Exploratory Meta-Analysis of Hypoxic Transcriptomes Using a Precise Transcript Reference Sequence Set

Yoko Ono[1] and Hidemasa Bono[1,2*]

[1] Laboratory of Genome Informatics, Graduate School of Integrated Sciences for Life, Hiroshima University, 3-10-23 Kagamiyama, Higashi-Hiroshima city, Hiroshima, 739-0046, Japan

[2] Laboratory of Bio-DX, Genome Editing Innovation Center, Hiroshima University, 3-10-23 Kagamiyama, Hi-gashi-Hiroshima city, Hiroshima, 739-0046, Japan

[*] Correspondence: bonohu@hiroshima-u.ac.jp; Tel.: +81-82-424-4013.

Running title: Exploratory Hypoxic Transcriptome Meta-Analysis

ABSTRACT

Gene expression studies are intrinsically biased with many being influenced by concomitant information such as the gene-disease association. However, this can be overcome using a data-driven analysis approach, which can provide insights without relying on ancillary information. This is exemplified by the FANTOM CAGE Associated Transcriptome (FANTOM-CAT) project, which provides a comprehensive meta-assembly of the human transcriptome using coding and non-coding genes. Various studies have shown that hypoxia has a significant effect on gene expression, with several of these reports suggesting that non-coding RNA (ncRNA) metabolism is downregulated in response to hypoxic stimuli. We aimed to interrogate the differential response of various transcripts to hypoxia by determining their hypoxia responsiveness score. We performed these evaluations using enrichment analysis, which revealed that several genes associated with ncRNA metabolism, especially those involved in ribosomal RNA processing, were significantly downregulated in response to hypoxia. Then, we used previously published information from the FANTOM-CAT project to create a novel reference file suitable for the meta-analysis of transcriptome sequencing data from both coding and ncRNAs and used this to evaluate the hypoxia responsiveness of target transcripts, while also evaluating the relationship between sense-antisense transcripts from the same locus. These evaluations revealed that these transcript pairs often exhibit similar regulatory patterns, but that this is largely reliant on their specific function. We believe that this study will facilitate the functional

38 annotation of various transcripts including ncRNAs, allowing for both sense and antisense and coding
39 and non-coding evaluations.
40
41 **INTRODUCTION**
42 Genomic research is inherently biased as it is heavily influenced by the known associations
43 between specific genes and their families and the disease of interest (Mihai et al. 2021). Meanwhile,
44 growing evidence from global molecular biology labs has facilitated the comprehensive functional
45 annotation of much of the genome. Therefore, further development requires a move away from those
46 familiar genes into the unknown using a data-driven approach. Our previous study revealed a clear
47 publication bias regarding gene expression variation in response to hypoxia and we identified a
48 specific group of genes that had not been investigated in previous reports (Ono and Bono 2021).
49
50 Oxygen is essential for the maintenance of vital functions, and cells have strict molecular
51 mechanisms designed to help them cope with hypoxia. The study of these responses was dramatically
52 advanced by the discovery of hypoxia-inducible factor-1 (HIF-1) in the 1990s (Semenza and Wang
53 1992; Wang and Semenza 1995; Wang et al. 1995). Under normoxic conditions, HIF is hydroxylated
54 by prolyl hydroxylase (PHD) and factor inhibiting HIF-1 (FIH-1) and then degraded by the ubiquitin-
55 proteasome system to inhibit transcriptional activation (Jaakkola et al. 2001; Mahon et al. 2001).
56 However, hypoxia reduces PHD and FIH-1 activity, allowing HIF to escape its hydroxylation and

57 promoting its interactions with the aryl hydrocarbon receptor nuclear translocator (ARNT) and
58 transcriptional co-factor, CREB, which in turn induces downstream gene expression (Ebert and Bunn
59 1998).

60 Our previous study revealed that genes related to the non-coding RNA (ncRNA) metabolic process
61 were responsive to hypoxia (Ono and Bono 2021).

62 Here, we evaluated the expression of genes related to ncRNA metabolism in response to hypoxia
63 despite a lack of functional data describing their specific effects. In addition, while there is a balance
64 in the data available for non-coding and coding RNAs we believe that some changes in the expression
65 of ncRNAs can only be revealed via data-driven analysis.

66 One tool facilitating these evaluations in the FANTOM5 project was designed to provide a reliable
67 5' human lncRNA data set for transcriptomic evaluations. This data was then combined with the data
68 from the FANTOM5 Cap Analysis of Gene Expression (CAGE) project to produce the FANTOM
69 CAGE-associated transcriptome (FANTOM-CAT) database. This dataset can be used to evaluate
70 patterns in ncRNA expression using data from these and various coding genes. FANTOM CAGE data
71 was used to identify approximately 28,000 lncRNA genes from the human genome with
72 approximately 20,000 of these transcripts likely to have some specific function (Hon et al. 2017).

73 Here we used FANTOM-CAT as a reference and comprehensively scored the hypoxic response of
74 various transcripts for both coding and non-coding genes. These evaluations revealed that the
75 antisense transcripts were affected by these conditions and showed that we were able to produce

overarching results for each chromosome. The data obtained in this study will help in the functional annotation of various transcripts, including ncRNAs, especially where there are fewer clues to their function.

RESULTS

Overview

Our data describe various analyses designed to evaluate hypoxia-responsive genes and transcripts based on two reference transcript sets (Fig. 1). This dataset and its construction were described in detail in our previous study (Ono and Bono 2021).

In the analyses using two reference transcripts, one describes the quantitative gene expression data annotated using GENCODE release 30 and focuses only on protein-coding genes, and the other describes the expression data for both protein-coding and non-coding transcripts using FANTOM-CAT.

Initial evaluations of the coding-only dataset revealed that hypoxic stimulation suppressed the ncRNA metabolic process, especially genes involved in rRNA processing. Given this, we went on to expand the scope of our evaluations to include both protein-coding and non-coding transcripts, and comprehensively evaluated the effects of hypoxia on both types of transcripts using a FANTOM-CAT data set as a reference.

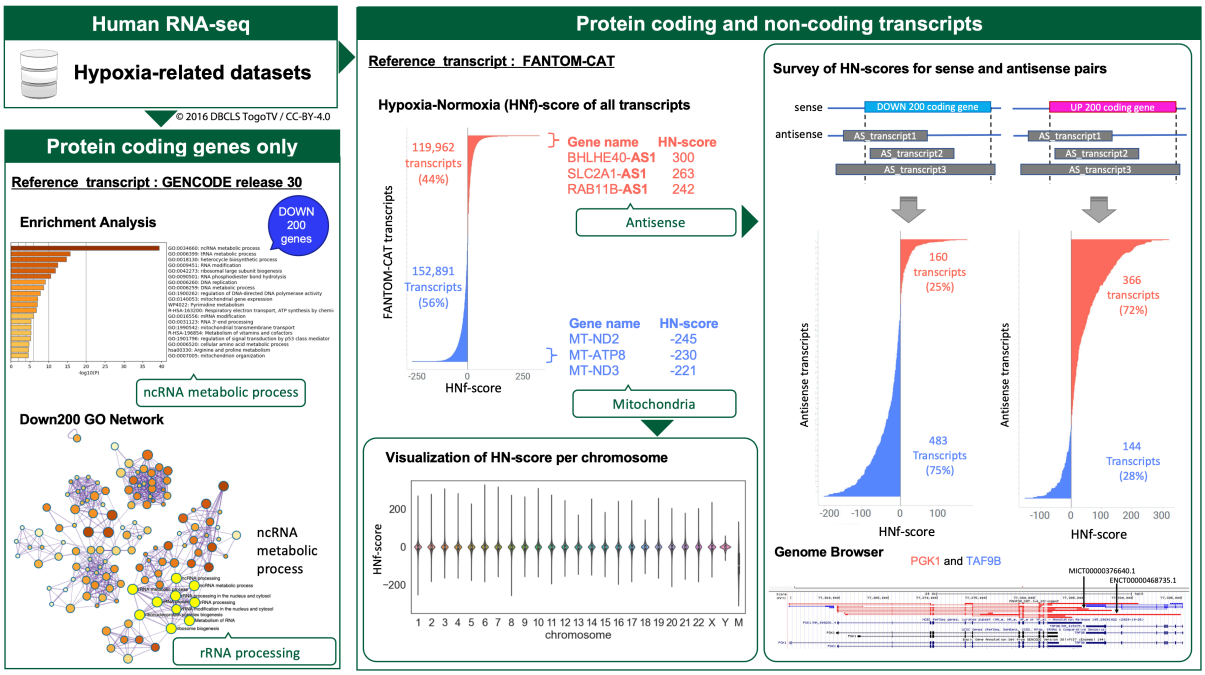


Figure 1. Schematic overview describing our exploratory study of hypoxia-responsive transcripts. Data source: human hypoxia RNA-seq data from the NCBI Gene Expression Omnibus database. Protein coding genes only: evaluated using GENCODE as the reference. Evaluation of the genes downregulated by hypoxia revealed that they were enriched for transcripts involved in rRNA processing. Protein coding and non-coding transcripts: Evaluated using FANTOM-CAT as reference. These evaluations revealed a decrease in mitochondrial DNA-derived transcripts while antisense analysis suggests a clear relationship between transcriptional regulation for both the sense and antisense genes (PGK1 and TAF9B).

Down-regulation of ncRNA metabolism-related genes by the hypoxic response

Enrichment analysis focusing only on protein-coding genes revealed that the top 200 genes identified by HNg-score, a measure of susceptibility to hypoxia, were enriched for hypoxia response-

related gene sets (Fig. 2B). However, enrichment analysis of the bottom 200 genes identified based on their HNg-score revealed suppression of expression of ncRNA metabolism- related genes (Fig. 2A). Detailed investigation showed that most of these genes were associated with rRNA processing.

The ncRNA metabolic process was then further subdivided into the following gene ontology terms: ncRNA processing, ribosome biogenesis, ribonucleoprotein complex biogenesis, rRNA processing, and rRNA metabolic process. This rRNA GO network analysis was then visualized using Upset plots which showed that the ncRNA metabolic process is heavily reliant on most of these rRNA metabolic genes.

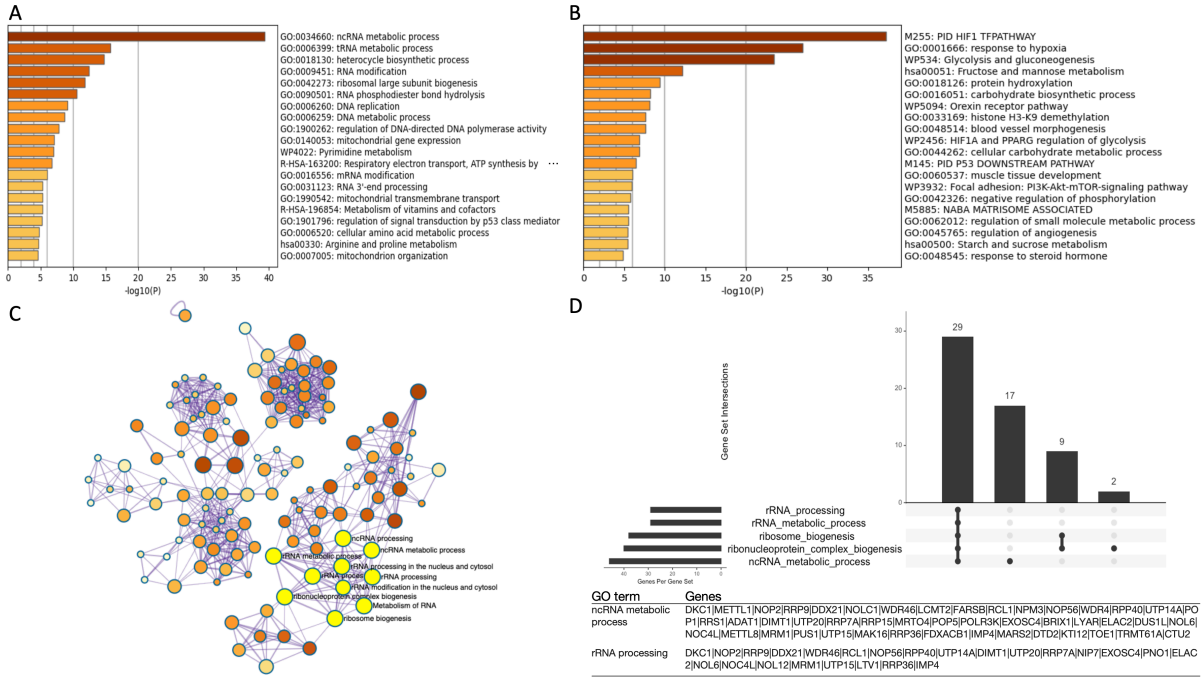


Figure 2. Hypoxic stimulation suppresses ncRNA metabolic process-related genes, and these genes are involved in rRNA processing. (A, B) Enrichment analysis for (A) the 200 DOWN regulated gene list and (B) the 200 UP regulated gene list. (C, D) GO network (C) and Upset plot (D) describing the relative changes in ncRNA metabolism associated genes in the downregulated gene list. In the GO

121 Network, the ncRNA metabolic process and rRNA processing focused on in this study are shown in

122 yellow.

123 **Survey of the number of papers referring FANTOM**

124 The number of articles in PubMed Central (PMC) that contained references to GENCODE,

125 FANTOM, and RNA-seq was calculated for each year, with the number of RNA-seq articles

126 exceeding 30,000 in 2021, while the number of articles referencing both FANTOM and RNA-seq was

127 0.5%, about one-tenth of the number referencing RNA-seq and GENCODE (Table 1).

128

129 **Table 1. Survey of the number of reports describing GENCODE, FANTOM and RNA-seq data**

Year	“RNA-seq”	“RNA-seq” &”GENCODE” (%)	“RNA-seq” &”FANTOM” (%)
2017	17,373	769 (4.4%)	123 (0.7%)
2018	20,400	996 (4.9%)	127 (0.6%)
2019	25,668	1,254 (4.9%)	130 (0.5%)
2020	33,607	1,593 (4.7%)	176 (0.5%)
2021	40,757	1,976 (4.8%)	210 (0.5%)
2022	9,420	406 (4.3%)	39 (0.4%)

130

131 **HNf-score evaluations of transcripts using FANTOM-CAT**

132 The HNf-score for each transcript was quantified using FANTOM-CAT as a reference. We then

133 went on to evaluate both the most increased and decreased transcripts using these values as a guide

134 (Fig. 3 A, B). A total of 25 transcripts for both the UP and DOWN data sets, as delimited by HNf-

135 score, are listed in Table 2 and a complete list of genes can be found within our archived figshare file

136 (https://doi.org/10.6084/m9.figshare.19679493.v1).

137 This investigation revealed the fact that there were several cases where hypoxia-responsive

138 transcripts with high HNf-scores were located within the antisense regions of genes with high HNg-

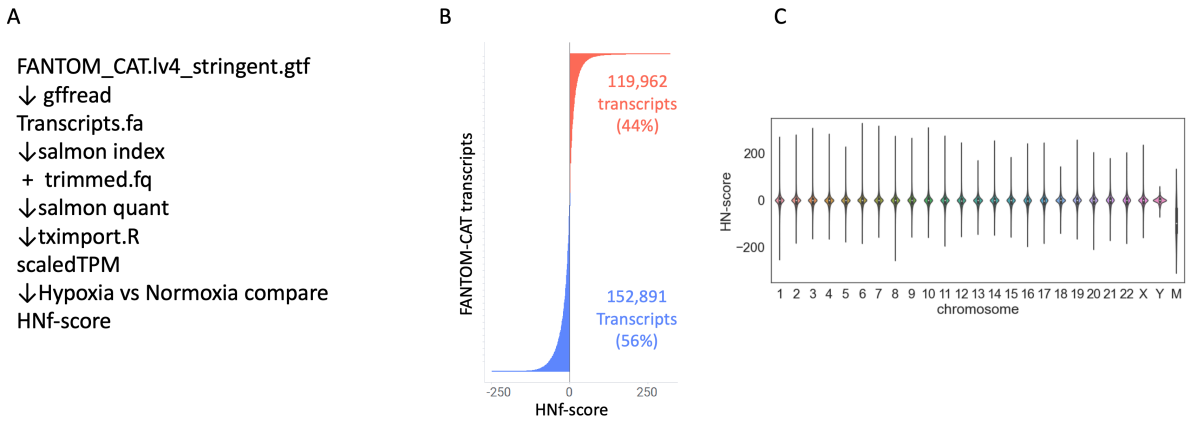
139 scores (*BHLHE40-AS1*, *SLC2A1-AS1*). In addition, these evaluations uncovered that the majority of

140 low HNf-score transcripts were associated with mitochondrial genes (*MT-ND2*, *MT-ATP8*). Hypoxic

141 stimuli were shown to affect mitochondria. Visualization of the HNf-score for each chromosome

142 showed that the median HNf-score was negative for mitochondria-derived transcripts (Fig. 3 C).

143



144

145 **Figure 3.** HNf-score evaluation of transcript using FANTOM-CAT**.** (A) Procedure for calculating

146 HNf-score using FANTOM-CAT as a reference. (B) HNf-score distribution of FANTOM-CAT

147 transcripts. (C) Violin plot of the HNf-score of each chromosome.

148

149 **Table 2. Top 25 UP and DOWN regulated transcripts identified using their HNf-score**

150

		Transcript ID	HNf-score	Chr	Gene Name	Gene Class	Antisense
Top 25 upregulated transcripts	1	ENST00000607600.1	321	chr6	RP1-261G23.7	short ncRNAs	AS_VEGFA
	2	ENST00000481651.1	309	chr7	RP11-61L23.2	pseudogenes	
	3	ENST00000307365.3	301	chr10	DDIT4	Protein-coding mRNAs	
	4	FTMT21000000269.1	300	chr3	BHLHE40-AS1	lncRNA, intergenic	AS_BHLHE40
	5	ENST00000368636.4	289	chr10	BNIP3	Protein-coding mRNAs	
	6	ENST00000335174.4	275	chr4	ANKRD37	Protein-coding mRNAs	
	7	ENST00000290573.2	272	chr2	HK2	Protein-coding mRNAs	
	8	FTMT24300004891.1	267	chr11	LDHA	Protein-coding mRNAs	
	9	HBMT00000242044.1	265	chr11	SBF2	Protein-coding mRNAs	AS_ADM
	10	FTMT24200000803.1	265	chr11	CATG00000004979.1	lncRNA, antisense	AS_LDHA
	11	ENST00000380629.2	265	chr8	BNIP3L	Protein-coding mRNAs	
	12	ENCT00000004975.1	263	chr1	SLC2A1-AS1	lncRNA, divergent	AS_SLC2A1
	13	ENST00000543445.1	257	chr11	LDHA	Protein-coding mRNAs	
	14	ENST00000378357.4	257	chr9	CA9	Protein-coding mRNAs	
	15	ENST00000453116.1	254	chr10	MXI1	Protein-coding mRNAs	
	16	ENST00000471240.1	253	chr10	DDIT4	Protein-coding mRNAs	
	17	ENST00000460806.1	251	chr3	BHLHE40	Protein-coding mRNAs	
	18	FTMT20200001505.1	250	chr1	SLC2A1	Protein-coding mRNAs	
	19	HBMT00000734702.1	250	chr19	CATG00000040757.1	lncRNA, divergent	AS_GPI
	20	FTMT21100041760.1	249	chr3	BHLHE40	Protein-coding mRNAs	
	21	ENST00000250457.3	246	chr14	EGLN3	Protein-coding mRNAs	
	22	HBMT00000734730.1	245	chr19	CATG00000040757.1	lncRNA, divergent	AS_GPI
	23	ENST00000426263.3	243	chr1	SLC2A1	Protein-coding mRNAs	
	24	HBMT00000727478.1	242	chr19	RAB11B-AS1	lncRNA, divergent	AS_ANGPTL4
	25	ENST00000534464.1	241	chr11	ADM	Protein-coding mRNAs	
Top 25 Downregulated transcripts	1	ENST00000320270.2	-248	chr8	RRS1	Protein-coding mRNAs	
	2	ENST00000368232.4	-246	chr1	GPATCH4	Protein-coding mRNAs	
	3	ENST00000361453.3	-245	chrM	MT-ND2	Protein-coding mRNAs	

	4	ENST00000361851.1	-230	chrM	MT-ATP8	Protein-coding mRNAs	
	5	ENST00000361227.2	-221	chrM	MT-ND3	Protein-coding mRNAs	
	6	ENST00000361390.2	-215	chrM	MT-ND1	Protein-coding mRNAs	
	7	ENCT00000264196.1	-200	chr20	IDH3B	Protein-coding mRNAs	AS_NOP56
	8	ENST00000362079.2	-194	chrM	MT-CO3	Protein-coding mRNAs	
	9	FTMT30000000001.1	-193	chrM	MT-ND5	Protein-coding mRNAs	AS_MT-ND6
	10	ENCT00000020267.1	-188	chr1	CATG00000042732.1	lncRNA, divergent	
	11	FTMT20300078215.1	-188	chr1	APOA1BP	Protein-coding mRNAs	AS_GPATCH4
	12	FTMT26100009680.1	-188	chr16	POLR3K	Protein-coding mRNAs	
	13	FTMT24100001922.1	-187	chr11	H2AFX	Protein-coding mRNAs	AS_HMBS
	14	MICT00000370386.1	-185	chrM	MT-TA	structural RNAs	AS_MT-ATP6 etc.
	15	ENST00000416718.2	-177	chr1	RP5-857K21.11	pseudogenes	
	16	HBMT00000533809.1	-176	chr16	ALG1	Protein-coding mRNAs	AS_EEF2KMT
	17	FTMT22300045607.1	-176	chr6	SRSF3	Protein-coding mRNAs	
	18	ENST00000585075.1	-175	chr17	RP11-649A18.12	lncRNA, divergent	AS_SLC25A19
	19	ENST00000295304.4	-175	chr2	CHAC2	Protein-coding mRNAs	
	20	MICT00000370388.1	-175	chrM	MT-TA	structural RNAs	AS_MT-ATP6 etc.
	21	ENST00000361899.2	-175	chrM	MT-ATP6	Protein-coding mRNAs	
	22	ENST00000458605.1	-174	chr22	RRP7B	pseudogenes	
	23	HBMT00000611042.1	-172	chr17	METTL23	Protein-coding mRNAs	
	24	ENST00000371538.3	-170	chr1	SELRC1	Protein-coding mRNAs	
	25	ENST00000293860.5	-169	chr16	POLR3K	Protein-coding mRNAs	

Investigating the HNf-score of antisense transcripts

The antisense transcripts for each of the UP200 and DOWN 200 hypoxia-responsive coding genes were classified into one of four categories based on their gene expression behavior (Fig. 4A, Table 3).

These evaluations revealed that approximately 75% of the antisense transcripts mirrored the behavior of their sense counterparts. However, there was still a small but significant number of sense-

157 antisense pairs presenting with opposite expression behaviors. This indicated that they were under
158 different transcriptional control.

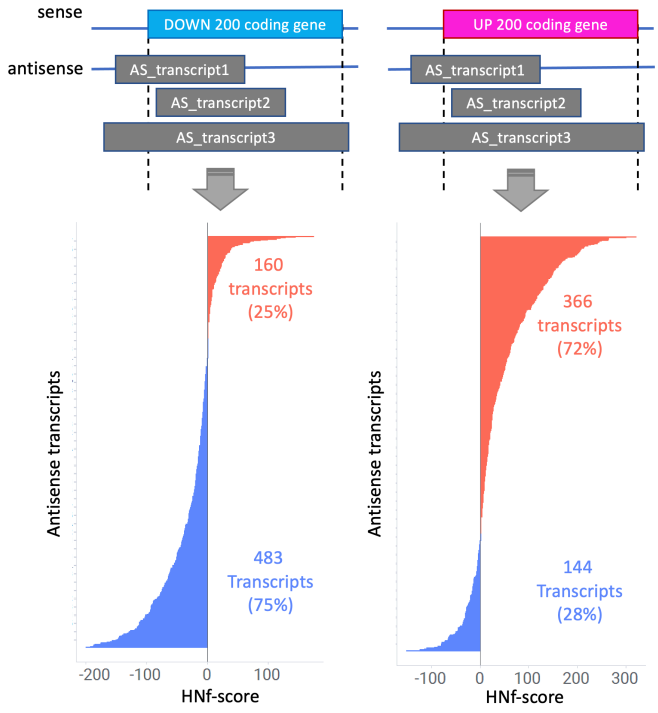
159 Further evaluation also revealed that the Top 10 antisense transcripts in the UP200 gene list were
160 dominated by both short and long ncRNAs (lncRNAs), while the bottom 10 transcripts from this
161 group were shown to be predominantly mRNA transcripts.

162 Interestingly, among the antisense transcripts from the DOWN200 gene set, phosphoglycerate
163 kinase 1 (*PGK1*), which was also included in the UP200 gene set, had the highest HNf-score.

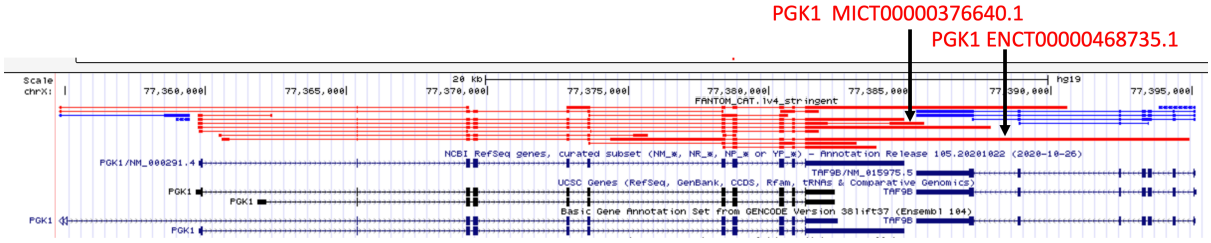
164 Given this, we homed in on *PGK1* and its antisense TATA-box binding protein associated factor 9b
165 (*TAF9B*), evaluating their chromosomal localization using UCSC Genome Browser and HN-score at
166 the gene level (Fig. 4B, C). This analysis revealed that while these two genes are not described as
167 having a sense-antisense relationship when evaluated using the UCSC or GENCODE datasets,
168 FANTOM-CAT-based analysis suggests that these two transcripts are actually in a sense-antisense
169 relationship. By using FANTOM-CAT as a reference, we were able to present the possibility that
170 these two genes affect the regulation of expression.

171

A



B



C

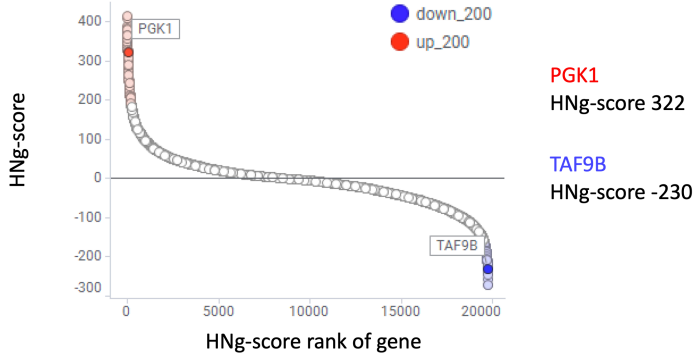


Figure 4. Evaluation of the HNf-score antisense transcripts in connection to their sense counterparts. (A) Distribution of the HNf-score for each of the antisense transcripts from each gene in the UP200 and DOWN200 gene lists, respectively. (B) Visualization of *PGK1* and *TAF9B* loci in the UCSC

178 genome browser. In FANTOM-CAT, *PGK1* transcripts are a sense-antisense pair for *TAF9B*, but

179 GENCODE does not show this relationship. (C) HNg-score for *PGK1* and *TAF9B*.

180

181 **Table 3. Top and bottom 10 antisense transcripts for the UP200 and DOWN200 HNg-score gene**

182 **sets as identified by their HNf-score**

			Transcript ID	HNf-score	Gene Name	Gene Class	Antisense
Antisense transcripts from the UP200 gene set	Top10	1	ENST00000607600.1	321	RP1-261G23.7	short ncRNAs	AS_VEGFA
		2	FTMT21000000269.1	300	BHLHE40-AS1	lncRNA, intergenic	AS_BHLHE40
		3	HBMT00000242044.1	265	SBF2	Protein-coding mRNAs	AS_ADM
		4	FTMT24200000803.1	265	CATG00000004979.1	lncRNA, antisense	AS_LDHA
		5	ENCT00000004975.1	263	SLC2A1-AS1	lncRNA, divergent	AS_SLC2A1
		6	HBMT00000734702.1	250	CATG00000040757.1	lncRNA, divergent	AS_GPI
		7	HBMT00000734730.1	245	CATG00000040757.1	lncRNA, divergent	AS_GPI
		8	HBMT00000727478.1	242	RAB11B-AS1	lncRNA, divergent	AS_ANGPTL4
		9	ENCT00000211059.1	241	RAB11B-AS1	lncRNA, divergent	AS_ANGPTL4
		10	HBMT00000734676.1	234	CATG00000040757.1	lncRNA, divergent	AS_GPI
	Bottom10	1	ENCT00000244163.1	-152	RP11-259N19.1	lncRNA, divergent	AS_HK2
		2	HBMT00000507967.1	-124	CATG00000025826.1	protein-coding mRNAs	AS_ISG20
		3	ENST00000309424.3	-115	CD3EAP	protein-coding mRNAs	AS_PPP1R13L
		4	FTMT21900007493.1	-111	SRFBP1	protein-coding mRNAs	AS_LOX
		5	ENST00000524270.1	-94	SPSB2	protein-coding mRNAs	AS_TPI1
		6	ENST00000357429.6	-87	C7orf50	protein-coding mRNAs	AS_GPR146
		7	ENST00000462901.1	-81	CGGBP1	protein-coding mRNAs	AS_ZNF654
		8	ENST00000439489.1	-81	DIABLO	protein-coding mRNAs	AS_B3GNT4
		9	MICT00000230070.1	-80	PPIL2	protein-coding mRNAs	AS_YPEL1
		10	ENST00000422285.2	-79	PDHA1	protein-coding mRNAs	AS_MAP3K15
Antisense transcripts from the DOWN200 gene set	Top10	1	MICT00000376640.1	175	PGK1	protein-coding mRNAs	AS_TAF9B
		2	ENCT00000468735.1	145	PGK1	protein-coding mRNAs	AS_TAF9B
		3	ENST00000229270.4	128	TPI1	protein-coding mRNAs	AS_SPSB2
		4	FTMT29100015042.1	120	PGK1	protein-coding mRNAs	AS_TAF9B
		5	ENST00000566326.1	114	MAP2K1	protein-coding mRNAs	AS_SNAPC5
		6	ENST00000329078.3	99	SPNS2	protein-coding mRNAs	AS_MYBBP1A
		7	ENST00000222256.4	87	RAB3A	protein-coding mRNAs	AS_MPV17L2
		8	ENCT00000059259.1	76	PDZD7	protein-coding mRNAs	AS_TWNK
		9	MICT00000139943.1	71	SPNS2	protein-coding mRNAs	AS_MYBBP1A
		10	FTMT25300024476.1	70	ZBTB25	protein-coding mRNAs	AS_MTHFD1
	Bottom10	1	ENCT00000264196.1	-200	IDH3B	protein-coding mRNAs	AS_NOP56

		2	FTMT30000000001.1	-193	MT-ND5	protein-coding mRNAs	AS_MT-ND6
		3	FTMT20300078215.1	-188	APOA1BP	protein-coding mRNAs	AS_GPATCH4
		4	FTMT24100001922.1	-187	H2AFX	protein-coding mRNAs	AS_HMBS
		5	MICT00000370386.1	-185	MT-TA	structural RNAs	AS_MT-ATP6
		6	MICT00000370386.1	-185	MT-TA	structural RNAs	AS_MT-ATP8
		7	MICT00000370386.1	-185	MT-TA	structural RNAs	AS_MT-CO3
		8	MICT00000370386.1	-185	MT-TA	structural RNAs	AS_MT-ND1
		9	MICT00000370386.1	-185	MT-TA	structural RNAs	AS_MT-ND2
		10	MICT00000370386.1	-185	MT-TA	structural RNAs	AS_MT-ND3

DISCUSSION

Human gene research is commonly biased towards known pathogenic genes and pathways that are fairly well established (Stoeger and Amaral 2022). However, given the development of novel data-driven tools, it is now possible to move beyond the targets identified in previous papers and into the realm of big data. Our previous studies have revealed a clear publication bias when evaluating gene expression changes in response to hypoxia and identified a novel set of largely uncharacterized hypoxia-responsive elements (Ono and Bono 2021). Despite this, the details of the "ncRNA metabolic process" related genes whose gene expression was suppressed in response to hypoxic conditions remain unclear.

Thus, we designed this study to further clarify the hypoxic response using a two-pronged approach. First, we completed a detailed investigation of the "ncRNA metabolic process" genes focused on only the coding genes in the dataset, and then we expanded this evaluation to any of the hypoxia-responsive genes annotated in the FANTOM-CAT database, which includes both coding and non-

197 coding elements (Fig. 1). All these data are available to everyone on figshare

198 (https://doi.org/10.6084/m9.figshare.c.5971218.v2). We believe that these data sets and this analysis

199 approach will help to further elucidate the mechanisms regulating the hypoxic response.

200 We previously reported that the expression of several genes associated with the "ncRNA metabolic

201 process" were downregulated by hypoxic stimuli. However, the specific metabolic processes remain

202 unknown. Given this, we started our evaluations by completing a gene ontology evaluation in an

203 effort to identify the likely functions of these target transcripts. Subsequent enrichment analysis of the

204 top and bottom 200 genes, identified using their HNg-score which was calculated using GENCODE,

205 revealed that the top 200 genes were commonly associated with hypoxia response-related transcripts,

206 while the bottom 200 genes were enriched for ncRNA metabolic processes (Fig. 2 A,B). Then GO

207 network analysis revealed a strong relationship between the ncRNA metabolic process and rRNA

208 processing genes (Fig. 2 C,D) (https://www.ebi.ac.uk/QuickGO/GTerm?id=GO: 0034660). This

209 suggests that amongst the most common downregulated transcripts most targeted or formed part of

210 the rRNA processing pathway.

211

212 rRNA is a critical part of the ribosome known to facilitate protein synthesis. rRNA has also been

213 linked to the degradation of cytoplasmic lncRNAs following their interaction with the ribosomes

214 (Carlevaro-Fita et al. 2016), and several other reports describe lncRNA mediated silencing of the

215 rRNA (Schmitz et al. 2010). All of which suggests an intimate relationship between these two sets of

216 RNA mediators. However, given our previous focus on only the coding genes, the effects of hypoxia
217 on the total transcript population, including the ncRNAs, remains unclear. Given this, we went on to
218 calculate the more comprehensive HNf-score, which evaluates all RNA transcripts including ncRNAs
219 via the application of a FANTOM-CAT reference.
220
221 The application of this dataset allows for a more comprehensive evaluation of changes in the gene
222 expression data from human experiments, expanding the evaluations to the ncRNA transcripts.
223 However, to the best of our knowledge, FANTOM-CAT is only published as a GTF file and is not
224 available in FASTA format, thus the underlying data is not publicly available. Therefore, we assumed
225 that this dataset is unlikely to be commonly applied as a reference point for RNA-seq. Therefore, we
226 investigated the number of articles referencing both "RNA-seq" and "FANTOM" in PMC.
227 The number of papers with both "RNA-seq" and "FANTOM" accounted for less than 1% of the total
228 that describes "RNA-seq" data even after 2017 when the FANTOM-CAT paper was first published
229 (Table 1). This data also indicated that there were only a few reports describing gene expression
230 quantification using FANTOM-CAT as a reference and that most studies continue to focus on more
231 well-established gene sets, potentially overlooking critical actors in the under-evaluated data (Ono and
232 Bono 2021). When we combined these observations, we decided that a comprehensive analysis,
233 including the ncRNAs, was likely worth the effort when discussing hypoxia. Given this, we went on
234 to complete a FANTOM-CAT-based evaluation of the hypoxia data described above.

235

236 To this end, we created a new FASTA file using the FANTOM-CAT genome annotation file in the
237 General Transfer Format (GTF) and then added the HNf-score to each of the FANTOM-CAT lv4
238 transcripts (Fig. 3 A,B). Subsequent analysis of this file then revealed that the bottom 25 genes in the
239 HNf-score dataset were predominantly mitochondrial DNA-derived transcripts. Visualization of this
240 patterning then revealed likely regulation of gene expression by chromosome (Fig. 3 C, Table 2), and
241 confirmed that these effects were most pronounced for mitochondrial transcripts. The downregulation
242 of mitochondrial genes such as *MT-ND2*, *MT-ATP8*, and *MT-ND* by hypoxic stimulation is already
243 well known (Arnaiz et al. 2021) validating our data. It has also been reported (Zhang et al. 2008) that
244 mitochondrial autophagy is commonly induced in response to hypoxia and that this process requires
245 hypoxia-dependent factor-1-dependent BNIP3 expression, which also supports our results. On the
246 contrary, the top 25 transcripts from the HNf-score dataset were dominated by antisense reads, such as
247 *BHLHE40-AS1* and *SLC2A1-AS1* (Table 2). This was an interesting observation, and its implications
248 were supported by the fact that the sense transcripts for both *BHLHE40* and *SLC2A1* are well-
249 established hypoxia response genes. In addition, RP1-261G23.7, which had the highest HNf-score, is
250 a short ncRNA localized to the antisense of vascular endothelial growth factor A (*VEGFA*), a well-
251 known hypoxia response gene. This short ncRNA has been reported to be functionally involved in the
252 transcriptional regulation of *VEGFA* during hypoxia (Nieminen et al. 2018). Based on these findings,

253 we thought that the transcripts from the antisense strands of known hypoxia-responsive genes were

254 likely worth further investigation.

255

256 Antisense transcripts are transcribed from the complementary strand of the DNA molecule and

257 usually overlap with other sense transcripts encoding both proteins and noncoding regulators. It is also

258 worth noting that approximately 30% of human (Ozsolak et al. 2010) and mouse (Katayama et al.

259 2005) transcripts originate from the antisense strand. Given this, we went on to select 200 hypoxia-

260 responsive genes from both the UP and DOWN regulated gene sets using their HNg-score. We then

261 evaluated the HNf-score of the antisense transcripts of these genes.

262 This revealed that 75% of the UP and DOWN regulated antisense transcripts presented with a similar

263 pattern of expression to their sense partner, while the other 25% of antisense transcripts were shown

264 to be regulated in opposition to their sense partner (Fig. 4A).

265 Further evaluation then revealed that the bottom 10 antisense transcripts in the DOWN200 gene set

266 were dominated by mitochondrial genes (Table 2, Fig. 3C). Interestingly, the Top 10 transcripts from

267 the DOWN200 gene set included *PGK1* and antisense *TAF9B* whose expression was suppressed in

268 response to hypoxia (Fig. 4C). This is of specific interest as *PGK1* catalyzes the reversible conversion

269 of 1,3 -diphosphoglycerate to 3-phosphoglycerate in the glycolytic system to generate ATP (Valentin

270 et al. 1998) and this gene is well known for its increased expression in response to HIF1A signaling

271 following hypoxic induction (Li et al. 1996). Meanwhile, TAF9B is one of the proteins associated

272 with the TFIID and plays an important role in transcriptional initiation (Tora 2002) (Frontini et al.
273 2005). TAF9B is also known to be upregulated in response to HIF1A knockdown and downregulated
274 under hypoxic conditions (Mathieu et al. 2011). Using FANTOM-CAT, the relationship between
275 *PGK1* and *TAF9B* was shown to be that of a sense-antisense pair. This relationship is not present in
276 either the UCSC gene or GENCODE annotations (Fig. 4B).
277 In addition, the regulation of genes by their antisense transcripts has been well studied (Pelechano
278 and Steinmetz 2013; Okada et al. 2008; Katayama et al. 2005), with most of these studies suggesting
279 that if the transcription is initiated from both the sense and antisense start sites at the same time,
280 factors involved in their transcription may collide and repress gene expression. These results suggest
281 that the 3' end of *PGK1* may affect the expression of *TAF9B* by elongation.
282 This data sets used in these evaluations were not derived from strand-specific RNA sequencing.
283 However, these findings suggest that strand-specific RNA-seq or knockdown experiments may
284 facilitate a better understanding of the hypoxic response and should be considered in the future.
285
286 Thus, this paper describes the use of FANTOM-CAT and other open data sources to facilitate the
287 exploratory analysis of existing data. In this setting, we used this approach to further characterize
288 hypoxia-responsive transcripts and overcome the native bias in genome research towards known
289 genes. Our study used the increased processing power of computers to enable the application of big
290 data analysis to drive discovery. We believe that this data-driven approach is increasingly necessary to

help advance genomic research. Our data suggest that FANTOM-CAT has not been widely utilized in research despite its obvious value. Thus, we designed this study to showcase the value of FANTOM-CAT derived transcript annotation as a means to uncover new targets for investigation and validation. Our evaluation of the hypoxia data uncovered a novel connection between *PGK1* and *TAF9B* which was not reported in any other evaluations despite their critical value in hypoxia.

Thus, we suggest that this approach can be applied to other questions to facilitate more nuanced, and less biased evaluations of large RNA-seq data sets and propose that FANTOM-CAT should be commonly evaluated as an indispensable tool in transcriptome evaluations.

METHODS

Evaluating the number of papers with descriptions of FANTOM

We used the following queries across our PMC web search (conducted on 2022.4.9)

RNA-Seq **:** “rna-seq”[MeSH Terms] OR “rna-seq”[All Fields] OR (“rna”[All Fields] AND “seq”[All Fields]) OR “rna seq”[All Fields].

RNA-Seq AND GENCODE : (“rna-seq”[MeSH Terms] OR “rna-seq”[All Fields] OR (“rna”[All Fields] AND “seq”[All Fields]) OR “rna seq”[All Fields]) AND “gencode”[All Fields]

RNA-Seq AND FANTOM : (“rna-seq”[MeSH Terms] OR “rna-seq”[All Fields] OR (“rna”[All Fields] AND “seq”[All Fields]) OR “rna seq”[All Fields]) AND “fantom”[All Fields] to identify any

310 relevant publications. Once identified we downloaded the relevant article information in MEDLINE

311 format and calculated the number of publications per year based on the Date of Publication.

312

313 **Public Gene Expression Data**

314 We then collected the RNA-seq data (https://doi.org/10.6084/m9.figshare.14141219.v1) from 495

315 pairs of human hypoxia-Normoxia samples evaluated in our previous report(Ono and Bono 2021).

316 However, since there were no FASTA-formatted files for this data recorded in the FANTOM-CAT

317 repository, we were forced to create a novel FASTA file based on the GTF file data using gffread

318 v0.12.1 and the following command:

319 > gffread FANTOM_CAT.lv4_stringent.only_lncRNA.gtf -g hg19.fa -w lv4.fa

320

321 **Quantifying Gene Expression**

322 We completed our RNA-seq data analysis as described in Fig. 3 A. We used salmon 0.14.0 for index

323 creation and quantification

324 (https://github.com/no85j/hypoxia_code/tree/master/salmon/salmon_v0.14.0), with some

325 modifications to Pitagora-cwl（https://github.com/pitagora-network/pitagora-

326 cwl/tree/master/tools/salmon）in order to match as closely as possible with the quantification method

327 used in our previous reports (Ono and Bono 2021; Bono and Hirota 2020). We then calculated

328 scaledTPM and used this as a representative value for gene expression(Soneson et al. 2016) and our

329 quantitative RNA-seq data are accessible at figshare

330 (https://doi.org/10.6084/m9.figshare.19679520.v1)

331

332 **HN-score**

333 The fold change of the [scaledTPM]+1 values for each of the 495 sample pairs was calculated and

334 differential expression was determined at a 1.5-fold cut-off. We then produced the Hypoxia-

335 Normoxia(HN)-score for this data by taking the number of samples demonstrating significant

336 upregulation and subtracting the number of samples with significant downregulation of their overall

337 expression profile. Here we added a second HN-score calculation method using different reference

338 transcripts to allow for a more nuanced evaluation. These scores were then referred to as HNg (coding

339 genes) and HNf (coding and non-coding gene), respectively. This first score was designed to evaluate

340 the coding gene and was calculated from the previously published data set

341 (https://doi.org/10.6084/m9.figshare.14141135.v1) and the previously described HN1.5 cut-off value.

342 This means that the HNf-score is effectively a descriptor of the HN-score for any FANTOM-CAT

343 transcripts. This HNf-score was calculated as described above except that FANTOM-CAT was used

344 as a reference. Transcript annotation was obtained from the FANTOM-CAT website

345 (https://fantom.gsc.riken.jp/cat/v1/#/genes) and the HNf-score data are accessible at figshare

346 (https://doi.org/10.6084/m9.figshare.19679508.v2)

347

Enrichment Analysis

We created UP200 and Down200 gene sets for the top and bottom 200 gene lists of the HNg-score respectively and performed enrichment analysis in Metascape (https://metascape.org/)(Zhou et al. 2019). We used the default parameters for Metascape settings.

Antisense transcript

We used the bedtools (v2.30.0) program to complete the following evaluations (Quinlan and Hall 2010), including the creation of a bed file in which all the regions of each gene were merged based on the transcript id of the UP200 and DOWN200 genes identified in the coding gene dataset. We then used this bed file to identify the transcripts in the reverse strand using the intersect command.

Visualization

We used TIBCO Spotfire Desktop version 11.5.0 (TIBCO Spotfire, Inc., Palo Alto, CA, USA) to produce all of our bar and scatter plots and we used the UpSetR (1.4.0) package from R (4.0.3) to create our upset plots. We then used the pandas (1.2.5), matplotlib (3.2.2), and seaborn (0.11.0) packages in Python (3.8.10) to complete our violin plotting and visualized the genome coordinates by adding "FANTOM5 summary Tracks" to the UCSC genome browser.

DATA ACCESS

Not applicable.

ACKNOWLEDGMENTS

The authors would like to thank Drs. Takeya Kasukawa and Masaki Suimye Morioka for their input and discussion.

COMPETING INTERESTS

Not applicable.

AUTHORS' 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.

REFERENCES

Arnaiz E, Miar A, Dias Junior AG, Prasad N, Schulze U, Waithe D, Nathan JA, Rehwinkel J, Harris AL. 2021. Hypoxia regulates endogenous double-stranded RNA production via reduced mitochondrial DNA transcription. *Front Oncol* **11**: 779739; doi:10.3389/fonc.2021.779739.

Bono H, Hirota K. 2020. Meta-analysis of hypoxic transcriptomes from public databases. *Biomedicines* **8**: 10; doi:10.3390/biomedicines8010010.

385 Carlevaro-Fita J, Rahim A, Guigó R, Vardy LA, Johnson R. 2016. Cytoplasmic long noncoding
386 RNAs are frequently bound to and degraded at ribosomes in human cells. *RNA* **22**: 867–882;
387 doi:10.1261/rna.053561.115.
388 Ebert BL, Bunn HF. 1998. Regulation of transcription by hypoxia requires a multiprotein complex
389 that includes hypoxia-inducible factor 1, an adjacent transcription factor, and p300/CREB
390 binding protein. *Mol Cell Biol* **18**: 4089–4096; doi: 10.1128/MCB.18.7.4089.
391 Frontini M, Soutoglou E, Argentini M, Bole-Feysot C, Jost B, Scheer E, Tora L. 2005. TAF9b
392 (formerly TAF9L) is a bona fide TAF that has unique and overlapping roles with TAF9. *Mol*
393 *Cell Biol* **25**: 4638–4649; doi:10.1128/MCB.25.11.4638-4649.2005.
394 Hon C-C, Ramilowski JA, Harshbarger J, Bertin N, Rackham OJL, Gough J, Denisenko E, Schmeier
395 S, Poulsen TM, Severin J, et al. 2017. An atlas of human long non-coding RNAs with accurate
396 5′ ends. *Nature* **543**: 199–204; doi:10.1038/nature21374.
397 Jaakkola P, Mole DR, Tian YM, Wilson MI, Gielbert J, Gaskell SJ, von Kriegsheim A, Hebestreit
398 HF, Mukherji M, Schofield CJ, et al. 2001. Targeting of HIF-alpha to the von Hippel-Lindau
399 ubiquitylation complex by O2-regulated prolyl hydroxylation. *Science* **292**: 468–472;
400 doi:10.1126/science.1059796.
401 Katayama S, Tomaru Y, Kasukawa T, Waki K, Nakanishi M, Nakamura M, Nishida H, Yap CC,
402 Suzuki M, Kawai J, et al. 2005. Antisense transcription in the mammalian transcriptome. *Science*
403 **309**: 1564–1566; doi:10.1126/science.1112009.

404 Li H, Ko HP, Whitlock JP. 1996. Induction of phosphoglycerate Kinase 1 gene expression by
405 hypoxia. *J Biol Chem* **271**: 21262–21267; doi:10.1074/jbc.271.35.21262.
406 Mahon PC, Hirota K, Semenza GL. 2001. FIH-1: a novel protein that interacts with HIF-1alpha and
407 VHL to mediate repression of HIF-1 transcriptional activity. *Genes Dev* **15**: 2675–2686; doi:
408 10.1101/gad.924501.
409 Mathieu J, Zhang Z, Zhou W, Wang AJ, Heddleston JM, Pinna CMA, Hubaud A, Stadler B, Choi M,
410 Bar M, et al. 2011. HIF Induces Human Embryonic Stem Cell Markers in Cancer Cells. *Cancer*
411 *Res* **71**: 4640–4652; doi: 10.1158/0008-5472.CAN-10-3320.
412 Mihai IS, Das D, Maršalkaite G, Henriksson J. 2021. Meta-analysis of gene popularity: less than half
413 of gene citations stem from gene regulatory networks. *Genes* **12**.doi: 10.3390/genes12020319.
414 Nieminen T, Scott TA, Lin F-M, Chen Z, Yla-Herttuala S, Morris KV. 2018. Long non-coding RNA
415 modulation of VEGF-A during hypoxia. *Non-Coding RNA* **4**: 34; doi:10.3390/ncrna4040034.
416 Okada Y, Tashiro C, Numata K, Watanabe K, Nakaoka H, Yamamoto N, Okubo K, Ikeda R, Saito R,
417 Kanai A, et al. 2008. Comparative expression analysis uncovers novel features of endogenous
418 antisense transcription. *Hum Mol Genet* **17**: 1631–1640; doi: 10.1093/hmg/ddn051.
419 Ono Y, Bono H. 2021. Multi-Omic meta-analysis of transcriptomes and the bibliome uncovers novel
420 hypoxia-inducible genes. *Biomedicines* **9**: 582; doi: 10.3390/biomedicines9050582.

421 Ozsolak F, Kapranov P, Foissac S, Kim SW, Fishilevich E, Monaghan AP, John B, Milos PM. 2010.
422 Comprehensive polyadenylation site maps in yeast and human reveal pervasive alternative
423 polyadenylation. *Cell* **143**: 1018–1029; doi: 10.1016/j.cell.2010.11.020.
424 Pelechano V, Steinmetz LM. 2013. Gene regulation by antisense transcription. *Nat Rev Genet* **14**:
425 880–893; doi: 10.1038/nrg3594.
426 Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features.
427 *Bioinformatics* **26**: 841–842; doi:10.1093/bioinformatics/btq033.
428 Schmitz K-M, Mayer C, Postepska A, Grummt I. 2010. Interaction of noncoding RNA with the rDNA
429 promoter mediates recruitment of DNMT3b and silencing of rRNA genes. *Genes Dev* **24**: 2264–
430 2269; doi:10.1101/gad.590910.
431 Semenza GL, Wang GL. 1992. A nuclear factor induced by hypoxia via de novo protein synthesis
432 binds to the human erythropoietin gene enhancer at a site required for transcriptional activation.
433 *Mol Cell Biol* **12**: 5447–5454; doi: 10.1128/mcb.12.12.5447-5454.1992.
434 Soneson C, Love MI, Robinson MD. 2016. Differential analyses for RNA-seq: transcript-level
435 estimates improve gene-level inferences. *F1000Res* **4**: 1521; doi:
436 10.12688/f1000research.7563.2.
437 Stoeger T, Amaral LAN. 2022. The characteristics of early-stage research into human genes are
438 substantially different from subsequent research. *PLoS Biol* **20:** e3001520; doi:
439 10.1371/journal.pbio.3001520.

440 Tora L. 2002. A unified nomenclature for TATA box binding protein (TBP)-associated factors
441 (TAFs) involved in RNA polymerase II transcription. *Genes Dev* **16**: 673–675;
442 doi:10.1101/gad.976402.
443 Valentin C, Birgens H, Craescu CT, Brødum-Nielsen K, Cohen-Solal M. 1998. A phosphoglycerate
444 kinase mutant (PGK Herlev; D285V) in a Danish patient with isolated chronic hemolytic anemia:
445 mechanism of mutation and structure-function relationships. *Hum Mutat* **12**: 280–287;
446 doi:10.1002/(SICI)1098-1004(1998)12:4<280::AID-HUMU10>3.0.CO;2-V.
447 Wang GL, Jiang BH, Rue EA, Semenza GL. 1995. Hypoxia-inducible factor 1 is a basic-helix-loop-
448 helix-PAS heterodimer regulated by cellular O2 tension. *Proc Natl Acad Sci U S A* **92**: 5510–
449 5514; doi: 10.1073/pnas.92.12.5510.
450 Wang GL, Semenza GL. 1995. Purification and characterization of hypoxia-inducible factor 1. *J Biol*
451 *Chem* **270**: 1230–1237; doi:10.1074/jbc.270.3.1230.
452 Zhang H, Bosch-Marce M, Shimoda LA, Tan YS, Baek JH, Wesley JB, Gonzalez FJ, Semenza GL.
453 2008. Mitochondrial autophagy is an HIF-1-dependent adaptive metabolic response to hypoxia. *J*
454 *Biol Chem* **283**: 10892–10903; doi: 10.1074/jbc.M800102200.
455 Zhou Y, Zhou B, Pache L, Chang M, Khodabakhshi AH, Tanaseichuk O, Benner C, Chanda SK.
456 2019. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets.
457 *Nat Commun* **10**: 1523; doi: 10.1038/s41467-019-09234-6.
458