*Article*

# Multi-omic meta-analysis of transcriptomes and the bibliome uncovers novel hypoxia-inducible genes

**Yoko Ono [1] and Hidemasa Bono [1,*]**

[1] Program of Biomedical Science, Graduate School of Integrated Sciences for Life, Hiroshima University, 3-10-23 Kagamiyama, Higashi-Hiroshima, Hiroshima 739-0046, Japan
* Correspondence: bonohu@hiroshima-u.ac.jp; Tel.: +81-82-424-4013

**Abstract:** Hypoxia is a condition in which cells, tissues, or organisms are deprived of sufficient oxygen supply. Aerobic organisms have a hypoxic response system, represented by hypoxia-inducible factor 1-$\alpha$ (HIF1A), to adapt to this condition. Due to publication bias, there has been little focus on genes other than well-known signature hypoxia-stimulating genes. Therefore, in this study, we performed a meta-analysis to identify novel hypoxia-responsive genes. We searched publicly available transcriptome databases to obtain hypoxia-related experimental data, retrieved the metadata, and manually curated it. We selected the genes that are differentially expressed by hypoxic stimulation, and evaluated their relevance in hypoxia by performing enrichment analyses. Next, we performed a bibliometric analysis using gene2pubmed data to examine genes that have not been well studied in relation to hypoxia. Gene2pubmed data that provides information about the relationship between gene and publication. We calculated and evaluated the number of reports and similarity coefficients of each gene to HIF1A, which is a representative gene in hypoxia studies. In this data-driven study, we report that several genes that were not known to be associated with hypoxia, including the G protein-coupled receptor 146 gene, are upregulated by hypoxic stimulation.

**Keywords:** hypoxia; RNA-seq; ChIP-seq; gene2pubmed; bibliome; meta-analysis; signature genes; *GPR146*; enrichment analysis

## 1. Introduction

The development of genetic engineering techniques such as genome editing have made it possible to test hypotheses that are constructed on the basis of the knowledge of each researcher, including literature information, by performing biological experiments. Hypothesis construction is triggered by literature surveys and the collective knowledge of researchers, and is influenced by publication bias. For example, when the frequency of about 600,000 publications annotated with about 20,000 human coding genes was calculated, more than 9,000 publications reported the p53 gene, whereas over 600 genes were not mentioned at all [1]. In contrast, the development of microarray and high-throughput sequencing technologies has enabled the comprehensive acquisition of thousands of gene expression profiles at a time. Published transcriptome data are archived in public databases such as the Gene Expression Omnibus (GEO) of the U.S. National Center for Biotechnology Information (NCBI) [2], ArrayExpress (AE) of the European Bioinformatics Institute [3], and Genomic Expression Archive (GEA) of the DNA Data Bank of Japan [4]. At present, although approximately 60,000 expression data series of human genes have been registered, we estimate that they are used in few studies. We believe that meta-analysis of these transcriptome data may provide additional insights into biology.

Among the various fields of biological research, the more well-studied the field, the greater the impact of publication bias. We have considered areas where meta-analysis based on gene expression data can uncover new findings buried in publication bias. The

analyses were performed by considering the following conditions: (1) hypoxia-inducible factor 1-$\alpha$ (HIF1A) is a representative transcription factor involved in hypoxic stimulus-response, and (2) categorization in Gene Ontology, a database where gene expression data can easily be obtained. In our previous study, we studied hypoxic stimulus-responses using public databases [5]. We performed a meta-analysis of gene expression data following the previous method. The present study is the continuation of our previous study to further build on the knowledge of hypoxic stimulus-response. In a previous report, we integrated meta-analyzed hypoxic transcriptome data with public ChIP-seq data of known human HIFs, HIF-1 and HIF-2, to gain insight into hypoxic response pathways involving direct binding of transcription factors. In this study, we added omic-scale analysis of bibliological data in PubMed, which is called 'bibliome', by using gene2pubmed that provides information about the relationship between gene and publication. Bibliometric analysis is a method of analysis of biological publications and is used in biomedical research [6].

Oxygen is essential for respiration in aerobic organisms. Hypoxia occurs in a variety of neurological conditions such as traumatic brain injury, Alzheimer's disease, and stroke, and is known to be responsible for some of their symptoms. In hyperbaric oxygen therapy, subjects are placed in a chamber filled with 100% oxygen gas at a pressure of 1 atm or higher, and is mainly used to treat hypoxia-related symptoms [7]. Oxygen is used to produce ATP by oxidative phosphorylation and electron transfer systems. Hypoxia is a condition in which cells, tissues, or organisms are deprived of sufficient oxygen supply as the amount of oxygen in the blood decreases. The tissues and cells of higher organisms including humans have a hypoxic response system, regulated by HIFs to adapt to this oxygen-deprived condition caused by hypoxic stimuli.

Under normoxic conditions, HIFs are hydroxylated by the $\alpha$-ketoglutarate–dependent dioxygenase factor inhibiting HIF-1 (FIH-1) and prolyl hydroxylases (PHDs), resulting in degradation by the ubiquitin-proteasome system and suppression of transcriptional activation. However, under hypoxic conditions, the activity of PHDs and FIH-1 is reduced, and HIFs escape hydroxylation and form a complex with aryl hydrocarbon receptor nuclear translocator (ARNT) and the transcriptional co-factor CREB-binding protein to induce expression of downstream genes.

In this study, we focused on hypoxic stimulus-response and aimed to identify novel genes by performing meta-analysis using public databases. We selected hypoxia-responsive genes based on the gene expression ratio of approximately 500 pairs of hypoxia-stimulated samples and evaluated the similarity coefficient of each gene for HIF1A, a representative factor in hypoxia research, using gene2pubmed, which shows the relationship between gene and publication.

## 2. Materials and Methods

### *2.1. Curation of Public Gene Expression Data*

To obtain the accession of hypoxia-related gene expression data from public databases, we used a graphical web tool called All of Gene Expression (AOE) [8,9]. AOE integrates metadata from not only GEO, AE, and GEA but also RNA sequencing (RNA-seq) data that are exclusively archived in Sequence Read Archive (SRA) [10]. We searched for a list of experimental data series related to hypoxia from GEO using the following search formulas: "hypoxia"[MeSH Terms] OR "hypoxia"[All Fields] AND "Homo sapiens"[porgn] AND "gse"[Filter]. We downloaded this list on August 17, 2020. We obtained the metadata from the Series Matrix files in GEO, from articles, or downloaded the metadata from SRA using the Python package pysradb (v 0.11.1). We then curated the data into the comparable sample pairs of hypoxia and normoxia (HN-pairs) [11]. The criteria for curation were that the specimens with the experimental conditions of hypoxia and normoxia should be in the same data series so as to set HN-pair and to derive RNA-seq reads from human cell lines or tissue specimens. All relevant data were adopted if the

metadata content was certain. We collected the SRA format data from the NCBI using prefetch (version 2.9.6).

### 2.2. Gene Expression Quantification

Since the downloaded data were in the SRA format, we used the fasterq-dump program in SRA Toolkit [12] to convert the data into the FASTQ formatted files for expression quantification. We then quantified single-end and paired-end RNA-seq reads using ikra (version 1.2.3) [13], an RNA-seq pipeline with default parameters. Ikra automates the RNA-seq data analysis process, which includes the quality control of reads (Trim Galore version 0.6.3) and transcript quantification (Salmon version 0.14.0 [14] with reference transcript sets in GENCODE release 30). We chose this workflow because it is suitable for quantifying gene expression from heterogeneous RNA-seq data obtained from various instruments and laboratories. In this study, the data acquisition and quality control processes required approximately 1 month to obtain the current dataset. Quantitative RNA-seq data are accessible at figshare [15].

### 2.3. Calculation of HN-ratio and HN-score

Based on the transcriptome data paired with hypoxia and normoxia, we calculated the ratio of expression value for each gene (termed as HN-ratio) [16]. HN-ratio $R$ is calculated using the following equation:

$$R = \frac{T_{hypoxia} + 1}{T_{normoxia} + 1} \quad (1)$$

where $T_{hypoxia}$ is the gene expression value called scaled transcripts per million (TPM) [17] under hypoxic conditions and $T_{normoxia}$ is TPM under normoxic conditions, paired with hypoxic conditions. To reduce the effect of small variations in extremely low gene expression values, we calculated HN-ratio by adding 1 to the denominator and numerator.

We then classified each gene into three groups based on its HN-ratio. When the HN-ratio was greater than the threshold, the gene was considered upregulated, and when the ratio was less than the inverse of the threshold, the gene was considered downregulated. If a gene was neither up- nor downregulated, it was classified as unchanged. For the classification of up- and downregulated genes, we tested 1.5-and 2-fold thresholds, of which we then selected the 1.5-fold threshold to classify upregulated and downregulated genes. For evaluating hypoxia-inducible genes, we calculated HN-score for each coding gene in humans. HN-score was calculated by subtracting the number of samples with downregulated genes from the number of samples with upregulated genes.

### 2.4. Enrichment Analysis

We used Metascape [18,19] and ChIP-Atlas [20,21] for Gene Set Enrichment Analysis. ChIP-Atlas is a comprehensive and integrated database for visualization and utilization of publicly available chromatin immunoprecipitation sequencing (ChIP-seq) data. In this study, we performed conventional "express analysis" using Metascape. In the ChIP-Atlas settings, we set the "Select dataset to be compared" item to "Refseq coding genes (excluding user data)" and set the other items to default.

### 2.5. Meta-Analysis of ChIP-Seq Data

Public ChIP-seq data were collected, curated, and pre-calculated for reuse in the ChIP-Atlas database [20,21]. In ChIP-Atlas, the category of proteins that bind to DNA is termed "antigens". We used the "Target Genes" function of ChIP-Atlas to obtain average Model-based Analysis of ChIP-seq (MACS2) scores for three antigens at a distance of ±5k from the transcription start site. These antigens were selected from the following genes that were shown to be related to the UP 100 genes by ChIP-Atlas enrichment analysis:

HIF1A, endothelial PAS domain-containing protein 1 (EPAS1, also known as hypoxia-inducible factor-2$\alpha$), and ARNT. We combined the ChIP-seq data with HN-score data described above, using the gene name to combine the two datasets.

### 2.6. Calculation of the Number of Publications for Each Gene and Similarity Coefficient for HIF1A

We calculated the number of publications and Simpson similarity coefficient per human gene using Python (version 3.8.6) [22]. The Simpson similarity coefficient $S$ was calculated using the following equation:

$$S(X,Y) = \frac{|X \cap Y|}{min(|X|,|Y|)} \quad (2)$$

where $|X \cap Y|$ is the number of PubMed IDs that overlaps between HIF1A and a gene in gene2pubmed, whereas $min(|X|,|Y|)$ is the number of PubMed IDs of HIF1A or the gene in gene2pubmed; whichever value is less. Gene2pubmed [23] and Gene Info [23], which were required for the above calculations, were downloaded on January 4, 2021. In order to select a gene to determine whether or not hypoxia-related studies have been reported, we calculated the similarity coefficients of each gene for EPAS1 and ARNT as well as HIF1A. After reviewing the results, we determined that HIF1A was suitable.

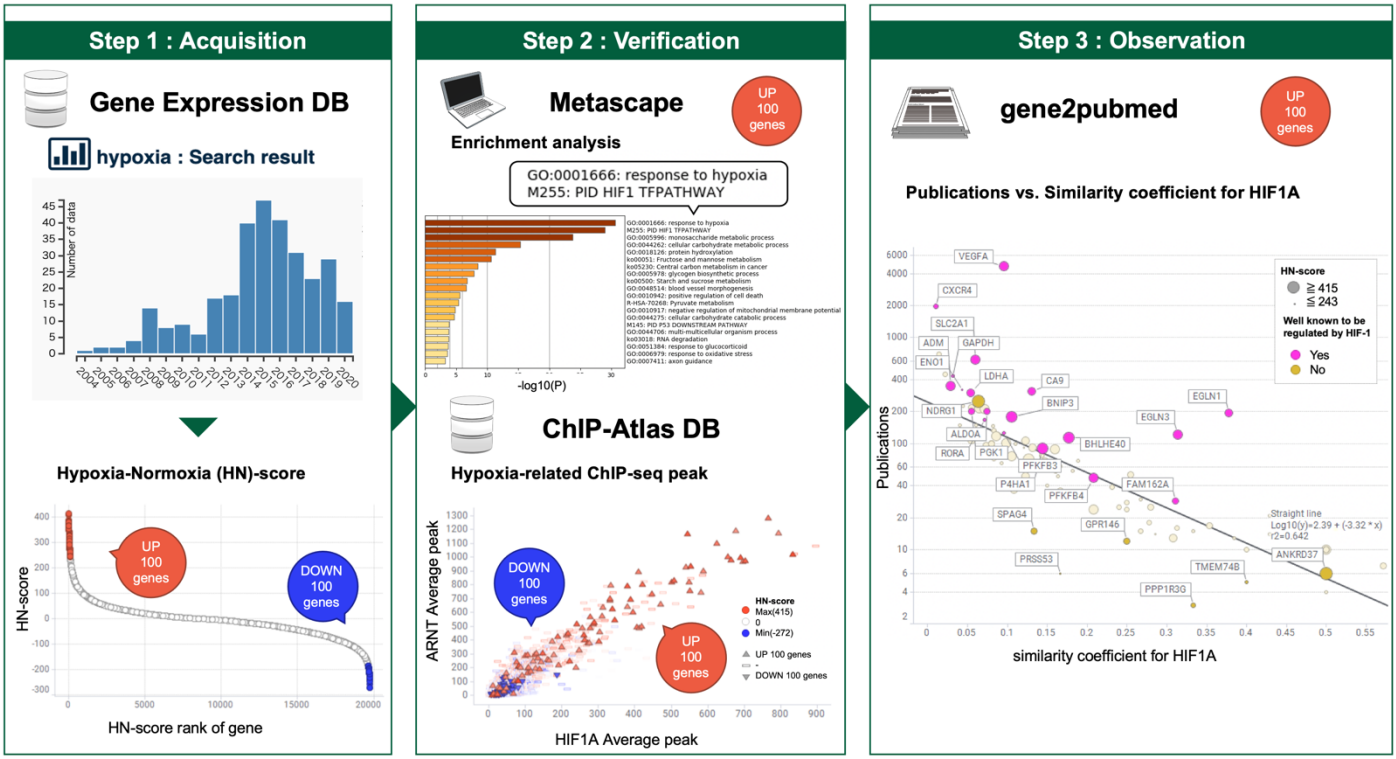
### 2.7. Visualization and Integrated Functional Analysis of Genes

To generate scatter and box plots, we used TIBCO Spotfire Desktop version 11.0.0 (TIBCO Spotfire, Inc., Palo Alto, CA, USA).

## 3. Results

### 3.1. Overview

This study was performed in three main steps (Figure 1). In step 1, we obtained the RNA-seq data related to hypoxic stimulation from public gene expression databases. After the manual curation of metadata, UP 100 and DOWN 100 gene lists related to hypoxia stimulation were obtained through the quantification of gene expression in the corresponding datasets. The UP 100 and DOWN 100 gene lists included the top 100 and bottom 100 genes in HN-score, respectively. In step 2, we verified hypoxia-related genes in UP 100 and DOWN 100 gene lists, not only by enrichment analyses, but also by visualizing the HIF-related ChIP-seq peak of each gene using HN-score. Finally, in step 3, we investigated novel hypoxia-responsive genes using bibliometric analysis of gene2pubmed.




© 2016 DBCLS TogoTV / CC-BY-4.0

**Figure 1.** Schematic view of hypoxic transcriptome meta-analysis. Step 1. Evaluation and listing of upregulation and downregulation of hypoxia-responsive genes. Step 2. Confirmation of known hypoxic stimulation-related genes. Step 3. Discovery of novel genes related to hypoxic stimulus−response.

### *3.2. Curation of Hypoxic Transcriptome Data in Public Databases*

Initially, we used AOE to check whether a large number of hypoxia-related data were registered. We then curated the metadata provided by the NCBI for further analysis. Next, we obtained the SRA IDs for 69 data series and 495 HN-pairs of GEO-drawn samples after integration with hypoxia-related data lists provided in previous reports and elimination of duplicates [11]. This resulted in a four-fold increase in HN-pairs compared to previous reports. In 495 HN-pair data, the hypoxic conditions ranged between 0.1% and 5% $O_2$ concentration, including some chemical hypoxic conditions where $CoCl_2$ was used to induce a hypoxia-related state under normoxic conditions. The treatment time ranged from 1 hour to a maximum of 3 months. The most common hypoxic condition among the data was 1% $O_2$ (266 HN-pairs, 53.7%) and 24 hours of treatment (234 HN-pairs, 47.3%). The most common cell type was cancer (324 HN-pairs, 65.5%), and the most common tissue among them was breast cancer (112 HN-pairs, 22.6%).

### *3.3. Meta-Analysis of Hypoxia-Responsive Genes*

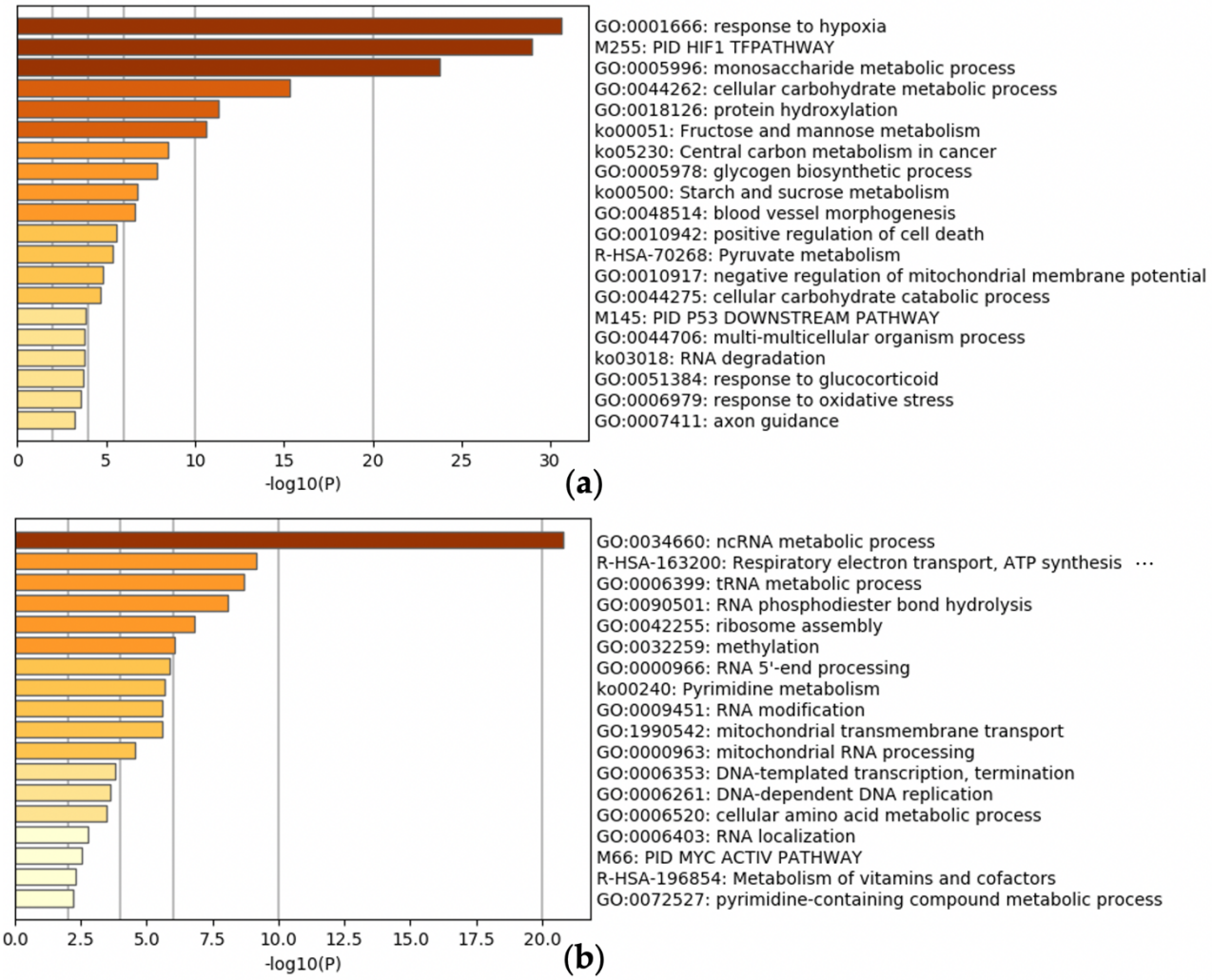
After the quantification of gene expression data obtained from RNA-seq, the number of conditions under which each gene was upregulated, downregulated, and remained unchanged was estimated on the basis of HN-ratio. The complete lists of meta-analyzed results are accessible from figshare [24]. We calculated the HN-score of each gene to evaluate responsiveness of each gene toward hypoxia. For example, in the case of vascular endothelial growth factor A *(VEGFA)*, the counts showed 406 UP, 25 DOWN, and 64 unchanged, and therefore, its HN-score was 381. Full lists of counts (upregulated/downregulated/unchanged) with HN-scores for all genes are accessible from figshare [24]. We selected the top 100 [25] and bottom 100 [26] HN-scored genes and used these genes as the UP 100 and DOWN 100 gene lists, respectively, for further analysis.

### *3.4. Evaluation of Hypoxia-Responsive Genes*

To confirm whether the UP 100 and DOWN 100 gene lists were affected by hypoxia-related regulators such as HIF1A, we performed enrichment analysis using ChIP-Atlas. Enrichment analysis showed that the UP 100 gene list was related to hypoxia-related antigens such as HIF1A, ARNT, and EPAS1, whereas the DOWN 100 gene list was related to epigenetic regulators such as Sin3A associated protein 30 (SAP30), histone deacetylase 1 (HDAC1), and MYC proto-oncogene, bHLH transcription factor (MYC), an oncogene involved in cell cycle progression (Table 1). Similarly, enrichment analysis by Metascape showed that hypoxia-related gene sets and genes related to the metabolic processes of non-coding RNAs (ncRNAs) were listed in the UP 100 and DOWN 100 gene lists, respectively (Figure 2 (a), (b)). Since it was important that the hypoxia-related genes were correctly evaluated by HN-score as intended in this analysis, we confirmed this not only by ChIP-Atlas enrichment analysis but also by visualizing the high HIF-related MACS2 score of each gene with high HN-score. The scatter plots of the mean MACS2 peaks of HIF1A, EPAS1, and ARNT, which were the top-ranked genes in the enrichment analysis of the UP 100 gene list, were plotted. The genes with higher MACS2 peaks also had higher HN-scores (Figure 2 (c), (d)). These results indicate that the genes regulated by hypoxia-related factors are listed in the UP 100 gene list, and genes regulated by SAP30, MYC, and HDAC1 are listed in the DOWN 100 gene list.

**Table 1.** Enrichment analysis of ChIP-Atlas.

| Input list | Antigen | ID | Log P-val | Log Q-val | Fold Enrichment |
|---|---|---|---|---|---|
| UP 100 gene list | HIF1A | SRX4802348 | -88.8246 | -84.064 | 35.6249 |
| | ARNT | SRX4802353 | -76.4303 | -72.3686 | 83.3136 |
| | EPAS1 | SRX3051209 | -73.1987 | -69.2831 | 34.9928 |
| DOWN 100 gene list | SAP30 | SRX116447 | -34.1844 | -30.0149 | 4.96916 |
| | MYC | SRX1497384 | -31.4158 | -27.5474 | 2.97453 |
| | HDAC1 | SRX186644 | -27.4205 | -24.1541 | 3.3231 |



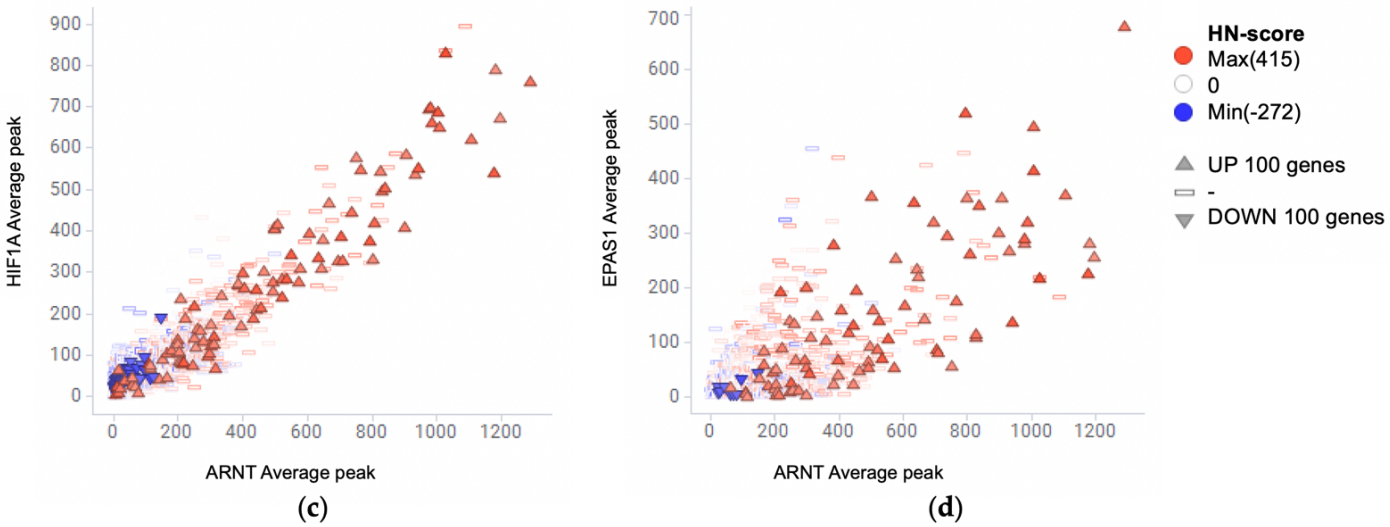



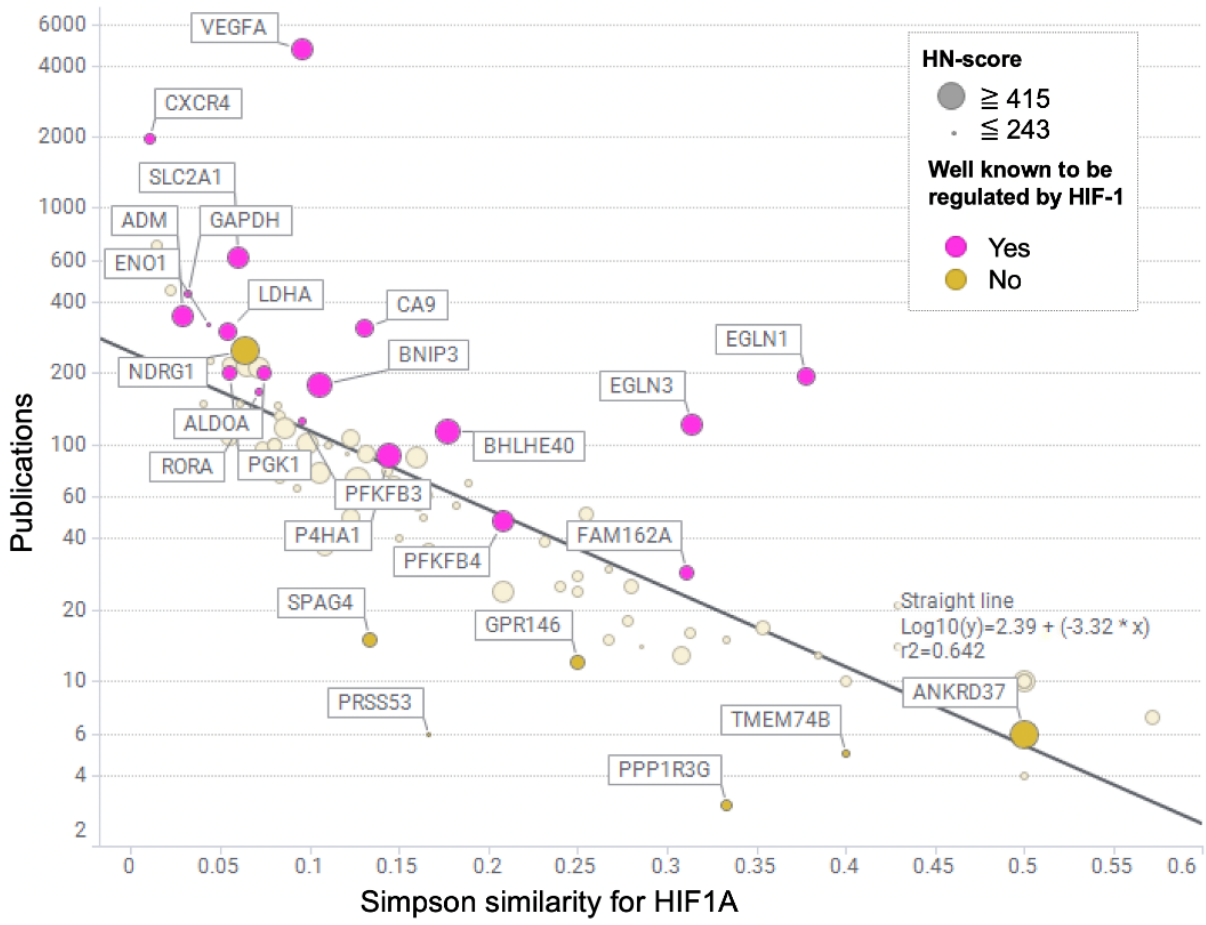

**Figure 2.** (a, b) Confirmation of known hypoxic stimulation-related genes. Enrichment analysis for (a) the UP 100 gene list and (b) the DOWN 100 gene list, (c, d) Scatter plot of ChIP-seq average peaks of hypoxic related antigens, (c) HIF1A vs. ARNT and (d) EPAS1 vs. ARNT colored by HN-score.

### 3.5. Evaluation of Simpson Similarity between HIF1A and Genes

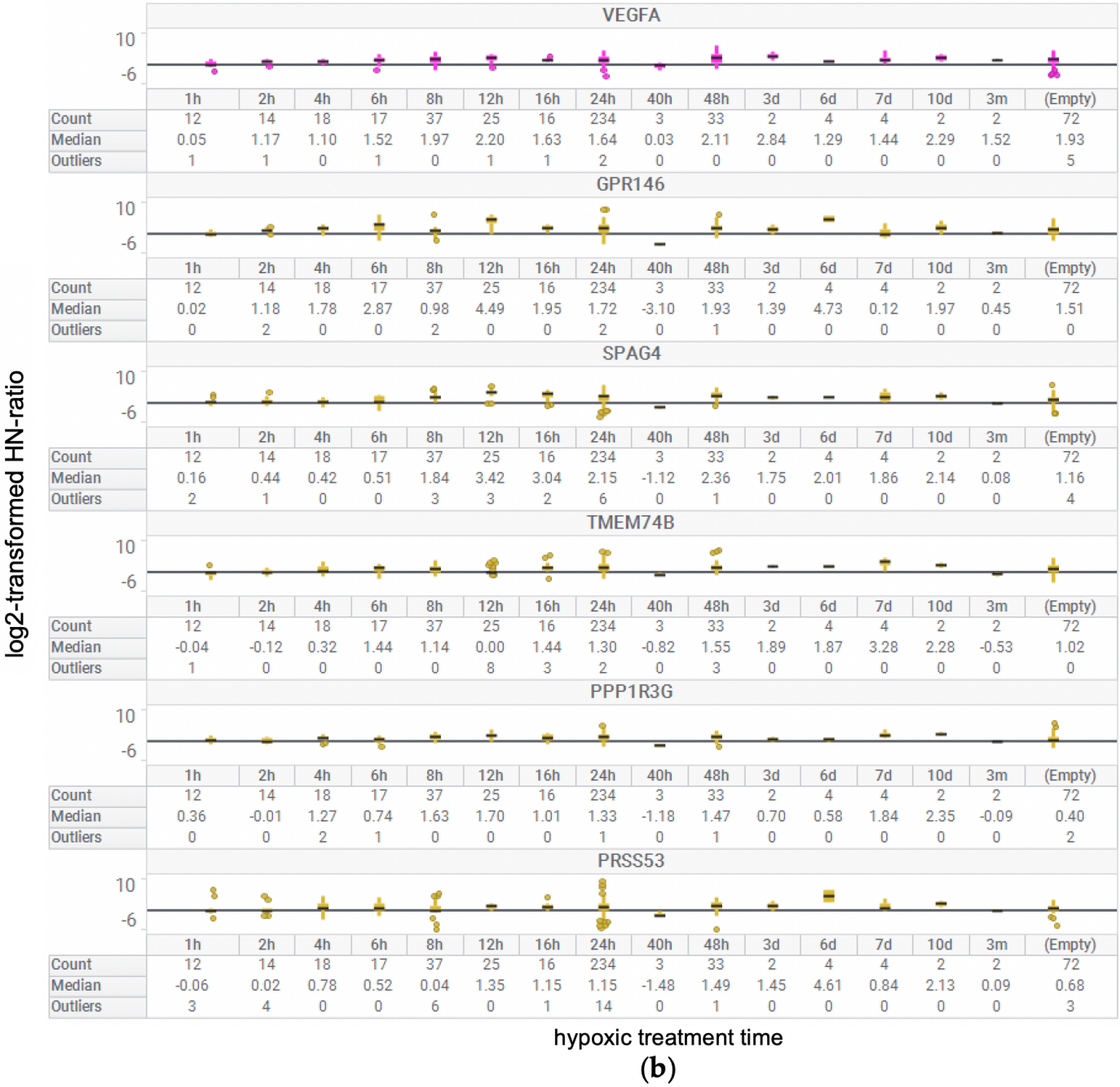
We performed bibliometric analysis, a method of analysis using biological publication information, to determine whether the genes of the UP 100 gene list had been previously studied in relation to hypoxia. We calculated the number of publications for each gene in the UP 100 gene list [27]. We also calculated the Simpson similarity coefficient between these genes and HIF1A [28]. The number of publications and Simpson similarity coefficients were visualized using the scatter plot (Figure 3 (a)). In addition, we defined genes that have already been reported to be involved with HIF-1 15 years ago [29] as "well known genes to be regulated by HIF-1", and are marked on this scatter plot. Interestingly, hypoxia-related genes were plotted on the upper side of the regression line.

### 3.6. Box Plot of HN-ratio by Treatment Time

To check the individual HN-ratio of genes with high HN-score based on the treatment time, we visualized base-2 logarithm (log2) transformed HN-ratio using the box plot (Figure 3 (b)). Genes in the box plot were selected from the HIF-1-regulated gene *VEGFA*, and some genes were plotted below the regression line, such as sperm associated antigen 4 (*SPAG4*), G protein-coupled receptor 146 (*GPR146*), protein phosphatase 1 regulatory subunit 3G (*PPP1R3G*), transmembrane protein 74B (*TMEM74B*), and serine protease 53 (*PRSS53*). These six genes visualized in the box plot showed positive HN-ratio values for the majority of treatment time.




(a)




| | 1h | 2h | 4h | 6h | 8h | 12h | 16h | 24h | 40h | 48h | 3d | 6d | 7d | 10d | 3m | (Empty) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **VEGFA** | | | | | | | | | | | | | | | | |
| Count | 12 | 14 | 18 | 17 | 37 | 25 | 16 | 234 | 3 | 33 | 2 | 4 | 4 | 2 | 2 | 72 |
| Median | 0.05 | 1.17 | 1.10 | 1.52 | 1.97 | 2.20 | 1.63 | 1.64 | 0.03 | 2.11 | 2.84 | 1.29 | 1.44 | 2.29 | 1.52 | 1.93 |
| Outliers | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 |
| **GPR146** | | | | | | | | | | | | | | | | |
| Count | 12 | 14 | 18 | 17 | 37 | 25 | 16 | 234 | 3 | 33 | 2 | 4 | 4 | 2 | 2 | 72 |
| Median | 0.02 | 1.18 | 1.78 | 2.87 | 0.98 | 4.49 | 1.95 | 1.72 | -3.10 | 1.93 | 1.39 | 4.73 | 0.12 | 1.97 | 0.45 | 1.51 |
| Outliers | 0 | 2 | 0 | 0 | 2 | 0 | 0 | 2 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| **SPAG4** | | | | | | | | | | | | | | | | |
| Count | 12 | 14 | 18 | 17 | 37 | 25 | 16 | 234 | 3 | 33 | 2 | 4 | 4 | 2 | 2 | 72 |
| Median | 0.16 | 0.44 | 0.42 | 0.51 | 1.84 | 3.42 | 3.04 | 2.15 | -1.12 | 2.36 | 1.75 | 2.01 | 1.86 | 2.14 | 0.08 | 1.16 |
| Outliers | 2 | 1 | 0 | 0 | 3 | 3 | 2 | 6 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 4 |
| **TMEM74B** | | | | | | | | | | | | | | | | |
| Count | 12 | 14 | 18 | 17 | 37 | 25 | 16 | 234 | 3 | 33 | 2 | 4 | 4 | 2 | 2 | 72 |
| Median | -0.04 | -0.12 | 0.32 | 1.44 | 1.14 | 0.00 | 1.44 | 1.30 | -0.82 | 1.55 | 1.89 | 1.87 | 3.28 | 2.28 | -0.53 | 1.02 |
| Outliers | 1 | 0 | 0 | 0 | 0 | 8 | 3 | 2 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 |
| **PPP1R3G** | | | | | | | | | | | | | | | | |
| Count | 12 | 14 | 18 | 17 | 37 | 25 | 16 | 234 | 3 | 33 | 2 | 4 | 4 | 2 | 2 | 72 |
| Median | 0.36 | -0.01 | 1.27 | 0.74 | 1.63 | 1.70 | 1.01 | 1.33 | -1.18 | 1.47 | 0.70 | 0.58 | 1.84 | 2.35 | -0.09 | 0.40 |
| Outliers | 0 | 0 | 2 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 2 |
| **PRSS53** | | | | | | | | | | | | | | | | |
| Count | 12 | 14 | 18 | 17 | 37 | 25 | 16 | 234 | 3 | 33 | 2 | 4 | 4 | 2 | 2 | 72 |
| Median | -0.06 | 0.02 | 0.78 | 0.52 | 0.04 | 1.35 | 1.15 | 1.15 | -1.48 | 1.49 | 1.45 | 4.61 | 0.84 | 2.13 | 0.09 | 0.68 |
| Outliers | 3 | 4 | 0 | 0 | 6 | 0 | 1 | 14 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 3 |

**Figure 3.** Discovery of novel genes associated with hypoxic stimulus-response. (a) Scatter plot of the number of publications vs. Simpson similarity coefficients for HIF1A in the UP 100 gene list. In this scatter plot, genes that have already been reported to be associated with HIF-1 15 years ago were marked as "known genes regulated by HIF-1". (b) Box plot of log2-transformed HN-ratio per hypoxic treatment time for some of the UP 100 genes.

## 4. Discussion

In this study, we identified hypoxia-responsive genes by performing metadata analysis. We reported that several genes that were not known to be associated with hypoxia, *GPR146*, *TMEM74B*, *PPP1R3G*, and *PRSS53*, were upregulated by hypoxic stimulation.

We analyzed 69 data series related to hypoxia, selected from the entire gene expression data registered in GEO, and obtained 495 HN-pairs, which is about four times more than previously reported data [5]. In this study, 65.5% of the gene expression data were derived from cancer experiments, reflecting the demand for these data in hypoxia-related

cancer research and ease of handling. We believe that each gene was evaluated without publication bias. Our study was possible because of the availability of data from public databases.

First, the Metascape and ChIP-Atlas enrichment analyses of the UP 100 gene list showed their association with hypoxia-related gene sets as expected, and hence, we concluded that HN-score helps select the hypoxia-responsive genes (Table 1, Figure 2). Thus, we decided to proceed with the analysis of the UP 100 gene list in subsequent analyses. On the other hand, the ChIP-Atlas enrichment analysis of the DOWN 100 gene list showed an association with MYC (Table 1). Some reports suggest that MYC is repressed by HIF. The regulation of MYC by HIF is mediated by the antagonistic inhibition of MYC protein complex formation [30], direct inhibition of MYC transcriptional activity by competition for DNA binding sites [31], and promotion of MYC proteasomal degradation [32]. We believe that these indirect effects of HIF have been reflected in this result. Metascape enrichment analysis of the DOWN 100 gene list suggested an association with the metabolic processes of ncRNA (Figure 2(b)). This suggests that hypoxic stimulation may also affect the expression of ncRNAs. In this study, only the protein-coding genes of GENCODE were quantitatively analyzed, and hence, ncRNAs were excluded from this analysis. Further studies are needed to identify ncRNAs involved in hypoxic stimulation.

Gene2pubmed has been used to determine the order of genes retrieved by RefEx [33]. In this study, we used gene2pubmed to evaluate the genes that are not well-known as hypoxia-related genes. We used gene2pubmed to calculate two variables: the number of publications per gene and Simpson similarity coefficient for HIF1A. The Simpson similarity coefficient measures the strength of the co-occurrence of PubMed IDs associated with HIF1A or with a gene. If the Simpson similarity coefficient is close to 1, the relationship between PubMed IDs and HIF1A is considered strong; however, the lower the number of PubMed IDs on the side being compared indicates a poor relationship. For this reason, we visualized the number of publications and Simpson similarity coefficient for the genes in the UP 100 gene list in the scatter plot. In addition, we also plotted HIF-1-regulated genes [29] in this plot (Figure 3 (a)). The Simpson similarity coefficient was lower in case of more publications, this might be because of the fact that more relationships were revealed, and thus the Simpson similarity coefficient is likely to be underestimated. The genes previously known to be regulated by HIF-1 scored high in the Simpson similarity coefficient and plotted on the upper side of regression line, despite a large number of publications. The genes with a small number of publications, such as ankyrin repeat domain 37, which was reported to be regulated by HIF1A [34], were also plotted above the regression line. Therefore, we hypothesized that the genes plotted below the regression line, such as *GPR146*, SP*AG4, TMEM74B, PPP1R3G and PRSS53* might be considered as candidates for novel hypoxia-responsive genes. The log2-transformed HN ratio of these genes was visualized as a box plot (Figure 3 (b)). These genes had positive HN-ratio values, as well as gene with known hypoxic responses, such as *VEGFA.* Although some reports state that *SPAG4* is related to hypoxia [35,36], no reports are available in PubMed search related to hypoxia-related genes other than *SPAG4*.

We focused on GPR146 because it is a G protein-coupled receptor and therefore worth considering as a drug target and because it has a higher HN-score than other genes. In the similarity coefficient analysis using gene2pubmed, three reports on the co-occurrence of PubMed IDs between HIF1A and GPR146 are available [37–39]. These were overall genetic analyses and did not focus on hypoxia. We confirmed that *GPR146* was upregulated, which was plotted below the regression line in the scatter plot (Figure 3 (a)), similar to other known hypoxia-responsive genes. *GPR146* was not reported to be associated with hypoxia in PubMed search; thus, we believe that *GPR146* is an important hypoxia-responsive gene that has not been focused on so far. In hypoxia studies using comprehensive gene expression data [40,41], *GPR146* is included in the list of genes upregulated by hypoxic stimulation. However, according to previous reports, *GPR146* is just a gene among many hypoxia-related genes. The inhibition of GPR146 is involved in the lowering

of cholesterol level [42]. C-peptide, a putative ligand of GPR146, inhibited low $O_2$-induced ATP release in erythrocytes [43]. The biological function of GPR146 is expected to be further elucidated in future studies.

We focused on the hypoxic response and visualized the number of research reports for each gene using gene2pubmed data. We identified the novel gene by using public data containing thousands of gene expression information, which can be obtained regardless of the researcher's interest. As shown in Figure 3(b), the gene expression variation at each time point was not constant. Also, the tissues and cells in each data set were different. In this study, we only evaluated a rough gene expression variation by using the collective intelligence from the public database. Further stratification analysis will be a future work. We plan to continue using public databases to discover similar new findings. For example, we believe that the ncRNA field, where hypothesis generation is difficult due to the lack of information, has a high potential for new discoveries. At present, the number of known human long ncRNA transcripts exceeds several hundred thousand [44] [45]. We plan to investigate the gene expression of ncRNAs during hypoxic conditions, based on our results of the effect of hypoxia on metabolic process-related genes of ncRNA (Figure. 2 (b)).

## 5. Conclusions

Multi-omics analysis of the transcriptome and bibliome revealed that several genes, which were not known to be associated with hypoxia, were upregulated under hypoxic conditions.

**Author Contributions:** Conceptualization, Y.O., H.B.; methodology, Y.O.; software, Y.O.; validation, Y.O., H.B.; formal analysis, Y.O., H.B.; investigation, Y.O.; resources, Y.O., H.B.; data curation, Y.O.; writing—original draft preparation, Y.O.; writing—review and editing, H.B.; visualization, Y.O.; supervision, H.B.; project administration, H.B.; funding acquisition, H.B. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was supported by the center of innovation for Bio-Digital Transformation (BioDX), program on open innovation platform for industry-academia co-creation (COI-NEXT), Japan Science and Technology Agency (JST, COI-NEXT, JPMJPF2010). This research was also supported by ROIS-DS-JOINT (010RP2020) and JSPS KAKENHI Grant Number 20H03776.

**Data Availability Statement:** The data presented in this study are openly available in figshare [46]. Source codes to replicate the study are also freely available at GitHub [22].

**Acknowledgments:** Authors would like to thank Late Mr. Yusuke Yamada for helpful discussions in the initial stage of this study.

**Conflicts of Interest:** The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

## References


1. Carter, A.J.; Kraemer, O.; Zwick, M.; Mueller-Fahrnow, A.; Arrowsmith, C.H.; Edwards, A.M. Target 2035: Probing the Human Proteome. *Drug Discovery Today* **2019**, *24*, 2111–2115, doi:10.1016/j.drudis.2019.06.020.
2. Barrett, T.; Wilhite, S.E.; Ledoux, P.; Evangelista, C.; Kim, I.F.; Tomashevsky, M.; Marshall, K.A.; Phillippy, K.H.; Sherman, P.M.; Holko, M.; et al. NCBI GEO: Archive for Functional Genomics Data Sets—Update. *Nucleic Acids Research* **2013**, *41*, D991–D995, doi:10.1093/nar/gks1193.
3. Athar, A.; Füllgrabe, A.; George, N.; Iqbal, H.; Huerta, L.; Ali, A.; Snow, C.; Fonseca, N.A.; Petryszak, R.; Papatheodorou, I.; et al. ArrayExpress Update – from Bulk to Single-Cell Expression Data. *Nucleic Acids Research* **2019**, *47*, D711–D715, doi:10.1093/nar/gky964.
4. Kodama, Y.; Mashima, J.; Kosuge, T.; Ogasawara, O. DDBJ Update: The Genomic Expression Archive (GEA) for Functional Genomics Data. *Nucleic Acids Research* **2019**, *47*, D69–D73, doi:10.1093/nar/gky1002.
5. Bono, H.; Hirota, K. Meta-Analysis of Hypoxic Transcriptomes from Public Databases. *Biomedicines* **2020**, *8*, 10, doi:10.3390/biomedicines8010010.
6. Cheung, W.A.; Ouellette, B.F.; Wasserman, W.W. Inferring Novel Gene-Disease Associations Using Medical Subject Heading Over-Representation Profiles. *Genome Med* **2012**, 4, 75, doi:10.1186/gm376.
7. Fischer, I.; Barak, B. Molecular and Therapeutic Aspects of Hyperbaric Oxygen Therapy in Neurological Conditions. *Biomolecules* **2020**, 10, doi:10.3390/biom10091247.
8. AOE Available online: https://aoe.dbcls.jp/ (accessed on 27 February 2021).
9. Bono, H. All of Gene Expression (AOE): An Integrated Index for Public Gene Expression Databases. *PLOS ONE* **2020**, *15*, e0227076, doi:10.1371/journal.pone.0227076.
10. Kodama, Y.; Shumway, M.; Leinonen, R.; on behalf of the International Nucleotide Sequence Database Collaboration The Sequence Read Archive: Explosive Growth of Sequencing Data. *Nucleic Acids Research* **2012**, *40*, D54–D56, doi:10.1093/nar/gkr854.
11. Ono, Y.; Information on Hypoxic Conditions and Samples in the Dataset Used for the Meta-Analysis **2021**, doi:10.6084/M9.FIGSHARE.14141219.V1.
12. Download SRA Sequences from Entrez Search Results Available online: https://www.ncbi.nlm.nih.gov/sra/docs/sradownload/ (accessed on 27 February 2021).
13. Yyoshiaki/Ikra: RNAseq Pipeline Centered on Salmon Available online: https://github.com/yyoshiaki/ikra (accessed on 27 February 2021).
14. Patro, R.; Duggal, G.; Love, M.I.; Irizarry, R.A.; Kingsford, C. Salmon Provides Fast and Bias-Aware Quantification of Transcript Expression. *Nature Methods* **2017**, *14*, 417–419, doi:10.1038/nmeth.4197.
15. Ono, Y.; Quantitative Gene Expression Data of Hypoxia Experimental Data Sets **2021**, doi:10.6084/M9.FIGSHARE.14141252.V1.
16. Ono, Y.; Ratio of Gene Expression under Hypoxic and Normoxic Conditions(HN-Ratio) **2021**, doi:10.6084/M9.FIGSHARE.14141126.V1.
17. Soneson, C.; Love, M.I.; Robinson, M.D. Differential Analyses for RNA-Seq: Transcript-Level Estimates Improve Gene-Level Inferences. *F1000Res* **2016**, *4*, 1521, doi:10.12688/f1000research.7563.2.
18. Metascape Available online: https://metascape.org/gp/index.html#/main/step1 (accessed on 27 February 2021).
19. Zhou, Y.; Zhou, B.; Pache, L.; Chang, M.; Khodabakhshi, A.H.; Tanaseichuk, O.; Benner, C.; Chanda, S.K. Metascape Provides a Biologist-Oriented Resource for the Analysis of Systems-Level Datasets. *Nature Communications* **2019**, *10*, 1523, doi:10.1038/s41467-019-09234-6.
20. ChIP-Atlas Available online: https://chip-atlas.org/ (accessed on 27 February 2021).
21. Oki, S.; Ohta, T.; Shioi, G.; Hatanaka, H.; Ogasawara, O.; Okuda, Y.; Kawaji, H.; Nakaki, R.; Sese, J.; Meno, C. ChIP-Atlas: A Data-Mining Suite Powered by Full Integration of Public ChIP-Seq Data. *EMBO reports* **2018**, *19*, e46255, doi:10.15252/embr.201846255.
22. No85j/Hypoxia_code Available online: https://github.com/no85j/hypoxia_code (accessed on 2 March 2021).
23. Index of /Gene/DATA Available online: https://ftp.ncbi.nlm.nih.gov/gene/DATA/ (accessed on 27 February 2021).
24. Ono, Y.; Score Based on the Ratio of Gene Expression between Hypoxic and Normoxic Conditions（HN-Score） **2021**, doi:10.6084/M9.FIGSHARE.14141135.V1.
25. Ono, Y.; Genelist_Top 100 Human Genes Up-Regulated under Hypoxic Conditions **2021**, doi:10.6084/M9.FIGSHARE.14141015.V1.
26. Ono, Y.; Genelist_Top 100 Human Genes Down-Regulated under Hypoxic Conditions **2021**, doi:10.6084/M9.FIGSHARE.14140997.V1.
27. Ono, Y.; Number of Publications per Human Gene **2021**, doi:10.6084/M9.FIGSHARE.14140850.V1.
28. Ono, Y.; Simpson Similarity of Each Human Gene to HIF1A Based on Gene2pubmed **2021**, doi:10.6084/M9.FIGSHARE.14140928.V1.
29. Hirota, K.; Semenza, G.L. Regulation of Angiogenesis by Hypoxia-Inducible Factor 1. *Critical Reviews in Oncology/Hematology* **2006**, *59*, 15–26, doi:10.1016/j.critrevonc.2005.12.003.
30. Gordan, J.D.; Bertout, J.A.; Hu, C.-J.; Diehl, J.A.; Simon, M.C. HIF-2$\alpha$ Promotes Hypoxic Cell Proliferation by Enhancing c-Myc Transcriptional Activity. *Cancer Cell* **2007**, *11*, 335–347, doi:10.1016/j.ccr.2007.02.006.
31. Koshiji, M.; Kageyama, Y.; Pete, E.A.; Horikawa, I.; Barrett, J.C.; Huang, L.E. HIF-1$\alpha$ Induces Cell Cycle Arrest by Functionally Counteracting Myc. *The EMBO Journal* **2004**, *23*, 1949–1956, doi:10.1038/sj.emboj.7600196.
32. Wong, W.J.; Qiu, B.; Nakazawa, M.S.; Qing, G.; Simon, M.C. MYC Degradation under Low O2 Tension Promotes Survival by Evading Hypoxia-Induced Cell Death. *Molecular and Cellular Biology* **2013**, *33*, 3494–3504, doi:10.1128/MCB.00853-12.

33. Ono, H.; Ogasawara, O.; Okubo, K.; Bono, H. RefEx, a Reference Gene Expression Dataset as a Web Tool for the Functional Analysis of Genes. *Scientific Data* **2017**, *4*, 170105, doi:10.1038/sdata.2017.105.
34. Benita, Y.; Kikuchi, H.; Smith, A.D.; Zhang, M.Q.; Chung, D.C.; Xavier, R.J. An Integrative Genomics Approach Identifies Hypoxia Inducible Factor-1 (HIF-1)-Target Genes That Form the Core Response to Hypoxia. *Nucleic Acids Research* **2009**, *37*, 4587–4602, doi:10.1093/nar/gkp425.
35. Shoji, K.; Murayama, T.; Mimura, I.; Wada, T.; Kume, H.; Goto, A.; Ohse, T.; Tanaka, T.; Inagi, R.; van der Hoorn, F.A.; et al. Sperm-Associated Antigen 4, a Novel Hypoxia-Inducible Factor 1 Target, Regulates Cytokinesis, and Its Expression Correlates with the Prognosis of Renal Cell Carcinoma. *Am J Pathol* **2013**, *182*, 2191–2203, doi:10.1016/j.ajpath.2013.02.024.
36. Knaup, K.X.; Monti, J.; Hackenbeck, T.; Jobst-Schwan, T.; Klanke, B.; Schietke, R.E.; Wacker, I.; Behrens, J.; Amann, K.; Eckardt, K.-U.; et al. Hypoxia Regulates the Sperm Associated Antigen 4 (SPAG4) via HIF, Which Is Expressed in Renal Clear Cell Carcinoma and Promotes Migration and Invasion in Vitro. *Molecular Carcinogenesis* **2014**, *53*, 970–978, doi:https://doi.org/10.1002/mc.22065.
37. Strausberg, R.L.; Feingold, E.A.; Grouse, L.H.; Derge, J.G.; Klausner, R.D.; Collins, F.S.; Wagner, L.; Shenmen, C.M.; Schuler, G.D.; Altschul, S.F.; et al. Generation and Initial Analysis of More than 15,000 Full-Length Human and Mouse CDNA Sequences. *Proc Natl Acad Sci U S A* **2002**, *99*, 16899–16903, doi:10.1073/pnas.242603899.
38. Gerhard, D.S.; Wagner, L.; Feingold, E.A.; Shenmen, C.M.; Grouse, L.H.; Schuler, G.; Klein, S.L.; Old, S.; Rasooly, R.; Good, P.; et al. The Status, Quality, and Expansion of the NIH Full-Length CDNA Project: The Mammalian Gene Collection (MGC). *Genome Res* **2004**, *14*, 2121–2127, doi:10.1101/gr.2596504.
39. Kimura, K.; Wakamatsu, A.; Suzuki, Y.; Ota, T.; Nishikawa, T.; Yamashita, R.; Yamamoto, J.; Sekine, M.; Tsuritani, K.; Wakaguri, H.; et al. Diversification of Transcriptional Modulation: Large-Scale Identification and Characterization of Putative Alternative Promoters of Human Genes. *Genome Res* **2006**, *16*, 55–65, doi:10.1101/gr.4039406.
40. Labrecque, M.P.; Takhar, M.K.; Nason, R.; Santacruz, S.; Tam, K.J.; Massah, S.; Haegert, A.; Bell, R.H.; Altamirano-Dimas, M.; Collins, C.C.; et al. The Retinoblastoma Protein Regulates Hypoxia-Inducible Genetic Programs, Tumor Cell Invasiveness and Neuroendocrine Differentiation in Prostate Cancer Cells. *Oncotarget* **2016**, *7*, 24284–24302, doi:10.18632/oncotarget.8301.
41. Qi, J.; Nakayama, K.; Cardiff, R.D.; Borowsky, A.D.; Kaul, K.; Williams, R.; Krajewski, S.; Mercola, D.; Carpenter, P.M.; Bowtell, D.; et al. Siah2-Dependent Concerted Activity of HIF and FoxA2 Regulates Formation of Neuroendocrine Phenotype and Neuroendocrine Prostate Tumors. *Cancer Cell* **2010**, *18*, 23–38, doi:10.1016/j.ccr.2010.05.024.
42. Yu, H.; Rimbert, A.; Palmer, A.E.; Toyohara, T.; Xia, Y.; Xia, F.; Ferreira, L.M.R.; Chen, Z.; Chen, T.; Loaiza, N.; et al. GPR146 Deficiency Protects Against Hypercholesterolemia and Atherosclerosis. *Cell* **2019**, *179*, 1276-1288.e14, doi:10.1016/j.cell.2019.10.034.
43. Richards, J.P.; Yosten, G.L.C.; Kolar, G.R.; Jones, C.W.; Stephenson, A.H.; Ellsworth, M.L.; Sprague, R.S. Low O2-Induced ATP Release from Erythrocytes of Humans with Type 2 Diabetes Is Restored by Physiological Ratios of C-Peptide and Insulin. *Am J Physiol Regul Integr Comp Physiol* **2014**, *307*, R862-868, doi:10.1152/ajpregu.00206.2014.
44. Fang, S.; Zhang, L.; Guo, J.; Niu, Y.; Wu, Y.; Li, H.; Zhao, L.; Li, X.; Teng, X.; Sun, X.; et al. NONCODEV5: A Comprehensive Annotation Database for Long Non-Coding RNAs. *Nucleic Acids Research* **2018**, 46, D308–D314, doi:10.1093/nar/gkx1107.
45. Ma, L.; Cao, J.; Liu, L.; Du, Q.; Li, Z.; Zou, D.; Bajic, V.B.; Zhang, Z. LncBook: A Curated Knowledgebase of Human Long Non-Coding RNAs. *Nucleic Acids Research* **2019**, 47, D128–D134, doi:10.1093/nar/gky960.
46. Ono, Y.; Multiomic Meta-Analysis of Hypoxic Stress from Public Databases. **2021**, doi:10.6084/m9.figshare.c.5323769.v1.