オープンデータを活用した低酸素応答新規パスウェイの探索手法の開発

Development of exploration methods for novel hypoxia response pathways using open data

小野擁子

2022 年 7 月

要旨

生命科学研究には、よく知られている分野や遺伝子ほどよく研究される出版バイアスが存在する。例えば、疾患との関連性の有無などの情報が既知の遺伝子ほど研究対象になりやすい。一方で、昨今のマイクロアレイ技術やハイスループットシーケンシング技術の発展により、研究者の興味対象の如何に関わらず一度の実験で生体内の数万種類の遺伝子発現プロファイルを網羅的に取得できるようになった。これらの網羅的な遺伝子発現情報を含む論文を投稿する際には、その再現性の担保のため、オープンデータベースに登録することが論文誌掲載の条件とされているのが一般的である。National Center for Biotechnology Information (NCBI)の gene expression omnibus（GEO）や European Bioinformatics Institute (EBI)の ArrayExpress に代表されるオープンデータベースには、2022 年現在ヒト遺伝子発現データはおよそ 6 万 6 千を超えるデータシリーズが登録されており、これらの遺伝子発現プロファイルデータを元にしたデータドリブンな解析により、遺伝子の注目度の影響を受けずに知見を見出すことが期待できる。

低酸素刺激応答は好気性生物において生命維持に重要な機能を担う。低酸素刺激下では転写因子である Hypoxia inducible factor -1（HIF-1）が下流の遺伝子発現を制御し、低酸素に対応するための厳密な分子機構が備わっていることが明らかになっている。低酸素に対する生体反応の研究は 1990 年台の HIF-1 の発見により劇的に進み、そのメカニズムを明らかにしたウィリアム・ケリン教授、ピーター・ラトクリフ教授、グレッグ・セメンザ教授は 2019 年にノーベル生理学・医学賞を受賞している。ノーベル賞を受賞するほどの研究分野であることから公共遺伝子発現データベースへの多くの低酸素刺激関連データの登録が期待できることと、低酸素応答研究においても出版バイアスの存在は否定できないことから、低酸素を研究対象とし低酸素応答新規パスウェイの探索手法の確立を本研究の目的とした。

本研究は新規低酸素応答遺伝子の探索と、non-coding RNA (ncRNA)を含めた転写産物の低酸素応答の評価方法構築の大きく二つの研究にて構成される。

新規低酸素応答遺伝子の探索では、まずはオープンデータベースである GEO および低酸素に関わる RNA sequencing (RNA-Seq) データのメタデータをマニュアルキュレーションにより精査したのちに解析対象のデータセットを選定し、Sequence Read Archive (SRA)から配列データを取得した。低酸素, 通常酸素の条件のおよそ 500 ペアとなるサンプルの遺伝子発現プロファイルを元に、遺伝子ごとに Hypoxia-Normoxia（HN）-score を算出した。高い HN-score を示す上位 100 遺伝子を元に

エンリッチメント解析を行い、想定通りに低酸素応答関連遺伝子群が濃縮されていることを確認した。加えて、SRA で公開されている既報の Chromatin Immunoprecipitaton sequencing (ChIP-Seq)データを元にした解析ツールである ChIP-Atlas を用いて、高い HN-score を示す上位 100 遺伝子の発現制御には HIF-1 を構成する遺伝子である HIF1A が関与していることを示した。

次に NCBI が提供している gene2pubmed と呼ばれる、gene ID とその遺伝子が研究報告されている PubMed ID が収載されたデータセットを用いて新規低酸素応答遺伝子を探索した。その結果、G Protein-Coupled Receptor 146 をはじめとするいくつかの遺伝子が、データドリブンな解析では低酸素応答遺伝子と判定されるにも関わらず今まで低酸素応答遺伝子として注目されていなかったことを明らかにした。

上記の研究では、HN-score の下位 100 遺伝子を元にしたエンリッチメント解析にて ncRNA metabolic process に関わる遺伝子の発現抑制が示された。しかしながら低酸素条件下での ncRNA metabolic process の詳細メカニズムについては明らかになっていない。新規低酸素応答遺伝子の探索の研究と同様にデータドリブンに ncRNA を含めた転写産物の低酸素応答の評価をすることが必要と考えた。

エンリッチメント解析に用いられた Gene Ontology を元に精査をしたところ、これらの発現が抑制された遺伝子群は ncRNA metabolic process の中でも ribosomal RNA (rRNA) processing に関わることが明らかになった。rRNA はリボソームを構成してタンパク質を合成する機能を持つ。リボソームと細胞質 long non-coding RNA (lncRNA)との分解の関係性や、ncRNA が rRNA のサイレンシングに寄与することを示した報告もあることから、ncRNA を含めたリファレンスを活用して転写産物の低酸素応答性を網羅的に評価した。FANTOM CAGE Associated Transcriptome (FANTOM-CAT)は FANTOM5 Cap Analysis of Gene Expression (CAGE)のデータを用いた信頼性の高い 5’末端をもつヒト lncRNA も含まれている転写産物のカタログである。このカタログをリファレンスとして活用することにより、ヒトのコーディング遺伝子以外も含めた遺伝子発現情報を活用することが可能と考えた。

転写産物の網羅的な低酸素応答性評価の結果、低酸素応答によりミトコンドリア DNA 由来の転写産物の発現抑制が顕著であること、低酸素応答遺伝子群のアンチセンスに着目した解析では、大半の転写産物は低酸素応答遺伝子群と同様の発現制御パターンを示す一方で、センス-アンチセンスで異なる発現制御がなされている遺伝子群があることを明らかにした。

以上、本研究ではデータドリブンに新規低酸素応答遺伝子を見出し、ncRNA を含めた転写産物の低酸素応答の評価方法を提示した。これらの研究は、低酸素応答研究のみならず新規パスウェイのデータドリブンな探索手法の提示の点で貢献したと考える。

略語

ANKRD37: ankyrin repeat domain 37

AOE: All Of gene Expression

ARNT: aryl hydrocarbon receptor nuclear translocator

ARNTL: aryl hydrocarbon receptor nuclear translocator like

ASA: American Statistical Association, アメリカ統計協会

BHLHE40-AS1: BHLHE40 antisense RNA 1

CA9: carbonic anhydrase 9

CAGE: Cap Analysis of Gene Expression

ChIP-Seq: Chromatin Immunoprecipitaton sequencing

ChIP: Chromatin Immuno Precipitation, クロマチン免疫沈降法

CREB: cAMP response element-binding protein

EBI: European Bioinformatics Institute

EGLN3: egl-9 family hypoxia inducible factor 3

EPAS1: endothelial PAS domain protein 1

FANTOM-CAT: FANTOM CAGE Associated Transcriptome

FIH-1: factor inhibiting HIF-1

GEA: DDBJ Genomic Expression Archive

GEO: gene expression omnibus

GPCR: G タンパク質共役受容体

GPR146: G protein-coupled receptor 146

GTF: gene feature format

HDAC1: histone deacetylase 1

HIF-1: Hypoxia inducible factor -1

HIF-2α: hypoxia-inducible factor-2α

HIF1A: hypoxia inducible factor 1 subunit alpha

HN-score: Hypoxia-Normoxia-score

HNf-score: FANTOM-CAT をリファレンスとした転写物の HN-score

HNg-score: GENCODE をリファレンスとしたコーディング遺伝子の HN-score

INSD: International Nucleotide Sequence Database, 国際塩基配列データベース

INSDC:International Nucleotide Sequence Database Collaboration, 国際塩基配列データベース協力体制

lncRNA: long non-coding RNA

MIAME: The Minimal Information About a Microarray Experiment

MINSEQE: The Minimal Information about a high throughput nucleotide Sequencing Experiment

MT-ATP8: mitochondrially encoded ATP synthase 8

MT-ND: mitochondrially encoded NADH dehydrogenase

MT-ND2: mitochondrially encoded NADH dehydrogenase 2

MYC: MYC proto-oncogene, bHLH transcription factor

NCBI: National Center for Biotechnology Information

NCBI, National Center for Biotechnology Information ： 米・国立生物工学情報センター

ncRNA: non-coding RNA

NDRG1: N-myc downstream regulated 1

NLM, National Library of Medicine: 米国立医学図書館

PGK1: phosphoglycerate kinase 1

PHD: prolyl hydroxylase

PMC: Pubmed Central

PPP1R3G: sperm-associated antigen 4

PRSS53: transmembrane protein 74B

pVHL：von Hippel-Lindau tumor suppressor

RefEx：Reference Expression dataset

RNA-Seq：RNA sequencing

rRNA: ribosomal RNA

SAP30: Sin3A associated protein 30

SRA: Sequence Read Archive

TATA-box binding protein associated factor 9b：TAF9B

TFIID: transcription factor II D

UCSC, University of California Santa Cruse: カリフォルニア大学サンタクルーズ校

VEGF: vascular endothelial growth factor

理研: 国立研究開発法人理化学研究所

目次

表目次

図目次

1 序論

1.1 出版バイアス

ゲノム編集技術などの遺伝子工学技術の発展は、生命科学研究において重要な技術である。遺伝子工学技術を用いて、各研究者は仮説を生物実験的に検証可能となった。しかしながら、仮説創出のきっかけは文献調査や各研究者の知識の蓄積に依存することが多く、その傾向には出版バイアスが影響している。例えば、ヒトのタンパク質コーディング遺伝子 2 万個にアノテーションされたおよそ 60 万件の出版物の頻度パターンを評価したところ、約 1 万件の出版物が TP53 をコードする遺伝子に紐づいている一方で、600 件以上の遺伝子は出版物がなかったとの報告がある（Figure 1）(Carter et al. 2019)。

生命科学の知見の蓄積において過去の知見をもとにした仮説検証は重要である。いくつかの重要な研究をきっかけに研究予算が配分され、その研究分野の発展が強化される。しかしながら、条件に恵まれた研究分野がさらに条件に恵まれるマタイ効果の影響は、特定の研究分野の進展に好影響をもたらす反面、着目されにくい遺伝子や、研究分野を生み出し得る(Mihai et al. 2021)。新しい知見を得るには、知見の蓄積の偏りを考慮に入れつつ、オミクスデータなどの社会的なバイアスを排除したデータを中心とした解析手法をもとに研究対象を理解する必要があると考える。

出版バイアスとは、研究の解釈のしやすさが影響してポジティブな結果ほど公表されやすいという偏りのことをいう。例えば、統計学的有意差が示されたことをもとに生物学的な意味もあるとみなして論文を出版する場合も出版バイアスに影響する。

統計学的有意差の判断に用いられる P 値についてはアメリカ統計協会(ASA, American Statistical Association) が 2016 年に「統計的優位性と P 値に関する声明」を公表している(Wasserstein and Lazar 2016)。この声明の中で、以下のように原則をあげている(佐藤 2018)。

1. P 値はデータと特定の統計モデル（訳注: 仮説も統計モデルの要素のひとつ）が矛盾する程度をしめす指標のひとつである.
2. P 値は調べている仮説が正しい確率や，データが偶然のみでえられた確率を測るものではない.
3. 科学的な結論や，ビジネス，政策における決定は P 値がある値（訳注:

有意水準）を超えたかどうかにのみ基づくべきではない.

4. 適正な推測のためには，すべてを報告する透明性が必要である.

5. P 値や統計的有意性は，効果の大きさや結果の重要性を意味しない.

6. P 値は，それだけでは統計モデルや仮説に関するエビデンスの，よい指標とはならない.

チェリーピッキングは、研究結果や実験結果を全て報告するのではなく、仮説や主張をもっともよく支持する結果を意図的に提示することである(Morse 2010)。すでに出版されている知見は重要なものである一方、研究者それぞれが研究結果の生物学的意義を吟味することが生命科学研究分野において重要であると考える。

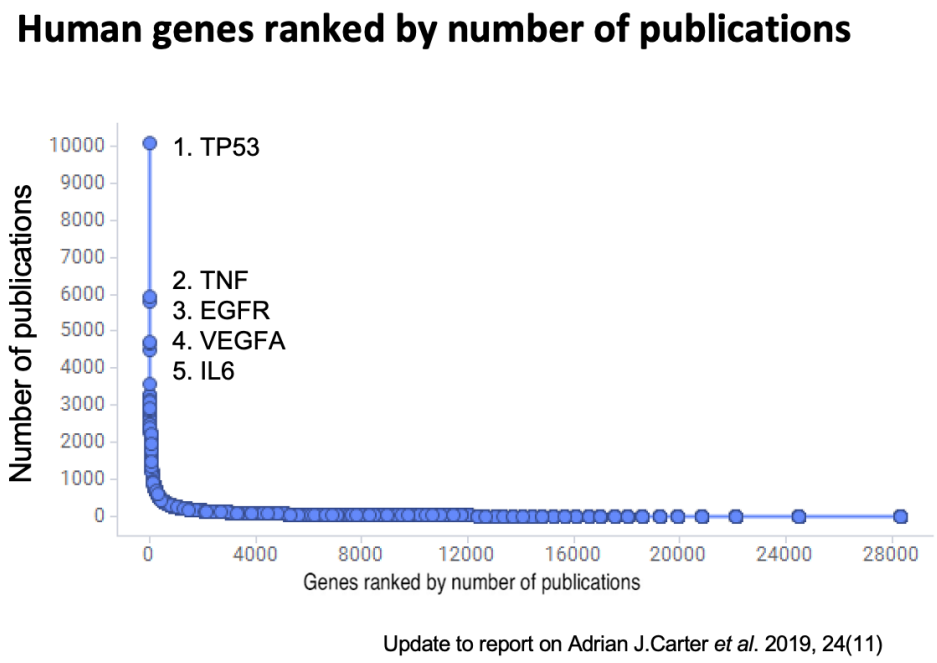


Figure 1 ヒト遺伝子の出版数順位

Adrian J.Carter et al. 2019, 24(11)の報告を 2021 年 1 月 4 日に NCBI よりダウンロードした gene2pubmed データをもとにアップデートした。もっともよく報告されている遺伝子は TP53 遺伝子であり、低酸素状態により遺伝子発現が亢進することが知られている VEGFA は上位4位だった。論文数が 1000 以上の遺伝子数は 532 遺伝子であった。

1.2 オープンデータ

オープンデータとは、誰もがアクセス、利用、共有することができるデータのことである。生命科学研究分野においては、文献情報であれば米国立医学図書館（NLM, National Library of Medicine）が運営している文献データベース PubMed (https://pubmed.ncbi.nlm.nih.gov/)や、PubMed の中でも全文がオープンアクセスとなっている論文を収録した Pubmed Central (PMC)(https://www.ncbi.nlm.nih.gov/pmc/)がオープンデータベースとして有名である。1.1 章「出版バイアス」で使用した、米・国立生物工学情報センター（NCBI, National Center for Biotechnology Information）が提供している gene2pubmed も利用制限のないオープンデータである（https://www.ncbi.nlm.nih.gov/home/about/policies/）。化合物情報であれば欧州バイオインフォマティクス研究所（EBI, European Bioinformatics Institute）の ChEMBL チームにより維持管理されている医薬品および開発化合物のデータベースである ChEMBL(https://www.ebi.ac.uk/chembl/), NCBI により維持管理されている低分子化合物の生物学的活性データベースである PubChem(https://pubchem.ncbi.nlm.nih.gov/)が該当する。また、塩基配列情報に関しては、国際塩基配列データベース協力体制（INSDC, International Nucleotide Sequence Database Collaboration）を構成している日本 DNA データバンク（DDBJ, DNA Data Bank of Japan）, NCBI, EBI が査定・収集したデータを、広く科学の発展のために再利用されるべき共有知的財産として国際塩基配列データベース(INSD, International Nucleotide Sequence Database）を構築している（https://www.ddbj.nig.ac.jp/about/insdc.html）。この INSD は，各データバンクが公開しているデータの全てを誰でも制限なしで利用できるという統一方針を共有している。

マイクロアレイ技術やハイスループットシーケンシング技術の発展により、一度の実験で生体内の数万種類の遺伝子発現プロファイルを網羅的に取得することができる。これらの網羅的な遺伝子発現情報を含む論文を投稿する際にはその研究結果の再現性の担保のため、NCBI の gene expression omnibus（GEO）や EBI の ArrayExpress、加えて、2018 年に始動した DDBJ Genomic Expression Archive (GEA) (Kodama et al. 2019)に代表される公共データベースに登録することが論文誌掲載の条件とされている場合が一般的である。The Minimal Information About a Microarray Experiment (MIAME) (https://www.fged.org/projects/miame)は 2001 年に出版されたマイクロアレイ

実験の結果を明確に解釈し、再現性の担保のために必要な最小限の情報を記述するためのガイドラインである(Brazma et al. 2001)。MIAME に続き、ハイスループットシーケンシング技術への拡張版として The Minimal Information about a high throughput nucleotide Sequencing Experiment (MINSEQE) (https://www.fged.org/projects/minseqe/) ガイドラインが 2008 年に提唱された(Rustici et al. 2021)。これらのガイドラインを遵守したデータについてはデータ横断的な統合解析の難易度を下げ、データのもつ価値の最大化が期待できる。しかしながら、INSD とは異なり、GEO と ArrayExpress の間で相互にデータ共有をしてはいない。かつて ArrayExpress は GEO からデータをインポートしていたが、2017 年にはそれを停止している（https://www.ebi.ac.uk/arrayexpress/help/GEO_data.html）。GEO のアーカイブデータは ArrayExpress からはまだ利用できるが、GEO にアーカイブされた新しいデータは ArrayExpress から利用できなくなった。つまりこれらのデータベースは独立してメンテナンスされており、ユーザーは個別に検索する必要がある。この課題に対し、公共遺伝子発現データベースの目次サイトである All Of gene Expression（AOE）(Bono 2020)は、2014 年に始動した(Figure 2)。AOE は GEO, ArrayExpress, GEA だけでなく、RNA-seq の配列情報としてのみ Sequence Read Archive (SRA)に登録されているトランスクリプトームデータもインデックスの対象としている。つまり AOE により遺伝子発現データベースを統合して把握するすることが可能となっている。AOE にて調査したところ、2022 年にはヒト遺伝子発現データはおよそ 6 万 6 千ものデータシリーズが登録されていた。しかしながら、多くの研究で網羅的な遺伝子発現プロファイリングは活用されているが、論文の主題となる遺伝子はそのうちの数遺伝子程度である場合は少なくない。

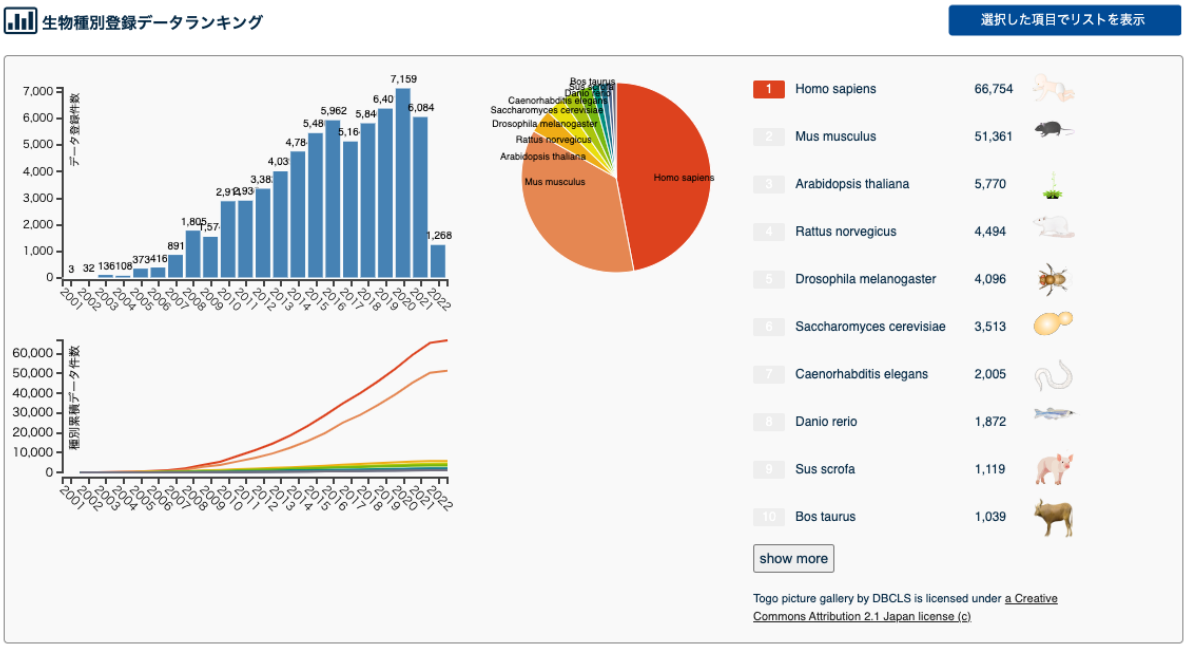


Figure 2 公共遺伝子発現データ登録数

公共遺伝子発現データベースの目次サイト AOE(https://aoe.dbcls.jp/)を用いて 2022 年 6 月 19 日に調査した。AOE により異なるデータベース由来の遺伝子発現データの統計情報が把握できる。

1.3 低酸素

好気性生物の呼吸に酸素は必須である。ヒトにおいては外傷性脳損傷、アルツハイマー病、脳卒中など様々な神経疾患では低酸素状態が発生し、それら症状の一因となることが知られている(Zhang and Le 2010; Ghajar 2000; Caplan and Hennerici 1998)。酸素は酸化的リン酸化や電子伝達系によって ATP を産生するために使われ、生命維持のエネルギー産生の中心的な働きをする。細胞や組織、生体が必要とする酸素量を供給されていない状態をハイポキシアという。ヒトをはじめとする高等生物の生体組織や細胞は、低酸素刺激による生命維持の危機を回避すべく Hypoxia inducible factor（HIF）に代表される低酸素応答のシステムを有している。HIF はαサブユニット(HIF-α)とβサブユニット(HIF-1β)のヘテロダイマーを形成して機能する転写因子である。また、HIF-αには HIF-1α, HIF2-α, HIF3-αがある。

低酸素に対する生体反応の研究は、1990 年代の HIF-1 の発見により劇的に進んだ(Semenza and Wang 1992; Wang and Semenza 1995; Wang et al. 1995)。この仕組みを明らかにしたウィリアム・ケリン教授、ピーター・ラトクリフ教授、グレッグ・セメンザ教授は 2019 年にノーベル生理学・医学賞を受賞している。正常酸素状態では、HIF-αの oxygen-dependent degradation ドメインに存在する二つのプロリン残基(P402, P564)は、酸素を補因子とした prolyl hydroxylase (PHD)や factor inhibiting HIF-1 (FIH-1)によって水酸化される(Figure 3)。その後、von Hippel-Lindau tumor suppressor (pVHL)などの E3 ユビキチンリガーゼが水酸化されたプロリンを認識し、ユビキチン-プロテアソーム系によって HIF-αが分解されて、転写活性化が抑制される(Jaakkola et al. 2001; Mahon et al. 2001)。一方で、低酸素状態では PHD や FIH-1 の活性が低下し、HIF は水酸化を免れて aryl hydrocarbon receptor nuclear translocator（ARNT）や転写コファクターである cAMP response element-binding protein (CREB)結合タンパク質と複合体を形成し、解糖系, 血管新生, 転移促進などに関連する下流の遺伝子発現を制御する(Ebert and Bunn 1998)。 例えば、血管新生に関係のある遺伝子として vascular endothelial growth factor (VEGF)があり、この遺伝子も HIF により発現が亢進されることが知られている (Semenza 2003)。

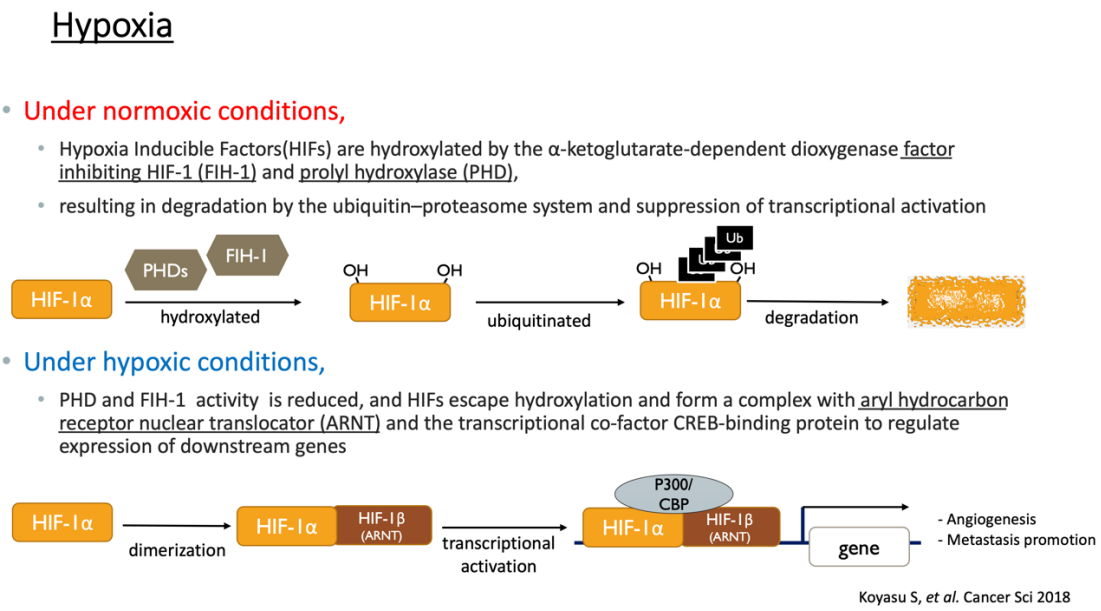


Figure 3 低酸素応答の模式図

通常酸素状態では HIF-1α は分解されるが、低酸素下では分解を免れて遺伝子発現を制御する。

1.4 本研究の目的

序論の 1.1 出版バイアスの章では、知見の蓄積の偏りを考慮に入れつつ、オミクスデータなどの社会的なバイアスを排除したデータを中心とした解析手法をもとに研究対象を理解する必要について述べた。また、1.2 オープンデータの章では、メタ解析が可能な公共データベースには多くの遺伝子発現データが存在していることについて述べた。また、1.3 低酸素の章では、低酸素に対する生体反応の研究はノーベル賞を受賞したほどの知名度があることと、低酸素状況下では、HIF が遺伝子発現の解糖系, 血管新生, 転移促進などに関連する下流の遺伝子発現を誘導することについて述べた。

低酸素に対する生体反応の研究分野は、1990 年代から研究分野として歴史があり、HIF という低酸素応答に関わる代表的な転写因子が存在し、オープンデータベースに多くの遺伝子発現データが存在するため、バイアスによって注目されなかった知見をデータドリブンにできるのではないかと考えた。以上より、私は本研究の目的を、「オープンデータを活用した低酸素応答新規パスウェイの探索手法の開発」とした。

2 方法

2.1 書誌情報解析

2.1.1 gene2pubmed

Hypoxia inducible factor 1 subunit alpha (HIF1A)は低酸素応答分野の代表的な遺伝子である(Supplemental figure 1)。NCBI が提供しているデータセットである gene2pubmed (Figure 4)を元にして、HIF1A と各ヒト遺伝子との類似度である Simpson 係数と論文数を算出した。gene2pubmed は、生物種ごとの Taxonomy ID、遺伝子の EntrezID とその遺伝子について言及されている論文の PubMed ID が記載されているデータセットである。Simpson 係数 S は、以下の式を使って算出した。

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

$|X \cap Y|$は HIF1A と任意の遺伝子のどちらにも該当する PubMed ID の数であり、$min(|X|,|Y|)$は HIF1A あるいは任意の遺伝子の PubMed ID の数のうち、小さい方の PubMed ID 数とする。Simpson 係数が 1 に近づくほど、HIF1A と類似であることを意味する。類似度を算出するに際には、Simpson 係数だけでなく、Jaccard 係数も検討した。Jaccard 係数より Simpson 係数の方が HIF1A と任意の遺伝子 X に該当する論文数の差の大きさによる影響を受けにくく、可視化に適するスコアと判断した（Supplemental figure 2)。

遺伝子ごとの出版物数の算出および Simpson 係数の算出には Python を使って算出した (https://github.com/no85j/hypoxia_code/tree/master/CodingGene)。上記の Simpson 係数の算出に必要な gene2Pubmed(ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/gene2pubmed.gz)と Gene Info (ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/gene_info.gz)は 2021 年 1 月 4 日にダウンロードした。

HIF1A 以外にも、endothelial PAS domain protein 1(EPAS1)や ARNT、aryl hydrocarbon receptor nuclear translocator like (ARNTL)といった低酸素との関連性が考えられる遺伝子との Simpson 類似度を任意の遺伝子 X に対し算出した(Supplemental figure 3)。HIF1A が低酸素研究分野の代表的な遺伝子と判断し、新規低酸素応答性遺伝子の探索には HIF1A との類似度を採用した。

Gene datasets provided by NCBI : Gene2pubmed

In [2]:
```
df = pd.read_csv('../data/gene2pubmed.gz', sep='\t')
df_hs =df.loc[df['#tax_id']==9606,] # human Tax_id
print('total row count : ' + str(len(df)))
print('human row count : ' + str(len(df_hs)))
df_hs.head()
```

```
total row count : 12339644
human row count : 1498491
```

Out[2]:

	#tax_id	GeneID	PubMed_ID
2707618	9606	1	2591067
2707619	9606	1	3458201
2707620	9606	1	3610142
2707621	9606	1	8889549
2707622	9606	1	12477932

Figure 4 gene2pubmed の内容

NCBI が提供している gene2pubmed を Jupyter Notebook を用いて内容を確認した。各生物種の遺伝子ごとに、その情報が記載された PubMed の ID が格納されたデータセットとなっている。全生物種で 1 千万レコード以上、ヒトの遺伝子に限定した場合はおよそ 150 万レコードの情報が含まれていた。

2.1.2 PMC の FANTOM, GENCODE 記述論文数

PubMed Central (PMC) 検索では、「RNA-Seq のみ」、「RNA-Seq と GENCODE」、「RNA-Seq と FANTOM」の記載がある論文数の調査のため、以下のクエリを使用した。(2022.4.9 実施)

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]

該当論文情報を MEDLINE format にてダウンロードし、DP (Date of Publication) をもとに各年あたりの出版数を算出した。出版数を算出するためのコマンドは検索結果ごとに以下の通りに実行した。

```
% grep -E 'DP  -'  PMC_result.txt | awk '{print $3}' | sort | uniq -c > output.txt
```

2.2 RNA-seq

2.2.1 データ調査

低酸素に関連する一連の実験データシリーズ一覧を GEO から以下の検索式 "hypoxia"[MeSH Terms] OR hypoxia[All Fields]) AND "Homo sapiens"[porgn] AND "gse"[Filter]で検索し、2020 年 8 月 17 日にダウンロードした。

Python パッケージ pysradb(v 0.11.1)を使って SRA からメタデータをダウンロードし、GEO 内の Series Matrix File(s) 、論文から該当 Series のメタデータを取得し、normoxia と hypoxia の比較可能なサンプルの対（HN-pair）になるデータをマニュアルキュレーションした。キュレーションの条件は、RNA-Seq のデータであることと同一実験セットの中に HN-pair となるサンプルがあることとした。キュレーション後、必要なデータセットを NCBI から prefetch コマンドにて収集した。

2.2.2 データ取得

低酸素に関連するデータを調査した後、対応する RUN データを DDBJ FTP サイト内の SRA からダウンロードした (ftp://ftp.ddbj.nig.ac.jp/)。ダウンロードした配列データは SRA フォーマットであったため、SRA Toolkit (https://www.ncbi.nlm.nih.gov/sra/docs/toolkitsoft/)の fasterq-dump プログラムを用いて、発現量定量用に FASTQ フォーマットファイルに変換した。シングルエンドリードとペアエンドリードの両方を解析対象とした。

2.2.3 リファレンス

コーディング遺伝子の解析には GENCODE release 30 をリファレンス配列として使用した。コーディング遺伝子および ncRNA の解析には FANTOM-CAT をリファレンス配列として使用した。

FANTOM CAGE-associated transcriptome (FANTOM-CAT)は FANTOM5 Cap Analysis of Gene Expression (CAGE)のデータを用いて作られた、信頼性の高い 5’ 完全ヒト lncRNA gene のオープンデータである。FANTOM-CAT のデータを使用するにあたり、国立研究開発法人理化学研究所 (理研)のリポジトリ
（https://fantom.gsc.riken.jp/5/suppl/Hon_et_al_2016/data/assembly/lv4_stringent/）にて提供されているデータを使用した。GTF ファイルをもとに gffread

v0.12.1 プログラムを使って以下のコマンドにて FASTA file を作成し、リファレンスとして使用した。

```
% gffread FANTOM_CAT.lv4_stringent.gtf -g hg19.fa -w lv4.fa
```

2.2.4 定量

コーディング遺伝子の発現定量には RNA-seq パイプラインである ikra (v1.2.3) (https://github.com/yyoshiaki/ikra)を使用した(Figure 5)。

ikra はリードの品質管理（Trim Galore version 0.6.3）や、GENCODE release 30 をリファレンスとした Salmon version 0.14.0(Patro et al. 2017)を使用した RNA-seq データの解析の自動化を可能にする。本研究では ikra のデフォルトの設定条件で解析した。salmon_tximport の設定では、ライブラリサイズをもとにスケーリングする scaledTPM を採用した。今回のデータセットでは、データ取得と品質管理の工程に 2.4 GHz 8 コア Intel Core i9, 64 GB メモリ搭載の MacBook Pro にて約 1 ヶ月を要した。また、RNA-seq データから処理した転写産物の定量結果は figshare にアップロードし公開されている（https://doi.org/10.6084/m9.figshare.14141252.v1）。

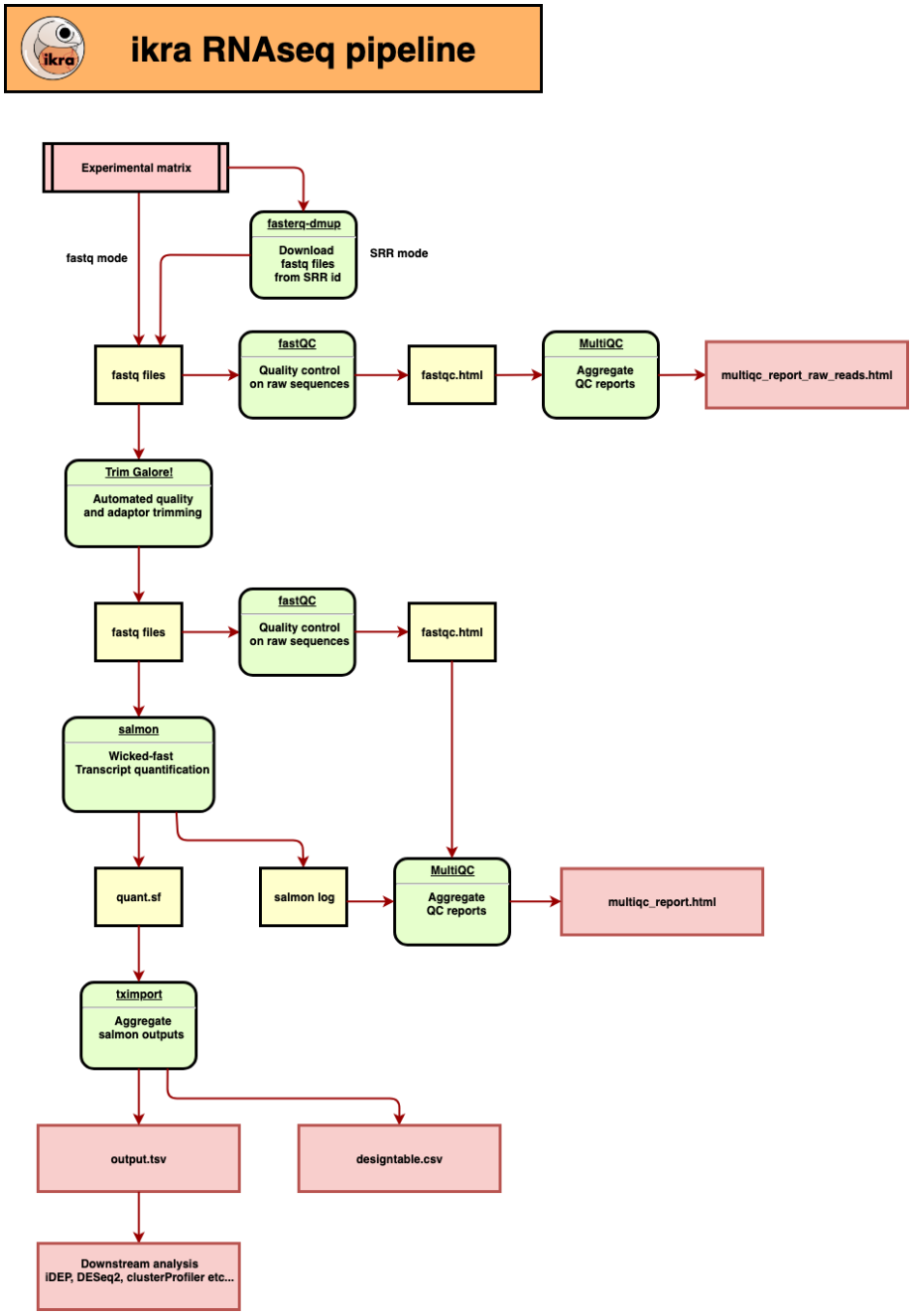


Figure 5 RNA-seq パイプライン ikra のワークフロー

図は https://github.com/yyoshiaki/ikra/blob/master/img/ikra_pipeline.png より引用した。FANTOM-CAT の転写物ごとの解析については Figure 14 A のフローの通りに計算した。Index 作成、定量の際には、コーディング遺伝子の発現定量の際に用いた ikra の定量条件を揃えるために、Pitagora-cwl（https://github.com/pitagora-network/pitagora-cwl/tree/master/tools/salmon）を一部改変し ikra の salmon のバージョンと同様に salmon 0.14.0 を使用した。

2.2.5 HN-score 計算

低酸素と通常酸素のペアごとに、遺伝子ごとに Hypoxia-normoxia ratio (HN-ratio), R を以下の式を用いて算出した。

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

T は各遺伝子や転写物の発現量 scaledTPM を示す。

次に、すべてのペアサンプルの HN-ratio の値を 1.5 倍の閾値によって up、down、unchanged の 3 群に分類した。閾値については 1.5 倍と 2 倍の条件について検討し、本研究には 1.5 倍を採用した。発現亢進したサンプル数から発現低下したサンプル数を引いた値を HN-score とした。これらのデータについては figshare からアクセス可能である (https://doi.org/10.6084/m9.figshare.14141135.v1)。

HN-ratio の算出の際には、分母が 0 になることを防ぐため、分母と分子にある一定数の値を加算することを事前に検討した。+1 あるいは、どの遺伝子定量結果よりも小さい値である＋1e-09 を加算した上で HN-score を算出し、その分布を確認した。どちらの結果も大きな違いはなかったが、発現量が非常に低い遺伝子のわずかな発現値の変動による影響を受けにくくなることを期待して、分母と分子に+1 を足して HN-ratio を算出した結果を採用した。また、HN-score を活用して、薬剤介入などの低酸素以外の実験介入が HN-pair どちらにも加わっている条件群の影響の確認や、当研究室の先行研究でオープンデータから収集済みであった実験サンプルと今回新規に収集したサンプルの HN-score の分布の比較を行った(Supplemental figure 4)。上位下位それぞれ HN-score の大きい順に 100 遺伝子を選抜した。これらの遺伝子をそれぞれ UP 100, DOWN100 gene list とした。

3.1 章「新規低酸素応答遺伝子の探索」にて記載している HN-score は GENCODE v30 をリファレンスとしたコーディング遺伝子に限定した HN-score であり、この HN-score は 3.2 章「ncRNA を含めた転写産物の低酸素応答の評価方法の構築」においては区別のため HNg-score と表記する。

3.2 章「ncRNA を含めた転写産物の低酸素応答の評価方法の構築」では、リファレンス転写物の違いのある二種類の HN-score を使用した。

HNg-score(GENCODE をリファレンスとしたコーディング遺伝子の HN-score) : 既報のデータセットの中から、低酸素により 1.5 倍の発現変動のあった

HN1.5 の値を HNg-score として使用した。このデータは figshare(https://doi.org/10.6084/m9.figshare.14141135.v1)にて公開した。

HNf-score (FANTOM-CAT をリファレンスとした転写物の HN-score)：FANTOM-CAT をリファレンスにした以外は HNg-score の算出方法に従って HNf-score を算出した。各転写物のアノテーション情報は https://fantom.gsc.riken.jp/cat/v1/#/genes から取得した。

2.3 エンリッチメント解析

エンリッチメント解析は、発現が変化した遺伝子群の機能の特徴を把握するための手法である。今回は、HN-score の高かった、あるいは低かった遺伝子群の特徴を把握するために解析した。HN-score エンリッチメント解析には Metascape (https://metascape.org) (Zhou et al. 2019) と ChIP-Atlas (https://chip-atlas.org/)(Oki et al. 2018)を使用した。Metascape では、対象の遺伝子群のリストをもとに、特定の生物学的プロセス、タンパク質の局在などの特徴によって定義された数千の遺伝子セットの情報をもとに解析することで、生物学的な洞察をもたらすことが期待できる。Metascape はデフォルト条件にて使用した。以下に述べる ChIP-Atlas のエンリッチメント解析では、解析対象の遺伝子群のリストをもとに、それらの遺伝子のゲノム座に結合する転写因子などのタンパク質を予測することが可能となる。ChIP-Atlas での設定では“Select data set to be compared”の項目を“Refseq coding genes (excluding user data)”とし、それ以外はデフォルトとした。

2.4 ChIP-Atlas 解析

クロマチン免疫沈降法（ChIP, Chromatin Immuno Precipitation）は、DNA 結合タンパク質に結合した DNA 配列を解析する技術である。ハイスループットシーケンサーを用いて配列解読する方法を ChIP-seq 法と呼ぶ。DNA 結合タンパク質にはヒストンや転写因子などが挙げられ、これらを認識する抗体を用い免疫沈降で回収した DNA 配列を明らかにすることで、ゲノム上の結合領域を把握することができる。ChIP-seq 解析では MACS2(Model-based Analysis for ChIP-Seq 2)プログラムを使って結合サイトを推定する。

UP100 gene list や DOWN 100 gene list の評価のために、SRA にて公開され

ている ChIP-seq データを収集、キュレーションしたデータを対象に解析できる Web ツールである ChIP-Atlas を活用した。各低酸素関連転写因子と、制御を受ける各遺伝子の転写開始点付近の DNA 配列との結合強度（MACS2 スコア）情報を取得した。ChIP-Atlas の“Target Genes”ツールを用いて、すべての遺伝子の平均 MACS2 スコアを取得した。対象抗原を HIF1A, hypoxia-inducible factor-2α（HIF-2α）としても知られる EPAS1 および ARNT として、転写開始点からの距離を意味する Distance from TSS パラメータを±5k と設定した。各遺伝子の HN-score と MACS2 スコアは、遺伝子名をキーにして結合した。

2.5 アンチセンス鎖に着目した解析

アンチセンス転写産物は、タンパク質コード遺伝子または非タンパク質コード遺伝子のセンス転写産物の反対側の鎖から転写されるものを意味する。今回の解析ではコーディング遺伝子の HN-score である HNg-score の高かった、あるいは低かった 200 遺伝子に対応するアンチセンス転写物に着目した（Figure 6)。アンチセンス転写物のリストを取得するために、Bedtools (v2.30.0)のプログラムを使い以下の処理を行った。HNg-score を元に選抜した UP 200 genes, DOWN 200 genes それぞれの transcript id をもとに、該当遺伝子ごとの全領域を merge した bed ファイルを作成した。この bed ファイルをもとに、Bedtools の intersect コマンドを使って、センス鎖の遺伝子と同じ配列領域がアンチセンス鎖に重複している転写物をリストアップし、それぞれの HNf-score と統合し、表にまとめた。

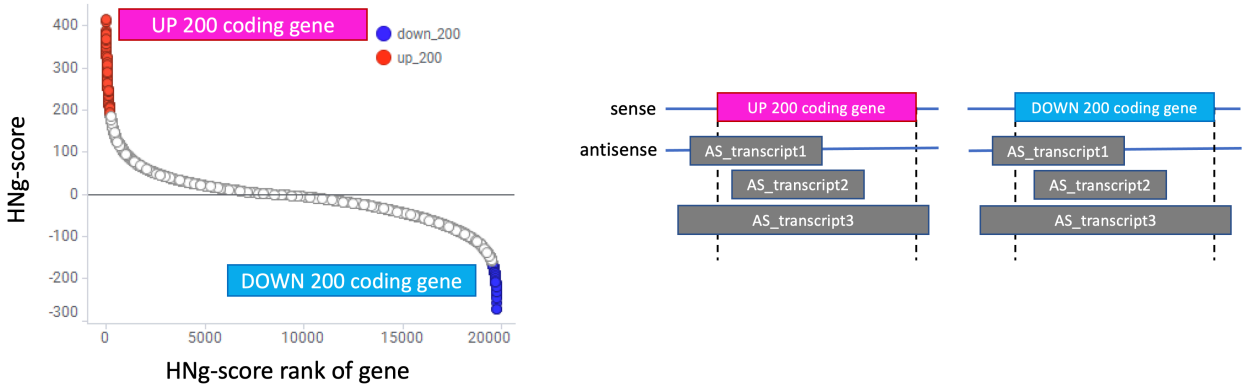


Figure 6 アンチセンス転写物リストの取得

HNg-score が高いあるいは低い値を示した UP 200 コーディング遺伝子, DOWN 200 コーディング遺伝子のアンチセンス鎖に位置する転写物をアンチセンス転写物と定義してリストアップした。

2.6 可視化

棒グラフや散布図には TIBCO Spotfire Desktop version 11.0.0 および 11.5.0 (TIBCO Spotfire, Inc., Palo Alto, CA, USA)を使用した。Spotfire はビジネス・インテリジェンス（BI）ツールであり、データのサブ集団を取り出した詳細解析であるドリルダウン解析が可能なことから探索的な可視化に適している。

集合の関係を示す代表的な可視化方法はベン図であるが、複数の集団の関係性を示すのには Upset plot が適している。エンリッチメント解析に使用した Gene Ontology ごとの該当遺伝子群の関係性を可視化をする際に Upset plot を使用した。Upset plotg は R (version 4.0.3)の UpSetR (1.4.0)パッケージを使用し作成した。

Violin plot とはデータの分布の密度を可視化する際に用いられる方法である。Violin plot の作図には Python(3.8.10)の pandas(1.2.5), matplotlib(3.2.2), seaborn(0.11.0)パッケージを使用した。

ゲノムブラウザとは、ゲノム上の位置情報に基づいて様々な情報を閲覧したりダウンロードしたりできるブラウザのことである。ゲノム座標の可視化にはカリフォルニア大学サンタクルーズ校（UCSC, University of California Santa Cruse）が維持管理している UCSC genome browser（https://genome-asia.ucsc.edu/index.html）を活用した。各種ゲノムのアノテーションのことを Track と呼ぶ。Track のアノテーション情報を含めこれらの情報はオープンデータであり誰でも無料で使用することができる。本解析では、この UCSC genome browser に“FANTOM5 summary Tracks”を追加して可視化した (Figure 7)。FANTOM-CAT のリファレンス配列を作成する際に使用したゲノムリファレンス hg19 とバージョンを揃える目的で、可視化に用いたリファレンスゲノムのバージョン（UCSC genome browser 上の表記は Human Assembly）も GRCh37/hg19 を使用した。リファレンスゲノムの違いによる影響がないことを確認するため、GRCh38/hg38 のリファレンスゲノムのバージョンを用いて、PGK1 と TAF9B のゲノム座位を可視化し、二つの遺伝子がセンス-アンチセンスの位置にあることを確認した。

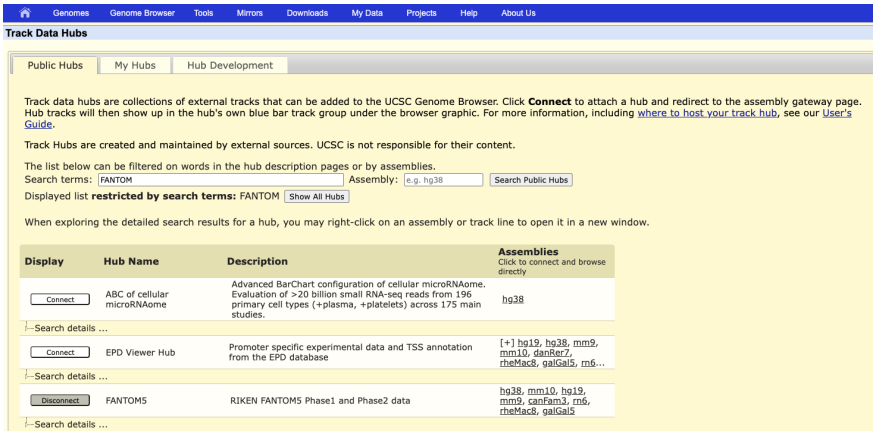


Figure 7 UCSC genome browser での FANTOM5 のデータの追加設定

Track Data Hubs にて Search terms にて”FANTOM5”を検索して Track を追加した。

3 結果

3.1 新規低酸素応答遺伝子の探索

本 3.1 章では低酸素刺激応答に着目し、オープンデータベースを元にメタ解析を活用し、出版バイアスにより見えにくくなっている知見を明らかにすることを目的とした。新規低酸素応答遺伝子の探索は大きく 3 つの工程に分けられる (Figure 8)。Step 1 では公開遺伝子発現データから低酸素刺激に関係する RNA-seq データを取得した。GEO よりメタデータを取得し、マニュアルキュレーション後に該当データセットの遺伝子発現の定量を経て低酸素刺激に関係する UP 100 gene list, DOWN 100 gene list を得た。Step 2 では UP 100 gene list, DOWN 100 gene list のバリデーションを行った。最後に Step 3 で gene2pubmed を用いた HIF1A の Simpson 類似度を低酸素応答研究での着目度の代替指標として用い、新規低酸素刺激応答遺伝子を探索した。

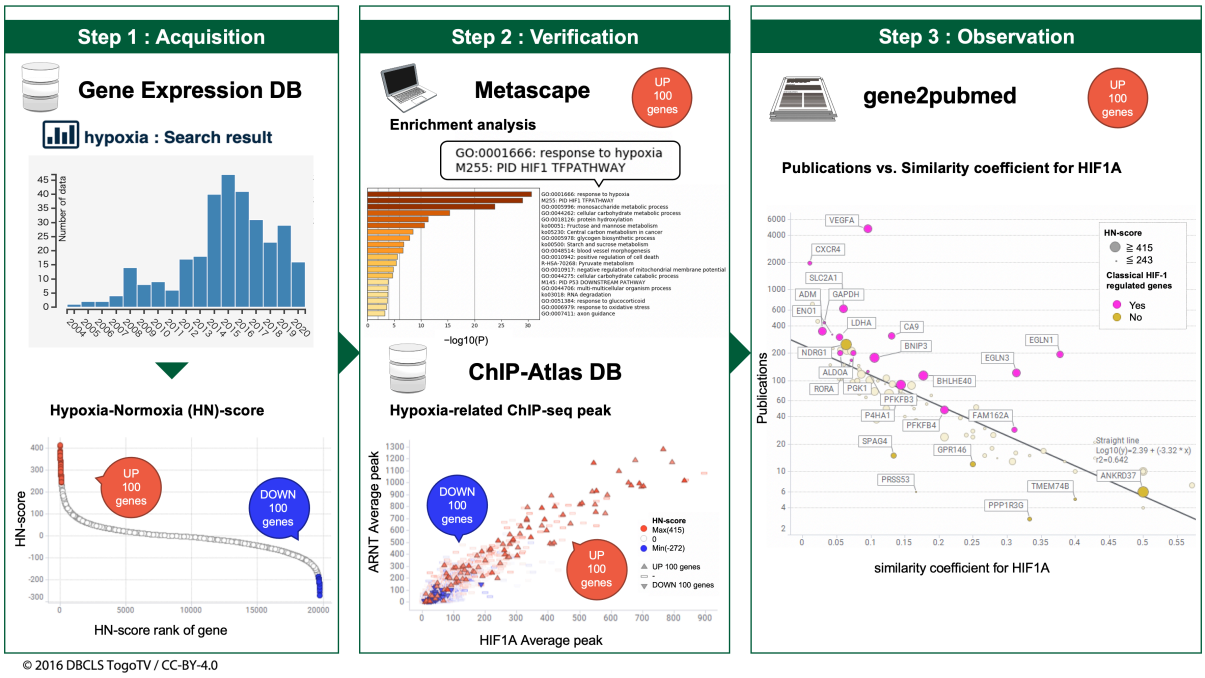


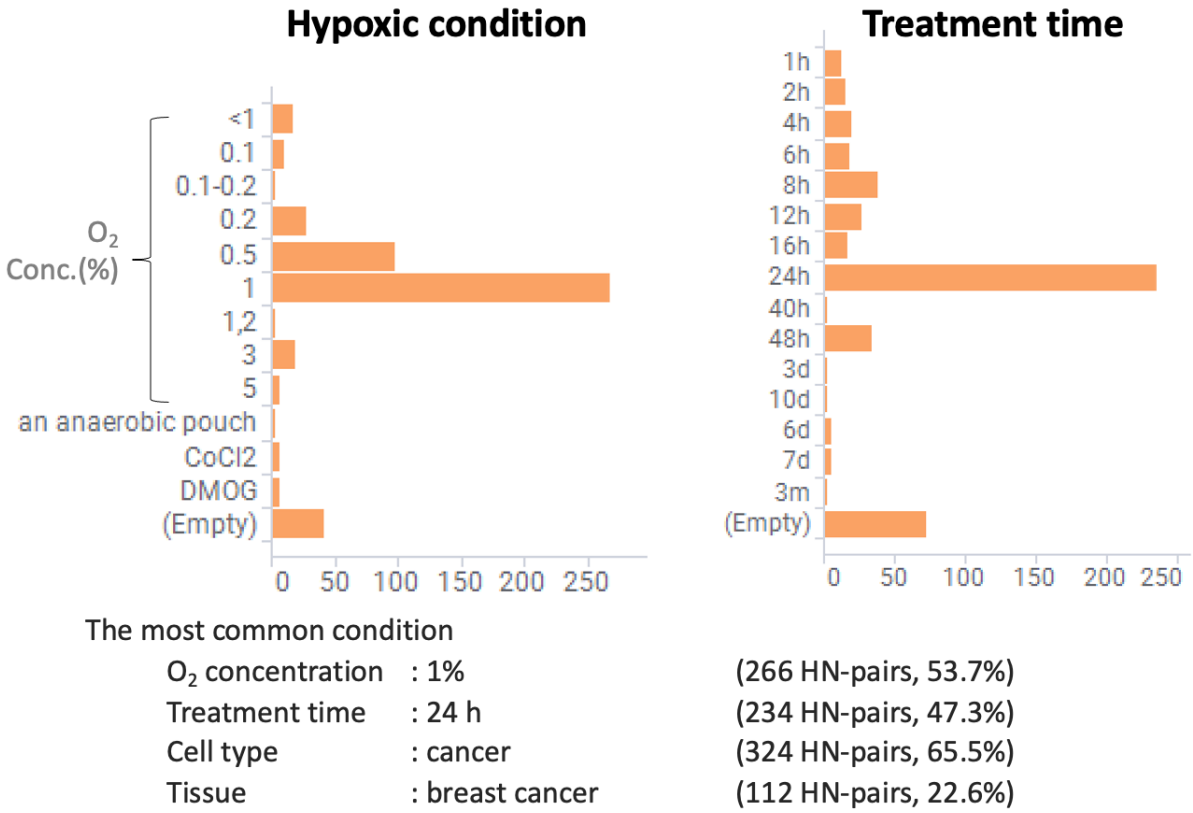
Figure 8 低酸素トランスクリプトームメタ解析の模式図

Step 1. オープンデータベースから低酸素応答性遺伝子のアップレギュレーションとダウンレギュレーションを評価し、リストアップした。Step 2. 既知の低酸素刺激関連遺伝子が選抜されていることを確認した。Step 3. 新規低酸素応答遺伝子を探索した。

3.1.1 解析対象データの要約

はじめに All Of gene Expression (AOE; https://aoe.dbcls.jp/) (Bono 2020) を用い、多くの低酸素に関連したデータが登録されていることを確認した。その後、NCBI の GEO から提供されているメタデータを元に解析対象データをキュレーションした。キュレーションする際の基準として、Hypoxia と Normoxia の実験条件の検体が同一データシリーズにあり HN-pair を設定できること、ヒトの細胞株あるいは組織検体由来の RNA-seq であることとした。次に既報から提供されている低酸素関連データリストとの統合および重複の排除を経て SRA から 69 データシリーズ、495 の HN pair となるサンプルの SRR numbers を取得した（https://doi.org/10.6084/m9.figshare.14141219.v1）。

解析対象のデータの低酸素条件について 495 サンプルについて要約した (Figure 9)。通常酸素状態については、一部記載のないサンプルも含まれたが一般的には 20% 酸素濃度である一方、低酸素条件は、酸素濃度の情報があるサンプルについては 0.1 から 5%の濃度、一部には $CoCl_2$ などのケミカルハイポキシアも該当した。処置時間は記載のあるものについては最短では 1 時間、最長では 3 ヶ月だった。データセット中、最も多い条件が、低酸素の条件が 1% 酸素濃度, 24 時間処理だった。全体の 65%ががん由来のサンプルでありその中でも最も多い由来組織は乳がん細胞だった。



The most common condition

O_2 concentration	: 1%	(266 HN-pairs, 53.7%)
Treatment time	: 24 h	(234 HN-pairs, 47.3%)
Cell type	: cancer	(324 HN-pairs, 65.5%)
Tissue	: breast cancer	(112 HN-pairs, 22.6%)

Figure 9 解析対象データの低酸素条件についての内訳

今回の対象データは、酸素濃度、処置時間、細胞種、由来組織について要約した。1%酸素濃度, 24 時間処置した乳がん細胞の条件が最も多かった。

3.1.2 HN-score を用いたデータセットの特徴

HN-score は、低酸素条件下の遺伝子発現変動を定性的に評価するスコアである。このスコアを用いることにより、RNA-seq にて定量された検体群の低酸素刺激応答遺伝子の低酸素応答性を評価可能と考えた。例えば *VEGFA* の場合は 406 UP, 25 DOWN, 64 unchange となり、その HN-score は 381 となる。この HN-score が高い順から 100 遺伝子、低い順から 100 遺伝子をそれぞれ UP 100 gene list, DOWN 100 gene list とした（Supplemental Table 1）。また、このリストは figshare からもアクセス可能である(UP100 gene list : https://doi.org/10.6084/m9.figshare.14141015.v1, DOWN 100 gene list : https://doi.org/10.6084/m9.figshare.14140997.v1）。

上述の UP 100 gene list, DOWN 100 gene list が HIF1A などの低酸素関連調節因子の影響を受けているかを確認すべく、ChIP-Atlas によるエンリッチメント解析を行った。UP 100 gene list では HIF1A, ARNT, EPAS1 などの低酸素関連因子がエンリッチメントされ、DOWN 100 gene list では Sin3A associated protein 30 (SAP30), histone deacetylase 1 (HDAC1)などのエピジェネティックな制御因子や、細胞周期の進行などに関与するがん遺伝子の MYC proto-oncogene, bHLH transcription factor (MYC)がエンリッチメントされた(Table 1)。同様に、Metascape にてエンリッチメント解析をしたところ、UP 100gene list では、Hypoxia 関連因子（GO:0001666: response to hypoxia, M255: PID HIF1 TFPATHWAY）がエンリッチメントした。DOWN 100 gene list では GO:0034660: ncRNA metabolic process が強くエンリッチメントされていた。(Figure 10 AB)

UP 100 gene list のエンリッチメント解析にて上位にエンリッチメントした HIF1A, EPAS1, ARNT について平均 MACS2 ピークの散布図をプロットした。MACS2 ピークの高い遺伝子は HN-score も高かった（Figure 10 C,D）。

Table 1　ChIP-Atlas でのエンリッチメント解析の結果

ChIP-Atlas のエンリッチメント解析ではゲノム上の遺伝子座や遺伝子に結合しているタンパク質を予測した。

UP 100 gene list をインプットとした結果

Antigen	ID	Log P-val	Log Q-val	Fold Enrichment
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 をインプットとした結果

Antigen	ID	Log P-val	Log Q-val	Fold Enrichment
SAP30	SRX116447	-34.1844	-30.0149	4.96916
MYC	SRX1497384	-31.4158	-27.5474	2.97453
HDAC1	SRX186644	-27.4205	-24.1541	3.3231

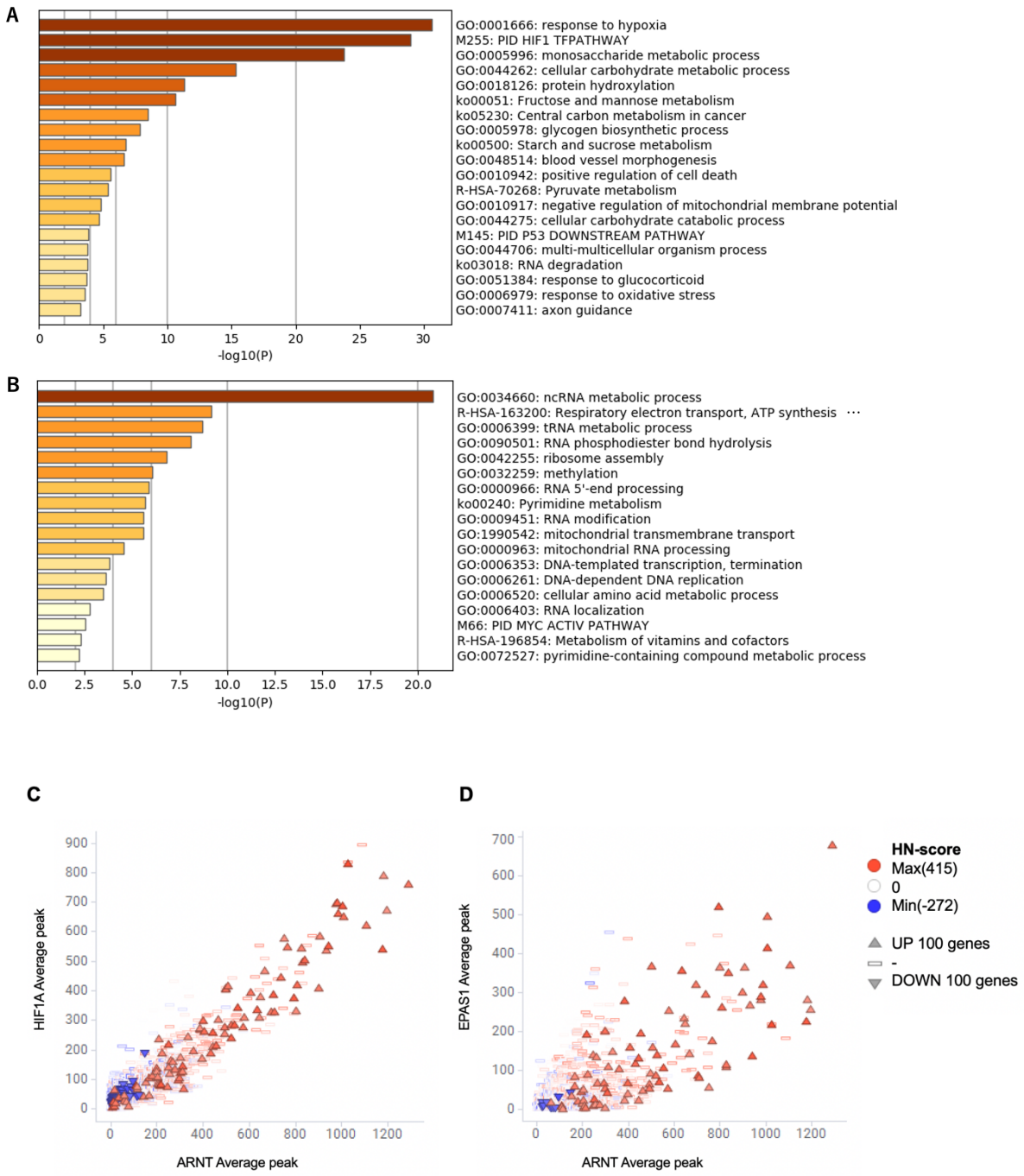


Figure 10 既知の低酸素刺激関連遺伝子の確認

(A–B) Metascape を用いたエンリッチメント解析 (A) UP 100 遺伝子をインプットにした結果 (B) DOWN 100 遺伝子をインプットにした結果。Figure の紙面の都合上省略した R-HAS-163200 の遺伝子セット名は次の通り。R-HAS-163200: respiratory electron transport, ATP synthesis by chemiosmotic coupling, and heat production by uncoupling proteins。(C–D) 低酸素関連抗原(HIF1A, ARNT, EPAS1)の ChIP-seq 平均ピークの散布図(C) HIF1A と ARNT の散布図 (D) EPAS1 と ARNT の散布図。色は HN-score による。

3.1.3 低酸素応答を対象としてビブリオーム解析

遺伝子と出版物の関係性の情報は gene2pubmed から取得可能である。UP 100 gene list の各遺伝子の出版数と、低酸素関連因子として著名な HIF1A との関係性を gene2pubmed のデータを元に Simpson 係数を算出することで、低酸素関連の研究がどれほどされているかを評価した(Figure 11A)。このような生物学的な出版情報を用いた解析方法をビブリオーム解析と呼ぶ。UP 100 gene list の中で HIF-1 による直接制御が知られている遺伝子群(Hirota and Semenza 2006)を、出版数と HIF1A との Simpson 係数の散布図で可視化した。HIF-1 による直接制御が知られている遺伝子群の中で興味のある遺伝子群 CA9, EGLN3, VEGFA と、今回のメタデータ解析で HN-score の高かった遺伝子群のうちの ankyrin repeat domain 37 (*ANKRD37*), N-myc downstream regulated 1 (*NDRG1*), G protein-coupled receptor 146 (*GPR146*)の処置時間ごとに log2 変換した HN-ratio を Box plot にて可視化した。(Figure 11B)

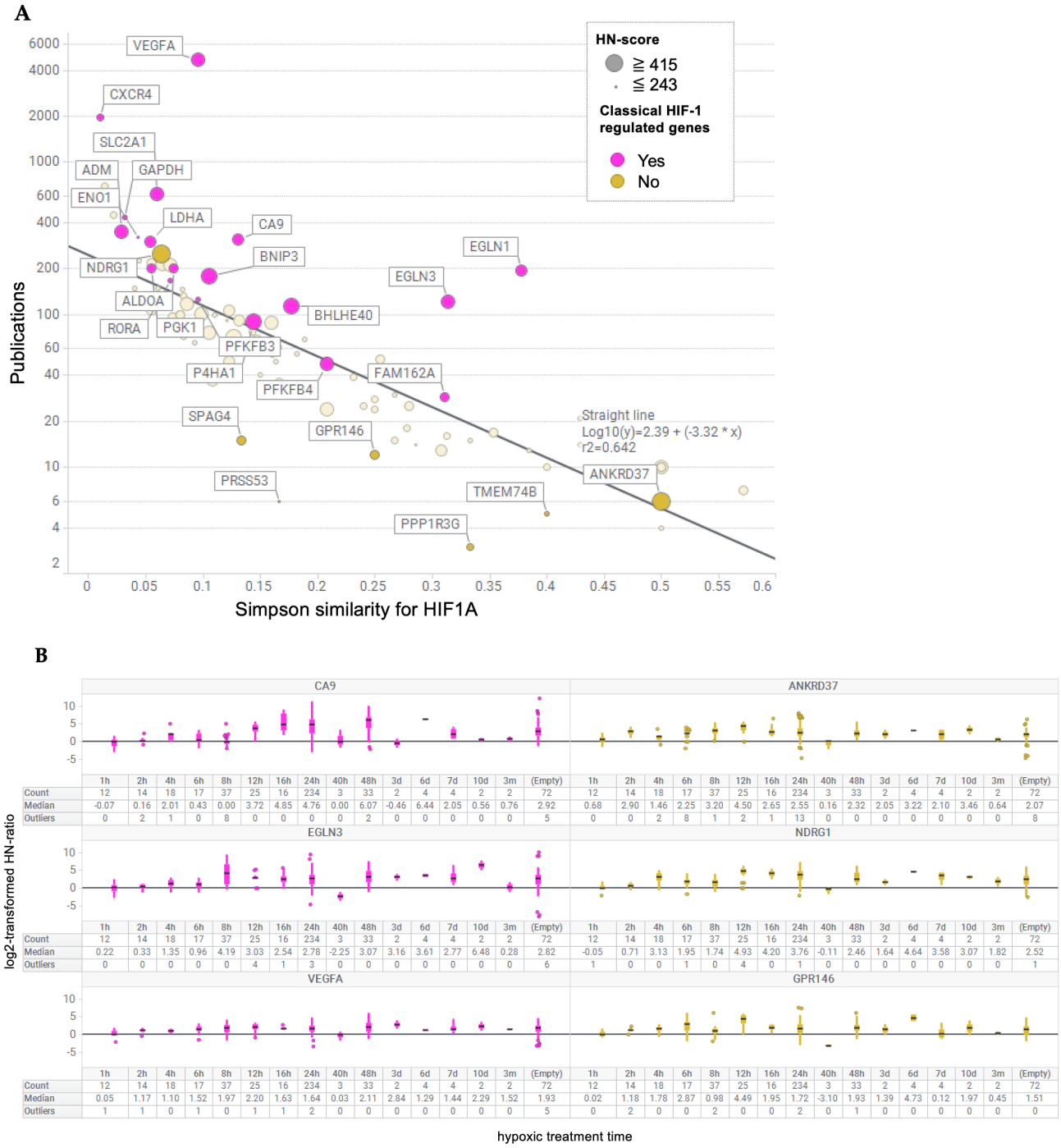


Figure 11 新規低酸素応答遺伝子の探索

(A)UP 100 gene list における HIF1A との gene2pubmed 情報を元にした Simpson 類似度と論文数 (B) UP 100 gene list の一部の遺伝子についての、低酸素処理時間当たりの HN-ratio を底を 2 とする Log 変換したボックスプロットにて可視化した。

3.1.4 考察

GEO に登録されている遺伝子発現データのうち、低酸素に関連する 69 データシリーズを解析対象とし、既報の約 4 倍となる 495 の HN-pair のデータを SRA から取得した。集めた遺伝子発現データの 65%ががん細胞由来であった。がん細胞由来の検体が多い理由には、低酸素に関係するがん研究が多くあることと、低酸素処置実験においてがん細胞株は実験に供しやすいことなどが考えられる。このデータセットは生体内の全てをバランスよく網羅しているわけではなく、データセットに偏りがあると考えられる。このキュレーションされたデータセットのサンプルの多くはがん細胞由来であるため、HN-score はがん細胞をメインに低酸素刺激反応を評価していることになる。

Up100 gene list を使った Metascape および ChIP-Atlas の解析結果で、想定通り Hypoxia 関連の gene set のエンリッチメントが確認できたことから、HN-score が想定通り低酸素刺激応答遺伝子を選抜できていると判断した（Table 1, Figure 10）

Metascape や ChIP-Atlas のエンリッチメント解析結果で up100 gene list で Hypoxia 関連の gene set がエンリッチメントされていたことから、以降の解析では UP100 gene list をメインに解析を進める方針とした。一方の DOWN 100 gene list では MYC がエンリッチメントされていた。HIF による MYC の抑制についてはいくつか報告がある。HIF による MYC の制御には MYC のタンパク質複合体形成のアンタゴニスティックな結合阻害(Gordan et al. 2007), DNA 結合部位の競合による MYC の転写活性の直接阻害(Koshiji et al. 2004), MYC のプロテアソーム分解促進(Zhang et al. 2007; Wong et al. 2013)を介したものが知られてることから、それらの HIF による間接的な影響が DOWN 100 gene list に反映していると考える。

DOWN 100gene list をインプットとした Metascape のエンリッチメント解析では ncRNA metabolic process との関連が示唆されていた。この結果は MYC による影響が関係している可能性があるが、詳細は明らかになっていない。そのため、ncRNA metabolic process に関係する研究は 3.2 章「ncRNA を含めた転写産物の低酸素応答の評価方法の構築」にて検討することとした。

Reference Expression dataset（RefEx）(https://refex.dbcls.jp/）は、正常な組織や細胞などの大規模な遺伝子発現データを収集し、EST, GeneChip, CAGE, RNA-seq と異なる 4 つの実験手法で得られた発現値を並列に比較を可能とする

ウェブツールである。RefEx で検索された遺伝子の表示順の決定には gene2pubmed が利用されている(Ono et al. 2017)。その活用を参考に、よく知られていない遺伝子を評価するために gene2pubmed を用いた。gene2pubmed を用いて二つの変数を作成した。一つ目は遺伝子ごとの出版数と、2 つ目は HIF1A との類似度を Simpson 係数である。Simpson 係数は、HIF1A と関連のある PubMed ID とある遺伝子と関連のある PubMed ID の共起する頻度の強さを示す類似度として使用した。Simpson 係数が 1 に近ければ HIF1A との関連が強いと推定されるが、比較される一方の PubMed ID の数が極端に少ない場合は、必ずしも関係性が強いと判断できない場合がある。

上記の理由により可視化して確認することが重要と考え、これらの出版数、Simpson 係数と、UP 100 遺伝子群の中で、HIF1A の制御を受けることがすでに知られている遺伝子群(Hirota and Semenza 2006)がどこにプロットされるかを可視化した（Figure 11A)。出版数が多い場合は Simpson 係数が低くなったが、それは研究される機会が増えたことにより関係が明らかになった遺伝子数が増え、その結果、Simpson 係数が過小評価されやすくなるためと考えられる。2006 年に報告された HIF1A によって制御されることが知られている遺伝子群は、論文数が多いにもかかわらず、Simpson 係数で高い値を示し、回帰直線上にプロットされた。一方、論文数が少ない遺伝子、例えば、少数ではあるが既に低酸素との関係が報告されている ANKRD37(Peng et al. 2018; Deng et al. 2020)などが回帰線の上にプロットされていた。そこで、回帰線の下にプロットされた sperm-associated antigen 4 (SPAG4), *GPR146,* protein phosphatase 1 regulatory subunit 3G (PPP1R3G), transmembrane protein 74B (TMEM74B) and serine protease 53 (PRSS53)などの遺伝子が、新規の低酸素応答性遺伝子の候補になるのではないかと考えた。PubMed や NCBI gene をはじめとする検索の結果、SPAG4 と低酸素に関係する報告はすでにあった(Knaup et al. 2014; Shoji et al. 2013)が、それ以外のこれらの遺伝子については該当文献はなかった。

Log2 変換した HN-ratio をボックスプロットにて可視化した（Figure 11B）。着目した遺伝子は、HIF-1 による直接制御が知られている遺伝子群から carbonic anhydrase 9 (*CA9*), egl-9 family hypoxia inducible factor 3 (*EGLN3)*, *VEGFA* と、今回のメタデータ解析で HN-score の高かった遺伝子群のうちの *ANKRD37, NDRG1, GPR146* とした。*ANKRD37*, *NDRG1* は本研究室の以前の研究(Bono and Hirota 2020)でも高い HN-score だった。これらの遺伝子は、40h などの一部の検体を除いて、低酸素応答が知られている遺伝子と同様に正の値の HN-ratio

を示していた。低酸素処置時間ごとの log2 変換した HN-ratio の結果では、40h で遺伝子発現変動が小さいか低下した。この処置時間の検体は全て NK 細胞由来で酸素濃度は 1%だった。検体数が 3 と少なく、NK 細胞由来の検体はこの 3 つ以外にはなかった。これは 40h 処理での影響よりむしろ NK 細胞ではこれらの遺伝子の低酸素応答は見られない可能性を示しているが、詳細については不明である。UP 100 遺伝子群の時間別の HN-ratio については Supplemental figure 5 にて可視化した。また、*CA9, VEGFA, GPR146* については Supplemental figure 6 にて細胞の条件ごとに可視化した。

G タンパク質共役受容体(GPCR)にはクラス A, B,C の分類がある。クラス A はロドプシン型とも呼ばれ、脂溶性の低分子リガンドやカテコラミンに結合する。クラス B はペプチドホルモンと結合し、N 末端に 100 から 300 アミノ酸残基の領域を持つ。クラス C は N 末端におよそ 300 アミノ酸残基の領域を持ち、グルタミン酸に代表されるアミノ酸などを結合する。*GPR146* 遺伝子は クラス A に属する GPCR である G protein-coupled receptor 146 タンパク質をコードする。Figure 11A の散布図では、GPR146 は出版数と、HIF1A との Simpson 係数との散布図では回帰線の下側にプロットされ、他の既知低酸素刺激応答遺伝子と同様に遺伝子発現亢進が確認された。一方で、PubMed や NCBI gene をはじめとする検索では、"hypoxia" "GPR146"をクエリとして文献調査したところ GPR146 は hypoxia との報告はされてはいなかったことから、GPR146 は今まで着目されていない重要な低酸素応答遺伝子であると考えている(https://pubmed.ncbi.nlm.nih.gov/?term=GPR146+hypoxia)。

網羅的な遺伝子発現データを用いた低酸素研究で、低酸素刺激によって発現亢進した遺伝子リストの中に GPR146 は含まれる(Qi et al. 2010; Labrecque et al. 2016)。しかしながら、今までの報告の内容では、GPR 146 は多くの低酸素刺激関連遺伝子群の中のただの一つの遺伝子にすぎない。GPR146 の抑制がコレステロールの低下に関与していたこと(Yu et al. 2019)や、Class A オーファン GPCR とは考えられているものの GPR146 のリガンド候補として考えられている C ペプチドが、赤血球の低酸素誘導 ATP 放出に抑制的に働いたこと(Richards et al. 2014)などが報告されており、GPR146 の生物学的機能解明はますます今後発展していくと考えられる。

PPP1R3G 遺伝子は G サブユニットである protein phosphatase 1 regulatory subunit 3 G をコードする。PPP1R3G は肝臓における食後のグルコースと脂質

のホメオスタシスに相関している(Zhang et al. 2014)。3T3L1 細胞では、PPP1R3G を過剰発現させるとグリコーゲンおよびトリグリセリドのレベルが上昇したとの報告があり、また PPP1R3G を全身で欠損させたマウスに高脂肪食を与えたところ、野生型と比較して体重及び脂肪組成が有意に減少したとのことだった。PPP1R3G を欠損すると酸素消費量および二酸化炭素生成量から測定される代謝速度が加速したとの報告がある(Zhang et al. 2017)。これらの結果を鑑みるに、低酸素応答により PPP1R3G の HN-score が高かった今回の結果は、低酸素状況下での生体内での代謝の速度を抑制する機構の可能性を示唆すると考える。

TMEM74B 遺伝子は transmembrane protein 74B をコードする遺伝子であり、C20orf46 という別名を持つ。論文報告数は少なく、乳がん細胞株と乳がん患者のコピー数変異（CNV）プロファイリングの結果、腫瘍のステージと関連する遺伝子として C20orf46 が提示されていた(Fatima et al. 2017)。低酸素との関係性については不明であるが、今回の研究に用いられたサンプルは乳がん由来のサンプルが多かったことから乳がんにおける TMEM74B の機能解析については今後の研究の発展が期待される。

PRSS53 遺伝子は serine protease 53 をコードする遺伝子である。頭髪および顔面の毛髪についてのラテンアメリカ人を対象にした genome-wide association scan の結果、髪の毛の形に影響を与えるとして PRSS53 の遺伝子座（Q30R）が同定されたとの報告がある(Adhikari et al. 2016)。しかしながら低酸素との関係性については未だ明らかにされていない。

今回の解析では ncRNA metabolic process が DOWN 100 gene list で顕著にエンリッチメントされていた(Figure 10B)。これは低酸素刺激が ncRNA の発現変動にも影響を及ぼしている可能性を示唆する。 今回の解析は GENCODE のタンパク質コーディング遺伝子のみを定量解析対象としたので、ncRNA の発現変動については解析の対象外となっている。低酸素刺激による ncRNA の遺伝子発現変動への影響については、さらなる研究が必要と考えたため、3.2 章「ncRNA を含めた転写産物の低酸素応答の評価方法の構築」にてその詳細を述べる。

本研究では低酸素応答に着目して、gene2pubmed データを元に各遺伝子の研究報告数を算出し可視化した。研究者の興味の如何に関わらず取得が可能な、数万の遺伝子発現情報を含む公共データを用いることにより、新規知見を明らかにすることを試みた。この研究によって、公共データベースを元にしたメタデータ

解析によるデータドリブンな解析により、多くの低酸素刺激実験において*GPR146*遺伝子の発現亢進が認められることを見出した。

3.2 ncRNA を含めた転写産物の低酸素応答の評価方法の構築

3.1 章「新規低酸素応答遺伝子の探索」を進める中で、低酸素状況下では ncRNA の代謝に関連のある遺伝子群の発現が低下していたことを明らかにした。しかしながらそれらの遺伝子群がどのように ncRNA の代謝に影響を与えるかについて詳細は不明なことが多く、コーディング遺伝子と比較して ncRNA についての知見は少ない。そこで、本研究の第 3.2 章では、ncRNA も含めた転写物に対しデータドリブンに低酸素刺激応答性を評価することを目的とした。

本章では、3.1 章で収集しキュレーションした、遺伝子発現オープンデータベース由来の低酸素刺激に関わるデータセットをもとに、二つのリファレンス配列を元に低酸素応答遺伝子や転写物を評価した(Figure 12)。一つはコーディング遺伝子にのみ着目し、GENCODE release 30 をもとにした遺伝子発現解析で、もう一つは FANTOM-CAT をリファレンスとし、コーディング遺伝子と ncRNA に対する発現定量データを用いた解析である。どちらの解析でも HN-score を算出したが、区別のため GENCODE を元に HN-score を算出した場合は HNg-score, FANTOM-CAT を元に HN-score を算出した場合は HNf-score と表記する。

コーディング遺伝子にのみ着目した解析では、低酸素刺激により ncRNA metabolic process、中でも ribosomal RNA (rRNA)に関与する遺伝子が抑制されていることを示した。次にコーディング遺伝子と ncRNA について視野を広げ、FANTOM-CAT を元に網羅的に低酸素による転写物への影響を評価し、染色体ごとの HNf-score の可視化やセンス-アンチセンス鎖に着目して解析した。

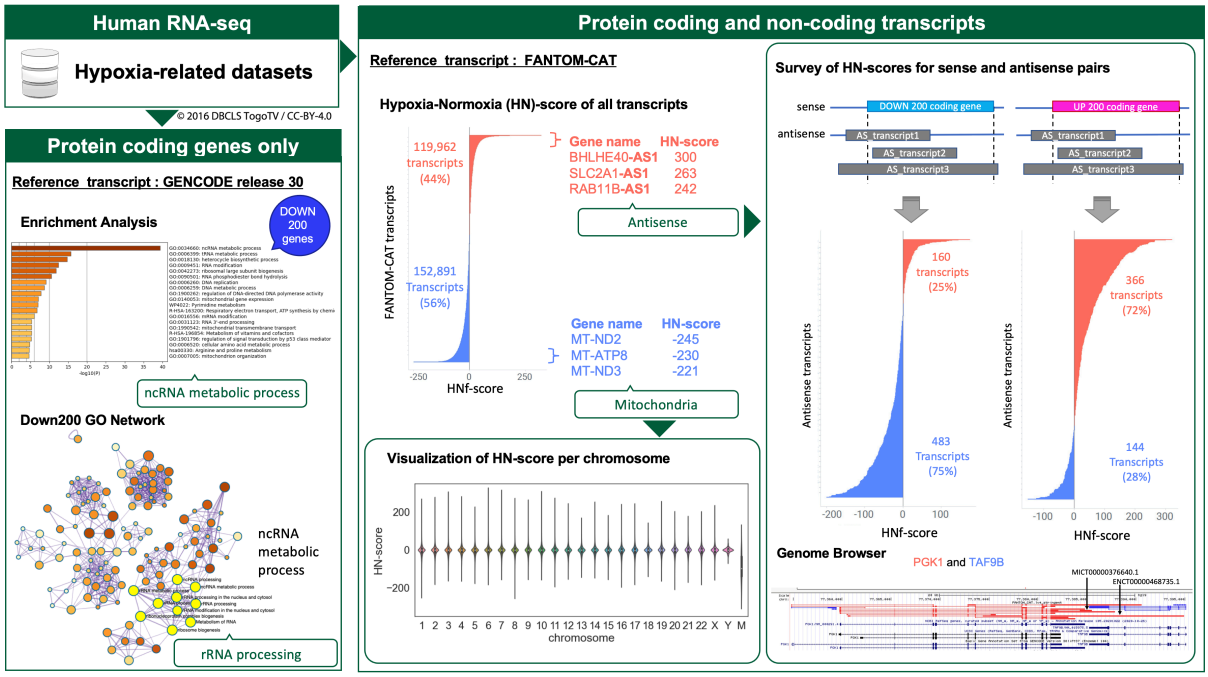


Figure 12 ncRNA を含めた転写産物の低酸素応答の評価のメタ解析

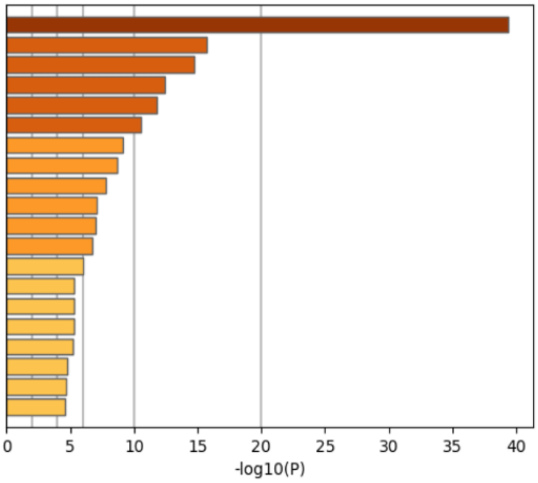
ヒトの低酸素 RNA-seq データを使用し、二つのリファレンスをもとに大きく二つの解析を行った。コーディング遺伝子のみに着目した解析では、GENCODE をリファレンスに使用した。低酸素による発現低下遺伝子が rRNA に関わる遺伝子がエンリッチメントされていた。コーディング遺伝子と ncRNA に着目した解析では、FANTOM-CAT をリファレンスに資料した。低酸素刺激下では、ミトコンドリア DNA 由来の転写物が低下していた。アンチセンスの解析では遺伝子発現制御（PGK1 と TAF9B）の関係性が示唆された。

3.2.1 エンリッチメント解析の詳細調査

3.1 章「新規低酸素応答遺伝子の探索」のコーディング遺伝子にのみ着目したエンリッチメント解析では、低酸素による影響されやすさの指標である HNg-score の上位 200 の遺伝子では低酸素応答関連の gene set がエンリッチメントされた（Figure 13A）。一方で HNg-score 下位 200 の遺伝子のエンリッチメント解析では、ncRNA metabolic process 関連遺伝子の発現抑制が明らかになった（Figure 13B）。そこで、まずは ncRNA metabolic process に関わる遺伝子群を精査するために、エンリッチメント解析で得られたこれらの遺伝子の詳細を調査することとした。

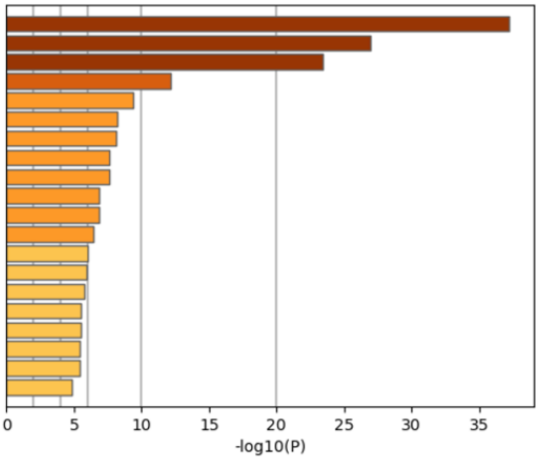
GO:0034660 ncRNA metabolic process について、より細分化したカテゴリの中では以下の項目（ncRNA processing, ribosome biogenesis, ribonucleoprotein complex biogenesis, rRNA processing, rRNA metabolic process など）が含まれた（Supplemental table 2）。GO network 解析（Figure 13C）と Upset plot の可視化（Figure 13D）により、ncRNA metabolic process に該当する遺伝子は rRNA に関わる遺伝子群で構成されていることを示した。

A



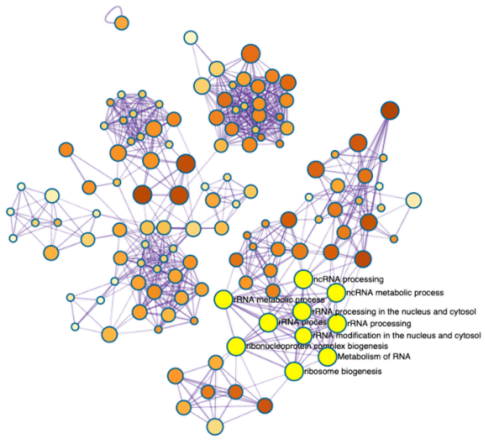
GO:0034660: ncRNA metabolic process
GO:0006399: tRNA metabolic process
GO:0018130: heterocycle biosynthetic process
GO:0009451: RNA modification
GO:0042273: ribosomal large subunit biogenesis
GO:0090501: RNA phosphodiester bond hydrolysis
GO:0006260: DNA replication
GO:0006259: DNA metabolic process
GO:1900262: regulation of DNA-directed DNA polymerase activity
GO:0140053: mitochondrial gene expression
WP4022: Pyrimidine metabolism
R-HSA-163200: Respiratory electron transport, ATP synthesis by ...
GO:0016556: mRNA modification
GO:0031123: RNA 3'-end processing
GO:1990542: mitochondrial transmembrane transport
R-HSA-196854: Metabolism of vitamins and cofactors
GO:1901796: regulation of signal transduction by p53 class mediator
GO:0006520: cellular amino acid metabolic process
hsa00330: Arginine and proline metabolism
GO:0007005: mitochondrion organization

B



M255: PID HIF1 TFPATHWAY
GO:0001666: response to hypoxia
WP534: Glycolysis and gluconeogenesis
hsa00051: Fructose and mannose metabolism
GO:0018126: protein hydroxylation
GO:0016051: carbohydrate biosynthetic process
WP5094: Orexin receptor pathway
GO:0033169: histone H3-K9 demethylation
GO:0048514: blood vessel morphogenesis
WP2456: HIF1A and PPARG regulation of glycolysis
GO:0044262: cellular carbohydrate metabolic process
M145: PID P53 DOWNSTREAM PATHWAY
GO:0060537: muscle tissue development
WP3932: Focal adhesion: PI3K-Akt-mTOR-signaling pathway
GO:0042326: negative regulation of phosphorylation
M5885: NABA MATRISOME ASSOCIATED
GO:0062012: regulation of small molecule metabolic process
GO:0045765: regulation of angiogenesis
hsa00500: Starch and sucrose metabolism
GO:0048545: response to steroid hormone

C



GO term	Genes
ncRNA metabolic process	DKC1|METTL1|NOP2|RRP9|DDX21|NOLC1|WDR46|LCMT2|FARSB|RCL1|NPM3|NOP56|WDR4|RPP40|UTP14A|POP1|RRS1|ADAT1|DIMT1|UTP20|RRP7A|RRP15|MRTO4|POP5|POLR3K|EXOSC4|BRIX1|LYAR|ELAC2|DUS1L|NOL6|NOC4L|METTL8|MRM1|PUS1|UTP15|MAK16|RRP36|FDXACB1|IMP4|MARS2|DTD2|KTI12|TOE1|TRMT61A|CTU2
rRNA processing	DKC1|NOP2|RRP9|DDX21|WDR46|RCL1|NOP56|RPP40|UTP14A|DIMT1|UTP20|RRP7A|NIP7|EXOSC4|PNO1|ELAC2|NOL6|NOC4L|NOL12|MRM1|UTP15|LTV1|RRP36|IMP4

D

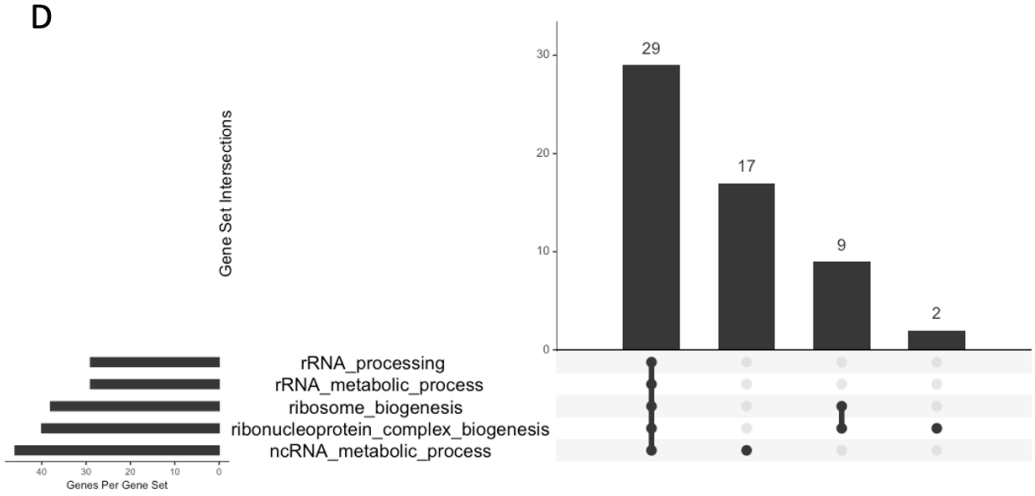


Figure 13 低酸素刺激応答 ncRNA metabolic process 遺伝子の解析

(A,B) Metascape によるエンリッチメント解析 (A) DOWN 200 gene list をインプットとしたエンリッチメント解析結果を示す。紙面の都合上省略した R-HAS-163200 の遺伝子セット名は以下の通り R-HAS-163200: respiratory electron transport, ATP synthesis by chemiosmotic coupling, and heat production by uncoupling proteins。(B) UP 200 gene list をインプットとしたエンリッチメント解析 (C,D) ncRNA metabolic process に類似している遺伝子セットの調査 (C) DOWN 200 gene set をインプットとしたエンリッチメント解析を元にした GO network 解析による ncRNA metabolic process と類似した遺伝子セットを可視化した。(D) ncRNA metabolic process と rRNA 関連遺伝子セットの集合の重複具合を Upset plot にて示した。

3.2.2 FANTOM-CAT と GENCODE の記述のある論文数調査

FANTOM-CAT や GENCODE を使用した RNA-seq 研究がどの程度なされているかを確認するために PMC にて検索した。PMC にはアブストラクトだけでなく論文全文が収録されており、Material and Method に記載されていることが想定されるリファレンスの情報も検索対象になる。そこで PMC を検索対象として、GENCODE, FANTOM と RNA-seq の記載のあった論文数を年毎に算出した。RNA-seq の記述のあった論文数は 2021 年には 4 万を超えていた一方で、FANTOM と RNA-seq 両者の記述がある論文はその 0.5%で、GENCODE のおよそ 10 分の 1 だった（Table 2)。

Table 2 FANTOM と RNA-seq の記述のある報告数の調査

year	“RNA-seq”	“RNA-seq” &”GENCODE” (%)	“RNA-seq” &”FANTOM” (%)
2017	17373	769 (4.4%)	123 (0.7%)
2018	20400	996 (4.9%)	127 (0.6%)
2019	25668	1254 (4.9%)	130 (0.5%)
2020	33607	1593 (4.7%)	176 (0.5%)
2021	40757	1976 (4.8%)	210 (0.5%)
2022	9420	406 (4.3%)	39 (0.4%)

3.2.3 FANTOM-CAT を用いた転写物の HNf-score の特徴

FANTON-CAT では、transcription initiation evidence score (TIEScore)と呼ばれるスコアを元に FANTOM-CAT 遺伝子を評価している。TIEScore の閾値ごとに遺伝子を permissive (n = 124,245), robust (n = 59,110), stringent (n = 31,520)に分類している。

公開されている FANTOM-CAT の GTF ファイルをもとに各 FASTA ファイル（raw, permissive, robust, stringent）を作成し、これをリファレンス配列として低酸素-通常酸素状態の 495 ペアの実験データセットの転写産物を定量し、FANTOM-CAT をリファレンスとした Hypoxia-Normoxia score (HNf-score)を算出した(Figure 14A)。Robust と stringent のリファレンス配列を用いて、低酸素条件により発現が亢進することが知られている lncRNA である MIR210HG(Ho et al. 2022)の HNf-score が正の値を示していたことを確認した（Supplemental figure 7)。また Robust と stringent も同様の HNf-score 分布であることを確認し、以降の解析では最も厳しい条件の TIEScore の閾値が使われた stringent の FANTOM-CAT のデータセットを活用することとした。

HNf-score の上位下位に含まれている転写物にどのようなものが含まれているかを精査した(Figure 14B, Table 3)。高い HNf-score を示す低酸素応答転写物の中には、高 HNg-score の遺伝子のアンチセンス鎖に位置している転写物がいくつかあった（BHLHE40-AS1, SLC2A1-AS1)。一方、低 HN-score の転写物はミトコンドリア遺伝子が多かった(MT-ND2, MT-ATP8)。

3.2.4 染色体ごとの HNf-score の要約

上位下位 25 のトランスクリプトの HNf-score について、Table 3 に掲載する。全データについては figshare に登録済みである（https://doi.org/10.6084/m9.figshare.19679493.v1)。染色体ごとに HNf-score を可視化した結果では、他の染色体とは異なりミトコンドリア DNA 由来の転写物では HNf-score の中央値は負の値を示した(Figure 14C, Supplemental table 3)。低酸素応答による染色体ごとの違いについて詳細解析を行ったが、今回の解析では染色体ごとの特徴的な影響は見出されなかった (Supplemental figure 8)。

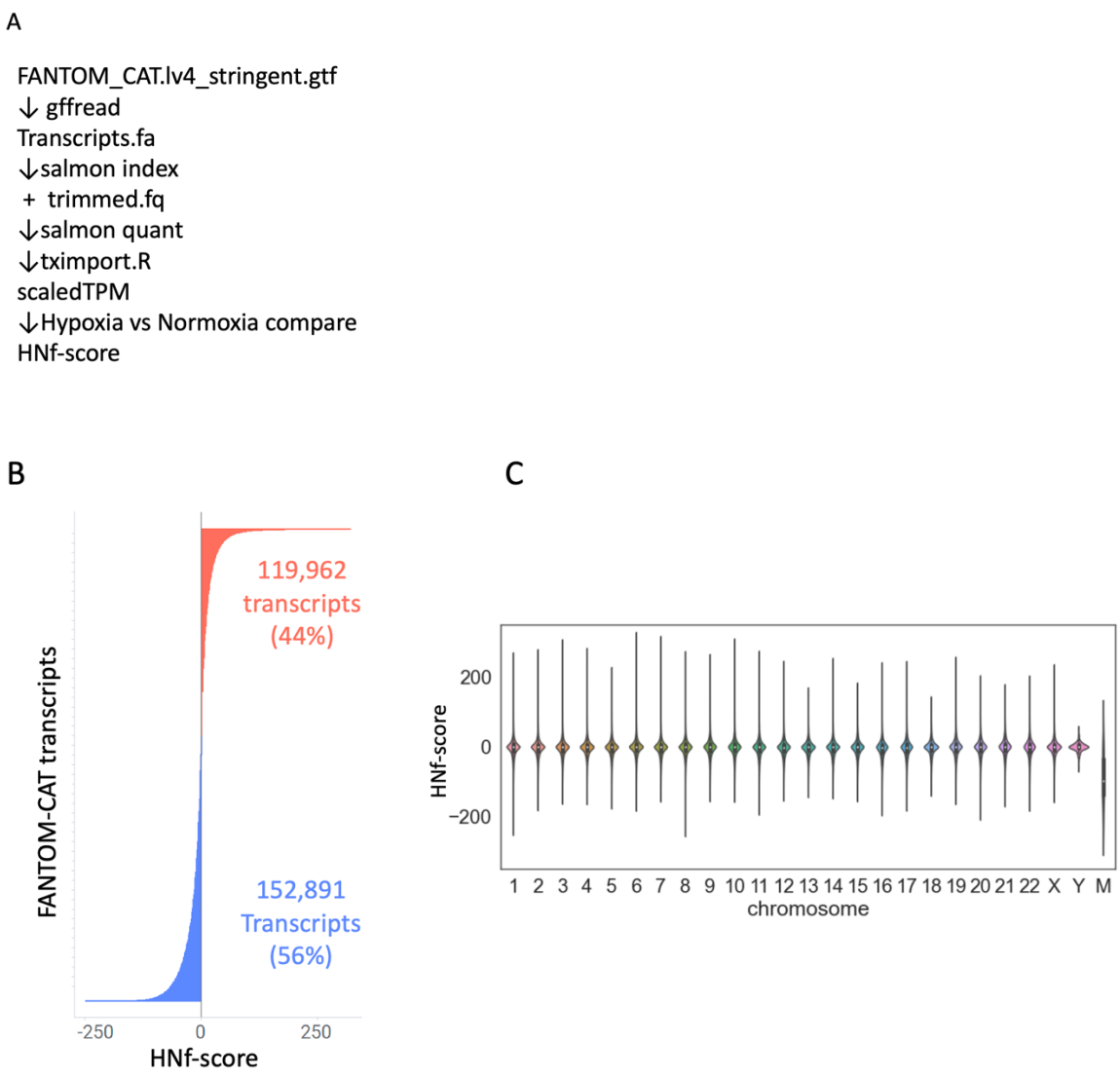


Figure 14 FANTOM-CAT を活用した転写物の HN-score 評価

(A)FANTOM-CAT をリファレンスにした HNf-score 算出手順 (B)FANTOM-CAT の転写物ごとの HNf-score 分布 (C)染色体ごとの HNf-score バイオリンプロット

Table 3 HNf-score の高かったあるいは低かった転写物の上位 25 リスト

高 HNf-score の転写物には BHLHE40 のアンチセンス鎖に位置する BHLHE40-AS1 など、AS (antisense)が遺伝子名に含まれていた。一方、低い HNf-score の転写物にはミトコンドリア由来の転写物がリストアップされた。

		Transcript id	HNf	Chr	gene name	gene class	antisense
UP 25 of all 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	
DOWN 25 of all 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	

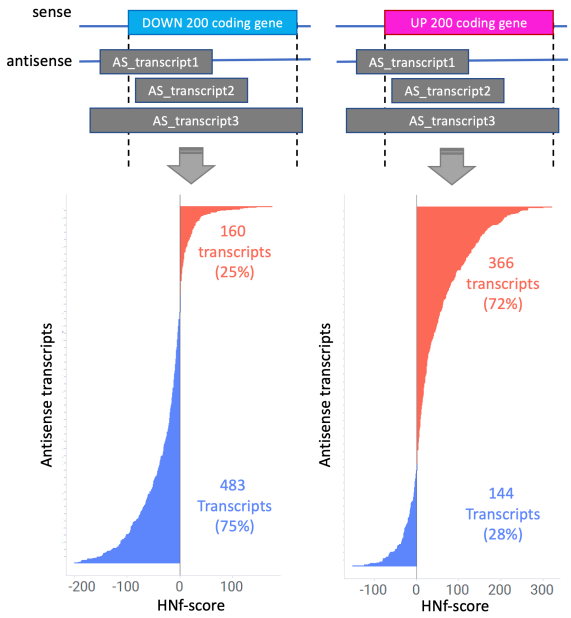
	20	MICT00000370388.1	-175	chrM	MT-TA	structural RNAs	AS_MT-ATP6 etc
	21	ENST00000361899.2	-175	chrM	MT-ATP6	protein coding mRNAs	
	22	ENST00000458605.1	-174	chr22	RRP7B	pseudogenes	
	23	HBMT00000611042.1	-172	chr17	METTL23	protein coding mRNAs	
	24	ENST00000371538.3	-170	chr1	SELRC1	protein coding mRNAs	
	25	ENST00000293860.5	-169	chr16	POLR3K	protein coding mRNAs	

3.2.5 センス-アンチセンス鎖に着目した調査

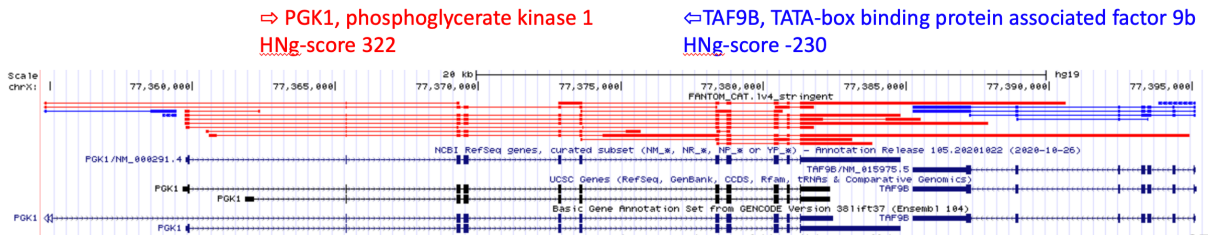
低酸素応答コーディング遺伝子 UP, DOWN 各 200 のアンチセンス（ncRNA 含む）の遺伝子発現の挙動ごとに 4 つのカテゴリに分類した（Figure 15A, Table 4)。センス鎖側の遺伝子と同様の挙動を示すアンチセンス鎖の転写物がおよそ 3/4 を占めていたが、逆向きの HNf-score を示したセンス-アンチセンス鎖ペアも存在した。UP200 gene set のアンチセンス鎖に対応する Top10 の転写物では short ncRNAs や lncRNA が該当し、Bottom10 では“protein coding mRNA”が多かった。Top10 では phosphoglycerate kinase 1 (PGK1)が低酸素によって発現抑制のあった TAF9B (TATA-box binding protein associated factor 9b)のアンチセンス側に該当した。

センス-アンチセンス鎖ペアの中で、TAF9B と PGK1 について着目し、UCSC ゲノムブラウザでその染色体上の位置と、遺伝子レベルでの HNg-score を確認した（Figure 15B, C)。UCSC genes や GENCODE の遺伝子情報ではこの二つの遺伝子は sense-antisense の関係とはならないが、FANTOM-CAT をもとにした解析ではこの二つの転写物は sense-antisense の位置に存在することが明らかとなった。

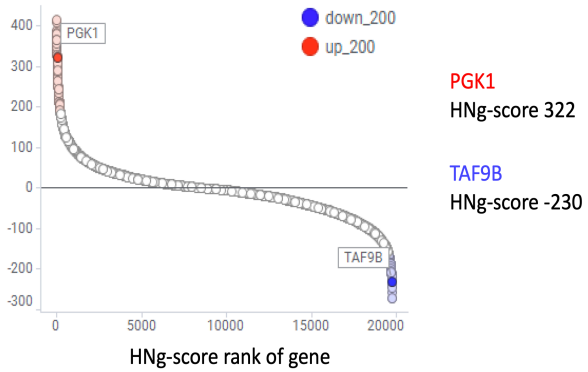
A



B



C



PGK1
HNg-score 322

TAF9B
HNg-score -230

Figure 15 アンチセンス鎖側に位置する転写物の HNf-score の調査

(A)UP200, DOWN200 それぞれの遺伝子のアンチセンス鎖に位置する転写物リストの取得と HNf-score の分布(B)GRCh38/hg38 のリファレンスゲノムのバージョンを用いて UCSC ゲノムブラウザによる PGK1 と TAF9B の座標の可視化 (C) PGK1 と TAF9B 遺伝子の HNg-score の可視化

Table 4 UP 200, DOWN 200 遺伝子群のアンチセンス鎖に位置する転写物

DOWN200 gene list 中の HNf-score Top10 では PGK1 が低酸素によって発現抑制のあった TAF9B アンチセンス側に該当した。

			Transcript id	HNf	gene name	gene class	antisense
Antisense transcripts of UP200 gene list	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
	Top10	1	MICT00000376640.1	175	PGK1	protein coding mRNAs	AS_TAF9B

Antisense transcripts of DOWN200 gene list		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

3.2.6 考察

ヒト遺伝子の研究は、過去によく研究されている遺伝子を中心に研究されて、知見の少ない遺伝子を研究対象にすることはあまりないことが知られている(Stoeger and Amaral 2022)。一方で、データドリブンな研究ができるようになった昨今は、知見を先人たちの論文からだけではなく、先人たちが取得したデータセットからも取得することができるようになった。第 3.1 章にて、低酸素応答による遺伝子発現変動について、出版バイアスがあることを示し、今まで見逃されていた低酸素応答遺伝子群を明らかにした(Ono and Bono 2021)。しかしながら、低酸素応答により遺伝子発現が抑制された”ncRNA metabolic process”関連遺伝子の詳細については不明である。

第 3.2 章の研究では、低酸素応答性をさらに明らかにする試みとして、二つのアプローチをとった。一つはコーディング遺伝子に着目した”ncRNA metabolic process“遺伝子の詳細調査、もう一つは ncRNA の情報も網羅した FANTOM-CAT をリファレンスした低酸素応答遺伝子の評価である（Figure 12)。

GENCODE を用いてコーディング遺伝子に着目して算出した HNg-score の上位下位の 200 遺伝子を対象としたエンリッチメント解析では、上位 200 遺伝子では低酸素応答関連 Gene set が該当した一方で、下位 200 遺伝子では ncRNA metabolic process がエンリッチメントされた（Figure 13 AB)。GO network 解析では ncRNA metabolic process と rRNA processing に関わるオントロジーが同様の遺伝子群で構成されていた（Figure 13 CD）(https://www.ebi.ac.uk/QuickGO/GTerm?id=GO:0034660)。つまり ncRNA の中でも rRNA processing に関わる遺伝子群への影響が大きいことが示された。

rRNA はリボソームを構成してタンパク質合成機能を持つ。リボソームと結合して細胞質に存在する lncRNA が分解されたとの報告があり(Carlevaro-Fita et al. 2016)、加えて ncRNA が rRNA をサイレンシングしたという報告もある(Schmitz et al. 2010)。

第 3.1 章の解析ではコーディング遺伝子のみに着目していたため、ncRNA も含めた網羅的な転写物への低酸素による影響については不明瞭であった。そこで、次に FANTOM-CAT を活用し ncRNA も含めた転写物の HNf-score を網羅的に算出することとした。

GENCODE グループは、ヒトとマウスのゲノムの機能アノテーションを行い、

そのデータを提供している。この遺伝子セットは、ENCODE コンソーシアムや他の多くのプロジェクト(eg. Genotype-Tissue Expression (GTEx), The Cancer Genome Atlas (TCGA), International Cancer Genome Consortium (ICGC), NIH Roadmap Epigenomics Mapping Consortium, Blueprint Epigenome Project, Exome Aggregation Consortium (EXAC), Genome Aggregation Database (gnomAD), 1000 Genomes Project and the Human Cell Atlas (HCA))に活用されている（https://www.gencodegenes.org/pages/gencode.html)。

FANTOM は、理研のマウスゲノム百科事典プロジェクトで収集された完全長 cDNA のアノテーションを行うことを目的として 2000 年に結成された国際研究コンソーシアムであり、その役割はマウスだけに留まらず哺乳類ゲノム，プロモーター，エンハンサー，ncRNA，マイクロ RNA など遺伝子発現に関わる分野を軸に拡大している。その中のプロジェクトのうちの一つである FANTOM5 プロジェクトは、トランスクリプトーム評価に信頼性の高い 5'ヒト lncRNA データセットである FANTOM-CAT データセットを設計、提供している(Figure 16)。FANTOM-CAT には、およそ 28,000 の human lncRNA genes with high-confidence 5′ ends、機能を持つ可能性のあるおよそ 20,000 の lncRNA の情報がある。GENCODE の情報よりも豊富であり FANTOM-CAT を活用することで、ヒトのコーディング遺伝子以外も含めた遺伝子発現情報をもとに遺伝子発現を定量することが可能となる。

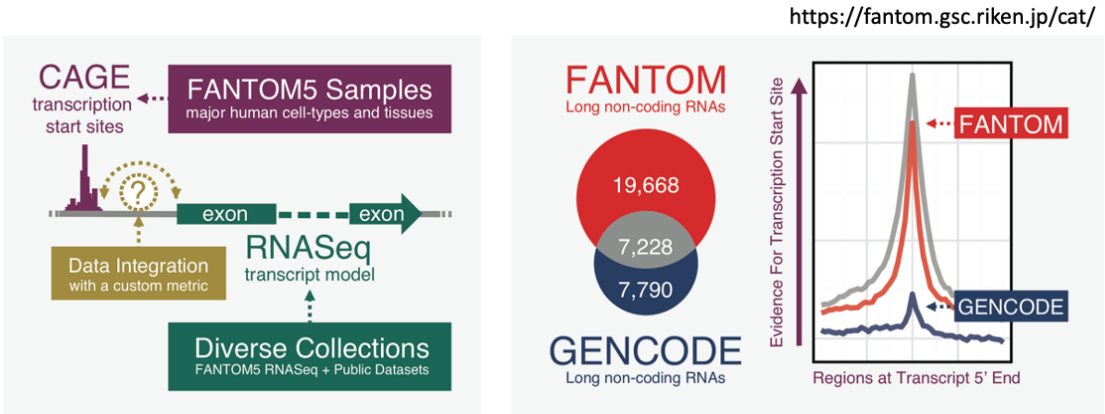


Figure 16 FANTOM-CAT

理研で開発された CAGE(Cap Analysis of Gene Expression)技術、複数の RNA-seq データを使って、ヒト long non-coding RNA の正確な 5' 末端をもつ FANTOM-CAT データセットが

作られた。この図は FANTOM-CAT のサイト(https://fantom.gsc.riken.jp/cat/)より引用した。

FANTOM-CAT のデータセットを活用することにより、ヒトのコーディング遺伝子以外も含めた遺伝子発現情報を定量することが可能である。13 のがん種に着目して遺伝子発現解析を行いがんの病因と進行に関与する可能性のある新規ノンコーディング遺伝子を同定した報告(Imada et al. 2020)や、原発性前立腺がんの進展に関係する可能性のある lncRNA を見出した報告(Imada et al. 2021)、FANTOM6 コンソーシアムによる新規 lncRNA 解析にも活用されている(Ramilowski et al. 2020)。

しかしながら、FANTOM-CAT はオリジナルサイトではゲノムアノテーション記述ファイルの gene feature format (GTF)形式のみで公開されており、FASTA フォーマットでは公開されていない。そのため RNA-seq のためのリファレンス配列としては頻繁には使用されていないことが推察された。そこで、PubMed Central を対象として"RNA-seq"と"FANTOM"両者が記載されている論文数を調査した。比較のため、"RNA-seq"のみが記載されている論文数と、"RNA-seq"と"GENCODE"両者が記載されている論文数も調査した。

"RNA-seq"と"FANTOM"両者を持つ論文数は FANTOM-CAT の論文が発表された 2017 年以降であっても"RNA-seq"の記述のある論文の 1%を下回る結果だった(Table 2)。つまりデータからも FANTOM-CAT をリファレンスとした遺伝子発現定量の報告はわずかであることが示された。よく知られているものほどよく研究され、見逃されている知見が残されていることは第 3.1 章で示したが、同様のことがリファレンスの活用状況にも反映されているのではないかと考えた。つまり FANTOM-CAT を用いた定量により新知見が得られることが期待で切ると考えた。

ncRNA を含めた網羅的な解析をするには FANTOM-CAT を活用して低酸素応答の評価は検討する価値があると考え、以下の解析をすることとした。FANTOM-CAT の GTF ファイルをもとに FASTA ファイルを作成し HNf-score を付加した（Figure 14 A)。GTF ファイルは TIEScore の閾値ごとに分類された lv1 raw, lv2 permissive, lv3 robust, lv4 stringent の 4 種類があり、本研究では lv3 robust と lv4 stringent の二つを用いて HNf-score を算出した。lv3 robust と lv4 stringent の二つのリファレンス配列で同様の HNf-score が算出できたことを確認した後に、lv4 stringent 由来の HNf-score で詳細解析を進めることとした

（Supplemental figure 7）。

HNf-score の下位 25 の遺伝子はミトコンドリア DNA 由来の転写物が多かったことから、次に染色体ごとの遺伝子発現制御を可視化した（Figure 14C, Table 3）。ミトコンドリア遺伝子に対する影響が他の染色体よりも顕著であることを示した。低酸素刺激による mitochondrially encoded NADH dehydrogenase 2 (*MT-ND2*), mitochondrially encoded ATP synthase 8 (*MT-ATP8*), mitochondrially encoded NADH dehydrogenase (*MT-ND*)などのミトコンドリア遺伝子の発現低下はすでに知られており(Arnaiz et al. 2021)、これらの結果と矛盾しない。加えて、ミトコンドリアオートファジーが低酸素によって誘導されること、このプロセスには HIF-1 存在下での BNIP3 の発現が必要なことが報告(Zhang et al. 2008)されており、その結果も今回の結果をサポートする。

一方、HNf-score の上位 25 の転写物の遺伝子名には、BHLHE40 antisense RNA 1 (BHLHE40-AS1)などの Antisense と関わりのある名前の遺伝子があった(Table 3)。遺伝子名の派生元の BHLHE40 は低酸素応答遺伝子として知られている遺伝子である。加えて、もっとも HNf-score の高かった RP1-261G23.7 は低酸素応答遺伝子として有名な VEGFA のアンチセンス鎖に位置した short ncRNA である。この short ncRNA は、低酸素時の VEGFA の転写調節に機能的に関与するとの報告がある(Nieminen et al. 2018)。これらより、低酸素応答遺伝子のアンチセンスにある転写物についてさらなる調査をする価値があると考えた。

アンチセンスの転写物とは、コーディング遺伝子やノンコーディング遺伝子の反対側の相補的なアンチセンス鎖から転写される転写物のことを意味する。ヒト(Ozsolak et al. 2010)とマウス(Katayama et al. 2005)ではおよそ 30%の転写物がアンチセンス鎖に転写物を有している。コーディング遺伝子の低酸素応答指標である HNg-score をもとに低酸素応答遺伝子を UP,DOWN それぞれ 200 遺伝子を選抜し、その遺伝子のアンチセンスの転写物の HNf-score を評価した。その結果、UP, DOWN それぞれのアンチセンス転写物の 3/4 はセンス鎖の遺伝子と同様の発現変動をすることが示された。一方で、逆向きの発現制御となる結果の転写物も存在した（Figure 15A）。

DOWN 200 位遺伝子群のアンチセンス鎖に位置する HNf-score の低い転写物には、HNf-score の低かった転写物の上位 25 リスト (Table 3)と同様にミトコンドリア DNA 由来の転写物が多かった。DOWN200 gene list の Top10 では興味

深いことにPGK1が低酸素によって発現抑制のあったTAF9Bアンチセンス側に該当した（Figure 15B)。PGK1遺伝子はphosphoglycerate kinase 1をコードする遺伝子である。PGK1 は解糖系において 1,3 -diphosphoglycerate から 3-phosphoglycerate への可逆的変換を触媒して 1ATP を生成する(Valentin et al. 1998)。PGK1はHIF1Aの制御の元、低酸素刺激により遺伝子発現が亢進することが20年以上前から知られている(Li et al. 1996)。

TAF9Bはtranscription factor II D (TFIID)を構成するタンパク質の一つであり、転写開始において重要な役割を果たす(Tora 2002; Frontini et al. 2005, 9)。また、TAF9BはHIF1Aのノックダウンで発現上昇、低酸素状況下で発現抑制があることが報告されている(Mathieu et al. 2011)。

FANTOM-CATを用いることにより、PGK1とTAF9Bの関係がセンスアンチセンスペアになることが示された。これはUCSC geneやGENCODEではその限りではない(Figure 4C)。また、PGK1 と TAF9B の位置は GRCh38/hg38 や T2T においても同様の位置関係であることを確認した（Supplemental figure 9)。アンチセンス転写物による遺伝子の制御は研究されており(Pelechano and Steinmetz 2013; Katayama et al. 2005; Okada et al. 2008)、重なっているセンス鎖とアンチセンス鎖の転写物がお互いの方向に向かって同時に転写が開始される場合は転写に関わる因子の衝突が起こり、遺伝子発現を抑制することが考えられる。これらの結果は、PGK1の3'末端が伸長によってTAF9Bの発現に影響を及ぼしている可能性を示している。

今回のデータ解析はStrand-specific RNA sequencing由来に限定していない。Strand-specific RNA-seqではない場合、誤ってセンス-アンチセンスの配列が区別されないまま定量されている可能性がある。今回着目したPGK1とTAF9BのHNf-scoreの変化は、増加と低下であり逆向きの変化である。そのため、Strand-specific sequencingに限定していないこの解析であっても、同じHNf-scoreに着目した場合と比較してその影響は小さいと考えている。しかしながら、さらなる研究にはStrand-specific RNA sequencingといったstrand specificな実験や、knock downやknock out実験等にて低酸素応答にまつわる新たな知見が蓄積されることを期待する。このデータセットや解析アプローチを活用することは更なる低酸素応答メカニズムの仮説の提示や解明の一助になると考える。

3.2章「ncRNAを含めた転写産物の低酸素応答の評価方法の構築」では、低酸素応答転写物の特徴をさらに把握するために FANTOM-CAT をはじめとするオ

ープンデータを活用して探索解析した。低酸素刺激により ribosomal RNA に関与する遺伝子群の発現低下が生じ、ミトコンドリア DNA 由来の転写物の発現量が低下することと、低酸素応答遺伝子群のアンチセンスに位置する転写物の低酸素応答性を HNf-score にて算出して提示した。

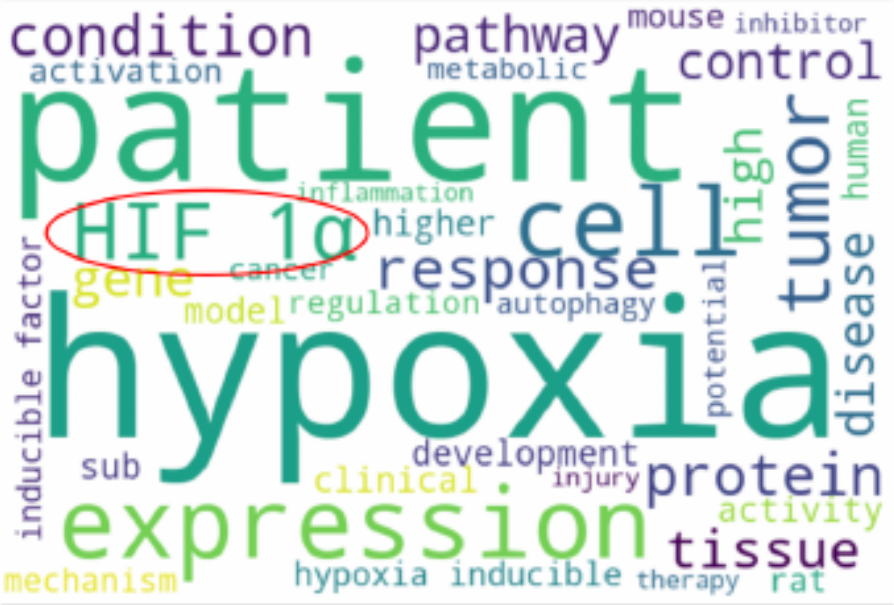
4 結論

遺伝子の注目度にはバイアスがあり、よく知られている遺伝子ほどよく理解される。昨今ではコンピュータの処理能力が向上し、活用できるデータ量も増えた。知見の少ない遺伝子についてはデータドリブンに解析する必要があると考える。

本研究ではデータドリブンに新規低酸素応答遺伝子を見出し、ncRNA を含めた転写産物の低酸素応答の評価方法を提示した。大規模な遺伝子発現データ解析を起点とする仮説の創出は、必要以上の実験操作を省略することが期待できる。特に以下の条件を満たす解析については検討の価値があると考える。(1)低酸素刺激応答の研究分野における HIF1A のような、当該研究分野での代表的な転写因子があること、(2)Gene Ontology にカテゴライズされる程よく研究されている研究分野であること、(3)遺伝子発現データを取得しやすい分野であることである。

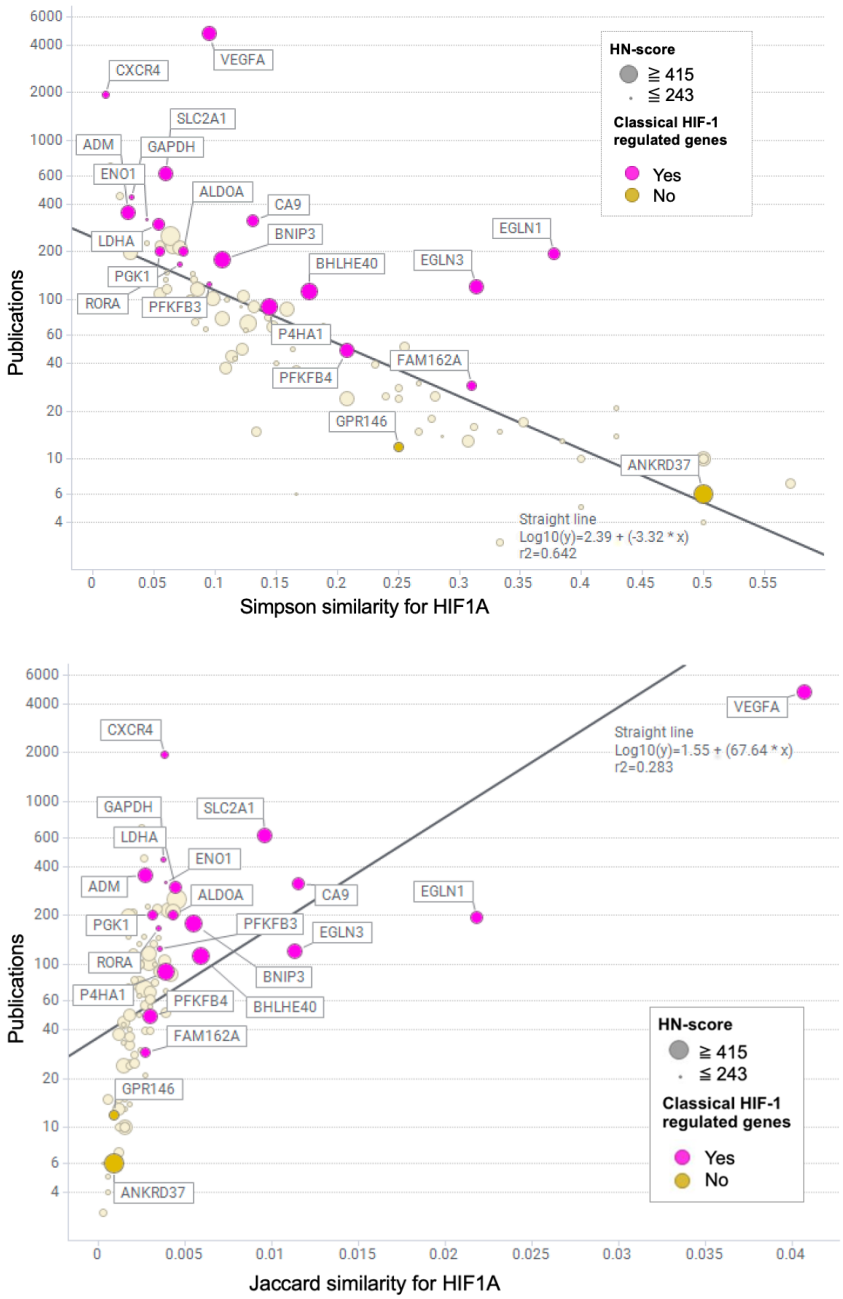
本研究については計算したデータは全て figshare にて誰でも利用可能な状態にて共有している 3.1 章「新規低酸素応答遺伝子の探索」については、https://doi.org/10.6084/m9.figshare.c.5323769.v1 にデータを登録しており、オープンアクセスなジャーナルに論文投稿済み(Ono and Bono 2021)である。および 3.2 章「ncRNA を含めた転写産物の低酸素応答の評価方法の構築」のデータは https://doi.org/10.6084/m9.figshare.c.5971218.v2 からアクセス可能であり、こちらは bioRxiv にプレプリントを投稿済みである(Ono and Bono 2022)。この解析手法は、この研究に限らずヒト RNA-Seq 解析において応用可能である。これらの研究は、低酸素応答研究のみならず新規パスウェイのデータドリブンな探索手法の提示の点で貢献したと考える。

5 付録



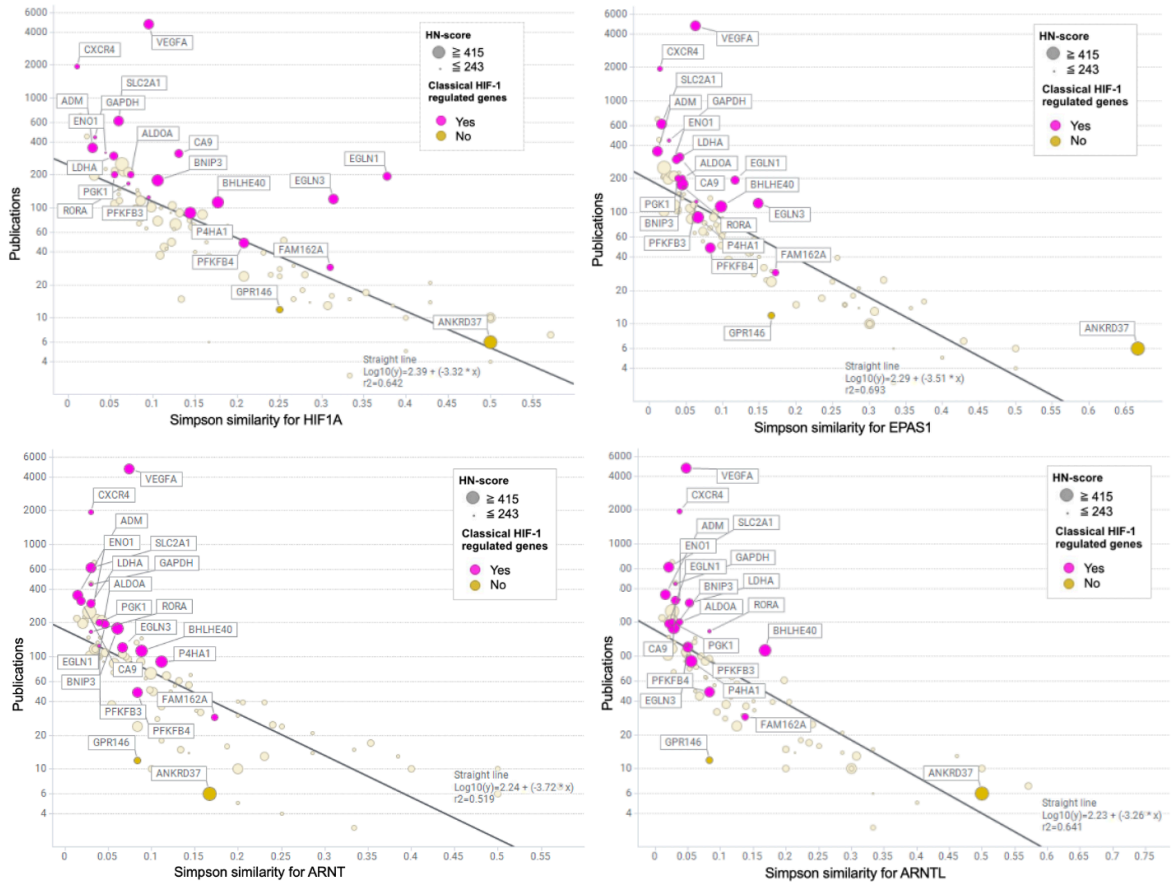
Supplemental figure 1 “hypoxia”をクエリとした論文情報をもとに作成した Word cloud

低酸素研究分野の代表的な遺伝子の評価には、word cloud による PubMed の 500 報のアブストラクトの可視化結果を参考とした。Biopython パッケージ（version 1.78）、scispacy パッケージ(version 0.4.0), spacy パッケージ（version 2.2.4), en-core-sci-sm-0.4.0 モデルを使用し、“hypoxia”をクエリとして頻出する単語を可視化した。解析は 2021 年 4 月 11 日に実行した。その結果、HIF 1α が唯一遺伝子名として表記された。解析の仕様上、コードを実行させるごとに可視化結果が変わるが、複数回検討したが同様に HIF 1α のみが遺伝子として表示された。低酸素を代表する遺伝子としてその Gene symbol である HIF1A を採用した。



Supplemental figure 2 bibliome 解析における HIF1A との Simpson 係数、Jaccard 係数の評価

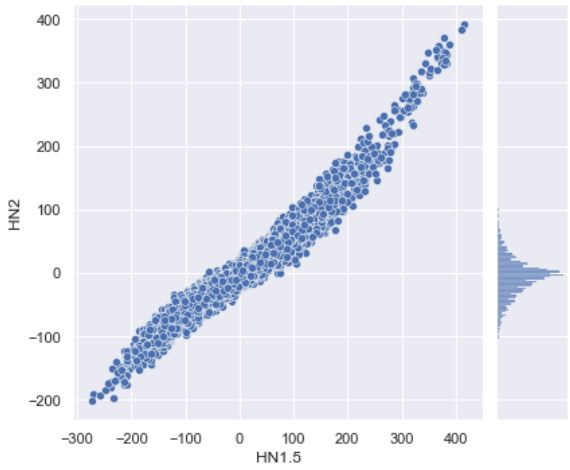
類似度を算出するに際には、Simpson 係数だけでなく、Jaccard 係数も検討した。Jaccard 係数は、比較対象の二つの集団の和集合を分母に持つため、今回のデータセットの場合は比較対象の HIF1A の遺伝子数の多さが影響して類似度全体の値が小さくなりすぎたため、可視化による選抜に適さないと判断した。Simpson 係数は各集合の共通要素を重視した類似度である。Jaccard 係数より Simpson 係数の方が HIF1A と任意の遺伝子 X に該当する論文数の差の大きさによる影響を受けにくく、可視化に適するスコアと判断した。



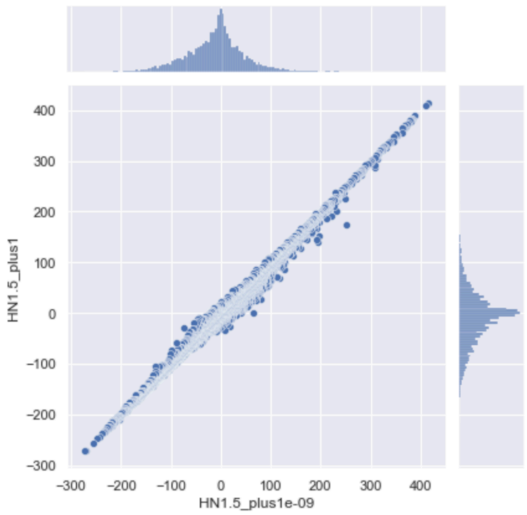
Supplemental figure 3 書誌解析における HIF1A,EPAS1, ARNT, ARNTL の Simpson 類似度の評価

HIF1A 以外にも、EPAS1 や ARNT、ARNTL といった低酸素との関連性が知られている遺伝子との Simpson 類似度を任意の遺伝子 X に対し算出した。縦軸に各遺伝子の論文数、横軸に各類似度を示した。いずれも同様の結果を示したが、HIF1A が低酸素研究分野の代表的な遺伝子であることから、新規低酸素応答性遺伝子の探索には HIF1A との類似度を採用した。（左上：HIF1A，右上：EPAS1，左下:ARNT，右下:ARNTL）

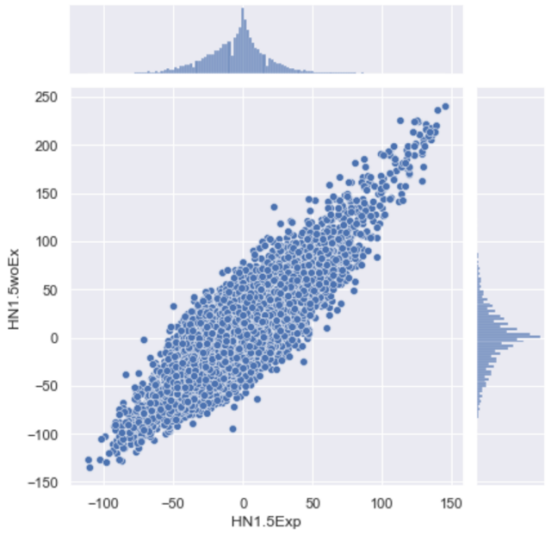
A


HN2
HN1.5
400
300
200
100
0
−100
−200
−300
−200
−100
0
100
200
300
400

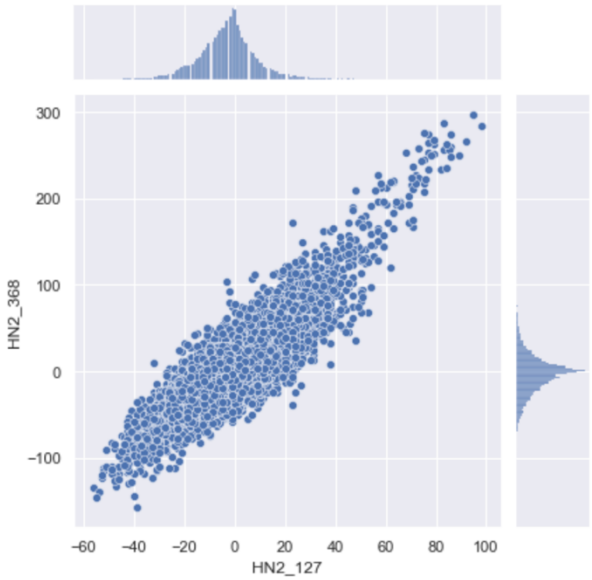
B


HN1.5_plus1
HN1.5_plus1e-09
400
300
200
100
0
−100
−200
−300
−300
−200
−100
0
100
200
300
400

C

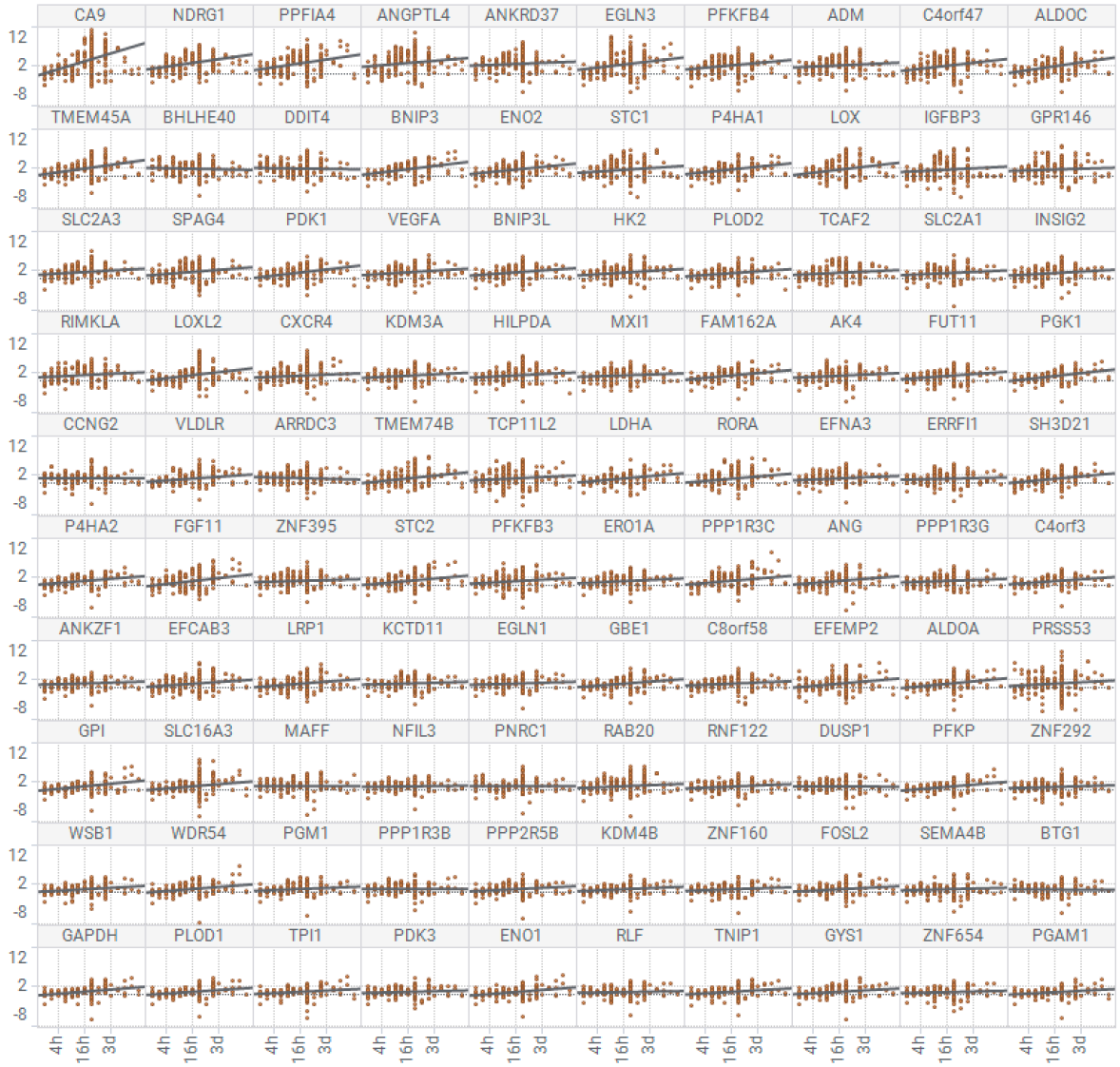

HN1.5woEx
HN1.5Exp
250
200
150
100
50
0
−50
−100
−150
−100
−50
0
50
100
150

D



Supplemental figure 4 HN-ratio, HN-score の条件検討

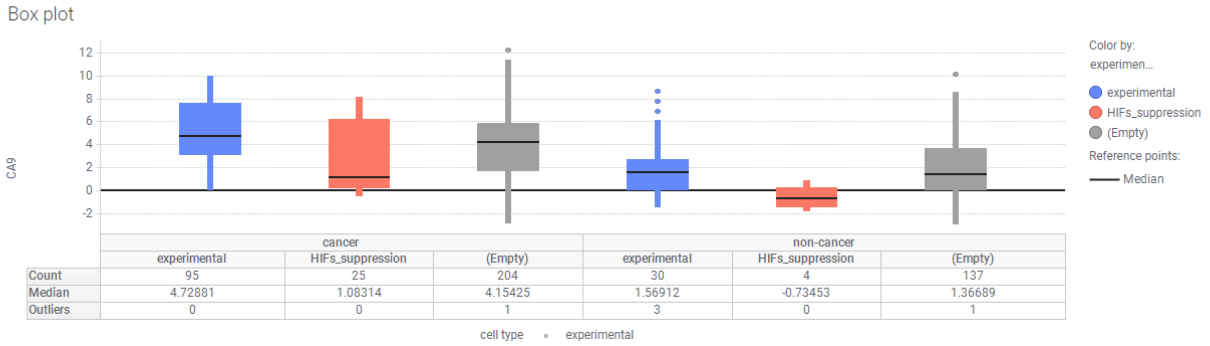
(A)遺伝子発現変動比の閾値(1.5 倍, 2 倍)の比較、(B)HN-ratio 算出の際の分母と分子に加算する値(＋1, ＋1e-09)の HN-score に対する影響の評価の結果を示す。(C)薬剤処置などの低酸素条件以外の実験介入の影響の評価。縦軸 HN1.5woEx は低酸素条件以外の実験介入がないサンプルの HN-score の各遺伝子の分布、横軸 HN1.5Exp は通常酸素条件、低酸素条件どちらのサンプルにも同様に薬剤処置などの実験介入があったサンプルの HN-score の各遺伝子の分布を示した。(D) 先行研究で蓄積されていたサンプルデータ (横軸, 127 HN-pairs)と、今回マニュアルキュレーションで新たに収集したサンプルデータ(縦軸, 368 HN-pairs)を元に算出した HN-score の分布の評価。A, B ではほぼ違いが見られないこと、C, D では異なるサンプル由来の HN-paird であるにもかかわらず同様のパターンであることを確認した。可視化には Jupyter-notebook を用いた。



Supplemental figure 5　UP100 gene list の低酸素処置時間別の HN-ratio

低酸素処置時間の記載があるサンプルに限定して、平均 HN-ratio 順に上段左から右に並べて表示した。HN-ratio は底 2 の Log 変換したものを使用した。

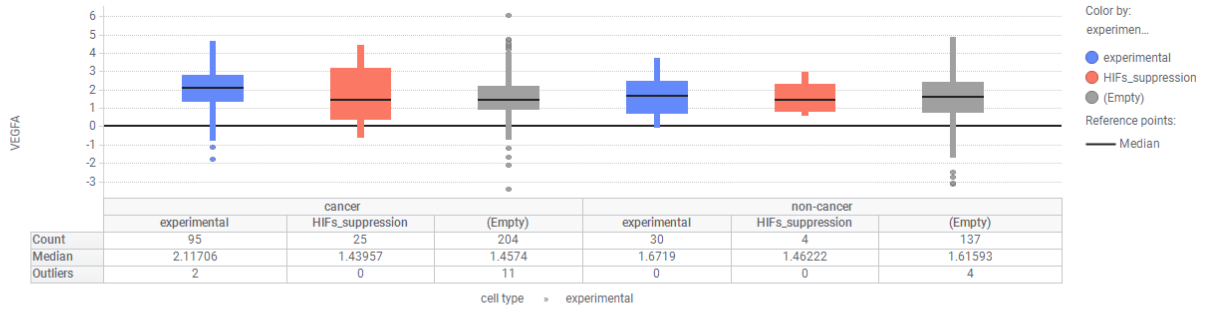
A　　*CA9*



	cancer			non-cancer		
	experimental	HIFs_suppression	(Empty)	experimental	HIFs_suppression	(Empty)
Count	95	25	204	30	4	137
Median	4.72881	1.08314	4.15425	1.56912	-0.73453	1.36689
Outliers	0	0	1	3	0	1

cell type » experimental

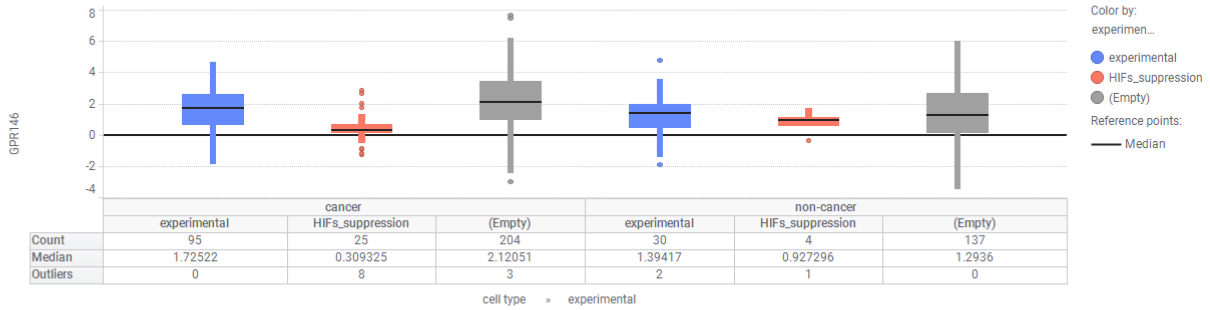
B　　*VEGFA*



	cancer			non-cancer		
	experimental	HIFs_suppression	(Empty)	experimental	HIFs_suppression	(Empty)
Count	95	25	204	30	4	137
Median	2.11706	1.43957	1.4574	1.6719	1.46222	1.61593
Outliers	2	0	11	0	0	4

cell type » experimental

C　　*GPR146*



	cancer			non-cancer		
	experimental	HIFs_suppression	(Empty)	experimental	HIFs_suppression	(Empty)
Count	95	25	204	30	4	137
Median	1.72522	0.309325	2.12051	1.39417	0.927296	1.2936
Outliers	0	8	3	2	1	0

cell type » experimental

Supplemental figure 6　HN-ratio（底 2 の Log 変換）の層別化解析

（A）*CA9*,（B）*VGEFA*,（B）*GPR146* のがん由来、非がん由来の違いと、実験介入あり（experimental），実験介入の中でも HIFs の抑制実験(HIFs_suppression)、実験介入なし（Empty）ごとの HN-ratio を可視化した。

Supplemental table 1 HNg-score を元に選抜した UP 200 gene および DOWN 200 gene list

UP 100 gene list および DOWN 100 gene list はこのリストの上位 100 位を採用した。

UP 200 gene list

	Symbol	description	HNg-score
1	ANKRD37	ankyrin repeat domain 37	415
2	NDRG1	N-myc downstream regulated 1	410
3	P4HA1	prolyl 4-hydroxylase subunit alpha 1	389
4	BNIP3L	BCL2 interacting protein 3 like	383
5	BHLHE40	basic helix-loop-helix family member e40	383
6	BNIP3	BCL2 interacting protein 3	382
7	VEGFA	vascular endothelial growth factor A	381
8	DDIT4	DNA damage inducible transcript 4	380
9	PPFIA4	PTPRF interacting protein alpha 4	379
10	SLC2A3	solute carrier family 2 member 3	377
11	ENO2	enolase 2	376
12	ADM	adrenomedullin	373
13	HK2	hexokinase 2	372
14	ANGPTL4	angiopoietin like 4	369
15	C4orf47	chromosome 4 open reading frame 47	367
16	SLC2A1	solute carrier family 2 member 1	366
17	ALDOC	aldolase, fructose-bisphosphate C	365
18	EGLN3	egl-9 family hypoxia inducible factor 3	365
19	PFKFB4	6-phosphofructo-2-kinase/fructose-2,6-biphosphatase 4	364
20	KDM3A	lysine demethylase 3A	356

21	PLOD2	procollagen-lysine,2-oxoglutarate 5-dioxygenase 2	353
22	MXI1	MAX interactor 1, dimerization protein	351
23	AK4	adenylate kinase 4	350
24	TMEM45A	transmembrane protein 45A	349
25	CA9	carbonic anhydrase 9	343
26	CCNG2	cyclin G2	339
27	EGLN1	egl-9 family hypoxia inducible factor 1	339
28	PDK1	pyruvate dehydrogenase kinase 1	336
29	STC1	stanniocalcin 1	335
30	LDHA	lactate dehydrogenase A	334
31	EFNA3	ephrin A3	329
32	INSIG2	insulin induced gene 2	328
33	FUT11	fucosyltransferase 11	328
34	SPAG4	sperm associated antigen 4	326
35	TCAF2	TRPM8 channel associated factor 2	326
36	FAM162A	family with sequence similarity 162 member A	323
37	NFIL3	nuclear factor, interleukin 3 regulated	322
38	PGK1	phosphoglycerate kinase 1	322
39	P4HA2	prolyl 4-hydroxylase subunit alpha 2	322
40	ERO1A	endoplasmic reticulum oxidoreductase 1 alpha	321
41	GPR146	G protein-coupled receptor 146	321
42	ALDOA	aldolase, fructose-bisphosphate A	319
43	ARRDC3	arrestin domain containing 3	317
44	ANKZF1	ankyrin repeat and zinc finger peptidyl tRNA hydrolase 1	316
45	SH3D21	SH3 domain containing 21	316

46	C4orf3	chromosome 4 open reading frame 3	314
47	ERRFI1	ERBB receptor feedback inhibitor 1	312
48	GBE1	1,4-alpha-glucan branching enzyme 1	311
49	ZNF395	zinc finger protein 395	311
50	C8orf58	chromosome 8 open reading frame 58	307
51	LOX	lysyl oxidase	306
52	VLDLR	very low density lipoprotein receptor	303
53	HILPDA	hypoxia inducible lipid droplet associated	301
54	ANG	angiogenin	298
55	KCTD11	potassium channel tetramerization domain containing 11	297
56	CXCR4	C-X-C motif chemokine receptor 4	293
57	PNRC1	proline rich nuclear receptor coactivator 1	293
58	IGFBP3	insulin like growth factor binding protein 3	291
59	PPP1R3G	protein phosphatase 1 regulatory subunit 3G	291
60	RIMKLA	ribosomal modification protein rimK like family member A	287
61	STC2	stanniocalcin 2	286
62	PGM1	phosphoglucomutase 1	285
63	MAFF	MAF bZIP transcription factor F	285
64	PFKP	phosphofructokinase, platelet	283
65	SLC16A3	solute carrier family 16 member 3	281
66	KDM4B	lysine demethylase 4B	280
67	LRP1	LDL receptor related protein 1	279
68	ZNF292	zinc finger protein 292	279
69	FGF11	fibroblast growth factor 11	278
70	RNF122	ring finger protein 122	277

71	GPI	glucose-6-phosphate isomerase	276
72	TNIP1	TNFAIP3 interacting protein 1	275
73	FOSL2	FOS like 2, AP-1 transcription factor subunit	274
74	LOXL2	lysyl oxidase like 2	274
75	ZNF160	zinc finger protein 160	274
76	WSB1	WD repeat and SOCS box containing 1	273
77	WDR54	WD repeat domain 54	273
78	EFCAB3	EF-hand calcium binding domain 3	273
79	PPP1R3C	protein phosphatase 1 regulatory subunit 3C	272
80	PFKFB3	6-phosphofructo-2-kinase/fructose-2,6-biphosphatase 3	270
81	PLOD1	procollagen-lysine,2-oxoglutarate 5-dioxygenase 1	270
82	TMEM74B	transmembrane protein 74B	267
83	DUSP1	dual specificity phosphatase 1	266
84	GAPDH	glyceraldehyde-3-phosphate dehydrogenase	266
85	TCP11L2	t-complex 11 like 2	264
86	TPI1	triosephosphate isomerase 1	262
87	BTG1	BTG anti-proliferation factor 1	260
88	RORA	RAR related orphan receptor A	260
89	SEMA4B	semaphorin 4B	256
90	RLF	RLF zinc finger	255
91	PPP2R5B	protein phosphatase 2 regulatory subunit B'beta	254
92	PPP1R3B	protein phosphatase 1 regulatory subunit 3B	253
93	PDK3	pyruvate dehydrogenase kinase 3	251
94	EFEMP2	EGF containing fibulin extracellular matrix protein 2	251
95	ZNF654	zinc finger protein 654	251

96	ENO1	enolase 1	249
97	RAB20	RAB20, member RAS oncogene family	248
98	PGAM1	phosphoglycerate mutase 1	246
99	PRSS53	serine protease 53	245
100	GYS1	glycogen synthase 1	243
101	B3GNT4	UDP-GlcNAc:betaGal beta-1,3-N-acetylglucosaminyltransferase 4	242
102	YEATS2	YEATS domain containing 2	239
103	PLEKHA2	pleckstrin homology domain containing A2	239
104	WFDC3	WAP four-disulfide core domain 3	239
105	PTPRH	protein tyrosine phosphatase receptor type H	238
106	STBD1	starch binding domain 1	237
107	UPK1A	uroplakin 1A	237
108	LRP2BP	LRP2 binding protein	237
109	APLN	apelin	235
110	CIART	circadian associated repressor of transcription	235
111	FAM47E-STBD1	FAM47E-STBD1 readthrough	235
112	JUN	Jun proto-oncogene, AP-1 transcription factor subunit	234
113	ISG20	interferon stimulated exonuclease gene 20	233
114	PAM	peptidylglycine alpha-amidating monooxygenase	232
115	VKORC1	vitamin K epoxide reductase complex subunit 1	232
116	FANK1	fibronectin type III and ankyrin repeat domains 1	232
117	AHNAK2	AHNAK nucleoprotein 2	232
118	NOG	noggin	231
119	SEC61G	SEC61 translocon subunit gamma	231

120	NCKIPSD	NCK interacting protein with SH3 domain	231
121	FOS	Fos proto-oncogene, AP-1 transcription factor subunit	229
122	SPRY1	sprouty RTK signaling antagonist 1	229
123	VPS37D	VPS37D subunit of ESCRT-I	229
124	BHLHE41	basic helix-loop-helix family member e41	228
125	KDM7A	lysine demethylase 7A	228
126	SERPINE1	serpin family E member 1	227
127	BMT2	base methyltransferase of 25S rRNA 2 homolog	227
128	ACAP1	ArfGAP with coiled-coil, ankyrin repeat and PH domains 1	226
129	ARFGEF3	ARFGEF family member 3	226
130	RRAGD	Ras related GTP binding D	226
131	RBPJ	recombination signal binding protein for immunoglobulin kappa J region	225
132	CDKN1C	cyclin dependent kinase inhibitor 1C	224
133	CSRNP1	cysteine and serine rich nuclear protein 1	224
134	LNPK	lunapark, ER junction formation factor	224
135	AMPD3	adenosine monophosphate deaminase 3	223
136	PRTN3	proteinase 3	223
137	YPEL1	yippee like 1	223
138	NYAP1	neuronal tyrosine phosphorylated phosphoinositide-3-kinase adaptor 1	223
139	SAP30	Sin3A associated protein 30	222
140	PPP1R13L	protein phosphatase 1 regulatory subunit 13 like	220
141	PTPRN	protein tyrosine phosphatase receptor type N	219
142	RNASE4	ribonuclease A family member 4	217
143	THAP8	THAP domain containing 8	217
144	CSRP2	cysteine and glycine rich protein 2	216

145	GPRC5A	G protein-coupled receptor class C group 5 member A	215
146	BCKDHA	branched chain keto acid dehydrogenase E1 subunit alpha	214
147	GNRH1	gonadotropin releasing hormone 1	214
148	SFXN3	sideroflexin 3	214
149	ITGA5	integrin subunit alpha 5	213
150	MPI	mannose phosphate isomerase	213
151	DPYSL4	dihydropyrimidinase like 4	213
152	NARF	nuclear prelamin A recognition factor	213
153	CREBRF	CREB3 regulatory factor	213
154	C2orf72	chromosome 2 open reading frame 72	213
155	PKM	pyruvate kinase M1/2	212
156	FAM13A	family with sequence similarity 13 member A	212
157	PFKL	phosphofructokinase, liver type	211
158	HERC3	HECT and RLD domain containing E3 ubiquitin protein ligase 3	211
159	RNF24	ring finger protein 24	211
160	TRIM9	tripartite motif containing 9	211
161	PLIN2	perilipin 2	209
162	HSF4	heat shock transcription factor 4	209
163	KDM4C	lysine demethylase 4C	209
164	KDM6B	lysine demethylase 6B	209
165	SNX33	sorting nexin 33	209
166	DPCD	deleted in primary ciliary dyskinesia homolog (mouse)	208
167	GALNT18	polypeptide N-acetylgalactosaminyltransferase 18	208
168	FOXD1	forkhead box D1	207
169	TNFRSF10D	TNF receptor superfamily member 10d	207

170	ARHGEF37	Rho guanine nucleotide exchange factor 37	207
171	RASSF7	Ras association domain family member 7	206
172	FAM110C	family with sequence similarity 110 member C	206
173	IER3	immediate early response 3	204
174	DNAJB2	DnaJ heat shock protein family (Hsp40) member B2	203
175	PTGS1	prostaglandin-endoperoxide synthase 1	203
176	S1PR4	sphingosine-1-phosphate receptor 4	203
177	ISM2	isthmin 2	203
178	ATF3	activating transcription factor 3	202
179	ALKBH5	alkB homolog 5, RNA demethylase	201
180	PNCK	pregnancy up-regulated nonubiquitous CaM kinase	200
181	MAP3K15	mitogen-activated protein kinase kinase kinase 15	200
182	S100A10	S100 calcium binding protein A10	199
183	RIOK3	RIO kinase 3	198
184	FAM210A	family with sequence similarity 210 member A	198
185	RCOR2	REST corepressor 2	197
186	DSP	desmoplakin	195
187	INHA	inhibin subunit alpha	195
188	PIK3IP1	phosphoinositide-3-kinase interacting protein 1	195
189	PRELID2	PRELI domain containing 2	195
190	NIM1K	NIM1 serine/threonine protein kinase	195
191	TMEM91	transmembrane protein 91	195
192	CLK3	CDC like kinase 3	194
193	QSOX1	quiescin sulfhydryl oxidase 1	194
194	TMEM158	transmembrane protein 158	194

195	CLK1	CDC like kinase 1	193
196	PHF21A	PHD finger protein 21A	193
197	NDUFA4L2	NDUFA4 mitochondrial complex associated like 2	193
198	ANGPTL6	angiopoietin like 6	193
199	TBC1D3L	TBC1 domain family member 3L	193
200	RSBN1	round spermatid basic protein 1	192

DOWN 200 gene list

	Symbol	description	HNg-score
1	RRS1	ribosome biogenesis regulator 1 homolog	-272
2	CHAC2	ChaC glutathione specific gamma-glutamylcyclotransferase 2	-271
3	GPATCH4	G-patch domain containing 4	-257
4	MT-ND2	mitochondrially encoded NADH dehydrogenase 2	-247
5	SFXN2	sideroflexin 2	-244
6	NOL6	nucleolar protein 6	-238
7	MARS2	methionyl-tRNA synthetase 2, mitochondrial	-237
8	MT-ATP8	mitochondrially encoded ATP synthase 8	-235
9	COA7	cytochrome c oxidase assembly factor 7	-234
10	MT-ND6	mitochondrially encoded NADH dehydrogenase 6	-232
11	TAF9B	TATA-box binding protein associated factor 9b	-230
12	DDX28	DEAD-box helicase 28	-227
13	SLC25A35	solute carrier family 25 member 35	-225
14	MON1A	MON1 homolog A, secretory trafficking associated	-225
15	CTU2	cytosolic thiouridylase subunit 2	-224

16	POP1	POP1 homolog, ribonuclease P/MRP subunit	-223
17	MPV17L2	MPV17 mitochondrial inner membrane protein like 2	-221
18	NOP2	NOP2 nucleolar protein	-221
19	RTN4IP1	reticulon 4 interacting protein 1	-220
20	MT-ND1	mitochondrially encoded NADH dehydrogenase 1	-219
21	TOR3A	torsin family 3 member A	-219
22	PNO1	partner of NOB1 homolog	-216
23	RRP9	ribosomal RNA processing 9, U3 small nucleolar RNA binding protein	-216
24	POLE2	DNA polymerase epsilon 2, accessory subunit	-215
25	CD3EAP	RNA polymerase I subunit G	-215
26	CTPS1	CTP synthase 1	-215
27	PRMT3	protein arginine methyltransferase 3	-213
28	TAP2	transporter 2, ATP binding cassette subfamily B member	-213
29	URB2	URB2 ribosome biogenesis homolog	-213
30	LTV1	LTV1 ribosome biogenesis factor	-213
31	POLR3K	RNA polymerase III subunit K	-212
32	HPDL	4-hydroxyphenylpyruvate dioxygenase like	-212
33	NOL12	nucleolar protein 12	-210
34	PNPT1	polyribonucleotide nucleotidyltransferase 1	-210
35	FDXACB1	ferredoxin-fold anticodon binding domain containing 1	-210
36	NOP16	NOP16 nucleolar protein	-209
37	PYCR3	pyrroline-5-carboxylate reductase 3	-209
38	SCLY	selenocysteine lyase	-209
39	TMEM138	transmembrane protein 138	-209
40	TOE1	target of EGR1, exonuclease	-209

41	WDR77	WD repeat domain 77	-209
42	UTP15	UTP15 small subunit processome component	-208
43	CDC25A	cell division cycle 25A	-208
44	FAM86C1	family with sequence similarity 86 member C1, pseudogene	-208
45	SLC5A6	solute carrier family 5 member 6	-207
46	TIPIN	TIMELESS interacting protein	-207
47	WDR4	WD repeat domain 4	-207
48	NDUFAF4	NADH:ubiquinone oxidoreductase complex assembly factor 4	-206
49	DOLK	dolichol kinase	-206
50	PSEN2	presenilin 2	-205
51	SNAPC5	small nuclear RNA activating complex polypeptide 5	-205
52	COQ3	coenzyme Q3, methyltransferase	-205
53	DCTPP1	dCTP pyrophosphatase 1	-205
54	EEF2KMT	eukaryotic elongation factor 2 lysine methyltransferase	-205
55	MT-CO3	mitochondrially encoded cytochrome c oxidase III	-204
56	METTL1	methyltransferase like 1	-204
57	MYBBP1A	MYB binding protein 1a	-203
58	GRWD1	glutamate rich WD repeat containing 1	-203
59	EIF2B3	eukaryotic translation initiation factor 2B subunit gamma	-202
60	MT-ATP6	mitochondrially encoded ATP synthase 6	-199
61	SLC43A2	solute carrier family 43 member 2	-198
62	NIP7	nucleolar pre-rRNA processing protein NIP7	-197
63	SCFD2	sec1 family domain containing 2	-196
64	ABCF2	ATP binding cassette subfamily F member 2	-196
65	HGH1	HGH1 homolog	-196

66	PDSS1	decaprenyl diphosphate synthase subunit 1	-195
67	PSME3	proteasome activator subunit 3	-195
68	UTP20	UTP20 small subunit processome component	-195
69	FARSB	phenylalanyl-tRNA synthetase subunit beta	-195
70	MT-ND3	mitochondrially encoded NADH dehydrogenase 3	-194
71	RIOX1	ribosomal oxygenase 1	-194
72	RRP7A	ribosomal RNA processing 7 homolog A	-194
73	NOP56	NOP56 ribonucleoprotein	-193
74	RPP40	ribonuclease P/MRP subunit p40	-193
75	GINS3	GINS complex subunit 3	-193
76	ATAD3A	ATPase family AAA domain containing 3A	-192
77	PPIL1	peptidylprolyl isomerase like 1	-192
78	PUS1	pseudouridine synthase 1	-192
79	ZNF689	zinc finger protein 689	-192
80	POLR1B	RNA polymerase I subunit B	-191
81	LSM10	LSM10, U7 small nuclear RNA associated	-191
82	PRMT6	protein arginine methyltransferase 6	-190
83	TPK1	thiamin pyrophosphokinase 1	-190
84	ZNHIT2	zinc finger HIT-type containing 2	-190
85	EXOSC4	exosome component 4	-190
86	PAK1IP1	PAK1 interacting protein 1	-189
87	PGAM5	PGAM family member 5, mitochondrial serine/threonine protein phosphatase	-189
88	SLC27A4	solute carrier family 27 member 4	-189
89	MCM10	minichromosome maintenance 10 replication initiation factor	-189
90	TMEM177	transmembrane protein 177	-188

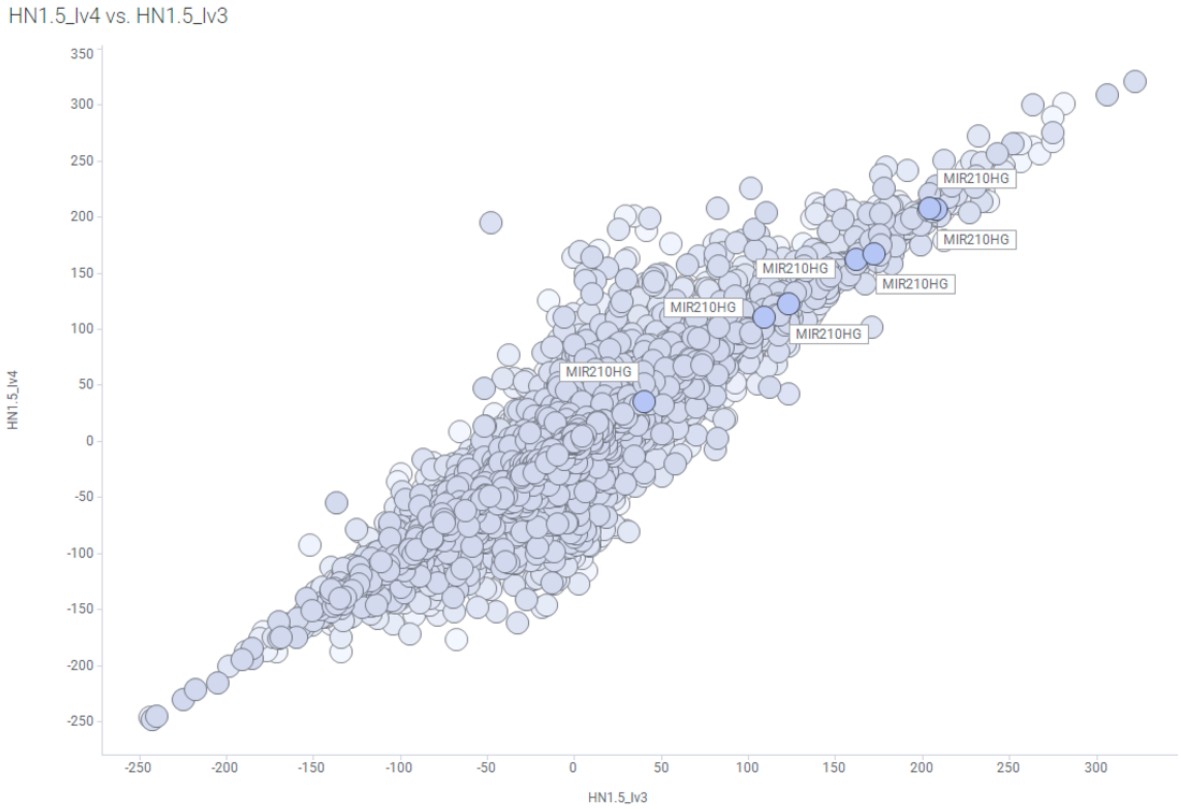
91	DSCC1	DNA replication and sister chromatid cohesion 1	-188
92	ATP5MC1	ATP synthase membrane subunit c locus 1	-187
93	AUNIP	aurora kinase A and ninein interacting protein	-187
94	UFSP1	UFM1 specific peptidase 1 (inactive)	-187
95	ODC1	ornithine decarboxylase 1	-186
96	TMEM201	transmembrane protein 201	-186
97	DIMT1	DIMT1 rRNA methyltransferase and ribosome maturation factor	-186
98	ELAC2	elaC ribonuclease Z 2	-186
99	PSMG1	proteasome assembly chaperone 1	-185
100	RUVBL1	RuvB like AAA ATPase 1	-185
101	SRM	spermidine synthase	-185
102	UBIAD1	UbiA prenyltransferase domain containing 1	-185
103	NOLC1	nucleolar and coiled-body phosphoprotein 1	-184
104	DAGLA	diacylglycerol lipase alpha	-184
105	GFM1	G elongation factor mitochondrial 1	-184
106	RFC3	replication factor C subunit 3	-183
107	TEDC2	tubulin epsilon and delta complex 2	-183
108	TOMM40L	translocase of outer mitochondrial membrane 40 like	-183
109	ZNF30	zinc finger protein 30	-183
110	E2F2	E2F transcription factor 2	-183
111	AIMP2	aminoacyl tRNA synthetase complex interacting multifunctional protein 2	-183
112	MCIDAS	multiciliate differentiation and DNA synthesis associated cell cycle protein	-182
113	CHRNA5	cholinergic receptor nicotinic alpha 5 subunit	-181
114	DPH2	diphthamide biosynthesis 2	-181
115	ALDH1B1	aldehyde dehydrogenase 1 family member B1	-181

116	FLAD1	flavin adenine dinucleotide synthetase 1	-181
117	DDX21	DExD-box helicase 21	-180
118	GTF2H2	general transcription factor IIH subunit 2	-180
119	PFAS	phosphoribosylformylglycinamidine synthase	-179
120	PIGW	phosphatidylinositol glycan anchor biosynthesis class W	-179
121	TELO2	telomere maintenance 2	-179
122	CDCA7	cell division cycle associated 7	-179
123	LYSMD2	LysM domain containing 2	-179
124	MAK16	MAK16 homolog	-179
125	MRPL20	mitochondrial ribosomal protein L20	-178
126	MRPL4	mitochondrial ribosomal protein L4	-178
127	TIMM8A	translocase of inner mitochondrial membrane 8A	-178
128	HERC6	HECT and RLD domain containing E3 ubiquitin protein ligase family member 6	-178
129	HMBS	hydroxymethylbilane synthase	-178
130	ADAT1	adenosine deaminase tRNA specific 1	-177
131	DTD2	D-aminoacyl-tRNA deacylase 2	-177
132	FOXRED2	FAD dependent oxidoreductase domain containing 2	-177
133	ABCG2	ATP binding cassette subfamily G member 2 (Junior blood group)	-177
134	METTL8	methyltransferase like 8	-177
135	SPSB2	splA/ryanodine receptor domain and SOCS box containing 2	-176
136	BRIX1	biogenesis of ribosomes BRX1	-176
137	TWNK	twinkle mtDNA helicase	-176
138	CDC45	cell division cycle 45	-176
139	JMJD4	jumonji domain containing 4	-176

140	LETM1	leucine zipper and EF-hand containing transmembrane protein 1	-176
141	POP5	POP5 homolog, ribonuclease P/MRP subunit	-175
142	BEND3	BEN domain containing 3	-175
143	SLC25A19	solute carrier family 25 member 19	-175
144	UTP14A	UTP14A small subunit processome component	-175
145	WDR62	WD repeat domain 62	-174
146	EXOG	exo/endonuclease G	-174
147	MRPL12	mitochondrial ribosomal protein L12	-173
148	ARMC7	armadillo repeat containing 7	-173
149	BAG2	BAG cochaperone 2	-173
150	SPRTN	SprT-like N-terminal domain	-173
151	KTI12	KTI12 chromatin associated homolog	-173
152	PTRH1	peptidyl-tRNA hydrolase 1 homolog	-172
153	TIMM10	translocase of inner mitochondrial membrane 10	-172
154	ACTRT3	actin related protein T3	-172
155	DHRS11	dehydrogenase/reductase 11	-172
156	MRPS12	mitochondrial ribosomal protein S12	-171
157	MRTO4	MRT4 homolog, ribosome maturation factor	-171
158	SLC35G1	solute carrier family 35 member G1	-171
159	EEF1E1	eukaryotic translation elongation factor 1 epsilon 1	-171
160	ALDH4A1	aldehyde dehydrogenase 4 family member A1	-171
161	GINS1	GINS complex subunit 1	-171
162	HNRNPAB	heterogeneous nuclear ribonucleoprotein A/B	-171
163	MTHFD1	methylenetetrahydrofolate dehydrogenase, cyclohydrolase and formyltetrahydrofolate synthetase 1	-170

164	PPARGC1B	PPARG coactivator 1 beta	-170
165	SLC19A1	solute carrier family 19 member 1	-170
166	TRMT61A	tRNA methyltransferase 61A	-170
167	ZNF239	zinc finger protein 239	-170
168	ADCK1	aarF domain containing kinase 1	-170
169	NSMCE4A	NSE4 homolog A, SMC5-SMC6 complex component	-169
170	C11orf98	chromosome 11 open reading frame 98	-169
171	CENPM	centromere protein M	-169
172	ANAPC7	anaphase promoting complex subunit 7	-169
173	PCNA	proliferating cell nuclear antigen	-168
174	RPS6KL1	ribosomal protein S6 kinase like 1	-168
175	RRP15	ribosomal RNA processing 15 homolog	-168
176	SPOUT1	SPOUT domain containing methyltransferase 1	-168
177	WDR46	WD repeat domain 46	-168
178	HSPBAP1	HSPB1 associated protein 1	-168
179	IMP4	IMP U3 small nucleolar ribonucleoprotein 4	-168
180	NPM3	nucleophosmin/nucleoplasmin 3	-167
181	POLR3H	RNA polymerase III subunit H	-167
182	RRP36	ribosomal RNA processing 36	-167
183	SFXN4	sideroflexin 4	-167
184	CLUH	clustered mitochondria homolog	-167
185	DKC1	dyskerin pseudouridine synthase 1	-167
186	LRR1	leucine rich repeat protein 1	-167
187	NOC4L	nucleolar complex associated 4 homolog	-166
188	OGFOD1	2-oxoglutarate and iron dependent oxygenase domain containing 1	-166

189	RCL1	RNA terminal phosphate cyclase like 1	-166
190	CCDC86	coiled-coil domain containing 86	-166
191	LYAR	Ly1 antibody reactive	-166
192	MRM1	mitochondrial rRNA methyltransferase 1	-165
193	POLR3B	RNA polymerase III subunit B	-165
194	WDR35	WD repeat domain 35	-165
195	ZNF593	zinc finger protein 593	-165
196	DUS1L	dihydrouridine synthase 1 like	-165
197	ELOVL6	ELOVL fatty acid elongase 6	-165
198	GEMIN5	gem nuclear organelle associated protein 5	-165
199	LCMT2	leucine carboxyl methyltransferase 2	-165
200	MCM6	minichromosome maintenance complex component 6	-165



Supplemental figure 7 FANTOM-CAT Robust (縦軸) と Stringent(横軸)の HNf-score の評価

MIR210HG のプロットについてラベルと色を強調して可視化した。いずれも HNf-score が正の値を示していること、各 HNf-score の分布が lv4 Robust と lv3 Stringent で大きく変わらないことを確認した。

Supplemental table 2　DOWN200 ncRNA metabolic process related group に該当する遺伝子

Metascape によるエンリッチメント解析結果を元に表を作成した。GO:0034660 ncRNA metabolic process を構成する member の Term の詳細を記した。

GroupID	Category	Term	Description	InTerm_InList	Symbols
1_Summary	GO Biological Processes	GO:0034660	ncRNA metabolic process	46/470	DKC1, METTL1, NOP2, RRP9, DDX21, NOLC1, WDR46, LCMT2, FARSB, RCL1, NPM3, NOP56, WDR4, RPP40, UTP14A, POP1, RRS1, ADAT1, DIMT1, UTP20, RRP7A, RRP15, MRTO4, POP5, POLR3K, EXOSC4, BRIX1, LYAR, ELAC2, DUS1L, NOL6, NOC4L, METTL8, MRM1, PUS1, UTP15, MAK16, RRP36, FDXACB1, IMP4, MARS2, DTD2, KTI12, TOE1, TRMT61A, CTU2, URB2, MYBBP1A, NIP7, NOP16, PAK1IP1, MRPL20, DDX28, MPV17L2, LTV1, RUVBL1, GEMIN5, GTF2H2, PSME3, PPIL1, PNO1, WDR77, NOL12, LSM10
1_Member	GO Biological Processes	GO:0034660	ncRNA metabolic process	46/470	DKC1, METTL1, NOP2, RRP9, DDX21, NOLC1, WDR46, LCMT2, FARSB, RCL1, NPM3, NOP56, WDR4, RPP40, UTP14A, POP1, RRS1, ADAT1, DIMT1, UTP20, RRP7A, RRP15, MRTO4, POP5, POLR3K, EXOSC4, BRIX1, LYAR, ELAC2, DUS1L, NOL6, NOC4L, METTL8, MRM1, PUS1, UTP15, MAK16, RRP36, FDXACB1, IMP4, MARS2,

					DTD2, KTI12, TOE1, TRMT61A, CTU2
1_Member	GO Biological Processes	GO:0034470	ncRNA processing	43/384	DKC1, METTL1, NOP2, RRP9, DDX21, NOLC1, WDR46, LCMT2, RCL1, NPM3, NOP56, WDR4, RPP40, UTP14A, POP1, RRS1, ADAT1, DIMT1, UTP20, RRP7A, RRP15, MRTO4, POP5, POLR3K, EXOSC4, BRIX1, LYAR, ELAC2, DUS1L, NOL6, NOC4L, METTL8, MRM1, PUS1, UTP15, MAK16, RRP36, FDXACB1, IMP4, KTI12, TOE1, TRMT61A, CTU2
1_Member	GO Biological Processes	GO:0042254	ribosome biogenesis	38/293	DKC1, NOP2, RRP9, DDX21, NOLC1, WDR46, URB2, RCL1, NPM3, MYBBP1A, NOP56, RPP40, UTP14A, RRS1, DIMT1, UTP20, RRP7A, RRP15, MRTO4, POP5, NIP7, NOP16, EXOSC4, PAK1IP1, MRPL20, BRIX1, LYAR, DDX28, NOL6, NOC4L, MRM1, UTP15, MAK16, MPV17L2, LTV1, RRP36, FDXACB1, IMP4
1_Member	GO Biological Processes	GO:0022613	ribonucleoprotein complex biogenesis	40/444	DKC1, NOP2, RUVBL1, RRP9, DDX21, NOLC1, WDR46, URB2, RCL1, NPM3, MYBBP1A, NOP56, RPP40, UTP14A, RRS1, GEMIN5, DIMT1, UTP20, RRP7A, RRP15, MRTO4, POP5, NIP7, NOP16, EXOSC4, PAK1IP1, MRPL20, BRIX1,

					LYAR, DDX28, NOL6, NOC4L, MRM1, UTP15, MAK16, MPV17L2, LTV1, RRP36, FDXACB1, IMP4
1_Member	GO Biological Processes	GO:0006364	rRNA processing	29/218	DKC1, NOP2, RRP9, DDX21, NOLC1, WDR46, RCL1, NPM3, NOP56, RPP40, UTP14A, RRS1, DIMT1, UTP20, RRP7A, RRP15, MRTO4, POP5, EXOSC4, BRIX1, LYAR, NOL6, NOC4L, MRM1, UTP15, MAK16, RRP36, FDXACB1, IMP4
1_Member	GO Biological Processes	GO:0016072	rRNA metabolic process	29/229	DKC1, NOP2, RRP9, DDX21, NOLC1, WDR46, RCL1, NPM3, NOP56, RPP40, UTP14A, RRS1, DIMT1, UTP20, RRP7A, RRP15, MRTO4, POP5, EXOSC4, BRIX1, LYAR, NOL6, NOC4L, MRM1, UTP15, MAK16, RRP36, FDXACB1, IMP4
1_Member	Reactome Gene Sets	R-HSA-8953854	Metabolism of RNA	39/673	DKC1, GTF2H2, METTL1, NOP2, RRP9, DDX21, WDR46, LCMT2, RCL1, PSME3, NOP56, WDR4, RPP40, UTP14A, POP1, ADAT1, GEMIN5, DIMT1, UTP20, RRP7A, POP5, NIP7, PPIL1, EXOSC4, PNO1, ELAC2, NOL6, NOC4L, WDR77, NOL12, MRM1, PUS1, UTP15, LTV1, LSM10, RRP36, IMP4, TRMT61A, CTU2
1_Member	Reactome Gene Sets	R-HSA-72312	rRNA processing	24/204	DKC1, NOP2, RRP9, DDX21, WDR46, RCL1, NOP56, RPP40, UTP14A, DIMT1, UTP20, RRP7A, NIP7,

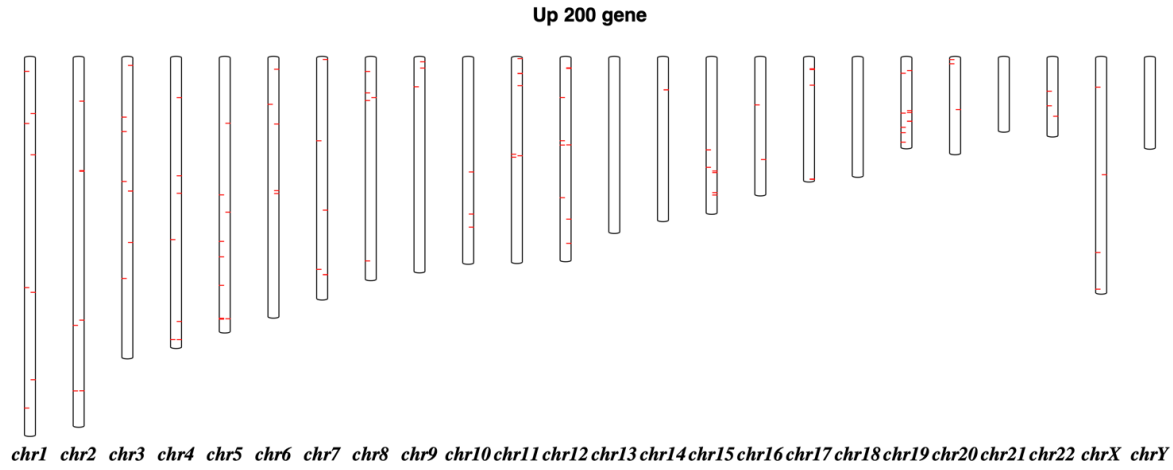
					EXOSC4, PNO1, ELAC2, NOL6, NOC4L, NOL12, MRM1, UTP15, LTV1, RRP36, IMP4
1_Member	Reactome Gene Sets	R-HSA-6790901	rRNA modification in the nucleus and cytosol	16/61	DKC1, NOP2, RRP9, WDR46, RCL1, NOP56, UTP14A, DIMT1, UTP20, RRP7A, PNO1, NOL6, NOC4L, UTP15, RRP36, IMP4
1_Member	Reactome Gene Sets	R-HSA-8868773	rRNA processing in the nucleus and cytosol	22/194	DKC1, NOP2, RRP9, DDX21, WDR46, RCL1, NOP56, RPP40, UTP14A, DIMT1, UTP20, RRP7A, NIP7, EXOSC4, PNO1, NOL6, NOC4L, NOL12, UTP15, LTV1, RRP36, IMP4
1_Member	Reactome Gene Sets	R-HSA-6791226	Major pathway of rRNA processing in the nucleolus and cytosol	19/184	RRP9, DDX21, WDR46, RCL1, NOP56, RPP40, UTP14A, UTP20, RRP7A, NIP7, EXOSC4, PNO1, NOL6, NOC4L, NOL12, UTP15, LTV1, RRP36, IMP4
1_Member	KEGG Pathway	hsa03008	Ribosome biogenesis in eukaryotes	11/109	DKC1, RCL1, NOP56, RPP40, UTP14A, POP1, RRP7A, POP5, NOL6, UTP15, IMP4

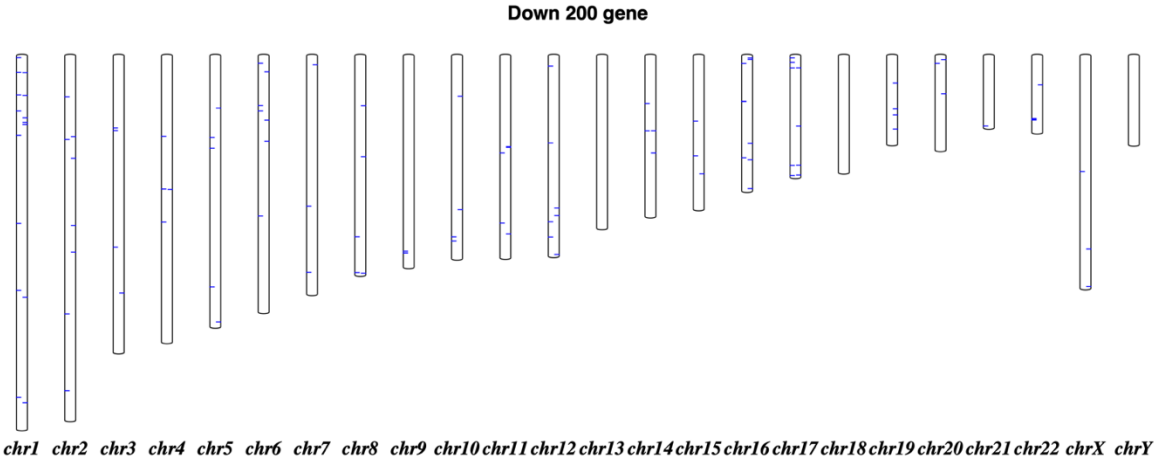
Supplemental table 3 染色体ごとの HNf-score の要約統計量

HNf-score はミトコンドリア DNA 由来の転写物において低値を示した。

Chr	Avg (HNf1.5)	StdDev (HNf1.5)	Max (HNf1.5)	Min (HNf1.5)	UniqueCount (transcript)	UniqueCount (gene)
chr1	-4.36	28.41	263	-246	25737	3080
chr2	-3.2	27.44	272	-175	19941	2183
chr3	-3.64	27.66	300	-156	16450	1723
chr4	-3.14	26.11	275	-157	10762	1265
chr5	-3.74	28	220	-169	12744	1533
chr6	-2.52	26.78	321	-176	14168	1778
chr7	-4.27	27.75	309	-149	12723	1516
chr8	-2.79	28.6	265	-248	9983	1171
chr9	-3.71	27.61	257	-148	10124	1218
chr10	-1.98	29.77	301	-149	10457	1244
chr11	-3.85	27.92	267	-187	16230	1664
chr12	-4	27.96	238	-147	15532	1643
chr13	-4.15	24.77	161	-136	4748	631
chr14	-4.14	26.69	246	-140	9596	1141
chr15	-3.4	26.94	175	-148	8819	988
chr16	-6.86	29.79	233	-188	11596	1267
chr17	-5.07	29.86	237	-175	16322	1746
chr18	-2.81	23.96	135	-132	4417	509
chr19	-2.28	27.48	250	-157	17788	1983
chr20	-3.69	28.15	195	-200	6810	888
chr21	-1.25	28.46	168	-160	3167	378

chr22	-4.51	28.38	194	-174	6064	729
chrX	-2.15	24.92	228	-151	8276	1057
chrY	-0.61	12.98	51	-63	269	47
chrM	-98.18	64.55	68	-245	54	34
(Empty)	-13.66	33.47	49	-128	76	0

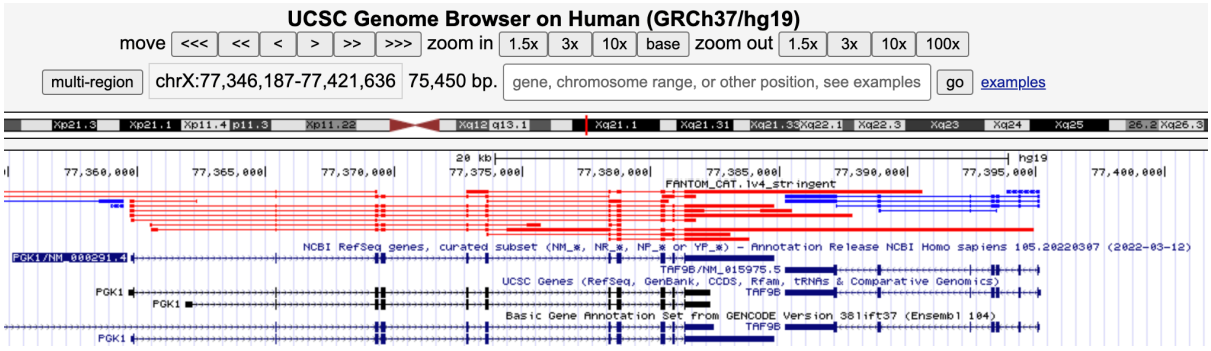




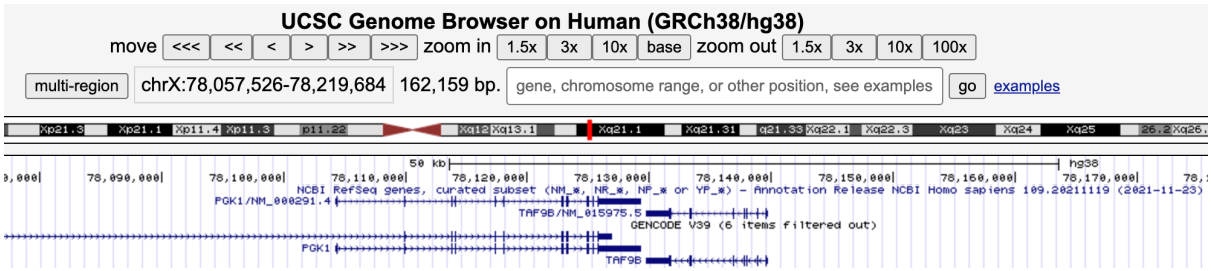
Supplemental figure 8 HNg-score による UP 200 遺伝子、DOWN200 遺伝子の染色体上の位置の可視化

13 番染色体、18 番染色体には該当遺伝子がなかったが、この二つの染色体にはもともと遺伝子数が少ない（Supplemental table2 参照）ため、低酸素応答性による特徴とは考えにくい。Python の Bio パッケージをもとに可視化した。

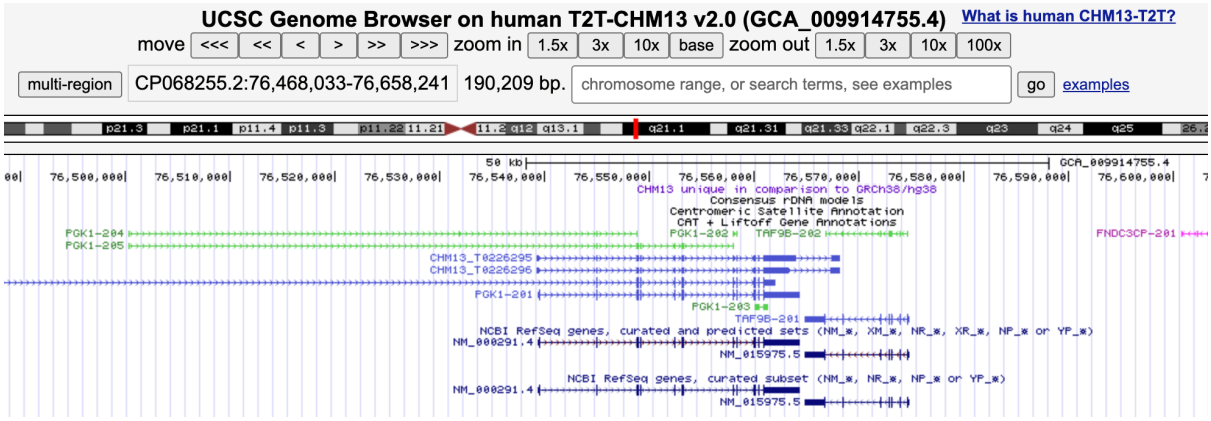
A



B



C



Supplemental figure 9 ゲノムリファレンスのバージョンごとの PGK1 と TAF9B の位置

UCSC ゲノムブラウザを用いて PGK1 と TAF9B の位置関係を確認した。異なるゲノムリファレンスにおいても PGK1 と TAF9B は同様の位置関係にあった。（A) GRCh37/hg19 (B) GRCh38/hg38 (C) T2T-CHM13 v2.0

参考文献

Adhikari K, Fontanil T, Cal S, Mendoza-Revilla J, Fuentes-Guajardo M, Chacón-Duque J-C, Al-Saadi F, Johansson JA, Quinto-Sanchez M, Acuña-Alonzo V, et al. 2016. A genome-wide association scan in admixed Latin Americans identifies loci influencing facial and scalp hair features. *Nat Commun* **7**: 10815. https://www.nature.com/articles/ncomms10815 (Accessed June 30, 2022).

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. *Frontiers in Oncology* **11**. https://www.frontiersin.org/article/10.3389/fonc.2021.779739 (Accessed March 29, 2022).

Bono H. 2020. All of gene expression (AOE): An integrated index for public gene expression databases. *PLOS ONE* **15**: e0227076. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0227076 (Accessed February 27, 2021).

Bono H, Hirota K. 2020. Meta-Analysis of Hypoxic Transcriptomes from Public Databases. *Biomedicines* **8**: 10. https://www.mdpi.com/2227-9059/8/1/10 (Accessed February 27, 2021).

Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC, et al. 2001. Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. *Nat Genet* **29**: 365–371. https://www.nature.com/articles/ng1201-365 (Accessed June 19, 2022).

Caplan LR, Hennerici M. 1998. Impaired clearance of emboli (washout) is an important link between hypoperfusion, embolism, and ischemic stroke. *Arch Neurol* **55**: 1475–1482.

Carlevaro-Fita J, Rahim A, Guigó R, Vardy LA, Johnson R. 2016. Cytoplasmic long noncoding RNAs are frequently bound to and degraded at ribosomes in human cells. *RNA* **22**: 867. https://www.ncbi.nlm.nih.gov/labs/pmc/articles/PMC4878613/ (Accessed March 10, 2022).

Carter AJ, Kraemer O, Zwick M, Mueller-Fahrnow A, Arrowsmith CH, Edwards AM. 2019. Target 2035: probing the human proteome. *Drug Discovery Today* **24**: 2111–2115. https://www.sciencedirect.com/science/article/pii/S1359644619301382 (Accessed February 27, 2021).

Deng M, Zhang W, Yuan L, Tan J, Chen Z. 2020. HIF-1a regulates hypoxia-induced autophagy via translocation of ANKRD37 in colon cancer. *Exp Cell Res* **395**: 112175.

Ebert BL, Bunn HF. 1998. Regulation of Transcription by Hypoxia Requires a Multiprotein Complex That Includes Hypoxia-Inducible Factor 1, an Adjacent Transcription Factor, and p300/CREB Binding Protein. *Mol Cell Biol* **18**: 4089–4096. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC108993/ (Accessed May 11, 2021).

Fatima A, Tariq F, Malik MFA, Qasim M, Haq F. 2017. Copy Number Profiling of MammaPrint™ Genes Reveals Association with the Prognosis of Breast Cancer Patients. *J Breast Cancer* **20**: 246–253.

Frontini M, Soutoglou E, Argentini M, Bole-Feysot C, Jost B, Scheer E, Tora L. 2005. TAF9b (Formerly TAF9L) Is a Bona Fide TAF That Has Unique and Overlapping Roles with TAF9. *Molecular and Cellular Biology* **25**: 4638–4649. https://journals.asm.org/doi/10.1128/MCB.25.11.4638-4649.2005 (Accessed April 10, 2022).

Ghajar J. 2000. Traumatic brain injury. *The Lancet* **356**: 923–929. https://www.sciencedirect.com/science/article/pii/S0140673600026891 (Accessed June 28, 2022).

Gordan JD, Bertout JA, Hu C-J, Diehl JA, Simon MC. 2007. HIF-2α Promotes Hypoxic Cell Proliferation by Enhancing c-Myc Transcriptional Activity. *Cancer Cell* **11**: 335–347. https://www.sciencedirect.com/science/article/pii/S1535610807000591 (Accessed February 27, 2021).

Hirota K, Semenza GL. 2006. Regulation of angiogenesis by hypoxia-inducible factor 1. *Critical Reviews in Oncology/Hematology* **59**: 15–26. https://www.sciencedirect.com/science/article/pii/S1040842806000138 (Accessed February 27, 2021).

Ho K, Shih C, Liu A, Chen K. 2022. Hypoxia-inducible lncRNA MIR210HG interacting with OCT1 is involved in glioblastoma multiforme malignancy. *Cancer Sci* **113**: 540–552. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8819343/ (Accessed July 3, 2022).

Imada EL, Sanchez DF, Collado-Torres L, Wilks C, Matam T, Dinalankara W, Stupnikov A, Lobo-Pereira F, Yip C-W, Yasuzawa K, et al. 2020. Recounting the FANTOM CAGE-Associated Transcriptome. *Genome Res* **30**: 1073–1081.

Imada EL, Sanchez DF, Dinalankara W, Vidotto T, Ebot EM, Tyekucheva S, Franco GR, Mucci LA, Loda M, Schaeffer EM, et al. 2021. Transcriptional landscape of PTEN loss in primary prostate cancer. *BMC Cancer* **21**: 856.

Jaakkola P, Mole DR, Tian YM, Wilson MI, Gielbert J, Gaskell SJ, von Kriegsheim A, Hebestreit HF, Mukherji M, Schofield CJ, et al. 2001. Targeting of HIF-alpha to the von Hippel-Lindau ubiquitylation complex by O2-regulated prolyl hydroxylation. *Science* **292**: 468–472.

Katayama S, Tomaru Y, Kasukawa T, Waki K, Nakanishi M, Nakamura M, Nishida H, Yap CC, Suzuki M, Kawai J, et al. 2005. Antisense Transcription in the Mammalian Transcriptome. *Science* **309**: 1564–1566. https://www.science.org/doi/10.1126/science.1112009 (Accessed April 2, 2022).

Knaup KX, Monti J, Hackenbeck T, Jobst-Schwan T, Klanke B, Schietke RE, Wacker I, Behrens J, Amann K, Eckardt K-U, et al. 2014. 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. *Mol Carcinog* **53**: 970–978.

Kodama Y, Mashima J, Kosuge T, Ogasawara O. 2019. DDBJ update: the Genomic Expression Archive (GEA) for functional genomics data. *Nucleic Acids Research* **47**: D69–D73. https://doi.org/10.1093/nar/gky1002 (Accessed February 27, 2021).

Koshiji M, Kageyama Y, Pete EA, Horikawa I, Barrett JC, Huang LE. 2004. HIF-1α induces cell cycle arrest by functionally counteracting Myc. *The EMBO Journal* **23**: 1949–1956. https://www.embopress.org/doi/full/10.1038/sj.emboj.7600196 (Accessed February 27, 2021).

Labrecque MP, Takhar MK, Nason R, Santacruz S, Tam KJ, Massah S, Haegert A, Bell RH, Altamirano-Dimas M, Collins CC, et al. 2016. The retinoblastoma protein regulates hypoxia-inducible genetic programs, tumor cell invasiveness and neuroendocrine differentiation in prostate cancer cells. *Oncotarget* **7**: 24284–24302. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5029701/ (Accessed February 27, 2021).

Li H, Ko HP, Whitlock JP. 1996. Induction of Phosphoglycerate Kinase 1 Gene Expression by Hypoxia. *Journal of Biological Chemistry* **271**: 21262–21267. https://linkinghub.elsevier.com/retrieve/pii/S0021925819746444 (Accessed April 2, 2022).

Mahon PC, Hirota K, Semenza GL. 2001. FIH-1: a novel protein that interacts with HIF-1alpha and VHL to mediate repression of HIF-1 transcriptional activity. *Genes Dev* **15**: 2675–2686.

Mathieu J, Zhang Z, Zhou W, Wang AJ, Heddleston JM, Pinna CMA, Hubaud A, Stadler B, Choi M, Bar M, et al. 2011. HIF Induces Human Embryonic Stem Cell Markers in Cancer Cells. *Cancer Research* **71**: 4640–4652. https://doi.org/10.1158/0008-5472.CAN-10-3320 (Accessed March 19, 2022).

Mihai IS, Das D, Maršalkaite G, Henriksson J. 2021. Meta-Analysis of Gene Popularity: Less Than Half of Gene Citations Stem from Gene Regulatory Networks. *Genes* **12**. https://www.ncbi.nlm.nih.gov/labs/pmc/articles/PMC7926953/ (Accessed February 25, 2022).

Morse JM. 2010. "Cherry picking": writing from thin data. *Qual Health Res* **20**: 3.

Nieminen T, Scott TA, Lin F-M, Chen Z, Yla-Herttuala S, Morris KV. 2018. Long Non-Coding RNA Modulation of VEGF-A during Hypoxia. *Non-Coding RNA* **4**. https://www.ncbi.nlm.nih.gov/labs/pmc/articles/PMC6315885/ (Accessed March 14, 2022).

Okada Y, Tashiro C, Numata K, Watanabe K, Nakaoka H, Yamamoto N, Okubo K, Ikeda R, Saito R, Kanai A, et al. 2008. Comparative expression analysis uncovers novel features of endogenous antisense transcription. *Hum Mol Genet* **17**: 1631–1640.

Oki S, Ohta T, Shioi G, Hatanaka H, Ogasawara O, Okuda Y, Kawaji H, Nakaki R, Sese J, Meno C. 2018. ChIP-Atlas: a data-mining suite powered by full integration of public ChIP-seq data. *EMBO reports* **19**: e46255. https://www.embopress.org/doi/full/10.15252/embr.201846255 (Accessed February 27, 2021).

Ono H, Ogasawara O, Okubo K, Bono H. 2017. RefEx, a reference gene expression dataset as a web tool for the functional analysis of genes. *Scientific Data* **4**: 170105. https://www.nature.com/articles/sdata2017105 (Accessed February 3, 2021).

Ono Y, Bono H. 2022. Exploratory Meta-Analysis of Hypoxic Transcriptomes Using a Precise Transcript Reference Sequence Set. 2022.05.01.489280. https://www.biorxiv.org/content/10.1101/2022.05.01.489280v1 (Accessed July 11, 2022).

Ono Y, Bono H. 2021. Multi-Omic Meta-Analysis of Transcriptomes and the Bibliome Uncovers Novel Hypoxia-Inducible Genes. *Biomedicines* **9**: 582. https://www.mdpi.com/2227-9059/9/5/582 (Accessed August 9, 2021).

Ozsolak F, Kapranov P, Foissac S, Kim SW, Fishilevich E, Monaghan AP, John B, Milos PM. 2010. Comprehensive Polyadenylation Site Maps in Yeast and Human Reveal Pervasive Alternative Polyadenylation. *Cell* **143**: 1018–1029. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3022516/ (Accessed April 2, 2022).

Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. 2017. Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods* **14**: 417–419. https://www.nature.com/articles/nmeth.4197 (Accessed February 27, 2021).

Pelechano V, Steinmetz LM. 2013. Gene regulation by antisense transcription. *Nat Rev Genet* **14**: 880–893. https://www.nature.com/articles/nrg3594 (Accessed February 2, 2022).

Peng C, Wu J, Zheng S. 2018. Hypoxia Induced ANKRD37 Promotes Angiogenesis and Reprograms Glucose Metabolism in Hepatocellular Carcinoma Following Liver Transplantation. *Transplantation* **102**: S683. https://journals.lww.com/transplantjournal/Abstract/2018/07001/Hypoxia_Induced_ANKRD37_Promotes_Angiogenesis_and.1098.aspx (Accessed June 29, 2022).

Qi J, Nakayama K, Cardiff RD, Borowsky AD, Kaul K, Williams R, Krajewski S, Mercola D, Carpenter PM, Bowtell D, et al. 2010. Siah2-Dependent Concerted Activity of HIF and FoxA2 Regulates Formation of Neuroendocrine Phenotype and Neuroendocrine Prostate Tumors. *Cancer Cell* **18**: 23–38. https://www.cell.com/cancer-cell/abstract/S1535-6108(10)00236-9 (Accessed February 27, 2021).

Ramilowski JA, Yip CW, Agrawal S, Chang J-C, Ciani Y, Kulakovskiy IV, Mendez M, Ooi JLC, Ouyang JF, Parkinson N, et al. 2020. Functional annotation of human long noncoding RNAs via molecular phenotyping. *Genome Res* **30**: 1060–1072. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7397864/ (Accessed July 3, 2022).

Richards JP, Yosten GLC, Kolar GR, Jones CW, Stephenson AH, Ellsworth ML, Sprague RS. 2014. Low O2-induced ATP release from erythrocytes of humans with type 2 diabetes is restored by physiological ratios of C-peptide and insulin. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology* **307**: R862–R868.

https://journals.physiology.org/doi/full/10.1152/ajpregu.00206.2014 (Accessed February 27, 2021).

Rustici G, Williams E, Barzine M, Brazma A, Bumgarner R, Chierici M, Furlanello C, Greger L, Jurman G, Miller M, et al. 2021. Transcriptomics data availability and reusability in the transition from microarray to next-generation sequencing. 2020.12.31.425022. https://www.biorxiv.org/content/10.1101/2020.12.31.425022v1 (Accessed June 19, 2022).

Schmitz K-M, Mayer C, Postepska A, Grummt I. 2010. Interaction of noncoding RNA with the rDNA promoter mediates recruitment of DNMT3b and silencing of rRNA genes. *Genes Dev* **24**: 2264–2269. http://genesdev.cshlp.org/content/24/20/2264 (Accessed April 9, 2022).

Semenza GL. 2003. Angiogenesis in ischemic and neoplastic disorders. *Annu Rev Med* **54**: 17–28.

Semenza GL, Wang GL. 1992. A nuclear factor induced by hypoxia via de novo protein synthesis binds to the human erythropoietin gene enhancer at a site required for transcriptional activation. *Mol Cell Biol* **12**: 5447–5454. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC360482/ (Accessed August 14, 2021).

Shoji K, Murayama T, Mimura I, Wada T, Kume H, Goto A, Ohse T, Tanaka T, Inagi R, van der Hoorn FA, et al. 2013. 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* **182**: 2191–2203.

Stoeger T, Amaral LAN. 2022. The characteristics of early-stage research into human genes are substantially different from subsequent research. *PLoS Biology* **20**. https://www.ncbi.nlm.nih.gov/labs/pmc/articles/PMC8769369/ (Accessed February 8, 2022).

Tora L. 2002. A unified nomenclature for TATA box binding protein (TBP)-associated factors (TAFs) involved in RNA polymerase II transcription. *Genes Dev* **16**: 673–675. http://genesdev.cshlp.org/content/16/6/673 (Accessed April 2, 2022).

Valentin C, Birgens H, Craescu CT, Brødum-Nielsen K, Cohen-Solal M. 1998. A phosphoglycerate kinase mutant (PGK Herlev; D285V) in a Danish patient with isolated chronic hemolytic anemia: mechanism of mutation and structure-function relationships. *Hum Mutat* **12**: 280–287.

Wang GL, Jiang BH, Rue EA, Semenza GL. 1995. Hypoxia-inducible factor 1 is a basic-helix-loop-helix-PAS heterodimer regulated by cellular O2 tension. *Proc Natl Acad Sci U S A* **92**: 5510–5514.

Wang GL, Semenza GL. 1995. Purification and Characterization of Hypoxia-inducible Factor 1 (∗). *Journal of Biological Chemistry* **270**: 1230–1237. https://www.jbc.org/article/S0021-9258(18)82990-8/abstract (Accessed August 14, 2021).

Wasserstein RL, Lazar NA. 2016. The ASA Statement on *p* -Values: Context, Process, and Purpose. *The American Statistician* **70**: 129–133. https://www.tandfonline.com/doi/full/10.1080/00031305.2016.1154108 (Accessed June 28, 2022).

Wong WJ, Qiu B, Nakazawa MS, Qing G, Simon MC. 2013. MYC Degradation under Low O2 Tension Promotes Survival by Evading Hypoxia-Induced Cell Death. *Molecular and Cellular Biology* **33**: 3494–3504. https://mcb.asm.org/content/33/17/3494 (Accessed February 27, 2021).

Yu H, Rimbert A, Palmer AE, Toyohara T, Xia Y, Xia F, Ferreira LMR, Chen Z, Chen T, Loaiza N, et al. 2019. GPR146 Deficiency Protects Against Hypercholesterolemia and Atherosclerosis. *Cell* **179**: 1276-1288.e14. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6889877/ (Accessed February 27, 2021).

Zhang H, Bosch-Marce M, Shimoda LA, Tan YS, Baek JH, Wesley JB, Gonzalez FJ, Semenza GL. 2008. Mitochondrial Autophagy Is an HIF-1-dependent Adaptive Metabolic Response to Hypoxia. *J Biol Chem* **283**: 10892–10903. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2447655/ (Accessed March 29, 2022).

Zhang H, Gao P, Fukuda R, Kumar G, Krishnamachary B, Zeller KI, Dang CV, Semenza GL. 2007. HIF-1 inhibits mitochondrial biogenesis and cellular respiration in VHL-deficient renal cell carcinoma by repression of C-MYC activity. *Cancer Cell* **11**: 407–420.

Zhang X, Le W. 2010. Pathological role of hypoxia in Alzheimer's disease. *Experimental Neurology* **223**: 299–303. https://www.sciencedirect.com/science/article/pii/S0014488609003057 (Accessed June 28, 2022).

Zhang Y, Gu J, Wang L, Zhao Z, Pan Y, Chen Y. 2017. Ablation of PPP1R3G reduces glycogen deposition and mitigates high-fat diet induced obesity. *Molecular and Cellular*

Endocrinology **439**: 133–140. https://www.sciencedirect.com/science/article/pii/S0303720716304476 (Accessed June 30, 2022).

Zhang Y, Xu D, Huang H, Chen S, Wang L, Zhu L, Jiang X, Ruan X, Luo X, Cao P, et al. 2014. Regulation of glucose homeostasis and lipid metabolism by PPP1R3G-mediated hepatic glycogenesis. *Mol Endocrinol* **28**: 116–126.

Zhou Y, Zhou B, Pache L, Chang M, Khodabakhshi AH, Tanaseichuk O, Benner C, Chanda SK. 2019. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nature Communications* **10**: 1523. https://www.nature.com/articles/s41467-019-09234-6 (Accessed February 27, 2021).

佐藤俊哉. 2018. ASA声明と疫学研究における*P*値. 計量生物学 **38**: 109–115.

謝辞

本研究の遂行にあたり指導教員であった広島大学大学院統合生命科学研究科ゲノム情報科学研究室特任教授 坊農秀雅先生に深謝いたします。また、統合生命科学研究科の井川武助教、今村拓也教授、浮穴和義教授、医系科学研究科谷本圭司准教授（順不同）には本論文の作成にあたり適切なご助言を賜りました。感謝申し上げます。

理化学研究所 生命医科学研究センター 生命医科学大容量データ技術研究チーム 粕川雄也先生 森岡勝樹先生には研究内容についてのご相談に親身なご対応を賜りました。感謝申し上げます。

また、コロナ禍のなか、リモートでの講義受講を可能としてくださった広島大学の皆様に感謝申し上げます。

サイエンスの議論や解析技術の研鑽を通じて基礎力を育ててくださった三島創薬勉強会 Mishima.syk コミュニティの皆さま、BioHackathon BH21.8 や SPARQLthon の関係者の皆さま、Bonohulab のラボメンバーの皆さまに厚く御礼申し上げます。

働きながらの学位取得を応援してくださった協和キリン株式会社の皆さまに御礼申し上げます。

多くのオープンデータやオープンソースソフトウェアの活用無くして本研究を遂行することは不可能でした。貴重なデータやソースコードを管理・公開してくださっている全てのオープンサイエンスコミュニティに感謝申し上げます。