Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2010, Article ID 878709, 11 pages
doi:10.1155/2010/878709

Methodology Report

Identification of Multiple Hypoxia Signatures in Neuroblastoma Cell Lines by l_1-l_2 Regularization and Data Reduction

Paolo Fardin,[1] Andrea Cornero,[1] Annalisa Barla,[2] Sofia Mosci,[2] Massimo Acquaviva,[1] Lorenzo Rosasco,[2, 3] Claudio Gambini,[4] Alessandro Verri,[2] and Luigi Varesio[1]

[1] *Laboratory of Molecular Biology, Gaslini Institute, 16147 Genoa, Italy*
[2] *Department of Computer and Information Science, University of Genoa, 16146 Genoa, Italy*
[3] *Center for Biological & Computational Learning, MIT, Cambridge, MA 02139, USA*
[4] *Human Pathology Section, Gaslini Institute, 16147 Genoa, Italy*

Correspondence should be addressed to Paolo Fardin, paolofardin@ospedale-gaslini.ge.it

Received 10 February 2010; Accepted 28 April 2010

Academic Editor: Xin-yuan Guan

Hypoxia is a condition of low oxygen tension occurring in the tumor and negatively correlated with the progression of the disease. We studied the gene expression profiles of nine neuroblastoma cell lines grown under hypoxic conditions to define gene signatures that characterize hypoxic neuroblastoma. The l_1-l_2 regularization applied to the entire transcriptome identified a single signature of 11 probesets discriminating the hypoxic state. We demonstrate that new hypoxia signatures, with similar discriminatory power, can be generated by a prior knowledge-based filtering in which a much smaller number of probesets, characterizing hypoxia-related biochemical pathways, are analyzed. l_1-l_2 regularization identified novel and robust hypoxia signatures within apoptosis, glycolysis, and oxidative phosphorylation Gene Ontology classes. We conclude that the filtering approach overcomes the noisy nature of the microarray data and allows generating robust signatures suitable for biomarker discovery and patients risk assessment in a fraction of computer time.

1. Background

Neuroblastoma is the most common pediatric solid tumor, deriving from immature or precursor cells of the ganglionic lineage of the sympathetic nervous system [1, 2] endowed with remarkable heterogeneity with regard to histology and clinical behavior [3, 4]. The neuroblastoma cell lines derived from the fresh tumors show various degrees of differentiation, chromosomal alterations, and morphology and consequently, a great variability in the gene expression profile. We studied the transcriptional response of neuroblastoma cell lines to hypoxia by microarray analysis [5].

Hypoxia is a condition of low oxygen tension that characterizes many pathological tissues and that is a critical determinant of tumor cell growth, susceptibility to apoptosis, and resistance to radio and chemotherapy [6–8]. The general response to hypoxia involves activation of biochemical pathways leading to alternative ways to generate energy that becomes scant in low oxygen [9]. Hypoxia modulates gene expression through the activation of several transcription factors, among which the hypoxia-inducible transcription factor-1α (HIF-1α) [7, 10], and -2α (HIF-2α) [11] are the most studied. Rapidly expanding neuroblastoma tumors present areas of hypoxia [12] and it has been reported that HIF-2α expression correlates with poor prognosis [13, 14] suggesting a central role of hypoxia in tumor progression. HIFs transactivate the hypoxia-responsive element (HRE) present in the promoter or enhancer elements of many genes encoding angiogenic, metabolic, and metastatic factors [8, 15, 16]. However, neuroblastoma cell lines respond differently to hypoxia and the nature of the modulated genes depends strongly on the type and genetic makeup of the cell [17]. Furthermore, amplification and/or overexpression of MYCN oncogene, occurring in poor prognosis tumors, influence the transcriptional response to hypoxia of neuroblastoma cell lines [5].

The identification of molecular markers capable of discriminating the hypoxic status of the tumor may result in the discovery of new risk factors for neuroblastoma patients' stratification and potential targets for tumor therapy. To this end, we were interested in identifying hypoxia signatures that discriminate the hypoxia status of neuroblastoma cell lines. Unsupervised analysis of gene expression profile could not be applied to this system because the overwhelming effect of MYCN amplification on the transcriptome masked the response to hypoxia [5]. The application of a supervised approach represented by regularization with double optimization on microarray data, an embedded feature selection technique proposed by Zou and Hastie [18] and studied by De Mol et al. [19], identified 11 probesets capable of reliably subdividing hypoxic and normoxic cell lines [5]. These results raise the question as to whether this signature is the only possible outcome of the l_1-l_2 regularization algorithm, and hence the only source of neuroblastoma hypoxia markers, or whether additional signatures, with similar characteristics of performances and robustness can be derived from the experimental data set. Hypoxia induces massive transcriptional changes in the cell [20–22] and it is possible that additional signatures may be found by the l_1-l_2 algorithm under appropriate conditions.

The l_1-l_2 regularization algorithm has to deal with heterogeneity of the response of each cell line and with the background noise that is enhanced by the high dimensionality of the system composed by a low number of samples ($n = 18$ in this work) relative to the large number of the expression values for each sample ($p = 54,613$). The $n \ll p$ scenario is a common issue in signal processing and machine learning [23, 24]. Furthermore, the strong response of each cell line to alteration of the genetic makeup (e.g., MYCN rearrangement) tends to overcame and mask the response to hypoxia. Here, we explore the possibility that l_1-l_2 feature selection algorithm may generate new hypoxia signatures following prior knowledge-based data filtering techniques as a preprocessing step to feature selection.

Most dimensionality reduction methods, such as PCA and other unsupervised learning methods [25], rely only on the input data and may be driven by strong concurrent signals which are unrelated with, and somehow hide, the problem under study. Alternative strategies of data filtering are based on some form of prior knowledge of the biology of the system. The information collected by Gene Ontology (GO), a project having the aim of classifying gene products in terms of their associated biological processes, cellular and molecular components [26] can help identifying the pathways related to hypoxia and restricts the analysis to smaller sets of data.

In this paper, we demonstrate that l_1-l_2 regularization applied separately to probesets representing genes belonging to selected GO ontologies has the capability to generate robust hypoxia signatures, equivalent to that generated by the whole data set yielding biologically relevant information in a fraction of computer time.

2. Materials and Methods

2.1. Microarray Experiments. Microarray data were downloaded from the Gene Expression Omnibus public repository at National Center for Biotechnology Information database (accession number GSE15583). These data represent the gene expression profile of nine neuroblastoma cell lines cultured under normoxic (20% O_2) or hypoxic (1% O_2) conditions for 18 hours as detailed in [5], to obtain a total of 18 samples. Affymetrix HG-U133 Plus 2.0 GeneChip (Affymetrix, SantaClara, CA) were used for this study. Gene expressions were then extracted from CEL files and normalized using the Robust Multichip Average (RMA) method [27] by running an R script using the Bioconductor [28] package *affy*.

Comparative analysis of hypoxic relative to normoxic expression profiles for each cell line was conducted on GeneSpring 7.3 software (Agilent Technologies). Gene expression data were normalized using "per chip normalization" and "per gene normalization" algorithms implemented in GeneSpring. First, each signal was normalized based upon the median signal in that chip ("per chip normalization"). We then performed a median centering using "per gene normalization" function by which each normalized value is corrected based upon the median of the measurements for that gene in all samples. Finally, only genes that were modulated by at least 2-fold between hypoxic and normoxic cells were considered differentially expressed.

2.2. Gene Ontology. The biological groups were obtained from the literature and they were divided into three main categories depending on the biological characteristics of our experimental system (see Table 2): (i) hypoxia related groups [9, 17, 29]; (ii) MYCN related groups [30–32]; (iii) neuroblastoma related groups [31, 33, 34]. The selected functional groups were then filtered to avoid overlapped or duplicated categories and were defined according to predetermined pathways and functional categories annotated by the Gene Ontology project [26].

2.3. Supervised Methods for Gene Selection: l_1-l_2 Regularization. The core of our approach is the l_1-l_2 regularization originally presented in [18] and further developed and studied in [35, 36]. To describe such method we first fix some notation in the learning framework. Assume we are given a collection of n examples/subjects, each represented by a p-dimensional vector x of gene expressions. Each sample is associated with a binary label Y, assigning it to a class (e.g., patient or control). The dataset is therefore represented by a $n \times p$ matrix X, where $p \gg n$ and Y is the n-dimensional labels vector. We consider a linear model $f(x) = \langle x, \beta \rangle$. Note that $\beta = \beta_1, \ldots, \beta_p$ is a vector of weight coefficients and each probeset is associated to one coefficient. A classification rule can be then defined taking sign $(f(x)) = \text{sign}(\langle x, \beta \rangle)$. If β is sparse, that is some of its entries are zero, then some genes will not contribute in building the estimator. The estimator defined by l_1-l_2 regularization solves the following optimization problem:

$$\beta_{l1l2} = \underset{\beta}{\operatorname{argmin}} \Big[||Y - X\beta||_2^2 + \tau \Big[||\beta||_1 + \varepsilon ||\beta||_2^2 \Big] \Big], \qquad (1)$$

where the least square error is penalized with the l_1 and l_2 norm of the coefficient vector. The least square term ensures fitting of the data whereas adding the two penalties allows avoiding overfitting. The relative weight of the two terms is controlled by the parameter ε. The role of the two penalties is different, the l_1 term (sum of absolute values) enforces the solution to be sparse while the l_2 term (sum of the squares) preserves correlation among the genes. This approach guarantees consistency of the estimator [19] and enforces the sparsity of the solution by the l_1 term, while preserving correlation among input variables with the l_2 term. Differently to [18] we follow the approach proposed in [36], where the solution β_{l1l2}, computed through the simple iterative soft thresholding, is followed by regularized least squares (RLSs) to estimate the classifier on the selected features. The parameter ε in the l_1-l_2 regularization is fixed a priori and governs the amount of correlation. By tuning ε in $(0, +\infty)$ we obtain a one-parameter family of solutions which are all equivalent in terms of prediction accuracy, but differ on the degree of correlation among the selected features. In practice, ε has an upper bound, $\varepsilon_{\max}$, such that for $\varepsilon > \varepsilon_{\max}$ selection does not change, because all correlated features were already selected with $\varepsilon = \varepsilon_{\max}$. By setting $\varepsilon = 100$, the maximal value, the maximal gene list, which is correlation aware, is obtained. Conversely, the minimal list is obtained for values of ε equal to or lower than 1.

The training for selection and classification requires the choice of the regularization parameters for both l_1-l_2 regularization and RLS denoted with λ^* and τ^*, respectively. Hence, statistical significance and model selection is performed within double-selection bias-free cross-validation loops (see [37] for details). The classification performance of the system is measured by the leave-one-out error that is the percentage of misclassified samples. In other words, leave-one-out error is equal to one minus accuracy. In order to assess a common list of probesets, it is necessary to choose an appropriate criterion [38]. We based ours on the *frequency*, that is, we decided to promote as relevant variables the most stable probesets across the lists. The complete validation framework comprising the l_1-l_2 regularization is implemented in MATLAB code.

2.4. Correlation Analysis. The correlation among the probesets selected by the l_1-l_2 algorithm was performed as previously described in [39]. Briefly, we build blocks of correlated probesets using a variation of well-known agglomerative clustering techniques based on Pearson distance. We first examine the minimal list, which genes are clustered via hierarchical clustering with correlation distance and average linkage. Since no objective algorithm, other than heuristics, is available for establishing the number of clusters, for each GO class the cut of the hierarchical graph determining the number of clusters is chosen following visual examination of the graph. In particular we set the cut at 0.75 of the maximum linkage value in the dendrogram. For each GO class the cut of the hierarchical graph determining the number of clusters is chosen following visual examination of the graph. Each probeset in the maximal list is then assigned to the cluster which average correlation with the given probeset is the highest. In this way we populate the clusters built from the minimal list with all the probesets coming from the maximal list. The correlation analysis was performed using MATLAB Statistic Toolbox.

2.5. HRE Analysis. We mapped the HRE elements in the promoter regions of the genes represented in the Affymetrix HG-U133 Plus 2.0 GeneChip. We downloaded the annotation file for the HG-U133 Plus 2.0 from NetAffx Analysis Center http://www.affymetrix.com/ and the dataset was restricted to the known mRNA sequences listed in the Ensembl database V56 [40]. The regulatory regions were retrieved from Ensembl database using Ensembl Perl APIs. We operationally defined as “promoter” the first 2,000 base pairs upstream the transcription initiation site and generated a dataset containing the promoters of the genes coding for the mRNAs spotted on the chip. The HRE matrix has been obtained from 69 experimental validated human HRE sequences [41] with MatDefine tool (Genomatix Software GmbH). HRE consensus elements [(G|C|A|T) (C|G|T|A) (G|C|A|T) (T|G|C|A) (A|G) (CGTG) (C|G|T|A) (G|C|A|T) (G|C|T|A) (C|G|T|A)] were searched in the promoter sequences with MatInspector software (Genomatix) with core similarity = 1 and optimized matrix similarity. About 33% of the promoters contain at least one HRE consensus element. χ^2 was used to evaluate the significance of the HRE frequency in the promoter regions of genes belonging to the different signatures. $P < .01$ was considered significant

3. Results and Discussion

We studied nine neuroblastoma cell lines [2] heterogeneous with respect to MYCN amplification and morphology (Table 1). The cell lines were cultured under normoxic and hypoxic conditions for 18 hours and the total RNA was tested for gene expression profiling using the Affymetrix HG-U133 Plus 2.0 platform. The response to hypoxia of each individual cell line was first analyzed by measuring the fold change as the ratio of the expression level between hypoxic and normoxic samples. We found that the response of each neuroblastoma cell line to hypoxia is characterized by a high number of modulated genes ranging from 855 to 1609 for the upregulated and from 758 to 1317 for the downregulated probesets (Table 1). However, the modulated genes changed from cell line to cell line (data not shown) and only the application of a strong feature selection technique, represented by the l_1-l_2 regularization, allowed to identify a single signature of 11 probesets (All-chip signature) discriminating the normoxic and the hypoxic status [5].

The large amount of hypoxia-modulated genes suggests that additional hypoxia signatures may be identified if we reduce the background noise of the system. To this end, we applied a data filtering strategy based on prior knowledge. We restricted our analysis to the genes known to be involved in the hypoxic response on the bases of our reading of the literature and comprised in the biological processes according to the Gene Ontology (GO) classification [13]. The selection of the GO classes was based on the reports of hypoxia modulated genes without attempting to distinguish the various cell types under investigation. 13 biological

Table 1: Modulation of gene expression by hypoxia in neuroblastoma cell lines.

cell line			gene expression[1]	
name	Morphology[2]	MYCN amplification[3]	up-regulated	down-regulated
ACN	neuroblast (N)	−	1400	1317
SHEP-2	epithelial (S)	−	1609	1043
GI-ME-N	neuroblast (N)	−	762	881
SK-N-F1	epithelial (S)	−	1206	1051
SK-N-SH	neuroblast/epithelial (I)	−	922	758
SK-N-BE(2)c	neuroblast/epithelial (I)	+	855	1273
IMR-32	neuroblast (N)	+	1000	1077
LAN-1	neuroblast (N)	+	1061	1016
GI-LI-N	neuroblast (N)	+	1516	1002

[1] Number of modulated probesets by hypoxia (1%O_2 for 18 hours). [2] N: neuroblast; S: substrate adherent; I: intermediate [2]. [3] For reference see [2].

processes that are involved in hypoxia response were selected (Table 2). We reasoned that this approach would restrict the analysis to the probesets that have a high impact on the hypoxic response while potentially eliminating the noisy features. To explore the potential interference from MYCN status in the classification process, we selected and tested 7 biological processes involved in MYCN activity (Table 2). Finally, we selected a third group of GO processes related to the neuroblastoma biology as a control. For each of the 38 classes shown in Table 2, the l_1-l_2 algorithm selected a list of hypoxia discriminating probesets and calculated the corresponding classification leave-one-out error. The output of the l_1-l_2 regularization algorithm depends on the parameter ε that governs the amount of correlation allowed among the probesets. We set $\varepsilon = 100$, the maximal value, to obtain the most comprehensive signature maximizing the number of correlated probesets to be included in the output [5].

The validation has been performed by leave-one-out cross-validation on the 18 samples. The 18 cross-validation loops produced 18 lists of probesets. Then, a unique list is obtained as the union of the probesets included in the 18 lists, with a frequency score calculated as the frequency of each probeset in the 18 lists generated by the cross validation loops. Stable probesets were defined as those characterized by a frequency score equal to, or higher than, 50% as previously reported in [5]. The use of cross validation allows the selection protocol to generate an unbiased and objective output [42] beyond the theoretical results that guarantee the robustness of the core algorithm [19]. The discriminatory power of the probeset lists is represented by the classification performance. A leave-one-out error of 20% was chosen as the cutoff level for the classification performance. The leave-one-out error of the All-chip signature is 17% [5].

The only classes characterized by a list of selected probesets capable of generating a leave one-out error lower than the 20% cutoff were apoptosis (17%), glycolysis (11%), and oxidative phosphorylation (11%) (Table 2), all of them belonging to the hypoxia biological group. These results demonstrate that, within each of the above classes, there is a list of probesets capable of discriminating the condition of the cell lines thereby defining three new neuroblastoma hypoxia signatures, named apoptosis signature, glycolisis signature, and oxidative phosphorylation signature. As expected, there were no GO classes belonging to the MYCN or neuroblastoma biological groups that generated hypoxia signatures, supporting the validity of our choice of hypoxia-related GO functional classes. Although MYCN represents a strong signal that drives major transcriptome difference in neuroblastoma cell lines [5], our results show that there are no enough discriminatory genes in the MYCN-related processes. These results demonstrate that the feature selection method applied is capable of revealing the differences occurring among hypoxic and normoxic neuroblastoma cell lines by filtering out strong competing signals, such as MYCN amplification status.

The list of the probesets comprising the 11 probesets (All-chip signature) [5] and the newly identified signatures is shown in Table 3 and consists of 10 probesets for apoptosis signature, 3 for glycolysis-signature, and 32 for the oxidative phosphorylation signature. The new signatures highlight 41 probesets that were not previously included in the All-chip signature and contribute to the discrimination of the hypoxic status. Furthermore, the 32 probesets of the oxidative phosphorylation signature does not overlap with the All-chip signature, demonstrating that the increased resolution generated by data filtering allows the identification of previously discarded relevant GO processes. The hypoxia signatures present in the literature show different sizes and gene composition [9, 43–46]. Since different cell types respond heterogeneously to hypoxia by modulating different set of genes, we decided to compare our results with the published hypoxic gene signatures obtained from neuroblastoma cell lines [47] (Table 4). In order to make the comparison feasible, the probesets constituting our signatures have been collapsed to gene symbol. The overlapping genes are underlined in bold in Table 4. While important differences among the signatures exist, the comparison highlights a general consistency. In fact, an overlap is present in All-chip (3/8 genes), apoptosis (2/4 genes), and glycolysis (2/2 genes) signatures. Interestingly, there is no overlap (0/24 genes) among the results published by Jögi et al. [47] and the oxidative phosphorylation signature.

About 33% of the genes spotted on the chip present a HRE sequence in the promoter region. We investigated

Table 2: Functional gene classes.

Biological group[1]	Functional class[2]	GO number[3]	no. of probesets[4]	error(%)[5]
Hypoxia	Angiogenesis	GO: 1525	257	39
	Apoptosis*	**GO: 6915**	**1366**	**17***
	Cell proliferation	GO: 8283	1406	28
	DNA repair	GO: 6281	567	44
	Glucose import	GO: 46323	10	89
	Glucose Transport	GO: 15758	48	33
	Glycolysis*	**GO: 6096**	**128**	**11***
	Iron ion homeostasis	GO: 6879	65	39
	Notch signaling pathway	GO: 7219	134	44
	Oxidative phosphorylation*	**GO: 6119**	**154**	**11***
	Oxygen transport	GO: 15671	38	28
	Regulation of pH	GO: 6885	41	72
	Response to hypoxia	GO: 1666	32	28
MYCN	G1-S transition of mitotic cell cycle	GO: 82	62	50
	Proteasomal ubiquitin-dependent protein catabolism	GO: 43161	33	50
	Protein folding	GO: 6457	601	22
	Ribosome biogenesis and assembly	GO: 42254	170	28
	Structural constituent of ribosome	GO: 3735	549	44
	Translational elongation	GO: 6414	56	56
	Translational initiation	GO: 6413	173	44
Neuroblastoma	Axon guidance	GO: 7411	109	50
	Axonal fasciculation	GO: 7412	4	100
	Cell cycle arrest	GO: 7050	210	39
	Dendrite morphogenesis	GO: 16358	12	72
	Glial cell migration	GO: 8347	2	100
	Inactivation of MAPK activity	GO: 188	50	44
	Nervous system development	GO: 7399	1284	22
	Neuron migration	GO: 1764	7	94
	Positive regulation of neuron differentiation	GO: 45666	6	100
	Regulation of axon extension	GO: 30516	14	100
	Regulation of G-protein coupled receptor protein signaling pathway	GO: 8277	88	39
	Regulation of neuronal synaptic plasticity	GO: 48168	8	100
	Regulation of neurotransmitter secretion	GO: 46928	11	61
	Synaptic vesicle transporter	GO: 48489	36	72
	Vesicle organization and biogenesis	GO: 16050	6	44

[1] Functional classes were clustered into three main biological groups depending on the characteristic of the experimental system and accordingly to the literature. [2] Defined according to the predetermined pathways and functional categories annotated by the Gene Ontology project [26]. [3] Gene Ontology ID [26]. [4] Number of probesets present in Affymetrix HG-U133 Plus 2.0 GeneChip belonging to the selected classes. [5] Leave-one-out error, as calculated by l_1-l_2 regularization by setting $\varepsilon = 100$ and frequency score = 50. *Functional classes with leave-one-out error <20%.

whether there was enrichment in HRE containing promoter in the genes composing our signatures. We found that all the signatures are significantly enriched ($P < .01$) in genes containing HRE (Table 4). In particular, all the genes included in All-chip, apoptosis, and glycolysis signatures contain at least one HRE, while HRE containing genes constitute 91% of the oxidative phosphorylation signature. These results support the idea that our signatures are associated with the hypoxia status.

The whole signature, rather than individual genes, is important for discriminating the hypoxic status. For example, VEGF is a gene whose expression is strongly related to hypoxia [45] and is part of the apoptosis and angiogenesis classes, both of which are part of the hypoxia biological group. However, the contribution of VEGF probesets is not sufficient to reach the discriminatory power required to generate a significant signature out of the angiogenesis class as opposed to the apoptosis class.

The strong discriminatory power of the signatures can be visualized by a 3-dimensional representation of the probesets projected on their 3 principal components. l_1-l_2 algorithm produces a multigene model but the multidimensional representation can be well approximated by the tridimensional picture when the number of probesets is

Table 3: Hypoxia signatures generated after data filtering.

			Signatures			
Probeset[1]	Gene Name	GeneBank[2]	Apo[3]	Gly[3]	OxP[3]	All[3]
201848_s_at	BNIP3	U15174	100	—	—	100
201849_at	BNIP3	NM_004052	83	—	—	100
210512_s_at	VEGF	AF022375	78	—	—	100
211527_x_at	VEGF	M27281	61	—	—	—
212171_x_at	VEGF	H95344	61	—	—	—
219232_s_at	EGLN3	NM_022073	61	—	—	—
210513_s_at	VEGF	AF091352	56	—	—	—
221478_at	BNIP3L	AL132665	56	—	—	—
221479_s_at	BNIP3L	AF060922	56	—	—	—
222847_s_at	EGLN3	AI378406	56	—	—	—
202022_at	ALDOC	NM_005165	—	100	—	100
1558365_at	PGK1	AK055928	—	72	—	—
228483_s_at	PGK1	BE856250	—	72	—	—
208972_s_at	ATP5G1	AF100741	—	—	100	—
222270_at	SMEK2	BF509069	—	—	100	—
1554847_at	ATP6V1B1	AY039759	—	—	94	—
218201_at	NDUFB2	NM_004546	—	—	94	—
203189_s_at	NDUFS8	NM_005006	—	—	89	—
203371_s_at	NDUFB3	NM_002496	—	—	89	—
218200_s_at	NDUFB2	NM_013387	—	—	89	—
203606_at	NDUFS6	NM_002494	—	—	83	—
204125_at	NDUFAF1	NM_001687	—	—	83	—
214241_at	NDUFB8	BE043477	—	—	78	—
230598_at	KIAA1387	AI742966	—	—	78	—
203190_at	NDUFS8	NM_002496	—	—	72	—
207335_x_at	ATP5I	NM_006294	—	—	72	—
208745_at	ATP5L	AF092131	—	—	72	—
203039_s_at	NDUFS1	NM_021074	—	—	67	—
203613_s_at	NDUFB6	NM_004553	—	—	67	—
208746_x_at	ATP5L	AA917672	—	—	67	—
210453_x_at	ATP5L	U33833	—	—	67	—
211752_s_at	NDUFS7	AL050277	—	—	67	—
228816_at	ATP6AP1L	AU153583	—	—	67	—
207573_x_at	ATP5L	NM_005176	—	—	61	—
226616_s_at	NDUFV3	AW241758	—	—	61	—
200096_s_at	ATP6V0E	BC005876	—	—	56	—
214923_at	ATP6V1D	AV717561	—	—	56	—
226209_at	NDUFV3	BC006215	—	—	56	—
200078_s_at	ATP6V0B	BC035703	—	—	50	—
210206_s_at	DDX11	AF061735	—	—	50	—
213378_s_at	DDX11	AV711183	—	—	50	—
214244_s_at	ATP6V0E	AA723057	—	—	50	—
218190_s_at	UCRC	NM_004549	—	—	50	—
241755_at	UQCRC2	BE467348	—	—	50	—
243498_at	ATP5J	BG010493	—	—	50	—
202887_s_at	DDIT4	NM_019058	—	—	—	94
223193_x_at	E2IG5	AF201944	—	—	—	94
224345_x_at	E2IG5	AF107495	—	—	—	89
225342_at	AK3L1	AK026966	—	—	—	78
226452_at	PDK1	AU146532	—	—	—	78
236180_at	—	W57613	—	—	—	61
235850_at	WDR5B	BF434228	—	—	—	50

[1] Probeset ID according to Affymetrix HG-U133 Plus 2.0 GeneChip. [2] GenBank mRNA accession number. [3] Frequency score as calculated by l_1-l_2 regularization for the selected probesets in the hypoxia signatures compared to the All-Chip signature. Apo: apoptosis; Gly: glycolisis; OxP: oxidative phosphorylation; All: All-chip.

Table 4: Hypoxia gene signatures overlap and HRE analysis.

Signature[(1)]	Gene Name[(2)]	HRE[(3)]
Apoptosis	BNIP3	9
	BNIP3L	5
	EGLN3	3
	VEGF	4
Glycolysis	**ALDOC**	3
	PGK1	1
Oxydative Phosphorylation	ATP5G1	2
	ATP5I	5
	ATP5L	0
	ATP6V0B	6
	ATP6V0E	2
	ATP6V0E	2
	ATP6V1B1	0
	ATP6V1D	1
	DDX11	1
	LOC92270	3
	NDUFAF1	1
	NDUFB2	1
	NDUFB3	1
	NDUFB6	1
	NDUFB8	2
	NDUFS1	3
	NDUFS6	9
	NDUFS7	2
	NDUFS8	3
	NDUFS8	3
	NDUFV3	7
	SMEK2	3
	UCRC	2
	UQCRC2	3
All-chip	AK3L1	4
	ALDOC	3
	BNIP3	9
	DDIT4	1
	E2IG5	5
	PDK1	4
	VEGF	4
	WDR5B	3

[(1)] Hypoxia gene signatures. [(2)] Multiple probesets were collapsed to single genes. The genes overlapping with Jögi et al. hypoxia signature are underlined in bold. [(3)] Number of HRE sequences found in the promoter region.

not too large. Figure 1 depicts the separation of probesets belonging to the glycolysis and shows that the two classes of normoxic and hypoxic cell lines are clearly separated in the multidimensional space.

In conclusion, we demonstrate that, upon data reduction the l_1-l_2 algorithm can identify new hypoxia signatures that have equivalent discriminatory power relative to that obtained by the analysis of the whole transcriptome. From the computational stand point, this process allows to reduce the computer time by approximately 10 times, from days to hours with an average machine, facilitating the analysis of the data. Finally, it is important to highlight the possibility of applying our method to different experimental settings by choosing appropriate selection of GO processes.

Our prior knowledge-based method produces nested lists of relevant probesets but does not highlight the correlation among them [39], and it should be completed by a postprocessing step depicting the correlation structure. The correlation within the oxidative phosphorylation signature is shown in Figure 2. We computed a distance matrix based

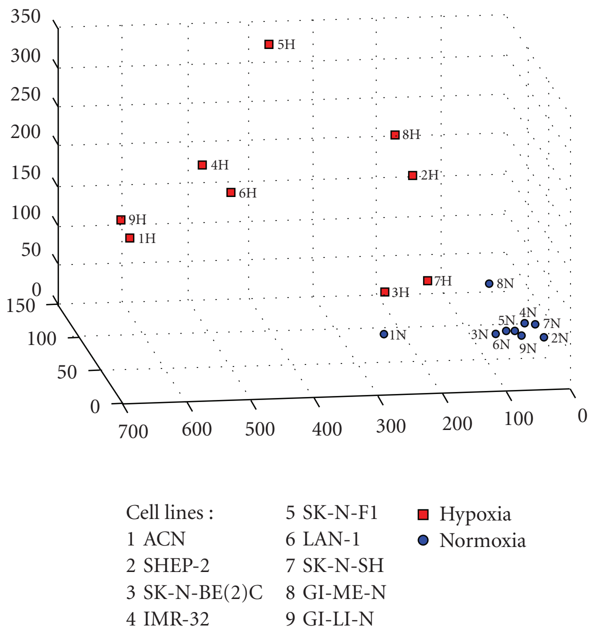

Figure 1: Three-dimensional representation of the principal components analysis of the Glycolysis GO process. Principal components representation of the multivariate analysis performed on the 9 cell lines by l_1-l_2 algorithm. This figure illustrates a 3-dimensional visualization of the dataset restricted to the selected probesets projected on their 3 principal components. Red squares (H) represent the cell lines in hypoxic status and the blue circles (N) the corresponding cell lines in normoxic status. The numbers indicate the cell lines.

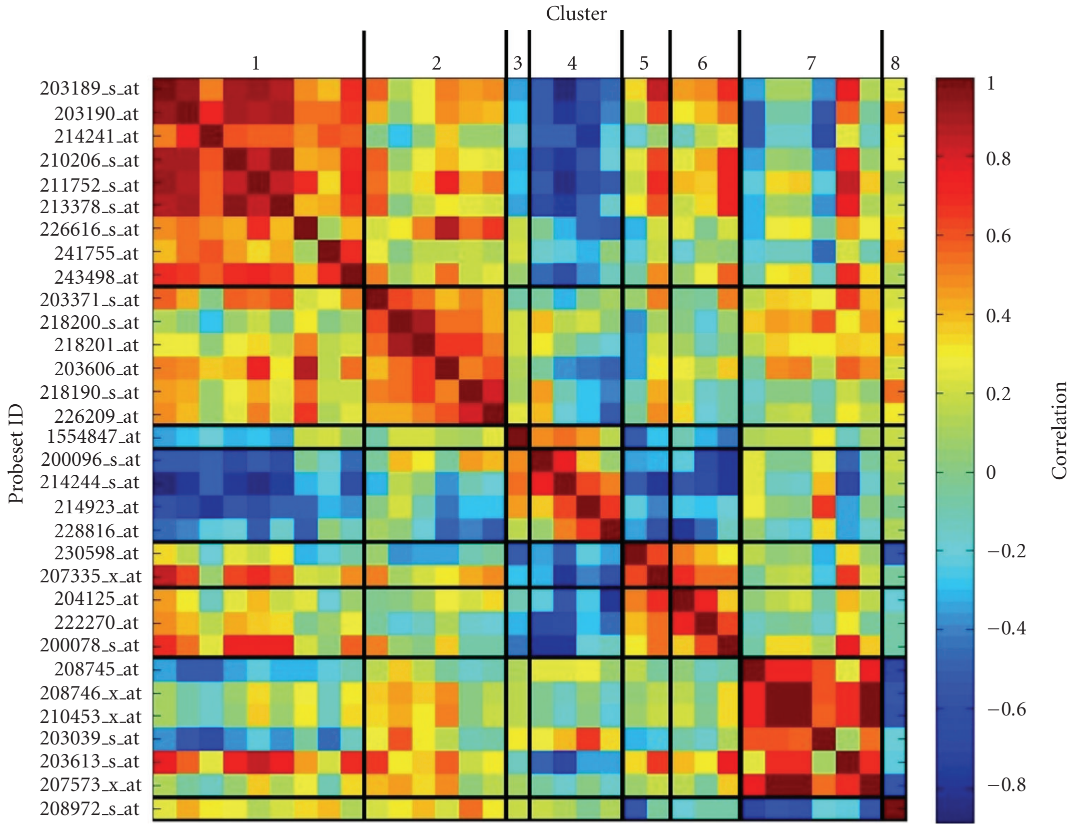

Figure 2: Symmetrical heatmap of the correlation analysis for oxidative phosphorylation-signature probesets. The correlation values among the 32 probesets selected by l_1-l_2 algorithm ($\varepsilon = 100$) for oxidative phosphorylation process are reported. The probesets, named accordingly to Affymetrix HG-U133 Plus 2.0 GeneChip platform, have been subdivided into 8 clusters according to the similarity among their correlation profiles by hierarchical clustering. Cluster ID is indicated at the top of the figure and correlation scale is reported on the right side.

Table 5: Tuning the oxidative phosphorylation signature by changing the ε parameter.

Cluster[(1)]	Probeset[(2)]	Gene Name	$\varepsilon = 1$[(3)]	$\varepsilon = 100$[(4)]
1	203189_s_at	NDUFS8	89	89
	203190_at	NDUFS8	56	72
	214241_at	NDUFB8	94	78
	210206_s_at	DDX11	—	50
	211752_s_at	NDUFS7	—	67
	213378_s_at	DDX11	—	50
	226616_s_at	NDUFV3	—	61
	241755_at	UQCRC2	—	50
	243498_at	ATP5J	—	50
2	203371_s_at	NDUFB3	56	89
	218200_s_at	NDUFB2	72	89
	218201_at	NDUFB2	89	94
	203606_at	NDUFS6	—	83
	218190_s_at	UCRC	—	50
	226209_at	NDUFV3	—	56
3	1554847_at	ATP6V1B1	100	94
4	200096_s_at	ATP6V0E	50	56
	214244_s_at	ATP6V0E	72	50
	214923_at	ATP6V1D	—	56
	228816_at	ATP6AP1L	—	67
5	230598_at	SMEK2	72	78
	207335_x_at	ATP5I	—	72
6	204125_at	NDUFAF1	67	83
	222270_at	SMEK2	100	100
	200078_s_at	ATP6V0B	—	50
7	208745_at	ATP5L	89	72
	208746_x_at	ATP5L	72	67
	210453_x_at	ATP5L	50	67
	203039_s_at	NDUFS1	—	67
	203613_s_at	NDUFB6	—	67
	207573_x_at	ATP5L	—	61
8	208972_s_at	ATP5G1	100	100

[(1)] Cluster number according to Figure 2. [(2)] Probeset ID according to Affymetrix HG-U133 Plus 2.0 GeneChip. [(3)] Frequency score (%) as calculated by l_1-l_2 regularization for the selected probesets by setting $\varepsilon = 1$. [(4)] Frequency score (%) as calculated by l_1-l_2 regularization for the selected probesets by setting $\varepsilon = 100$.

on the expression values of the probesets and subdivided it into 8 modules by hierarchical clustering. These modules represent subgroups of correlated probesets that are positively or negatively associated to the hypoxic status. This information is important to pick the correct probesets in order to assess the expression of these markers in the *in vivo* setting. Furthermore, these data lend themselves to the tuning of the ε parameter that is part of the l_1-l_2 algorithm.

The output of the l_1-l_2 regularization algorithm depends on the free parameter ε that governs the amount of correlation allowed among the probesets and selects the amount of probesets to be included in the signature. By setting $\varepsilon = 100$, the maximal value, we can obtain a comprehensive signature more descriptive of the biology of the system. By setting $\varepsilon = 1$, we can obtain an equally discriminating signature with fewer genes thereby more effective in identifying critical biomarkers for diagnostic applications [5]. We analyzed the effects of tuning ε on the oxidative phosphorylation signature. The results are shown in Table 5, where the probesets selected by l_1-l_2 regularization with both $\varepsilon = 1$ and $\varepsilon = 100$ for oxidative phosphorylation are listed. The results demonstrated that the reduction in ε is associated with a smaller signature (from 32 to 16 probesets) as expected by the fact that correlated probesets tend to be discarded.

4. Conclusions

The identification of signatures discriminating the hypoxic status of the tumor cell may be important for our understanding of the biology of neuroblastoma tumors and for the stratification of the patients in risk groups. One way to generate a robust and reliable hypoxic signature is the application of a supervised approach represented by l_1-l_2 regularization that generates an 11 probesets signature discriminating the hypoxic status of our panel of nine neuroblastoma cell lines.

Here, we demonstrate that l_1-l_2 feature selection algorithm generates new and robust hypoxia signatures following prior knowledge-based data filtering techniques as a preprocessing to feature selection. These new signatures have the same discriminatory power as that generated by the whole data set and yield biologically relevant information in a fraction of computer time.

The data filtering is based upon the use of the prior information contained in GO and the literature, and it allows restricting the analysis to smaller data sets. This process filters out not only many noisy probesets but also the probesets selected from the all-chip analysis whose strong relation with hypoxia hid some weaker but important genes. l_1-l_2 regularization algorithm following data filtering selects probesets that were not the first chosen when all the probesets were considered. The prior knowledge utilized in setting up the filter, comes from the current literature from which we derived the molecular pathways that are important for the response of the cell to the hypoxic environment. These pathways were gathered in the hypoxia biological group. Interestingly, the new signatures were found only in this group and not in other collections of GO pathways like those related to the effects of the MYCN oncogene or to the neuroblastoma biology. In general, the identification of the GO classes related to the phenomenon under investigation may be an empirical, but effective way to target the potential source of signatures to be fed to the l_1-l_2 regularization. We speculate that this approach could be used to address questions that go beyond the hypoxic status and may find signatures characterizing other pathophysiological situations provided that there is a relevant cellular model and there are sufficient insights in the underlying molecular mechanisms.

The nested structure of the selected gene lists allows the choice of the desired level of complexity, which is the magnitude of signature, maintaining all the information extracted from the data. For example, the minimal list may be preferable when interested in finding biomarkers to be used on large-scale diagnostic tests due to potential constrains on time, cost, and resources.

Finally, working on a limited number of probesets has a major impact on the computational time required for the analysis that changes from days to hours, thereby allowing more leeway to the study of the dataset.

Acknowledgment

This paper was supported by Fondazione Italiana per la Lotta al Neuroblastoma, the Italian Association for Cancer Research (AIRC), the Italian Health Ministry, the EU Integrated Project Health-e-Child IST-2004-027749 and Compagnia di San Paolo Project 4998- ID/CV 2007.0887.

References

[1] K. De Preter, J. Vandesompele, P. Heimann, et al., "Human fetal neuroblast and neuroblastoma transcriptome analysis confirms neuroblast origin and highlights neuroblastoma candidate genes," *Genome Biology*, vol. 7, no. 9, article R84, 2006.

[2] J. Maris, M. Hogarty, R. Bagatell, and S. Cohn, "Neuroblastoma," *The Lancet*, vol. 369, no. 9579, pp. 2106–2120, 2007.

[3] C. J. Thiele, "Neuroblastoma," in *Human Cell Culture*, J. R. W. Master and B. Palsson, Eds., pp. 21–22, Kluwer Academic, London, UK, 1999.

[4] J. L. Weinstein, H. M. Katzenstein, and S. L. Cohn, "Advances in the diagnosis and treatment of neuroblastoma," *Oncologist*, vol. 8, no. 3, pp. 278–292, 2003.

[5] P. Fardin, A. Barla, S. Mosci, L. Rosasco, A. Verri, and L. Varesio, "The l_1 -l_2 regularization framework unmasks the hypoxia signature hidden in the transcriptome of a set of heterogeneous neuroblastoma cell lines," *BMC Genomics*, vol. 10, p. 474, 2009.

[6] G. L. Semenza, "Targeting HIF-1 for cancer therapy," *Nature Reviews Cancer*, vol. 3, no. 10, pp. 721–732, 2003.

[7] G. L. Semenza, "HIF-1 and tumor progression: pathophysiology and therapeutics," *Trends in Molecular Medicine*, vol. 8, no. 4, pp. S62–S67, 2002.

[8] P. Carmeliet, Y. Dor, J.-M. Herber, et al., "Role of HIF-1α in hypoxiamediated apoptosis, cell proliferation and tumour angiogenesis," *Nature*, vol. 394, no. 6692, pp. 485–490, 1998.

[9] A. L. Harris, "Hypoxia—a key regulatory factor in tumour growth," *Nature Reviews Cancer*, vol. 2, no. 1, pp. 38–47, 2002.

[10] L. Carta, S. Pastorino, G. Melillo, M. C. Bosco, S. Massazza, and L. Varesio, "Engineering of macrophages to produce IFN-γ in response to hypoxia," *Journal of Immunology*, vol. 166, no. 9, pp. 5374–5380, 2001.

[11] K. L. Talks, H. Turley, K. C. Gatter, et al., "The expression and distribution of the hypoxia-inducible factors HIF-1α and HIF-2α in normal human tissues, cancers, and tumor-associated macrophages," *American Journal of Pathology*, vol. 157, no. 2, pp. 411–421, 2000.

[12] K. K. Matthay, J. G. Villablanca, R. C. Seeger, et al., "Treatment of high-risk neuroblastoma with intensive chemotherapy, radiotherapy, autologous bone marrow transplantation, and 13-cis-retinoic acid," *The New England Journal of Medicine*, vol. 341, no. 16, pp. 1165–1173, 1999.

[13] A. Jögi, I. ∅ra, H. Nilsson, et al., "Hypoxia alters gene expression in human neuroblastoma cells toward an immature and neural crest-like phenotype," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 10, pp. 7021–7026, 2002.

[14] L. Holmquist-Mengelbier, E. Fredlund, T. Löfstedt, et al., "Recruitment of HIF-1α and HIF-2α to common target genes is differentially regulated in neuroblastoma: HIF-2α promotes an aggressive phenotype," *Cancer Cell*, vol. 10, no. 5, pp. 413–423, 2006.

[15] G. Melillo, T. Musso, A. Sica, L. S. Taylor, G. W. Cox, and L. Varesio, "A hypoxia-responsive element mediates a novel pathway of activation of the inducible nitric oxide synthase promoter," *Journal of Experimental Medicine*, vol. 182, no. 6, pp. 1683–1693, 1995.

[16] G. Melillo, E. A. Sausville, K. Cloud, T. Lahusen, L. Varesio, and A. M. Senderowicz, "Flavopiridol, a protein kinase inhibitor, down-regulates hypoxic induction of vascular endothelial growth factor expression in human monocytes," *Cancer Research*, vol. 59, no. 21, pp. 5433–5437, 1999.

[17] E. Fredlund, M. Ovenberger, Å. Borg, and S. Påhlman, "Transcriptional adaptation of neuroblastoma cells to hypoxia," *Biochemical and Biophysical Research Communications*, vol. 366, no. 4, pp. 1054–1060, 2008.

[18] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society. Series B*, vol. 67, no. 2, pp. 301–320, 2005.

[19] C. De Mol, E. De Vito, and L. Rosasco, "Elastic-net regularization in learning theory," *Journal of Complexity*, vol. 25, no. 2, pp. 201–230, 2009.

[20] M. C. Bosco, M. Puppo, C. Santangelo, et al., "Hypoxia modifies the transcriptome of primary human monocytes: modulation of novel immune-related genes and identification of CC-chemokine ligand 20 as a new hypoxia-inducible gene," *Journal of Immunology*, vol. 177, no. 3, pp. 1941–1955, 2006.

[21] A. Ricciardi, A. R. Elia, P. Cappello, et al., "Transcriptome of hypoxic immature dendritic cells: modulation of chemokine/receptor expression," *Molecular Cancer Research*, vol. 6, no. 2, pp. 175–185, 2008.

[22] J.-T. Chi, Z. Wang, D. S. A. Nuyten, et al., "Gene expression programs in response to hypoxia: cell type specificity and prognostic significance in human cancers," *PLoS Medicine*, vol. 3, no. 3, pp. 395–409, 2006.

[23] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information," *IEEE Transactions on Information Theory*, vol. 52, no. 2, pp. 489–509, 2006.

[24] D. L. Donoho, "For most large underdetermined systems of equations, the minimal l_1-norm near-solution approximates the sparsest near-solution," *Communications on Pure and Applied Mathematics*, vol. 59, no. 7, pp. 907–934, 2006.

[25] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, New York, NY, USA, 2001.

[26] M. Ashburner, C. A. Ball, J. A. Blake, et al., "Gene ontology: tool for the unification of biology," *Nature Genetics*, vol. 25, no. 1, pp. 25–29, 2000.

[27] R. A. Irizarry, B. M. Bolstad, F. Collin, L. M. Cope, B. Hobbs, and T. P. Speed, "Summaries of affymetrix GeneChip probe level data," *Nucleic Acids Research*, vol. 31, no. 4, article e15, 2003.

[28] R Development Core Team, *R: A Language and Environment for Statistical*, R Foundation for Statistical Computing, 2004.

[29] S. Patiar and A. L. Harris, "Role of hypoxia-inducible factor-1α as a cancer therapy target," *Endocrine-Related Cancer*, vol. 13, no. 1, pp. S61–S75, 2006.

[30] K. Boon, H. N. Caron, R. Van Asperen, et al., "N-myc enhances the expression of a large set of genes functioning in ribosome biogenesis and protein synthesis," *EMBO Journal*, vol. 20, no. 6, pp. 1383–1393, 2001.

[31] P. Warnat, A. Oberthuer, M. Fischer, F. Westermann, R. Eils, and B. Brors, "Cross-study analysis of gene expression data for intermediate neuroblastoma identifies two biological subtypes," *BMC Cancer*, vol. 7, p. 89, 2007.

[32] E. Bell, J. Lunec, and D. A. Tweddle, "Cell cycle regulation targets of MYCN identified by gene expression microarrays," *Cell Cycle*, vol. 6, no. 10, pp. 1249–1256, 2007.

[33] S. Asgharzadeh, R. Pique-Regi, R. Sposto, et al., "Prognostic significance of gene expression profiles of metastatic neuroblastomas lacking MYCN gene amplification," *Journal of the National Cancer Institute*, vol. 98, no. 17, pp. 1193–1203, 2006.

[34] M. Fischer, A. Oberthuer, B. Brors, et al., "Differential expression of neuronal genes defines subtypes of disseminated neuroblastoma with favorable and unfavorable outcome," *Clinical Cancer Research*, vol. 12, no. 17, pp. 5118–5128, 2006.

[35] A. Destrero, S. Mosci, C. De Mol, A. Verri, and F. Odone, "Feature selection for high-dimensional data," *Computational Management Science*, vol. 6, no. 1, pp. 25–40, 2009.

[36] C. De Mol, S. Mosci, M. Traskine, and A. Verri, "A regularized method for selecting nested groups of relevant genes from microarray data," *Journal of Computational Biology*, vol. 16, no. 5, pp. 677–690, 2009.

[37] A. Barla, S. Mosci, L. Rosasco, and A. Verri, "A method for robust variable selection with significance assessment," in *Proceedings of the European Symposium on Artificial Neural Networks (ESANN '08)*, Bruges, Belgium, April 2008.

[38] G. Jurman, S. Merler, A. Barla, S. Paoli, A. Galea, and C. Furlanello, "Algebraic stability indicators for ranked lists in molecular profiling," *Bioinformatics*, vol. 24, no. 2, pp. 258–264, 2008.

[39] S. Mosci, A. Verri, A. Barla, and L. Rosasco, "Finding structured gene signatures," in *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW '08)*, pp. 158–165, Philadelphia, Pa, USA, November 2008.

[40] T. J. P. Hubbard, B. L. Aken, S. Ayling, et al., "Ensembl 2009," *Nucleic Acids Research*, vol. 37, supplement 1, pp. D690–D697, 2009.

[41] R. H. Wenger, D. P. Stiehl, and G. Camenisch, "Integration of oxygen signaling at the consensus HRE," *Science's STKE*, vol. 2005, no. 306, article re12, 2005.

[42] C. Ambroise and G. J. McLachlan, "Selection bias in gene extraction on the basis of microarray gene-expression data," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 10, pp. 6562–6566, 2002.

[43] S. M. Mense, A. Sengupta, M. Zhou, et al., "Gene expression profiling reveals the profound upregulation of hypoxia-responsive genes in primary human astrocytes," *Physiological Genomics*, vol. 25, no. 3, pp. 435–449, 2006.

[44] H. Kim, D.-K. Lee, J.-W. Choi, J.-S. Kim, S. C. Park, and H.-D. Youn, "Analysis of the effect of aging on the response to hypoxia by cDNA microarray," *Mechanisms of Ageing and Development*, vol. 124, no. 8-9, pp. 941–949, 2003.

[45] Y. Jiang, W. Zhang, K. Kondo, et al., "Gene expression profiling in a renal cell carcinoma cell line: dissecting VHL and hypoxia-dependent pathways," *Molecular Cancer Research*, vol. 1, no. 6, pp. 453–462, 2003.

[46] D. J. Manalo, A. Rowan, T. Lavoie, et al., "Transcriptional regulation of vascular endothelial cell responses to hypoxia by HIF-1," *Blood*, vol. 105, no. 2, pp. 659–669, 2005.

[47] A. Jögi, J. Vallon-Christersson, L. Holmquist, H. Axelson, Å. Borg, and S. Påhlman, "Human neuroblastoma cells exposed to hypoxia: induction of genes associated with growth, survival, and aggressive behavior," *Experimental Cell Research*, vol. 295, no. 2, pp. 469–487, 2004.