Meta-analysis of hypoxic transcriptomes from public databases

Hidemasa Bono
Database Center for Life Science (DBCLS), Joint Support-Center for Data Science Research, Research Organization of Information and Systems, 1111 Yata, Mishima, Shizuoka 411-8540, Japan

Abstract

Hypoxia is an insufficient level of oxygen supply in the cell, and hypoxia-inducible factor is a central regulator of oxygen homeostasis. In order to elucidate functional insights in hypoxic response in data-driven way, we attempted meta-analysis of hypoxic transcriptome for public expression data which have been archived as microarray and RNA-seq data in public databases, NCBI Gene Expression Omnibus (GEO) and EBI ArrayExpress.

While various hypoxic conditions (oxygen concentration and duration of hypoxia) and cell lines are taken in the stored data, we manually curated possible pairs of transcriptome before and after hypoxic stress from microarray and RNA-seq data. As a result, we got 37 pairs in human and 53 pairs in mouse from microarray and 23 pairs in human from RNA-seq. We counted the number of experiments for all genes and classified into three categories, which are up-regulated, down-regulated, and unchanged. We then compared human and mouse in microarray, and microarray and RNA-seq in human. Genes up-regulated in all records contained well-studied hypoxia responsive genes, while the number of down-regulated genes were smaller and most of them included unfamiliar ones in hypoxia.

Meta-analysis approach to public gene expression database is useful for selecting candidate genes from gene expression profiles of various experimental conditions. The data produced in this study can be a good resource for hypoxic researches.

Introduction

After the invention of microarray, which can measure gene expression profile of genes (transcriptome) at a time, an attempt to archive transcriptome data to the public database has been conducted over two decade. After the completion of genome sequencing of human and mouse, the continuous effort to develop the high-throughput sequencing technology. It enables us to sequence tens of millions reads at a time, and this method, which we calls RNA-seq, is now one of major method to measure transcriptomes for samples of interest. By counting the number of sequences, we can measure the intensity of gene expression, which is now called RNA-seq.

Those data are archived to two large public repositories for transcriptome data, which consist of NCBI Gene Expression Omnibus (GEO) [1] and EBI ArrayExpress [2]. Total size of these archives is now over two million by samples and near a hundred thousand by series. They are open access and thus ready to be utilized for the data-driven research. Nevertheless, large-scale comparison among archived transcriptomes has not been conducted because it is not easy to compare gene expression data. The hardness of the comparison is that there are several microarray platforms. Even if the same platform is used, it is difficult to interpret data from different laboratories as the experimental protocols are slightly different.

Hypoxia is an insufficient level of oxygen supply in the cell, and hypoxia-inducible factor is a central regulator of oxygen homeostasis. Experimental design of archived transcriptome data is different, but they often include gene expression data for before and after hypoxic stress. Hypoxic stress gives very large changes in gene expression, and thus we thought it could be a model stimuli for comparison of gene expression data from different data sources.

Thus, we attempted meta-analysis approach for the gene expression data of hypoxic stress to elucidate key genes in data-driven way. For that purpose, we first curated hypoxic transcriptome data from NCBI GEO and EBI ArrayExpress to make a list of hypoxia-normoxia transcriptome pairs. We then did systematic transcriptome quantification procedures for the data sets we collected. By introducing a new metric, counting samples by genes where expression of a gene is up-regulated, we can see that gene is up-regulated in most of experimental conditions. First, we compared such values from hypoxia transcritomes of human and mouse by microarray. We then compared those values from human hypoxic transcriptomes by microarray and RNA-seq. These comparisons reveal actually up-regulated genes in data-driven way.

Results

Curation of hypoxic transcriptome data in public database

Hypoxia related gene expression data were sought in the public database utilizing a web tool, All of gene expression (AOE; http://aoe.dbcls.jp/), which has been maintained as the index of EBI ArrayExpress including NCBI GEO. The conventional keyword search by ‘hypoxia’ showed that the number of experiments was very few in most model organisms except human and mouse.

As a first step, the human and mouse data of those were investigated. In order to reduce the noise raising from the comparison among different microarray platforms, only Affymetrix GeneChip data was considered for further analyses in this case. We curated all data manually to

make pairs of data before and after hypoxic stress. The complete list of those pairs in Affymetrix GeneChip platform is accessible from figshare.

- human microarray https://doi.org/10.6084/m9.figshare.5811786.v1
- mouse microarray https://doi.org/10.6084/m9.figshare.5811735.v1

The pair of samples before and after hypoxic stress was made after careful curation of description of dataset. From GeneChip data, we can get 37 pairs from 11 data series in human and 53 pairs from 8 data series in mouse.

For RNA-seq samples, similar curation of data is possible and we made pairs of data before and after hypoxic stress similarly. The complete list of pairs in RNA-seq data by Illumina sequencers is also accessible from figshare (https://doi.org/10.6084/m9.figshare.5811987.v1). As a result, we got 23 pairs from 7 data series in human as of June 2017. Interestingly, no mouse RNA-seq data can be found currently. Overall procedure of this study is depicted in Figure1.

Meta-analysis of hypoxia responsive genes

After the quantification of gene expression from microarray and RNA-seq, the number of conditions was counted by up-regulated, down-regulated, and unchanged after the normalization of microarray data. The reason for categorizing these three groups is that data is highly series specific owing to various cell lines and experimental conditions. Complete lists of the meta-analysis result are accessible from figshare with following DOIs.

- Human microarray https://doi.org/10.6084/m9.figshare.5812695.v1
- Mouse microarray https://doi.org/10.6084/m9.figshare.5812701.v1
- Human RNA-seq https://doi.org/10.6084/m9.figshare.5812710.v1

In order to visualize differentially expressed genes, values called totaldiff (the number of UP counts minus the number of DOWN counts) were calculated for all genes. And then, corresponding genes between human and mouse were related utilizing orthologous gene table generated from Ensembl Biomart [3]. Figure 2 shows a scatter plot of totaldiff values for human and mouse. Genes located in upper-right are those with up-regulated after hypoxia both in human and mouse. Genes in that area contained contained previously reported typical hypoxia responsive genes for example *DEC1*(*BHLHE40*), *VEGFA* and *EGLN3*. This observation indicates that our meta-analysis does not draw erroness conclusion. On the other hand, genes down-regulated both in human and mouse included unfamiliar ones. A set of down-regulated genes from human microarrays was used in the analysis and was revealed to be involved in DNA damage recognition and repair genes [4].

As only human data was available in RNA-seq currently, the comparison of totaldiff values for microarray and RNA-seq was drawn in the scatterplot (Figure 3). Although the number of samples is small in RNA-seq, the correlation was more clear than that of microarray between human and mouse. It also can be a good reference for the hypoxic research.

In order to highlight an observation of genes by manual inspection, gene set enrichment analysis (GSEA) was done using Metascape (http://metascape.org/) [12]. Metascape clearly revealed that some genes up-regulated in many samples are well-known hypoxia responsive genes

while some are not (Figure 4). The latter can be a novel candidate of hypoxia responsive genes to study signaling pathway in the hypoxia research.

Discussion

While over two millions of transcriptome data by samples have been archived in public databases (NCBI Gene Expression Omnibus (GEO) and EBI ArrayExpress), these data are reported from various laboratories and thus they are very different in measured platforms and sampling conditions. These matters make it difficult to compare them each other.

We investigated the number of database entries in the public database for gene expression by the web tool All Of gene Expression (AOE; http://aoe.dbcls.jp/). It revealed that there were many platforms for microarray and RNA-seq, but the number of entries in database was biased. Indeed most platforms had few entries and a few platforms which were widely used have so many entries. These widely used platforms were Affymetrix Human Genome U133 Plus 2.0 Array (GPL570) and Affymetrix Mouse Genome 430 2.0 Array (GPL1261) for microarray, and Illumina platform for RNA-seq. Thus, we decided to curate data in these three platforms from the database for our first attempt to do meta-analysis on hypoxic transcriptome in the public database.

We manually curated data by checking metadata in the record and made the pairs of hypoxia-normoxia data. Currently we could only find human data for RNA-seq. For microarray data, we could collect 37 pairs for human and 53 pairs for mouse, and 23 pairs for RNA-seq for human.

It is often troublesome to handle ratio data which contains so many columns. Thus we tentatively set the threshold for up-regulation and down-regulation, and reduced the ratio information for all samples into the count of up-regulation and down-regulation. This conversion made it dramatically easy to interpret genes biologically. The threshold we used in this study was 1.5 fold, and it has been frequently used in the microarray analysis of Affymetrix GeneChip.

Our conventional analysis on genes where the count of up-regulation was high revealed that genes involved in enriched HIF-1 alpha transcription factor network (GSEA Gene Set: PID_HIF1_TFPATHWAY) and response to oxygen levels (Gene Ontology: GO:0070482) were significantly enriched (Figure 4). This is of course a reasonable result and shows the power of data-driven way of biological analysis.

One of future works is detailed analyses for co-upregulated genes across much more species, which will be possible after the compilation of such transcriptome data. Another future work is the inclusion of transcriptome data by other platforms because there are some hypoxia-related microarray data by different microarray platforms (by Agilent Oligoarray and Illumina Beadsarray).

All data described in the manuscript are archived in figshare as a collection (Bono, H. figshare https://doi.org/10.6084/m9.figshare.c.3983880 (2018)). Source codes to replicate our study is also available from GitHub (https://github.com/bonohu/chypoxia).

Materials and methods

Curation of public gene expression data

A web tool All Of gene Expression (AOE; http://aoe.dbcls.jp/) was mainly used for collecting hypoxia related gene expression data. AOE gives graphical web interface to search EBI ArrayExpress, which is the public database of gene expression data. Gene expression data in NCBI Gene Expression Omnibus (GEO) was continuously imported into ArrayExpress until 2017, and thus we can get all data deposited to GEO from ArrayExpress. Conventional search by keywords 'hypoxia' or 'hypoxic' was adopted to scan the databases.

In order to minimize the noise from different microarray platform, it is needed to collect only the data by the most used microarray platform. These platforms are Affymetrix Human Genome U133 Plus 2.0 Array (GPL570) for human and Affymetrix Mouse Genome 430 2.0 Array (GPL1261) for mouse. These platforms were used in this analysis because these were the most used platforms in hypoxia research in human and mouse respectively.

The process to pair 'hypoxia' and 'normoxia' experiment was done by manual curation. The complete list of these paired data is available from figshare.

- human microarray https://doi.org/10.6084/m9.figshare.5811786.v1
- mouse microarray https://doi.org/10.6084/m9.figshare.5811735.v1
- human RNA-seq https://doi.org/10.6084/m9.figshare.5811987.v1

Gene expression quantification

The expression values of the genes were calculated from original image files (CEL file). CEL files were individually downloaded from the GEO web site, and processed with robust multi-array averaging (RMA) normalization [5] by the affy package (version 1.50.0) [6] in R (version 3.3.3)/BioConductor (version 3.3) [7]. The R code used in this step is available from Github (https://github.com/bonohu/chypoxia/blob/master/RMA.r). Log ratio data before and after hypoxic stress for human and mouse is uploaded to figshare.

- Human microarray https://doi.org/10.6084/m9.figshare.5812668.v1
- Mouse microarray https://doi.org/10.6084/m9.figshare.5812698.v1

For RNA-seq data, hypoxia related records were sought utilizing AOE and then corresponding sequence reads in the Sequence Read Archive (SRA) [8] were retrieved from DDBJ FTP site (ftp://ftp.ddbj.nig.ac.jp/). Because the retrieved data was compressed in SRA format, these files were converted to FASTQ files by fastq-dump program in SRA Toolkit (https://www.ncbi.nlm.nih.gov/sra/docs/toolkitsoft/). For faster data processing in real time, paralleled implementation of fastq-dump, called pfastq-dump, was adopted (https://github.com/inutano/pfastq-dump/). Only Illumina reads were taken, and both single and pair-end reads were re-used for the analysis.

The expression values from RNA-seq data were quantified from sequence reads (FASTQ file) obtained from SRA. Raw FASTQ files in SRA were all processed with Trim Galore! which is a wrapper script to automate quality and adapter trimming as well as quality control (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/). RNA-seq reads were then processed

using RNA-Seq by Expectation-Maximization (RSEM) (version1.2.31) [9]. As a reference mapping tool, Bowtie2 (version 2.3.2) was used from RSEM [10]. For differential expression analysis, EBSeq was used by following a short tutorial (https://github.com/bli25ucb/RSEM_tutorial) and "PostFC" value was used for further analyses [11].
Fold change data before and after hypoxic stress from RNA-seq was also uploaded to figshare (https://doi.org/10.6084/m9.figshare.5812704.v1).

For up/down regulated gene selection, tentative threshold (1.5 fold) was adopted in this study. The number of experiments in which gene was up/down regulated was counted for all genes in human. Same calculation was also done in mouse for microarray data. Orthologous gene mapping between human and mouse and the corresponding annotation of genes were obtained using Ensembl Biomart (http://www.ensembl.org/biomart) [3]. Complete lists of counts (up/down/unchanged) after hypoxic stress are uploaded to figshare.

- Human microarray https://doi.org/10.6084/m9.figshare.5812695.v1
- Mouse microarray https://doi.org/10.6084/m9.figshare.5812701.v1
- Human RNA-seq https://doi.org/10.6084/m9.figshare.5812710.v1

All codes and scripts in bash, R and Perl to process data is available from GitHub at https://github.com/bonohu/chypoxia/ .

Visualization and functional analysis of genes

For the visualization of data obtained, TIBCO Spotfire Desktop (version 7.6.0) with TIBCO Spotfire's "Better World" program license (TIBCO Spotfire, Inc., Palo Alto, CA, USA) (http://spotfire.tibco.com/better-world-donation-program/) was used to produce scatter plots.

Metascape (http://metascape.org/) was used for a gene set enrichment analysis with default parameters. For the extraction of a gene list for metascape input, the score called totaldiff was calculated for all genes. The totaldiff score was calculated by count of human microarray UP]-[count of human microarray DOWN] + [count of human RNA-seq UP] - [count of human RNA-seq DOWN]. A list of gene was generated by extracting genes whose totaldiff score was over 15, where roughly 1% of all genes could be listed.

Acknowledgement

Author thanks Dr. Hiromasa Ono for microarray data processing protocol in R/BioConductor and Mr. Tazro Ohta for the recommendation for the use of parallel implementation of fastq-dump program. The computing resource was partly provided by super computer system at National Institute of Genetics (NIG), Research Organization of Information and Systems (ROIS), Japan.

Funding

This work was supported by the National Bioscience Database Center of the Japan Science and Technology Agency.

References

1. Barrett, T. et al. NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res*. **41**, D991–D995 (2013).
2. Kolesnikov, N. et al. ArrayExpress update--simplifying data submissions. *Nucleic Acids Res*. **43**, D1113-1116 (2015).
3. Kinsella RJ, Kähäri A, Haider S, Zamora J, Proctor G, Spudich G, Almeida-King J, Staines D, Derwent P, Kerhornou A, Kersey P, Flicek P. Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database (Oxford)*. **2011**, bar030 (2011)
4. Nakamura H, Bono H, Hiyama K, Kawamoto T, Kato Y, Nakanishi T, Nishiyama M, Hiyama E, Hirohashi N, Sueoka E, Poellinger L, Tanimoto K. Differentiated Embryo Chondrocyte plays a crucial role in DNA damage response via transcriptional regulation under hypoxic conditions. *PLOS ONE,* **13**(2): e0192136 (2018).
5. Irizarry, R.A. et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics.* **4**, 249-264 (2003).
6. Gautier, L., Cope, L., Bolstad, B.M., Irizarry, R.A. affy--analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics.* **20**, 307-315 (2004).
7. Gentleman, R.C. et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*. **5**, R80 (2004).
8. Kodama, Y., Shumway, M. & Leinonen, R. International Nucleotide Sequence Database Collaboration. The Sequence Read Archive: explosive growth of sequencing data. *Nucleic Acids Res*. **40**, D54–D56 (2012).
9. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*. **12**, 323 (2011).
10. Langmead B, Salzberg S. Fast gapped-read alignment with Bowtie 2. *Nature Methods.* **9**:357-359 (2012).
11. Leng N, Dawson JA, Thomson JA, Ruotti V, Rissman AI, Smits BM, Haag JD, Gould MN, Stewart RM, Kendziorski C. *Bioinformatics.* **29**, 1035-43 (2013).
EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments.
12. Tripathi et al. Cell Host & Microbe (2015), 18: 723-735.

Figures

Figure 1

Schematic view of presenting meta-analysis.

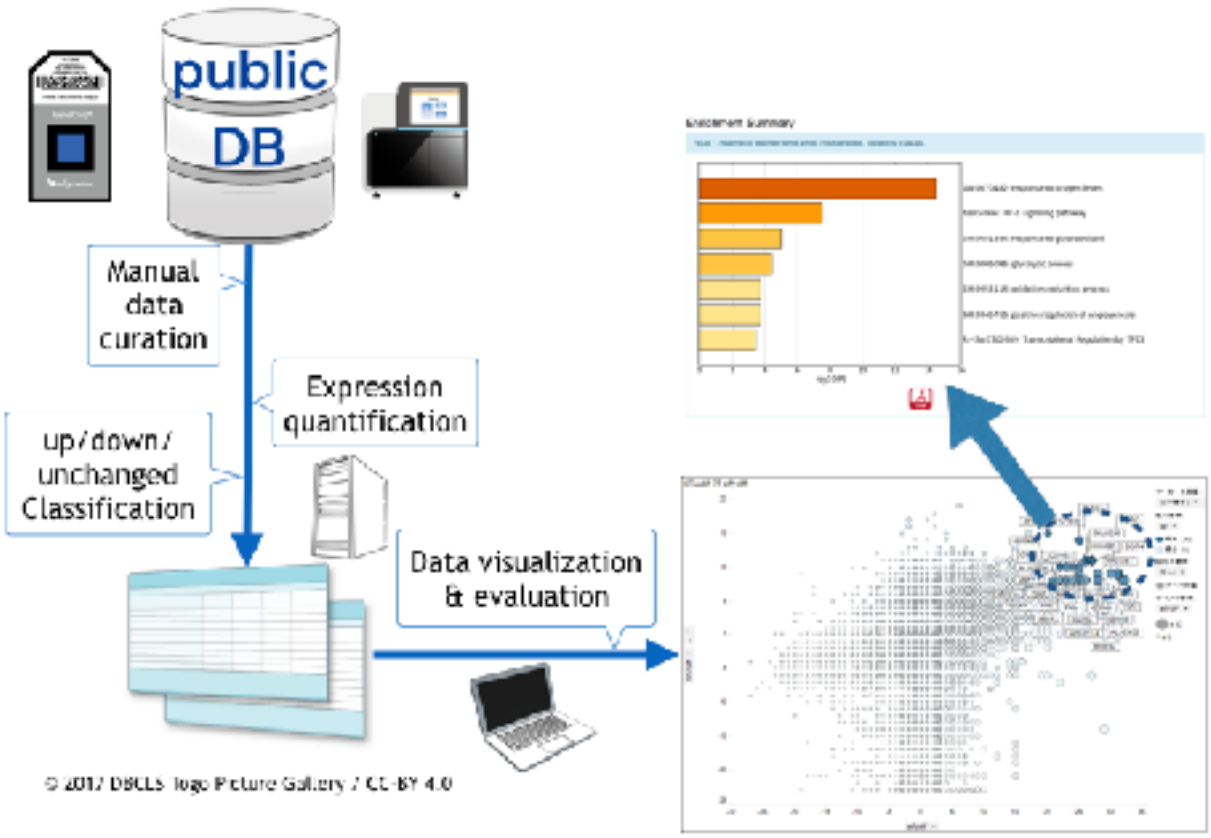


Figure 2

Comparison of meta-analysis of hypoxic transcriptomes in human and mouse by microarray. [count of human microarray UP]-[count of human microarray DOWN] in X axis and [count of mouse microarray UP]-[count of mouse microarray DOWN] in Y axis.

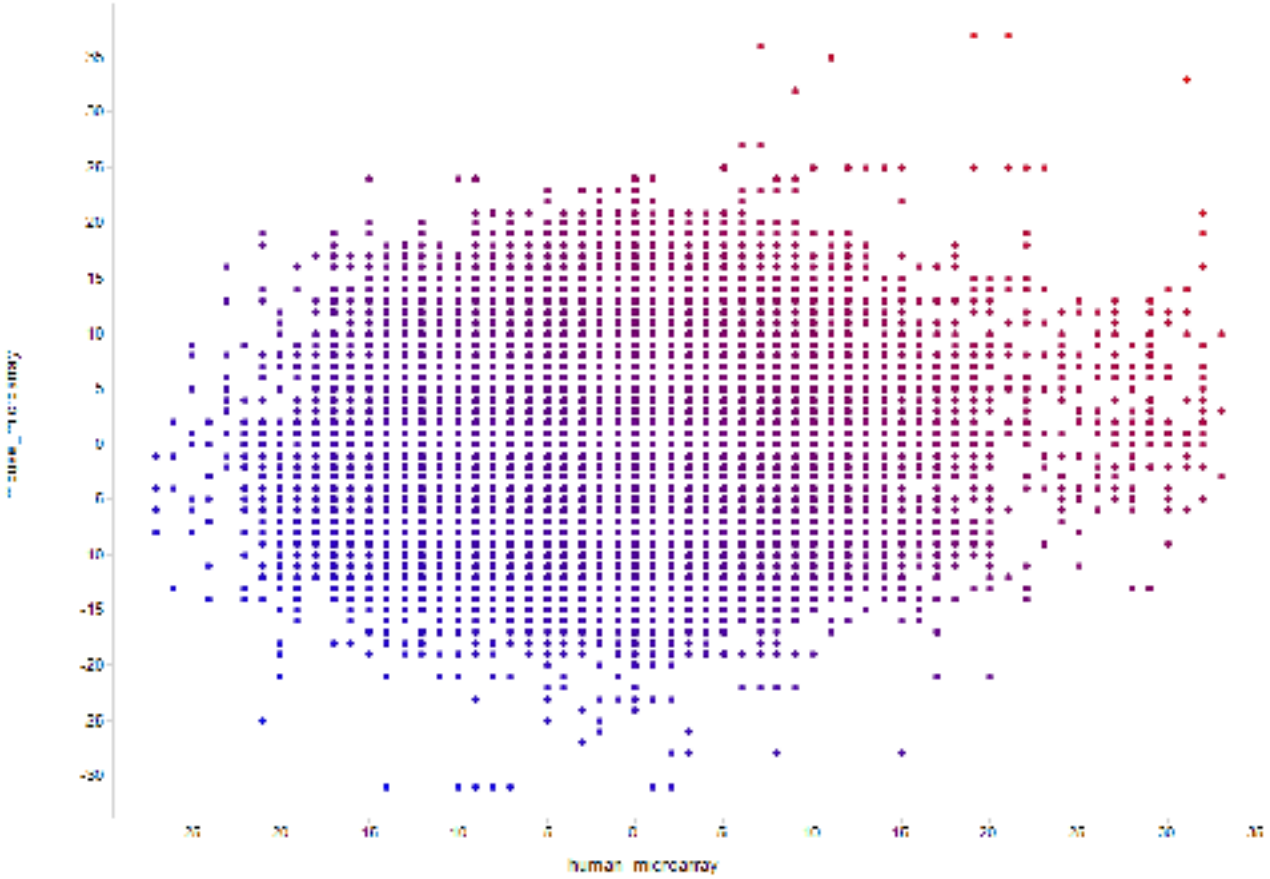


Figure 3

Comparison of meta-analysis of hypoxic transcriptomes in human measured by microarray and RNA-seq. [count of human microarray UP]-[count of human microarray DOWN] in X-axis and [count of human RNA-seq UP] - [count of human RNA-seq DOWN] in Y-axis.

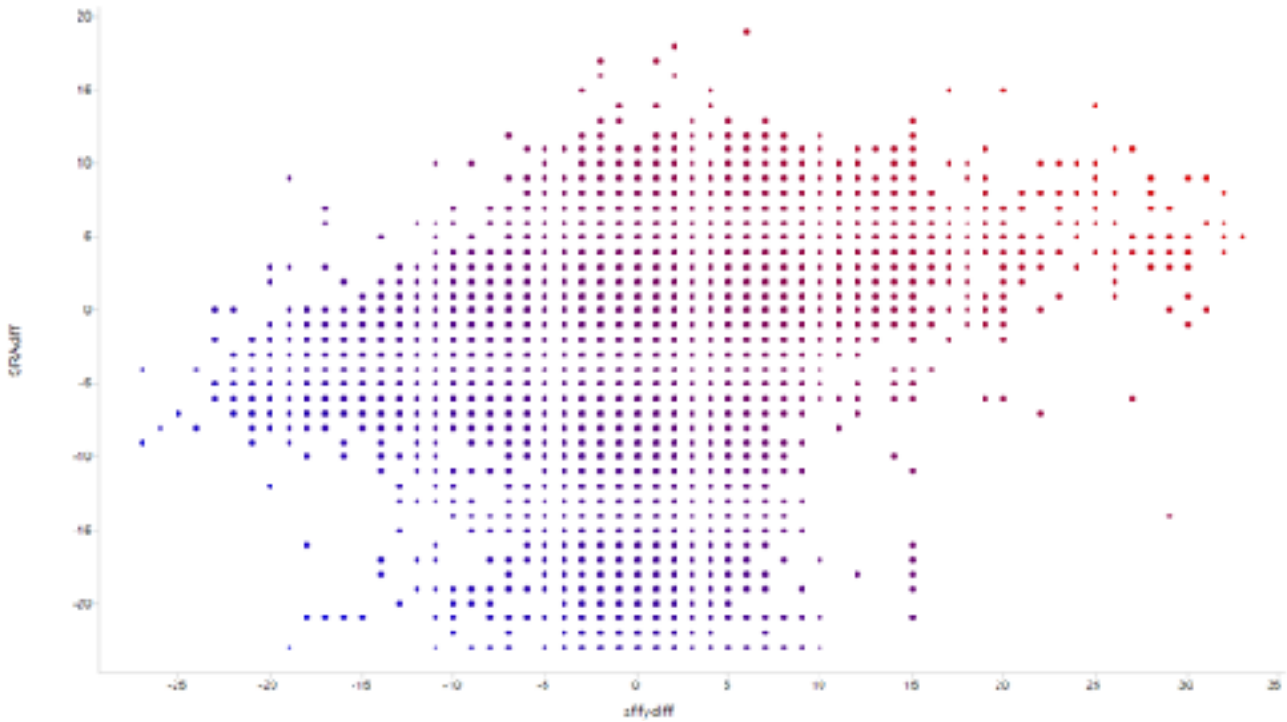
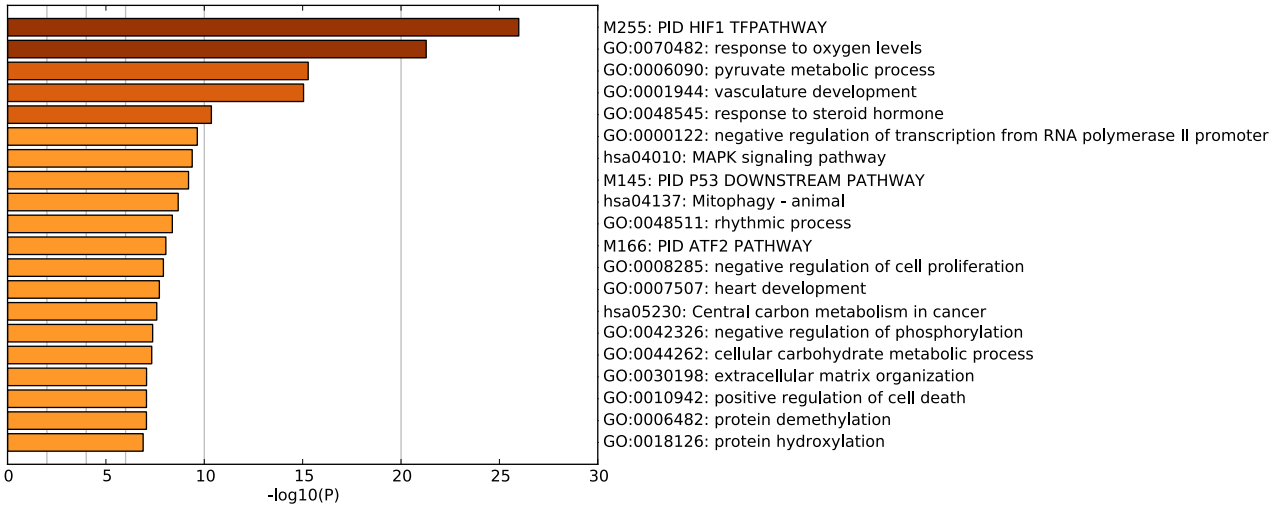


Figure 4

Metascape analysis of top genes up-regulated after hypoxic stress. For the extraction of a gene list for metascape input, the score called totaldiff was calculated for all genes. The totaldiff score was calculated by [count of human microarray UP]-[count of human microarray DOWN] + [count of human RNA-seq UP] - [count of human RNA-seq DOWN]. A list of gene was generated by extracting genes whose totaldiff score was over 15, where roughly 1% of all genes could be listed.