OMICs Data. Frontiers in Genetics. 2019;10:452. doi:10.3389/fgene.2019.00452.

21. Hinton GE, Zemel R. Autoencoders, Minimum Description Length and Helmholtz Free Energy. In: Advances in Neural Information Processing Systems. vol. 6. Morgan-Kaufmann; 1993.

22. Vincent P, Larochelle H, Lajoie I, Bengio Y, Manzagol PA. Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion;.

23. Emdadi A, Eslahchi C. Auto-HMM-LMF: feature selection based method for prediction of drug response via autoencoder and hidden Markov model. BMC bioinformatics. 2021;22(1):33. doi:10.1186/s12859-021-03974-3.

24. Liu D, Huang Y, Nie W, Zhang J, Deng L. SMALF: miRNA-disease associations prediction based on stacked autoencoder and XGBoost. BMC bioinformatics. 2021;22(1):219. doi:10.1186/s12859-021-04135-2.

25. Kingma DP, Welling M. Auto-Encoding Variational Bayes. 2013;.

26. Lotfollahi M, Naghipourfar M, Luecken MD, Khajavi M, Büttner M, Wagenstetter M, et al. Mapping single-cell data to reference atlases by transfer learning. Nature Biotechnology. 2022;40(1):121–130. doi:10.1038/s41587-021-01001-7.

27. Barlaud M, Guyard F. Learning a Sparse Generative Non-Parametric Supervised Autoencoder. In: ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2021. p. 3315–3319.

28. Stoeckius M, Zheng S, Houck-Loomis B, Hao S, Yeung BZ, Mauck WM, et al. Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. Genome Biology. 2018;19(1):224. doi:10.1186/s13059-018-1603-1.

29. Hill AJ, McFaline-Figueroa JL, Starita LM, Gasperini MJ, Matreyek KA, Packer J, et al. On the design of CRISPR-based single-cell molecular screens. Nature Methods. 2018;15(4):271–274. doi:10.1038/nmeth.4604.

30. Barlaud M, Guyard F. Learning sparse deep neural networks using efficient structured projections on convex constraints for green AI. In: 2020 25th International Conference on Pattern Recognition (ICPR); 2021. p. 1566–1573.

31. Frankle J, Carbin M. The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks. International Conference on Learning Representations. 2019.

32. Wen W, Wu C, Wang Y, Chen Y, Li H. Learning Structured Sparsity in Deep Neural Networks. In: Advances in Neural Information Processing Systems. vol. 29. Curran Associates, Inc.; 2016.

33. Zhou H, Lan J, Liu R, Yosinski J. Deconstructing Lottery Tickets: Zeros, Signs, and the Supermask. In: Advances in Neural Information Processing Systems. vol. 32. Curran Associates, Inc.; 2019.

34. Hastie T, Rosset S, Tibshirani R, Zhu J. The Entire Regularization Path for the Support Vector Machine. Advances in Neural Information Processing Systems 17, MIT Press. 2004.

35. Huber PJ. Robust Statistics. In: Lovric M, editor. International Encyclopedia of Statistical Science. Berlin, Heidelberg: Springer; 2011. p. 1248–1251.

36. Yuan M, Lin Y. Model Selection and Estimation in Regression with Grouped Variables. Journal of the Royal Statistical Society Series B: Statistical Methodology. 2006;68(1):49–67. doi:10.1111/j.1467-9868.2005.00532.x.

37. Barlaud M, Chambolle A, Caillau JB. Classification and feature selection using a primal-dual method and projection on structured constraints. In: 2020 25th International Conference on Pattern Recognition (ICPR); 2021. p. 6538–6545.

38. Condat L. Fast projection onto the simplex and the ℓ_1 ball. Mathematical Programming. 2016;158(1):575–585. doi:10.1007/s10107-015-0946-6.

39. Perez G, Barlaud M, Fillatre L, Régin JC. A filtered bucket-clustering method for projection onto the simplex and the ℓ_1 ball. Mathematical Programming. 2020;182(1):445–464. doi:10.1007/s10107-019-01401-3.

40. Kingma DP, Ba LJ. Adam: A Method for Stochastic Optimization. 2015;.

41. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. NIPS'17. Red Hook, NY, USA: Curran Associates Inc.; 2017. p. 4768–4777.

42. Kim Jw, Tchernyshyov I, Semenza GL, Dang CV. HIF-1-mediated expression of pyruvate dehydrogenase kinase: a metabolic switch required for cellular adaptation to hypoxia. Cell Metabolism. 2006;3(3):177–185. doi:10.1016/j.cmet.2006.02.002.

43. Raval RR, Lau KW, Tran MGB, Sowter HM, Mandriota SJ, Li JL, et al. Contrasting properties of hypoxia-inducible factor 1 (HIF-1) and HIF-2 in von Hippel-Lindau-associated renal cell carcinoma. Molecular and Cellular Biology. 2005;25(13):5675–5686. doi:10.1128/MCB.25.13.5675-5686.2005.

44. Dong P, Xiong Y, Yue J, Hanley SJB, Kobayashi N, Todo Y, et al. Long Non-coding RNA NEAT1: A Novel Target for Diagnosis and Therapy in Human Tumors. Frontiers in Genetics. 2018;9:471. doi:10.3389/fgene.2018.00471.

45. Amodio N, Raimondi L, Juli G, Stamato MA, Caracciolo D, Tagliaferri P, et al. MALAT1: a druggable long non-coding RNA for targeted anti-cancer approaches. Journal of Hematology & Oncology. 2018;11(1):63. doi:10.1186/s13045-018-0606-4.

46. Papalexi E, Mimitou EP, Butler AW, Foster S, Bracken B, Mauck WM, et al. Characterizing the molecular regulation of inhibitory immune checkpoints with multimodal single-cell screens. Nature Genetics. 2021;53(3):322–331. doi:10.1038/s41588-021-00778-2.

47. Arun G, Aggarwal D, Spector DL. MALAT1 Long Non-Coding RNA: Functional Implications. Non-coding RNA. 2020;6(2):22. doi:10.3390/ncrna6020022.

48. Gutschner T, Hämmerle M, Eißmann M, Hsu J, Kim Y, Hung G, et al. The Noncoding RNA MALAT1 Is a Critical Regulator of the Metastasis Phenotype of Lung Cancer Cells. Cancer Research. 2013;73(3):1180–1189. doi:10.1158/0008-5472.CAN-12-2850.

49. Adriaens C, Rambow F, Bervoets G, Silla T, Mito M, Chiba T, et al. The long noncoding RNA NEAT1_1 is seemingly dispensable for normal tissue homeostasis and cancer cell growth. RNA (New York, NY). 2019;25(12):1681–1695. doi:10.1261/rna.071456.119.

50. Tamang S, Acharya V, Roy D, Sharma R, Aryaa A, Sharma U, et al. SNHG12: An LncRNA as a Potential Therapeutic Target and Biomarker for Human Cancer. Frontiers in Oncology. 2019;9.

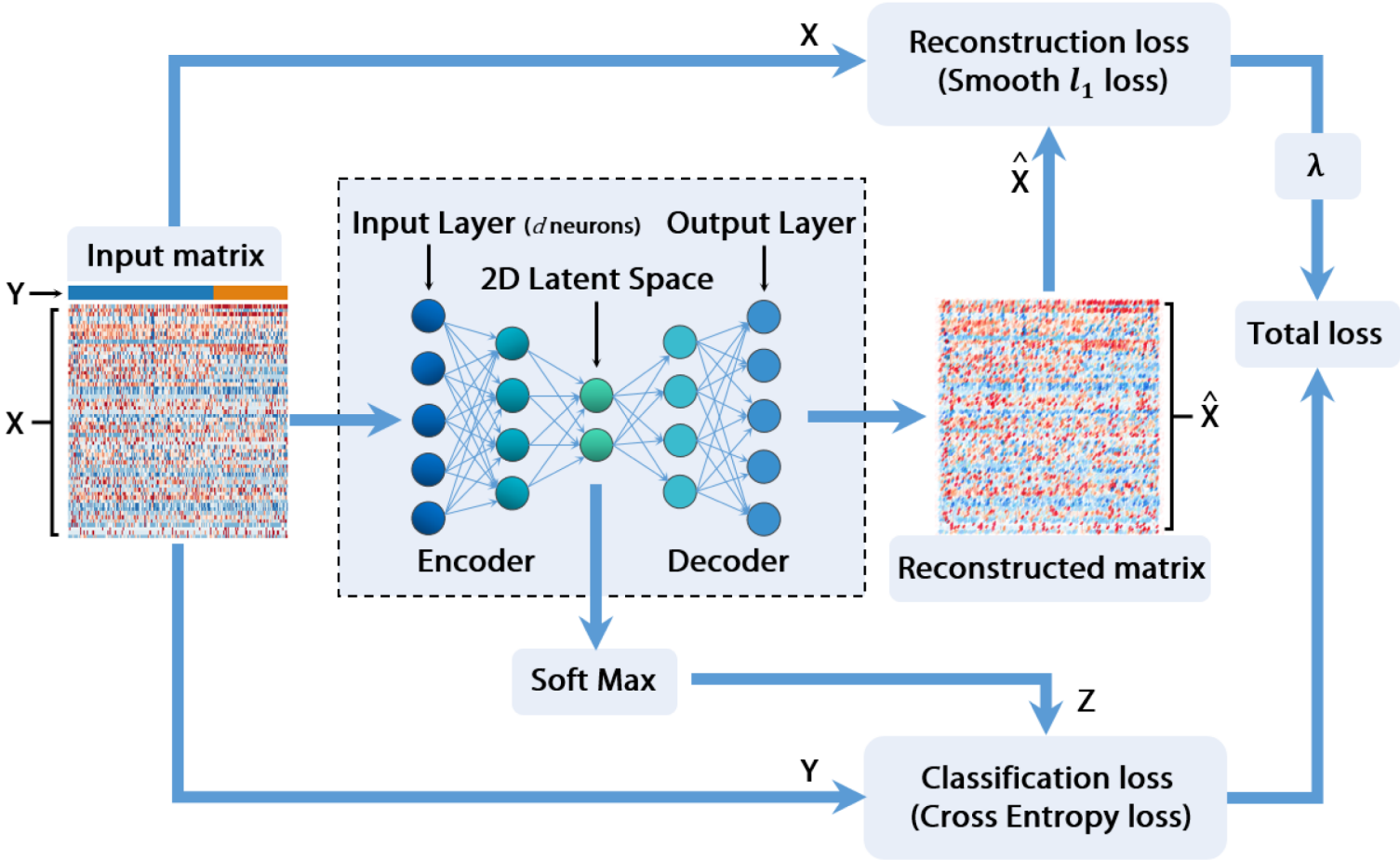

Fig 1. Supervised autoencoder framework

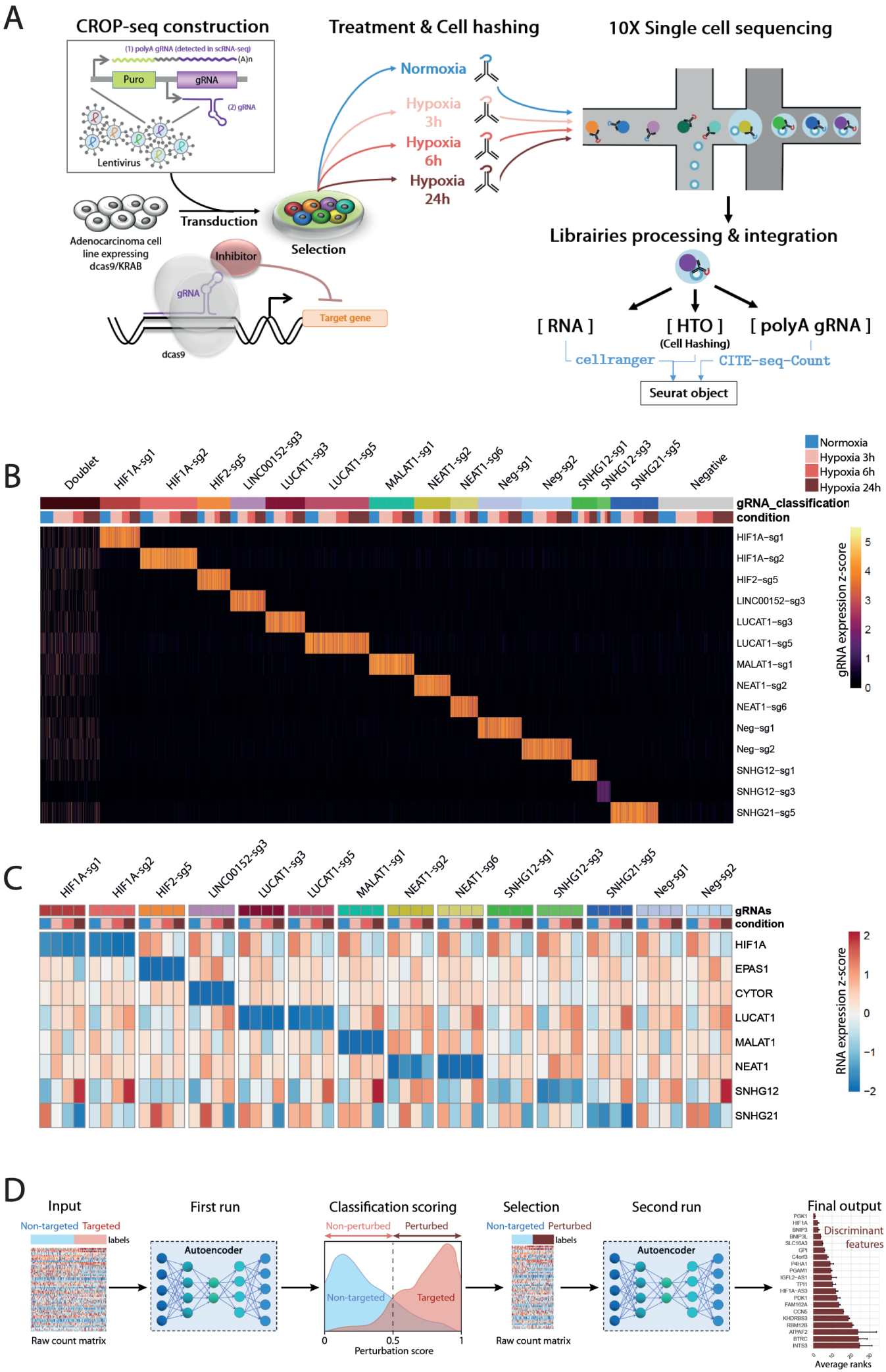

Fig 2. Single-cell CRISPRi screening A: Design of CROP-seq experiment. B: Heatmaps of gRNA counts or target gene RNA in each cell, labelled according to assigned gRNA and condition after demultiplexing. C: Heatmap of target gene RNA in each cell, labelled according o assigned gRNA and condition after demultiplexing. D: Supervised autoencoder classification workflow

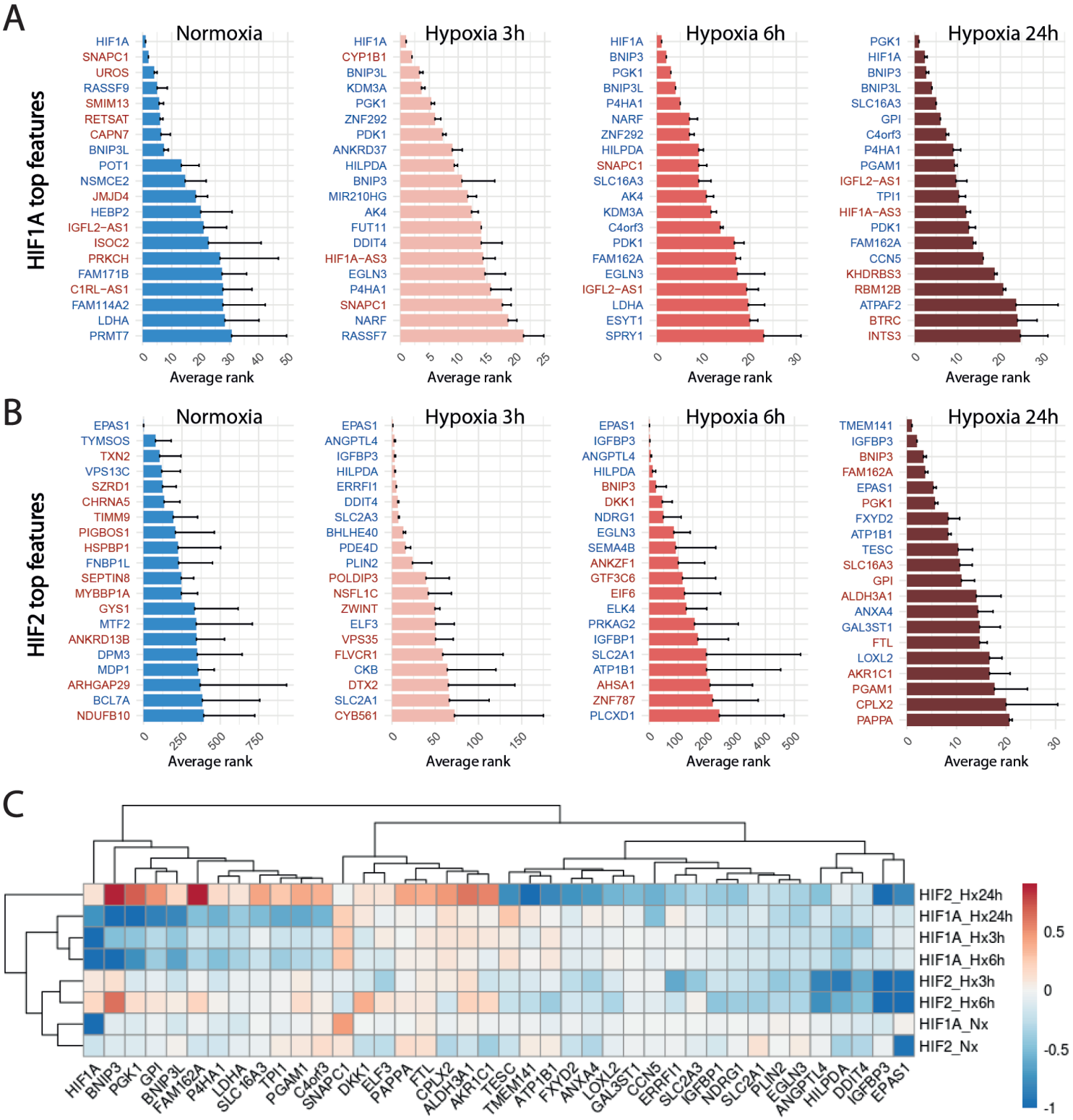

Fig 3. Knock-down of HIF1A and HIF2 differentially modulate the hypoxicresponse A: Top 20 discriminant features between perturbed and control cells for HIF1A for each treatment. Upregulated or downregulated genes are written in red or blue respectively. B: Top 20 discriminant features between perturbed and control cells for HIF2/EPAS1 for each treatment. C: Differentially expressed genes between perturbed and control cells for HIF1A and HIF2 for each treatment.

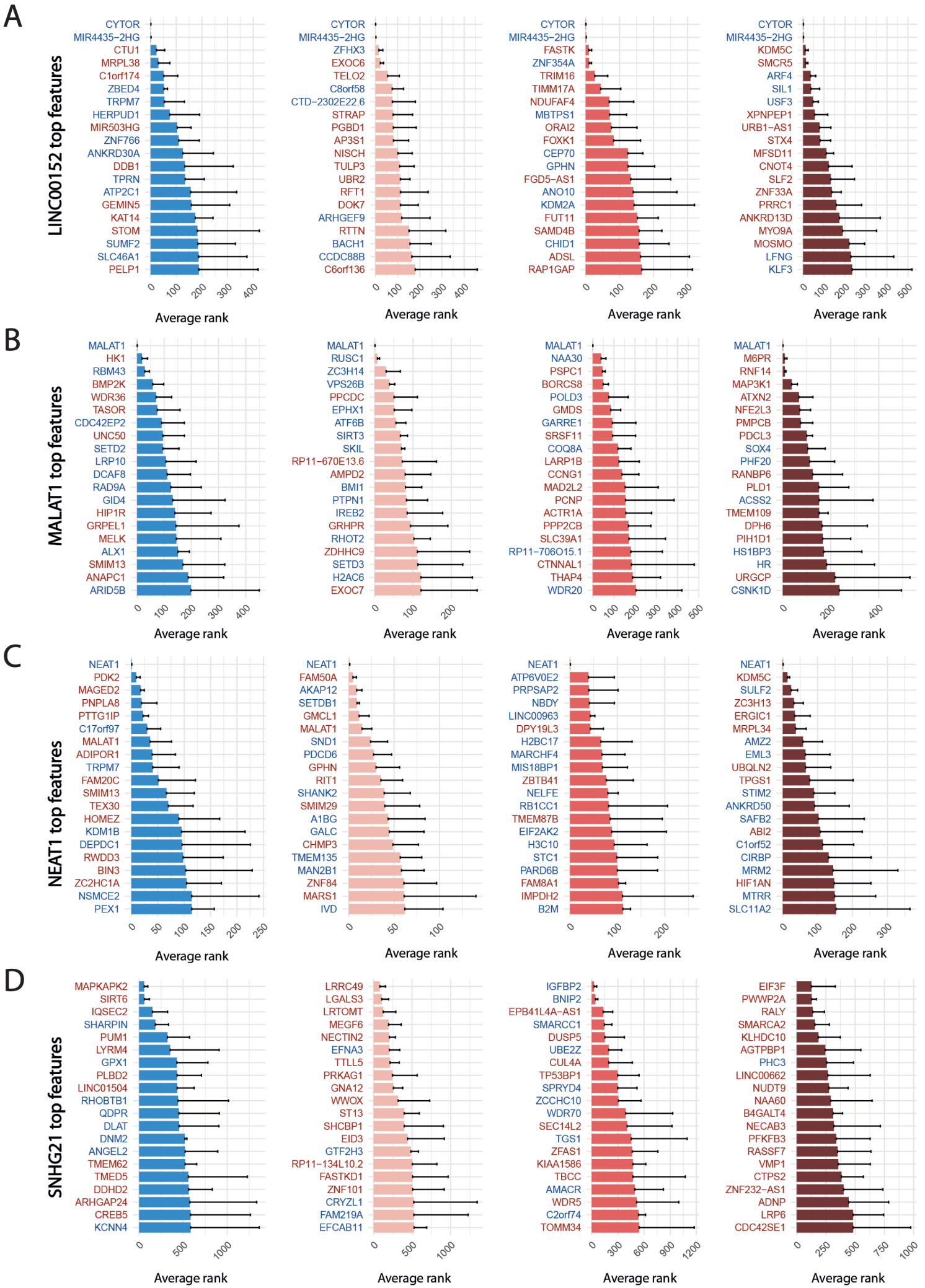

Fig 4. Top 20 discriminant features between perturbed and control cells for LINC00152 (A), MALAT1 (B), NEAT1 (C), and SNHG21 (D) for each treatment. Upregulated or downregulated genes are written in red or blue respectively.

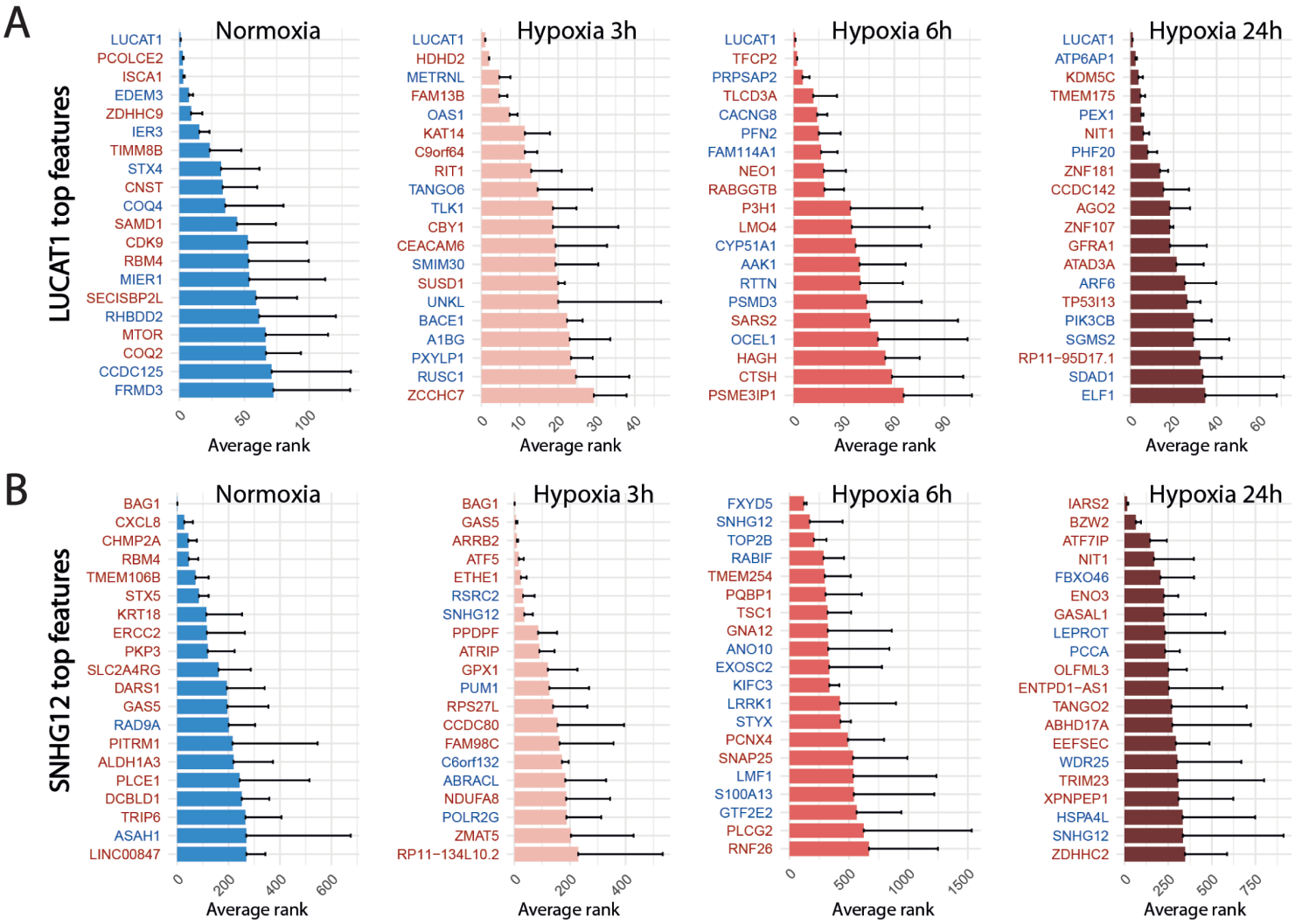

Fig 5. Top 20 discriminant features between perturbed and control cells for LUCAT1 (A) and SNHG12 (B) for each treatment. Upregulated or downregulated genes are written in red or blue respectively.

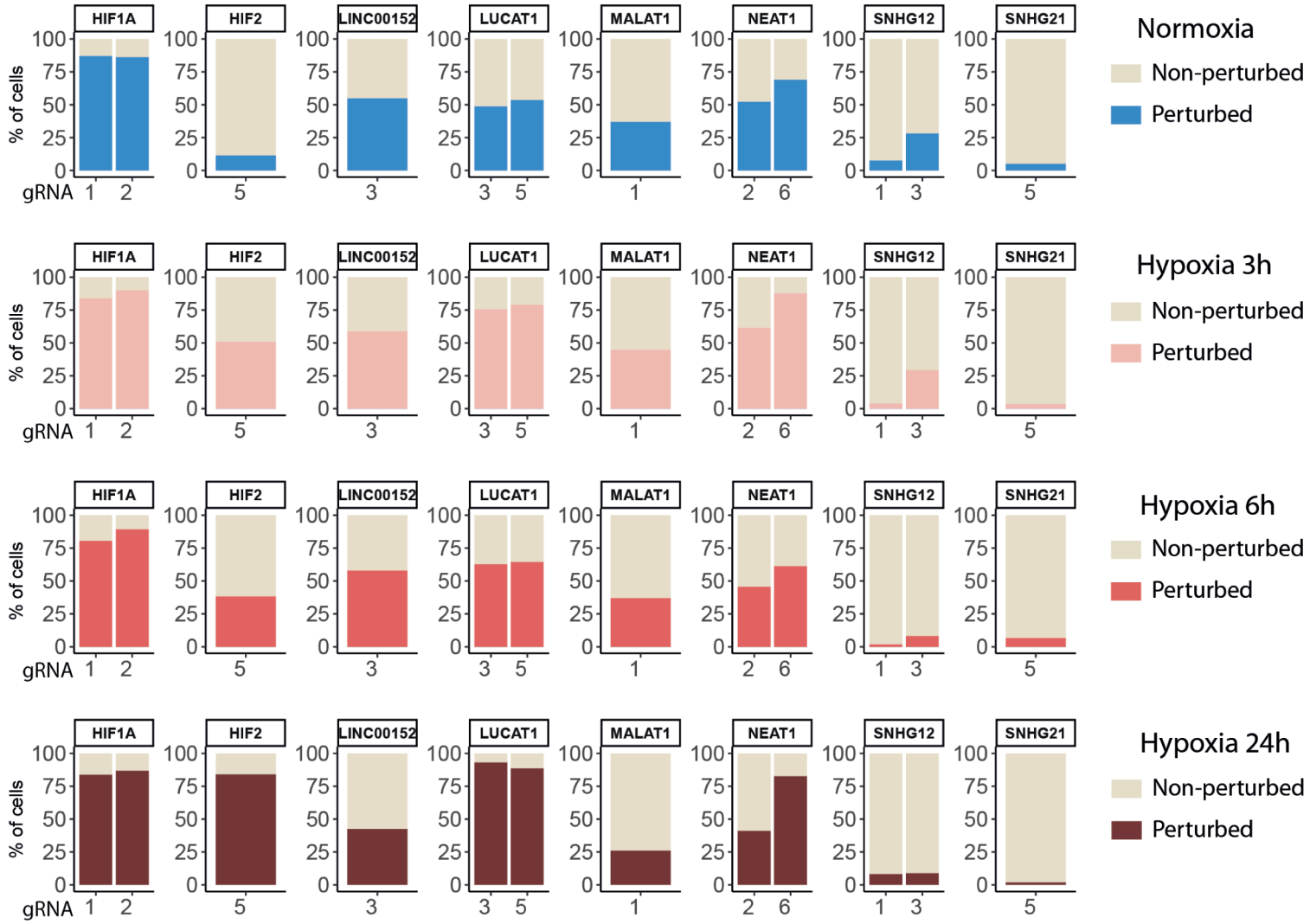

Fig 6. Percentages of targeted cells classified as perturbed or non-perturbed for each gRNA in each condition.

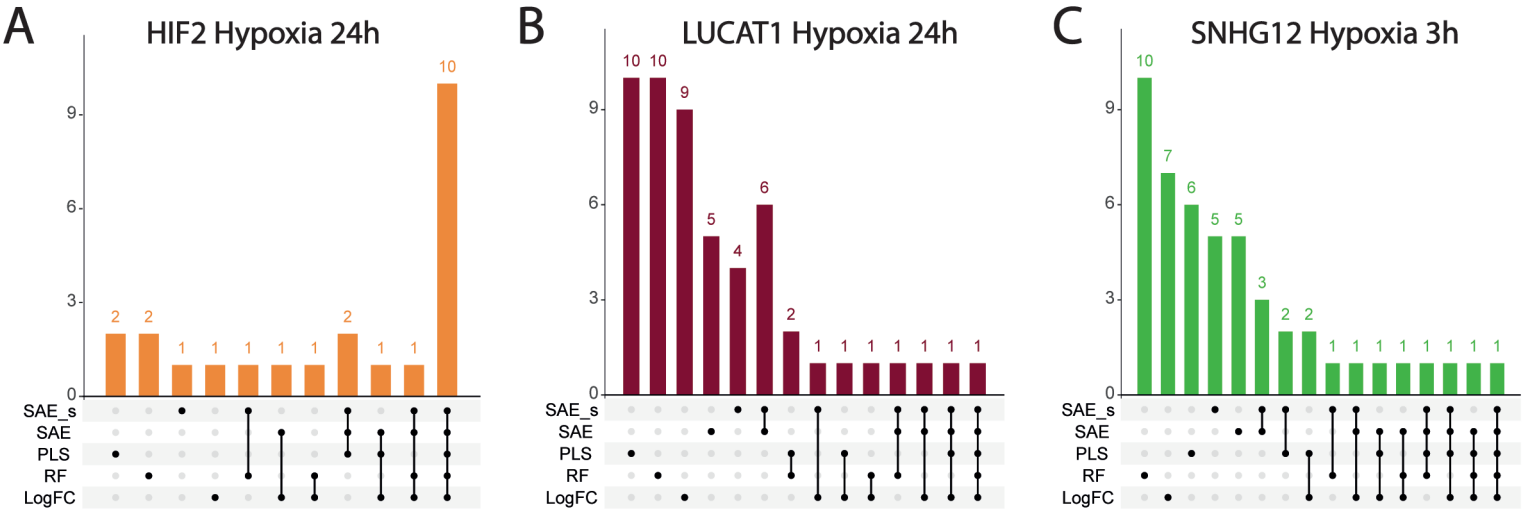

Fig 7. Intersection of the 15 first selected features lists between SAE with selection, SAE without selection, PLS, Random Forest and LFC for indicated datasets.