Published in final edited form as:

Int J Comput Biol Drug Des. 2020 ; 13(1): 124–143. doi:10.1504/ijcbdd.2020.10026794.

Modelling of hypoxia gene expression for three different cancer cell lines

Babak Soltanalizadeh¹, Erika Gonzalez Rodriguez², Vahed Maroufy¹, W Jim Zheng³, Hulin Wu^{1,*}

¹Department of Biostatistics & Data Science, University of Texas Health Science Center at Houston, Houston, TX, USA

²Center for translational Injury Research, Department of Surgery, McGovern Medical School, UT Houston, Houston, TX, USA

³School of Biomedical Informatics, University of Texas Health Science Center at Houston, Houston, TX, USA

Abstract

Gene dynamic analysis is essential in identifying target genes involved pathogenesis of various diseases, including cancer. Cancer prognosis is often influenced by hypoxia. We apply a multi-step pipeline to study dynamic gene expressions in response to hypoxia in three cancer cell lines: prostate (DU145), colon (HT29), and breast (MCF7) cancers. We identified 26 distinct temporal expression patterns for prostate cell line, and 29 patterns for colon and breast cell lines. The module-based dynamic networks have been developed for all three cell lines. Our analyses improve the existing results in multiple ways. It exploits the time-dependence nature of gene expression values in identifying the dynamically significant genes; hence, more key significant genes and transcription factors have been identified. Our gene network returns significant information regarding biologically important modules of genes. Furthermore, the network has potential in learning the regulatory path between transcription factors and the downstream genes. In addition, our findings suggest that changes in genes BMP6 and ARSJ expression might have a key role in the time-dependent response to hypoxia in breast cancer.

Keywords

Gene expression; Hypoxia; colon cancer; breast cancer; significant genes

1. Introduction

Hypoxia occurs when cells are deprived of adequate oxygen supply for a certain period. In tumors, hypoxia is prompted by rapid growth and expansion of a mass without appropriate angiogenesis that can supply the new oxygen demands (Vaupel 2004). Hypoxia is a common feature of human cancers and presents a highly dynamic spatial and temporal distribution in cancer patients and affects normal physiological processes such as embryonic development,

*Corresponding author: hulin.wu@uth.tmc.edu.

metabolism, and stem cell regulation, as well as pathophysiological conditions such as infarction, stroke, and cancer (Semenza 2000; Harris 2002). The existence of hypoxia is also related to the poor prognosis in different cancer types, such as head and neck, esophageal, gastric, liver, lung, breast (Schindl et al., 2002; Trastour et al., 2007), bladder and prostate cancers.

Falling oxygen tensions place a selective pressure on cells to either adapt or die and initiate rapid transcriptional and post-transcriptional responses (Giatromanolaki et al., 2011). Hypoxia is therefore linked to several of the key hallmarks of cancer, including sustained angiogenesis, metastasis and invasion, evasion of apoptosis, suppression of the immune response and genomic instability. Hypoxic cells are about three-fold more resistant to the effects of sparsely ionizing radiation. Furthermore, hypoxia can also affect the success of conventional chemotherapy. Therefore, studying hypoxia has been the focus of many significant studies, especially in the radiotherapy field (Gray et al., 1953). The negative impact of hypoxia on efficacy of conventional chemotherapy begs for a good approach to assess the hypoxia for finding the tumors that are related to the therapeutic strategies targeting hypoxia.

There are several methods for measuring the level of oxygen or the tumor's molecular response to hypoxia. They can be divided into two main group: first, by identifying genes that show significant expression in cancer cells cultured under hypoxic versus normoxic conditions. Second, by applying the previously characterized hypoxia-regulated genes/ biomarkers defining hypoxic tumors and identifying other genes that are over- or under-expressed in hypoxic tumors.

Determining a global transcriptional response to hypoxia by using gene expression microarrays has been considered in many distinguished papers. This method groups genes that show a significantly unregulated or exceed a defined threshold from comparing to normoxic case which can be referred to as a "hypoxia gene expression signature" or "hypoxia signature". Microarray expression has been used previously to study the effects of hypoxia on the vasculature (Peters et al., 2006; Ning et al., 2004), the immune system (Bosco et al., 2006), and on tumor cells (Jogi et al., 2004; Aprelikova et al., 2006).

There is significant overlap in the genes that have been found by different methods, however, to the best of our knowledge, there are not any genes identified in all signatures. Nevertheless, it is important to notice that the two different approaches may not result in the same output, even if the starting point is the same published signature. Thus, we feel there is a significant demand for a more efficient method to identify significant genes which can incorporate all the temporal dependence, since identification of genes or common pathways in hypoxia signatures can improve the current therapeutic procedures or provide new therapeutic targets. Hypoxia signatures play an important roles in finding new hypoxia-induced genes and related cellular pathways (Blick et al., 2013; Masiero et al., 2013). In (Buffa et al., 2010) a continuous score has been used to rank the clinical samples within a cohort from the least to the most hypoxic. Toustrup et.al trained a weighted classifier using an external measure of hypoxia (Toustrup et al., 2011). Authors of (Chi et al., 2006; Fardin et al., 2010) presented the log2 gene expression of the signature and K-means clustering

score to classify patients to the “more hypoxic” and “less hypoxic” groups, also an ensemble classification technique which gives the best stability and most reproducible size of effect has been presented by (Fox et al., 2014).

The hypoxia-related gene sets, from both in vitro and in vivo data, that demonstrate prognostic power have been discussed by several papers (Chi et al., 2006; Fardin et al., 2010; Fox et al., 2014). Furthermore, (Liang et al., 2008) has derived the in vitro hypoxia signatures tested for prognostic value from a single data set when a series of cell lines were exposed to two different hypoxia cell lines in several time points and has concluded that a general hypoxia signature derived from this series had prognostic value. It has also been shown in (Seigneuric et al., 2007) that, the induced genes at early time-points had better prognostic value. In (Elvidge et al., 2006; Sorensen et al., 2010), the hypoxia-regulated gene sets from in vitro microarray data have been presented, but these gene sets were not evaluated in patient data to measure their prognostic value. (Hu et al., 2009; Winter et al., 2007; Buffa et al., 2010) provided the in vivo derived hypoxia signatures from a single gene and a set of known hypoxia up-regulated genes. Starmans et al (Starmans et al., 2012) have investigated the prognostic value of in vitro and in vivo derived hypoxia gene-expression data in a more comprehensive manner and conclude that even with the inclusion of the new signatures where none of the in vitro derived signatures consisting of hypoxia-induced genes are prognostic, but signatures derived from in vitro hypoxia repressed genes exhibit prognostic power. Also, (Harris Starmans et al 2015) provided a summary of several published hypoxia signatures and discussed their commonalities, differences and highlighted their strengths and limitations. (Koritzinsky Starmans et al 2013) concluded that since hypoxia causes ER stress, suggesting a role for oxygen in protein folding, then the existence of two phases of disulfide bond formation in living mammalian cells, with differential requirements for oxygen has been demonstrated, i.e., disulfide bonds introduced rapidly during protein synthesis can occur without oxygen, whereas those introduced during post-translational folding or isomerization are oxygen dependent.

In this paper, we focus on the hypoxia transcriptomic time-resolved data in three different cancer cell lines in 8 time points versus the normoxia situation. The methods in the literature often use statistical models with strong underlying assumptions such as linearity and independence which do not hold here and are not capable of exploiting the time dependence among expression values and dependence among gene's temporal behavior. To account for that, we use a high-dimensional ordinary differential equation (ODE) model to construct the genome-wide dynamics of the gene regulatory networks (GRN) for three cancer cell lines. The model has also been implemented in many studies leading to interesting findings. (Lu Starmans et al 2011) used a set of ordinary differential equations (ODE), coupled with dimension reduction by clustering and mixed-effects modeling techniques, to model the dynamic gene regulatory network (GRN). They also proposed a several-steps procedure including clustering, smoothing, regulation identification, parameter estimates refining and function enrichment analysis to identify the ODE-based dynamic GRN. Wu et.al. (Wu et al., 2014), applied this method to reconstruct high-dimensional dynamic GRNs from genome-wide time course gene expression data in mouse lungs with influenza infection. In (Sun Starmans et al 2016), the controllability and stability analyses are conducted for the reconstructed gene response networks of 17 influenza infected subjects based on ODE

models. Liang and Wu proposed the parameter estimation methods for ODE models based on the local smoothing approach and a pseudo-least squares principle under a framework of measurement error in regression models (Liang Starmans et al 2008).

The method used in this paper implements recent tools developed in statistics and data science and improves the earlier methods in multiple ways. First, it exploits the dynamic behavior of gene expression as a sample of a curve to identify significant genes, instead of considering a few consecutive up/down-regulated values. Second, it clusters the dynamically significant genes, which is both computationally and geologically interpretable, since genes undertake significant actions in groups. Finally, the dynamic gene network, in addition to identify highly regulated and highly regulating genes, has a promising potential in revealing regulation pathways between transcription factors and downstream genes.

2. Analytical methods

In the experimental study, the transcriptome profiling was performed for 3 cell lines derived from diverse tumor-types after hypoxic exposure for a total of 8 time points. The details of the experiment and the corresponding data set can be found in Gene Expression Omnibus (GEO) database (accession number: GSE23961). We introduce the pipeline technique to study the behavior of genome-wide time course gene expression data and outputs of the dynamic GRN which has been discussed in (Carey et al., In press). The pipeline process includes the following steps: detection of temporally differentially expressed (DE) genes; clustering differential genes into co-expressed modules; identification of network structure; parameter estimate refinement; and functional enrichment analysis. We have applied several statistical techniques which include the functional principal component analysis (FPCA) with hypothesis testing, time course gene expression clustering, nonparametric mixed-effects modeling, and parameter estimation and statistical inference for ODE models. The first step is the pipeline analyzing. Therefore, we apply the pipeline method to analyze the time course data available in data series GSE41491 of the Gene Expression Omnibus (GEO) to be able to identify differentially expressed genes and the gene regulatory network of those genes. The analysis is focused on the 3 experimental Cell lines with 8 time points which will be mentioned by DU145, HT29 and MCF7. Our approach includes obtaining the data, preprocessing and analyzing for the identification of dynamic response genes which are basically the genes that exhibit significant changes across time, the clustering of these and the discovery of a gene regulatory network between these clusters.

Let us define $X_{i,j}(t)$ as the centered expression profile of the i^{th} gene, belonging to subject j , which is a smooth function over time, also the time course gene expression measurements are discrete observations from this smooth function, which have been distorted by noise, i.e., $Y_{i,j}(t_k) - \mu_{i,j} = X_{i,j}(t_k) + \epsilon_{i,j}(t_k)$, for $i = 1, \dots, n$, $j = 1, \dots, N$ and $k = 1, \dots, K_{i,j}$, where n is the number of genes, N is the number of subjects, and $K_{i,j}$ is the number of time points observed for the i^{th} gene.

By using the spline smoothing method, the pipeline obtains the functional entity $X_{i,j}(t)$. Spline smoothing involves minimizing the following penalized residual sum of squares

$$(Y_{i,j}(t_k) - \mu_{i,j} - X_{i,j}(t_k))^2 + \lambda \int [D^2 X_{i,j}(t) dt]^2. \quad (1)$$

Since the process of smoothing each individual gene is computationally expensive and is likely to obtain biased estimates as our data does not contain replicates and is expected to contain a low signal-to-noise ratio, then the pipeline pools the genes together to obtain an estimate of the amount of regularity required.

In fact, many of the genes have a relatively flat expression profile in the time estimating, so the smoothing parameter by pooling all the genes together will not be a good idea because it will tend to over smooth the data. Therefore, to solve this problem, the pipeline picks a subset of the genes that exhibit time course patterns that have relatively smooth trajectories and do not fluctuate widely. So, these genes will be ranked by their interquartile range, then the pipeline will select the top genes for the estimation subset.

The next step is a hypothesis testing to detect the dynamic response genes (DRG) which helps us to identify the genes with the most significant expression changes over the different time periods with the biological process of interest or a particular stimulus. Thus, we apply the following statistical test to identify the DRGs:

$$H_0: X_{i,j}(t) = 0, H_a: X_{i,j}(t) \neq 0. \quad (2)$$

The test statistic is the conventional F statistic and it is comparing the goodness-of-fit of the null model to the alternative model which can be presented by

$$F_{i,j} = \frac{(RSS_{i,j}^0 - RSS_{i,j}^1)}{RSS_{i,j}^1}, \quad (3)$$

where $RSS_{i,j}^0$ and $RSS_{i,j}^1$ are residual sum of squares under the null and the alternative models for the i -th gene, belonging to subject j , respectively. The genes with large F -ratios can be considered as exhibiting notable changes with respect to time. The F statistic is calculated for each gene as

$$F_i = \frac{(RSS_{i,j}^0 - RSS_{i,j}^1)/(df - 1)}{(RSS_{i,j}^1)/(n - df)}, \quad (4)$$

where df is the degrees of freedom obtained from the smoothing of the gene expression data by penalized least squares. Our criteria is selecting a gene as dynamic response gene if its F statistic F_i is such that $F_i > t_{\alpha, df-1}$. where $d_1 = df - 1$ and $d_2 = n - df$. We consider the top 3,000 of these genes as DRGs because it is more convenient to have a consistent number of DRGs across all subjects in the comparing process of several subjects associated to a single study.

Next, we cluster the genes into groups of genes which have similar gene expression profile over time which will be called temporal gene response modules since a large number of the

DRGs have similar expression patterns over time. To this end, we will use the Iterative Hierarchical Clustering (IHC) method which can identify in-homogeneous clusters, capture both the large and very small clusters and provides an automated selection of the optimal number of clusters and the process can be given by the following steps: The first section is initialization which includes clustering the gene expression curves by applying the hierarchical clustering. The second section is merging that is treating each of the cluster centers (exemplars) as new genes and then using the same rule which we have used in the first step to merge the exemplars into new clusters. The next step is pruning which means if the correlation between the cluster center and gene is less than the Spearman correlation, then remove the corresponding gene. The next section is repeat sections 2 and 3 until either the index of clusters fully or partially converges, or the number of iterations reaches the maximum number of iterations. The last section is repeating section 2 until the between-cluster correlations are less than the Spearman correlation.

The other step is discovering the gene regulation network (GRN). Differential equations have many applications in various fields of science (Sarmast et al., 2014; Soltanalizadeh et al., 2018, Sarmast et al., 2018). Ordinary differential equation (ODE) modelling has been applied widely to construct high-dimensional GRN and the main goal of GRN is considering the impact of different gens on the expression of other ones and can be formulated as the following rate equations:

$$\frac{dM_{q,j}}{dt} = \alpha_{0,q,j} + \sum_{p=1}^P \alpha_{p,q,j} M_{p,j}, \quad (5)$$

where $\alpha_{0,q,j}$ is the intercept of the q^{th} gene response module, belonging to subject j and the coefficients $\alpha_{p,q,j}$ quantify the regulation effects of the p^{th} gene response module on the instantaneous rate of change in q^{th} gene response module. Since only a few gene response modules effect the instantaneous rate of change in q^{th} gene response module, thus only a few of the $[\alpha_{p,q,j}]_{p=1}^Q$ will be non-zero. First, a model selection which determines which $[\alpha_{p,q,j}]_{p=1}^Q$ are non-zero can be applied and then their coefficients can be approximated to determine the regulation effects. Although the current model size is comparably smaller than the original model, simultaneous model selection and parameter optimization of ODE parameters are still computationally very expensive. So, a smoothing-based estimation method can be applied which converts the system of differential equations into a set of pseudo-regression models which avoids numerically solving the differential equations directly and does not require the initial or boundary conditions of the state variables and also allows us to use model selection and parameter estimation for one equation at a time, which significantly reduces the computational cost. In order to reduce the computational cost of the ODE models and also facilitate the variable selection procedure, we apply the two-stage smoothing-based estimation method. By the way, this method does not give us accurate estimations of $\widehat{M_{q,j}}$ and $\widehat{\frac{dM_{q,j}}{dt}}$. Also converting the system of differential equations into a set of pseudo-regression models only accounts for the direct effect which is how the q^{th} gene response module is regulated by the p^{th} gene response module $q \rightarrow p$ and does not account for the indirect effect which is how the r^{th} gene response module is regulated by the

q^{th} gene response module which in turn regulates the p^{th} gene response module ($r \rightarrow q \rightarrow p$) which is a fundamental property of any system of differential equations. Thus, to solve this issue, we apply the nonlinear least squares (NLS) method which refines the parameter estimates for the selected ODE model. Therefore, we use the parameter estimates from the two-stage method for estimating the structure of the estimates only which helps us to figure out which parameters are non-zero.

3. Results

The number of DRGs, and number GRMs for all three cell lines are given in table 1. The number of DRGs under HT29 and MCF7 increased by 21% and 30%, respectively, compared to that under DU145. Also, the number of GRMs under DU145 is slightly smaller than that under the other two cell lines, which may be the influence of the smaller number of the DRGs. Frequency of DRGs across all analyzed cell lines are listed in Table1. Authors of (Carey et al., 2016) discussed the importance of large size GRM over small size GRMs and show that the large size modules are related to performing various fundamental biological functions and have a consistent pattern across individuals under the same experimental cell line representing the state-transition patterns in response to the stimuli. On the other hand, small sized and single gene modules tend to perform homeostasis and have an oscillatory temporal response. Hence, we divided the GRMs into four groups: single-gene modules (SGM), small-size modules (SSM) contain between 2–10 genes in each cluster, medium-size modules (MSM) include 11–99 genes in each of the clusters, and finally the large-size modules (LSM) contain over 100 genes in each cluster. Also, corresponding mean curves for each group are presented for all three cell lines in Fig.3. The gene regularity networks for all three cell lines have been provided for the cell lines. A high network density indicates strong network connectivity.

The plots of expression of all genes, smooth expression of all genes, dynamic response genes and top rank DRGs for cell line MCF7 have been given by Figure 1 sections a, b, c and d, respectively. The clustered of the DRGs into temporal GRMs has been presented in 29 modules by Figure 2 and the GRMs' mean expression curves grouped into four categories by cluster size have been shown by Figure 3. Also, Figure 4 shows the gene regulatory network. The plots of expression of all genes, smooth expression of all genes, dynamic response genes, top rank DRGs and the gene regularity network for the cell lines DU145 and HT29 have been provided in Supplementary data. The clustered of the DRGs into temporal GRMs has been presented in 26 modules for DU145 by Figure 5 and the GRMs' mean expression curves grouped into four categories by cluster size has been shown by Figure 6. Also, the clustered of the DRGs into temporal GRMs has been presented in 29 modules for HT29 by Figure 7 and the GRMs' mean expression curves grouped into four categories by cluster size has been shown by Figure 8.

Compared to the (Starmans et al., 2012) which identifies 1436 genes with significant behavior under at least two conditions, we have identified 6802 genes (Fig 9). We identified more than 96% (1385 genes) of the genes found by the original method. However, pipeline found 5417 additional significant genes, which means that the original method only covered 22% of pipeline's DRGs (Fig 10). Furthermore, we identified 1424 genes which are significant

under all three conditions, but not identified by the original (Starmans et al., 2012). These findings demonstrate the efficiency of our pipeline as it inspects the global behavior of genes instead of only one pick as a result of at least two consecutive up/down-regulation. The complete lists of significant genes in each cell line has been presented in supplementary Tables. Figure 11 shows the temporal behavior of a selected sample of genes which we identified as significant but were missed by Starmans. Evidently, there is a clear up/down and significant trend in each of these genes, which their exclusion from any results may cause a significant bias in the final results. Among these genes, 6225 of them are associated with a GO term and gives us the following results. The GO term of GO:0042975 with the function of peroxisome proliferator activated receptor binding, the GO term of GO:0000977 with the function of RNA polymerase II regulatory region sequence-specific DNA binding and the GO term of GO:0001012 with the function of RNA polymerase II regulatory region DNA binding. Next, we apply the GA-analysis for the 5417 genes which our pipeline found them as significant genes but the original method could not find. After applying the GO analysis, we saw that 4939 of these genes were associated with a GO term. Also, among these 5417 genes, we found a set of transcription factor genes by comparing them to the complete list of transcription genes. Finally, we have found 494 transcription factor genes that the authors of (Starmans et al., 2012) did not find them among their significant genes related to the hypoxia. Actors high on betweenness have the potential to influence others near in a network through both direct and indirect pathways. A node with high betweenness centrality can potentially influence the spread of information through the network, by facilitating, hindering, or even altering the communication between others. The most important modules for MCF7, DU145 and HT29 can be seen in Tables .2, 3 and 4. We can see that the modules 21 and 25 are among the five important modules in the all in-degree, out-degree and betweenness groups, modules 3 and 7 are among the five important modules in both in-degree and betweenness groups, modules 23, 26 and 29 are among the 5 important modules in the out-degree and betweenness groups, module 27 in in-degree and module 5 in betweenness.

4. Discussion

The main aim of this study was to elucidate the power and advantage of a sophisticated statistical and computational tool in gene dynamic analyses, compared to the well-known generic statistical methods. In (Starmans et al., 2012) significant difference in comparing hypoxia to normoxia is characterized by at least 2-fold change in two or more consecutive time-points. This criterion is much of a local expression detection in a temporal sense, it does not exploit the total change or trend throughout all time points; hence, it leaves out potentially significant genes and their corresponding biological achievements. In contrast, our identification of the dynamic response genes is based on the significance of the genes throughout the entire curve. The extraction of dynamic gene lists in (Elvidge et al., 2006) is based on finding the genes that show differential expression across the time series and then minimizing the log2-transformed fold-change in at least two consecutive time-points relative to normoxia and this is the reason that their method was unable to recognize many genes that are actually significant.

With the method applied here, we were able to identify a set of transcription factor genes that was not previously reported in (Starmans et al., 2012). This is important because transcription factors are one of the groups of proteins that read and interpret the genetic “blueprint” in the DNA. They bind to the DNA and help initiate a program of increased or decreased gene transcription. Therefore, they are essential for many essential cellular processes and an important subject of research. We found a set of transcription factor genes that Starmans and colleagues (Starmans et al., 2012) did not find among their significant genes related to the hypoxia. Some of the transcription factors identified by the pipeline are involved in metabolic processes, cell signaling, immune response, and protein metabolism [35]. The difference in the number of genes identified by our method in comparison to previous studies could be attributed to changes in gene expression. Since the pipeline considers more than two time points, it is more sensitive to changes in gene expression over time, which is a characteristic of the gene-expression response to hypoxia.

By considering the network plot between the modules, we find the most important modules contributing to the network. The modules that have more contribution will induce the genes in other modules more. For a directed graph, the out-degree is the number of arcs directed away from the node and the in-degree is the number of arcs directed towards the node. The betweenness centrality quantifies the number of times a node acts as a bridge along the shortest path between two other nodes. Finding and reporting the most important modules based on the node statistics technique also is another achievement of this method which can give us information regarding to the important genes which regulates others. We were able to identify two modules (21 and 25) that are among the top five important modules when considering the in-degree, out-degree and betweenness together. Only three genes were identified in these two modules (two genes in module 21 and one in module 25). One of the genes in module 21 is BMP6 (bone morphogenetic protein 6), a breast cancer tumor suppressor. This gene is an important regulator of cell growth and apoptosis in cancer, and its reduced expression is associated to a drug resistance phenotype in breast cancer cells (Liu et al., 2014; Wiegmann et al., 2013). Our findings suggest that changes in BMP6 expression might also have a key role in the time-dependent response to hypoxia in breast cancer.

Gene ARSJ was also identified by the pipeline as a gene of high importance. This gene encodes for an enzyme that belongs to the family of sulfatases, which are enzymes that hydrolyze sulfate esters in glycosaminoglycan and have important roles in hormone regulation, cell signaling and cellular degradation. Mutations of some of these enzymes have been linked to the development of rare neurodegenerative conditions such as lysosomal storage diseases (Wiegmann et al., 2013). We found that the expression of ARSJ changes significantly over time suggesting a new role for this gene in the response to hypoxia in breast cancer cells. However, more studies are needed to determine how its expression is regulated and its mechanisms in the response to hypoxia in breast cancer cells.

Module 27 is ranked at the top of the highest in-degree list of modules. This module contained one gene, RBM8A, which codes for RNA-binding motif protein 8A. This protein helps to transport mRNA, with important roles in protein translation. Previous research demonstrated that this RNA-binding motif protein interacts with OVAC, a tumor suppressor gene associated with breast and ovarian carcinoma [39]. In studies in hepatocellular

carcinoma, Liang R et al found that the high expression of RBM8A is a factor associated with poor prognosis and tumor progression. RBM8A promoted tumor progression through the activation of the epithelial-mesenchymal transition signaling pathway and is necessary for tumor cell proliferation and anti-apoptosis. Our findings suggest that this gene could play a role in the response to hypoxia in breast cancer cells. We can speculate that a change in the expression of this gene could also result in poor prognosis in the context of breast cancer in part because hypoxic conditions result in resistance to chemotherapy.

Interestingly, one of the most important modules for the HT29 (colorectal cancer) contained only one gene, FOCAD (KIAA1797). Previous research identified that loss of function mutations of this gene could predispose to colorectal adenoma [40]. Our findings provide evidence of a potential role of this gene in the hypoxia signature in colorectal cancer.

The non-receptor tyrosine- protein kinase Src, a proto-oncogene, promotes survival, proliferation and invasion. Its dysregulation has been linked to the development of certain types of cancer including breast and colon cancer. MPP2 is a member of the member of the Membrane-Associated Guanylate Kinase family, that interacts with Src and suppresses its expression [41]. MPP2 was the only gene in module 20 (line DU145), which ranked among the highest in all three categories (in-degree, out-degree and betweenness). The altered expression of MPP2 could lead to a dysregulated overactivity of Src, a mechanism that could lead to malignant cell transformation. Furthermore, according to our results, MPP2 seems to play a key role in the hypoxia signature in prostate cancer. However, more studies are needed to validate these findings.

Finally, we should mention that all the efficacy and power of the new methodology is not cost free. First, the statistical testing needs the data with more than 7 time points to efficiently distinguish DRGs from the rest of the gens. Moreover, the power of any statistical hypothesis test depends on the sample size, and a very big sample size returns any small discrepancy as statistically significant. In addition, the number of GRMs under different cell lines or for different cell lines is confounded with the number of DRGs under each cell line. Therefore, to account for that we compare biological results based on the fixed number 3000 top ranked genes as the significant ones. We should also clarify that, existing noise in the recorded time resolved expressions is not an issue here since in the process of selecting DRGs, the noisy genes are usually filtered.

5. Conclusion

In this study, we presented a novel algorithm to find the significant genes by considering the whole curve of the corresponding time course data. We have considered a different version of clustering these significant genes into several modules based on the similarity of the plots of their expression. By studying the gene expression profiles of hypoxia in three cancer cell lines, we have demonstrated and validated that its performance is superior to previously reported methods and allows the identification of a significant number of genes that had not been identified previously.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work is partially supported by Cancer Prevention and Research Institute of Texas (CPRIT) P170668 and NIH/ NIAID RO1 AI087135. Authors thank Mr. Nan Deng and other GEO-Big Data Project members for their valuable comment.

References

- [1]. Vaupel P (2004) 'The Role of Hypoxia-Induced Factors in Tumor Progression', *The oncologist*, Vol. 9, No. 5, pp.10–17.
- [2]. Semenza GL. (2000) 'HIF-1: using two hands to flip the angiogenic switch', *Cancer Metastasis Rev*, Vol. 19 No.1e2, pp.59–65. [PubMed: 11191064]
- [3]. Harris AL (2002) 'Hypoxia e a key regulatory factor in tumour growth', *Nat Rev Cancer*, Vol. 2, No. 1, pp.38–47. [PubMed: 11902584]
- [4]. Schindl M, Schoppmann SF., Samonigg H., et al. (2002) 'Overexpression of hypoxiainducible factor 1alpha is associated with an unfavorable prognosis in lymph node-positive breast cancer', *Clin Cancer Res*, Vol. 8, No. 6, pp.1831–1837. [PubMed: 12060624]
- [5]. Trastour C, Benizri E, Ettore F, et al. (2007) 'HIF-1alpha and CA IX staining in invasive breast carcinomas: prognosis and treatment outcome', *Int J Cancer*, Vol. 120, No.7, pp.1451–1458. [PubMed: 17245699]
- [6]. Giatromanolaki A, Harris AL., (2011) 'Tumour hypoxia, hypoxia signaling pathways and hypoxia inducible factor expression in human cancer', *Anticancer Res*, vol. 21 No.6B, pp.4317–4324.
- [7]. Gray LH., Conger AD., Ebert M., et al. (1953) 'The concentration of oxygen dissolved in tissues at the time of irradiation as a factor in radiotherapy', *Br J Radiol* Vol. 26, No.312, pp. 638–648. [PubMed: 13106296]
- [8]. Peters DG., Ning W., Chu TJ., et al. (2006) 'Comparative SAGE analysis of the response to hypoxia in human pulmonary and aortic endothelial cells', *Physiol Genomics*, Vol. 26, No. 2, pp.99–108. [PubMed: 16595741]
- [9]. Ning W, Chu TJ., Li CJ., et al. (2004) 'Genome-wide analysis of the endothelial transcriptome under short-term chronic hypoxia', *Physiol Genomics*, Vol. 18, No.1, pp.70–78. [PubMed: 15100389]
- [10]. Bosco MC., Puppo M., Santangelo C., et al. (2006) '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' *J Immunol*, Vol.177 No.3, pp.1941–1955. [PubMed: 16849508]
- [11]. Jogi A, Vallon-Christersson J, Holmquist L, et al. (2004) 'Human neuroblastoma cells exposed to hypoxia: induction of genes associated with growth, survival, and aggressive behavior'. *Exp Cell Res*, Vol. 295, No. 2, pp.469–487. [PubMed: 15093745]
- [12]. Aprelikova O, Wood M, Tackett S, et al. (2006) 'Role of ETS transcription factors in the hypoxia-inducible factor-2 target gene selection', *Cancer Res*, Vol. 66, No. 11, pp.5641–5647. [PubMed: 16740701]
- [13]. Blick C, Ramachandran A, Wigfield S, et al. (2013) 'Hypoxia regulates FGFR3 expression via HIF-1alpha and miR-100 and contributes to cell survival in non-muscle invasive bladder cancer', *Br J Cancer*, Vol. 109, No.1, pp.50–59. [PubMed: 23778527]
- [14]. Masiero M, Simoes FC., Han HD., et al. (2013) 'A core human primary tumor angiogenesis signature identifies the endothelial orphan receptor ELTD1 as a key regulator of angiogenesis', *Cancer Cell*, Vol. 24, No. 2, pp.229–241. [PubMed: 23871637]
- [15]. Buffa FM., Harris AL., West CM., et al. (2010) 'Large meta-analysis of multiple cancers reveals a common, compact and highly prognostic hypoxia metagene', *Br J Cancer* Vol. 102, No. 2, pp.428–435. [PubMed: 20087356]

- [16]. Toustrup K, Sorensen BS., Nordsmark M., et al. (2011) 'Development of a hypoxia gene expression classifier with predictive impact for hypoxic modification of radiotherapy in head and neck cancer', *Cancer Res*, Vol. 71, No. 17, pp.5923–5931. [PubMed: 21846821]
- [17]. Chi JT., Wang Z., Nuyten DS., et al. (2006) 'Gene expression programs in response to hypoxia: cell type specificity and prognostic significance in human cancers', *PLoS Med* Vol. 3, No.3, pp.e47. [PubMed: 16417408]
- [18]. Fardin P, Barla A, Mosci S, et al. (2010) 'A biology-driven approach identifies the hypoxia gene signature as a predictor of the outcome of neuroblastoma patients', *Mol Cancer*, Vol. 9, No. 1, pp. 185–200 [PubMed: 20624283]
- [19]. Fox NS., Starmans MH., Haider S., et al. (2014) 'Ensemble analyses improve signatures of tumour hypoxia and reveal interplatform differences', *BMC Bioinformatics*, Vol. 15, pp170. [PubMed: 24902696]
- [20]. Seigneure R, Starmans MH., Fung G., et al. (2007) 'Impact of supervised gene signatures of early hypoxia on patient survival', *Radiother Oncol*, Vol. 83, No. 3, pp.374–82. [PubMed: 17532074]
- [21]. Hu Z, Fan C, Livasy C, et al. (2009) 'A compact VEGF signature associated with distant metastases and poor outcomes', *BMC Med*, Vol. 7, No.9.
- [22]. Winter SC., Buffa FM., Silva P., et al. (2007) 'Relation of a hypoxia metagene derived from head and neck cancer to prognosis of multiple cancers', *Cancer Res*, Vol. 67, pp.3441–3449. [PubMed: 17409455]
- [23]. Liang H, Wu H, (2008) 'Parameter Estimation for Differential Equation Models Using a Framework of Measurement Error in Regression Models', *Journal of the American Statistical Association*, Vol. 103, No. 484, pp.1570–1583 [PubMed: 19956350]
- [24]. Elvidge GP., Glenny L., Appelhoff RJ., Ratcliffe PJ., Ragoussis J., Gleadle JM., (2006) 'Concordant regulation of gene expression by hypoxia and 2-oxoglutarate-dependent dioxygenase inhibition: the role of HIF-1 α , HIF-2 α , and other pathways', *J Biol Chem* Vol. 281, pp.15215–15226.
- [25]. Sorensen BS., Toustrup K., Horsman MR., Overgaard J., Alsner J., (2010) 'Identifying pH independent hypoxia induced genes in human squamous cell carcinomas in vitro', *Acta Oncol* Vol. 49, pp.895–905. [PubMed: 20429727]
- [26]. Starmans MH, Chu KC, Haider S et al. (2012) 'The prognostic value of temporal in vitro and in vivo derived hypoxia gene-expression signatures in breast cancer', *Radiother Oncol*, Vol. 102, No. 3, pp.436–443. [PubMed: 22356756]
- [27]. Harris BHL., Barberis A., West CML., Buffa FM., (2015) 'Gene expression signatures as biomarkers of tumour hypoxia', *Clin Oncol*, Vol. 27, No.10, pp.547–560.
- [28]. Koritzinsky M, Levitin F, van den Beucken T, (2013) 'Rumantir RA et al., Two phases of disulfide bond formation have differing requirements for oxygen', *J Cell Biol*, Vol. 203, No. 4, pp.615–27 [PubMed: 24247433]
- [29]. Lu T, Liang H, Li HZ, Wu H (2011) 'High-dimensional ODEs coupled with mixed-effects modeling techniques for dynamic gene regulatory network identification' *Journal of the American Statistical Association*, Vol. 106, No. 496, pp.1242–1258. [PubMed: 23204614]
- [30]. Wu S, Liu Z-P, Qiu X, & Wu H (2014). Modeling genome-wide dynamic regulatory network in mouse lungs with influenza infection using high dimensional ordinary differential equations. *PLOS ONE*, Vol.9, No.5, pp.952–976.
- [31]. Sun X, Fang H, Wu S, Qiu X, Linel P, Wu H, (2016) 'Controllability and stability analysis of large transcriptomic dynamic systems for host response to influenza infection in human', *Infectious Disease Modelling*, Vol. 1, No.1, pp52–70. [PubMed: 29928721]
- [32]. Sarmast Z, Soltanalizadeh B, Boubaker K, (2014) 'A new numerical method to study a second-order hyperbolic equation', *South Asian Journal of Mathematics*, Vol. 4, No. 6, pp.285–296.
- [33]. Soltanalizadeh B, Esmalifalak H, Hekmati R, Sarmast Z, Shabani S, (2018) 'Numerical analysis of the one-dimensional wave equation subject to a boundary integral specification', *Walailak J Sci & Tech*, Vol. 15, No.6, pp.421–437.
- [34]. Sarmast Z, Safi MR., Farzaneh E, (2018) 'Post optimality analysis in fuzzy multi objective linear programming' *Int. J. Math. Appl*, Vol. 6, No. 1, pp.925–37.

- [35]. Carey M, Wu S, Gan G, Wu H, (2016) 'Correlation-based iterative clustering methods for time course data: The identification of temporal gene response modules for influenza infection in humans', *Infectious Disease Modelling*, Vol. 1, No. 1, pp.28–39. [PubMed: 29928719]
- [36]. <https://reactome.org>
- [37]. Liu G, Liu YJ., Lian WJ., Zhao ZW., Yi T., Zhou HY., (2014) 'Reduced BMP6 expression by DNA methylation contributes to EMT and drug resistance in breast cancer cells', *Oncol Rep*, Vol. 32, No. 2, pp.581–588. [PubMed: 24890613]
- [38]. <https://www.ncbi.nlm.nih.gov/gene/654>
- [39]. Wiegmann EM, Westendorf E, Kalus I, Pringle TH, Lübke T, Dierks T, Arylsulfatase K, (2013) 'A Novel Lysosomal Sulfatase', *The journal of biological chemistry*, Vol. 288, No.42, pp.30019–30028. [PubMed: 23986440]
- [40]. <https://www.ncbi.nlm.nih.gov/pubmed/11013075>
- [41]. <https://www.ncbi.nlm.nih.gov/pubmed/25219767>
- [42]. <https://www.ncbi.nlm.nih.gov/pubmed/19665017>
- [43]. Carey M, Ramirez JC., Wu S., Wu H., (2018) 'A Big Data Pipeline: Identifying Dynamic Gene Regulatory Networks from Time Course GEO Data with Applications to Influenza Infection', *SAGE*, Vol. 27, No.7, pp.1930–1955.

Fig 1:
Expression of genes in MCF7

Fig 2:
Expression of the gene response modules for MCF7

Fig 3:
Mean curves for MCF7

Fig 4:
Gene regularity network in MCF7

Fig 5:
Expression of the gene response modules for DU145

Fig 6:
Mean curves for DU145

Fig 7:
Expression of the gene response modules for DU145

Fig 8:
Mean curves for HT29

Figure 9:
Significant genes obtained from two methods

Figure 10:
Comparison of the significant genes found by two methods

Figure 11:
Plot of significant genes in three cell lines where x-axis is time and y-axis is gene expression

Table 1:

Result statistics from the 3 Cell lines analyzed in series GSE29641.

Cell line	# of time points	# of DRGs	# of top DRGs clustered	# of GRMs
DU145	8	6325	3000	26
HT29	8	7651	3000	29
MCF7	8	8169	3000	29

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table. 2.

The most important modules for MCF7

Highest In-degree	Highest Out-degree	Highest Betweenness
M27	M21	M5
M25	M26	M3
M7	M5	M21
M21	M25	M25
M3	M15	M7

Table. 3.

The most important modules for DU145

Highest In-degree	Highest Out-degree	Highest Betweenness
M20	M23	M3
M11	M26	M20
M23	M22	M1
M3	M3	M10
M10	M20	M5

Table. 4.

The most important modules for HT29

Highest In-degree	Highest Out-degree	Highest Betweenness
M4	M27	M9
M27	M9	M5
M17	M29	M27
M21	M24	M14
M6	M22	M2

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript