# 複数細胞種における核酸結合タンパク質の発現制御遺伝子と機能を予測する手法の開発

大里直樹　佐藤健吾

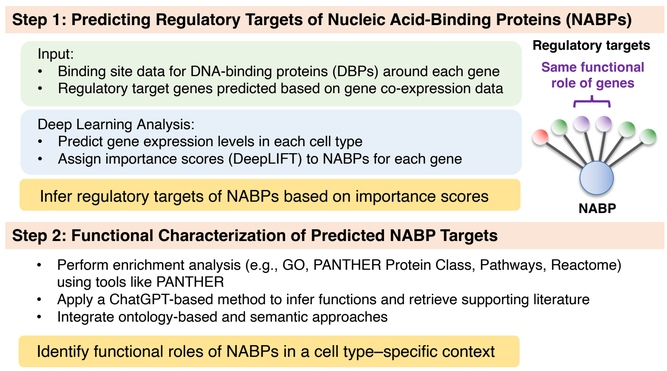
東京科学大学 生命理工学院

## 1. 要旨


- **核酸（DNA/RNA）結合タンパク質（Nucleic acid-binding proteins; NABPs）**は、細胞種特異的な制御や機能に関わるが、その**制御標的遺伝子**や**生物学的役割**は十分に解明されていない。
- **遺伝子の共発現相関**を利用した**深層学習の手法**を開発し、**NABPの結合位置やモチーフの情報を用いずに**、NABPの標的遺伝子を予測し、その機能を推定した。
- 共発現に基づく標的遺伝子を用いることにより、遺伝子発現予測の精度が向上した。
- 予測されたNABPの標的遺伝子は、**ChIP-seqとeCLIPの実験による標的遺伝子と高い整合性を示**し、ランダムに選んだ遺伝子よりも優れた性能を示した。また標的遺伝子を含む共発現遺伝子は、**DNA結合タンパク質のノックアウト実験結果と高い整合性を示した。**
- **機能アノテーション解析、ChatGPT、GSEA** (Gene set enrichment analysis)**を用いた機能推定**により、PKMによる解糖系制御など、**生物学的な意味のある機能**が同定された。
- 深層学習・共発現ネットワーク・大規模言語モデルを組み合わせることで、**既知および新規のNABP機能を細胞種特異的に体系的に解析**できる。


## 2. 背景

- 核酸結合タンパク質は、**DNA結合タンパク質（DBPs）**や**RNA結合タンパク質（RBPs）**を含み、**細胞種特異的な制御や機能**があるが、網羅的な解析は困難である。
- ChIP-seq や eCLIP などの実験解析は高品質な抗体を必要とし、**多数のタンパク質や細胞種に適用することは難しい。**
- **RBP** においては、**明確なRNA結合モチーフ配列を示さない場合**があり、またディスオーダー領域（IDR: Intrinsically Disordered Regions）を介した相互作用が知られており、**標的遺伝子の同定が困難**である。
- 本研究では、遺伝子共発現データを用いて、**細胞種ごとのNABPの標的遺伝子と機能を予測する。**




## 3. 方法

- 深層学習を用いて、複数のヒト細胞種（例：HFF, HMEC, NPC, HepG2, K562）における**DNA結合タンパク質のDNA結合位置のデータから遺伝子発現量を予測**した。
- DNA結合位置は、転写開始点（TSS）から±1 Mb以内の遠位エンハンサーを含み、ChIP-seqデータ（GTRD）により同定され、eQTLデータにより遺伝子と対応づけされた。
- プロモーターおよびエンハンサーを統合し、1,310種類のDBPsに対応する230ビンの入力マトリクスを構築し、DeepLIFTを用いたContribution scoreの解析を行った。
- Contribution scoreが低いDBPsを、**共発現データベース**（COXPRESdb）**から取得した核酸（DNA/RNA）結合タンパク質（NABPs）に置き換えた。**
- 共発現は Pearson相関係数の*z*スコア > 2 を選び、共発現遺伝子のTSSから -2 kbまたは0 bpの位置にNABPsの結合を仮定した。
- 上位あるいは下位にランクされたNABP–遺伝子ペアを選び、その標的遺伝子群についてPANTHERにより機能エンリッチメント解析を行なった。
- さらに、ChatGPTを活用した手法により、遺伝子セットの機能推定および関連文献の抽出を行なった。手動による文献検索も行った。
- 実験データによる検証として、DBPの場合はChIP-seq、RBPの場合はeCLIPデータ（ENCODEおよびGEO）を用い、高スコア遺伝子との一致を評価した。DBPのノックアウト実験により発現量が変動した遺伝子と共発現遺伝子を比較した。

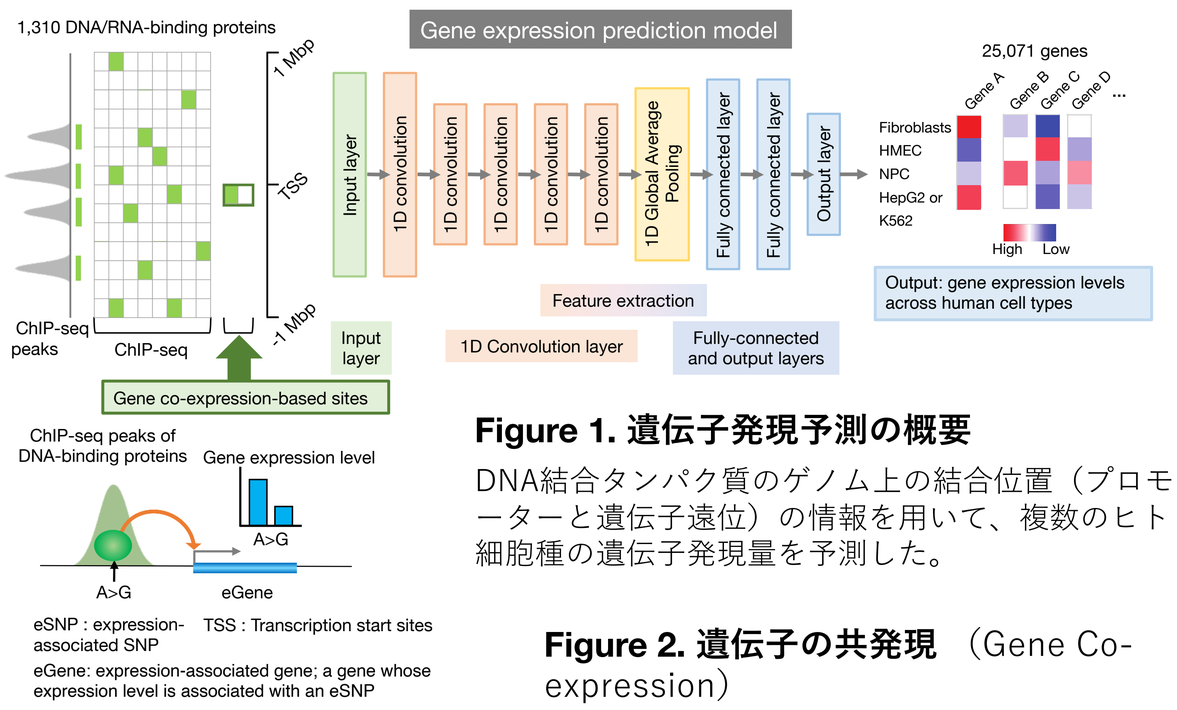
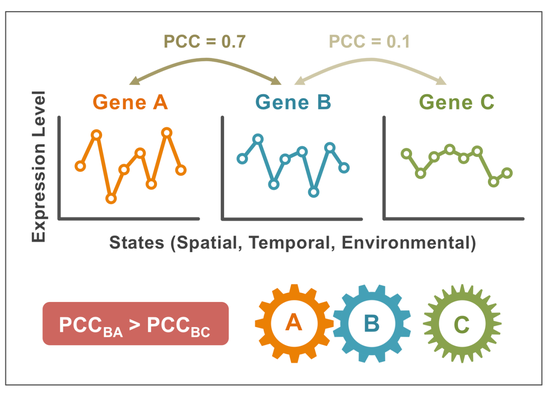



**Figure 1. 遺伝子発現予測の概要**

DNA結合タンパク質のゲノム上の結合位置（プロモーターと遺伝子遠位）の情報を用いて、複数のヒト細胞種の遺伝子発現量を予測した。




Pearson's Correlation Coefficient (PCC)
http://alcodb.jp/coexpression/tutorial

**Figure 2. 遺伝子の共発現**（Gene Co-expression）

遺伝子の共発現は、２つの遺伝子間の発現プロファイルの類似性を示す。

機能的に関連する遺伝子や同じ調節機構により制御されている遺伝子は、一般に細胞や組織間で同様に発現変動する傾向がある。

ある核酸結合タンパク質（NABP）をコードする遺伝子が他の遺伝子と類似した発現パターンを示す場合、遺伝子発現が核酸結合タンパク質により制御され、機能的に関連する可能性がある。

## 4. 結果

- Contribution scoreの低いDNA結合タンパク質（DBPs）を遺伝子共発現データに置き換えると、遺伝子発現予測の精度が向上した（DBPの場合：R = 0.70 → 0.80、RBPの場合：R = 0.70 → 0.81） **(Figures 1 and 2)**。
- NABPに対してChIP-seqまたはeCLIPで実験的に同定された結合位置をもつ遺伝子は、**Contribution scoreの上位または下位にランクされる傾向**を示し、遺伝子発現制御においてより強い活性または抑制の役割が示唆された **(Figure 3a)**。
- Contribution scoreの高い遺伝子は、ランダムな遺伝子よりも**高い割合でChIP-seqおよびeCLIPピークと共局在する**ことが確認された**(Figure 3b)**。
- 遺伝子共発現データに基づくDBPの標的遺伝子のほぼすべてが、DBPのノックアウト実験により発現量が有意に変動した遺伝子に含まれた。
- **12種類のタンパク質について予測されたNABPの制御標的遺伝子の機能エンリッチメント解析では、UniProtのアノテーションと一致するGene Ontologyとの関連が示された (Table 1)** 。
- ChatGPTを用いた機能予測もこれらの役割と同様の結果を示し、さらに**新規の関連性を同定した。**文献により確認した**(Table 1)** 。
- 深層学習の学習後には、各細胞の機能に関連する機能アノテーションが特異的に得られた**(Table 1)** 。GSEAにより発現制御や機能に関わる遺伝子のスコアの分布を示した。
- 共発現データに基づく深層学習とAIによる文献解析を組み合わせた本手法は、**細胞種特異的なNABPの制御的役割や機能を同定できる(Figures 4 and 5)** 。

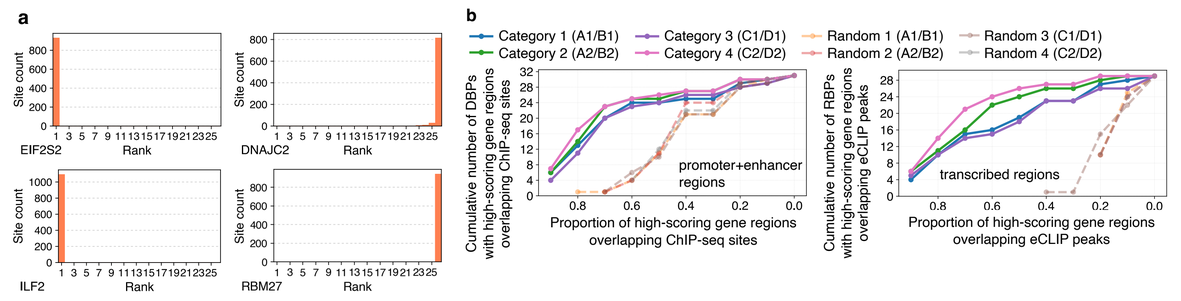



**Figure 3. 予測された制御標的遺伝子とChIP-seqおよびeCLIP結合データの比較**

**ZBTB34 (HepG2 cells)**
C2H2 zinc finger transcription factor
RNA polymerase II transcription regulatory region sequence-specific DNA binding
regulation of transcription by RNA polymerase II
**C2H2 zinc-finger–directed regulation of RNA polymerase II transcription (0.88)**

**ZC3H4 (HepG2 cells)**
chromatin/chromatin-binding, or -regulatory protein
RNA processing factor
basal RNA polymerase II transcription machinery binding
**Chromatin-regulated RNA polymerase II transcription and RNA processing (0.89)**

**PKM (K562 cells)**
Hypoxia response via HIF activation
glucose 6-phosphate metabolic process
Disorders of transmembrane transporters
Membrane Trafficking
**Hypoxia/HIF-1–orchestrated metabolic reprogramming with cytoskeletal–membrane remodeling (0.80)**

**PKM (NPC cells)**
Cellular response to hypoxia
KEAP1-NFE2L2 (NRF2) pathway
actin or actin-binding cytoskeletal protein
**Hypoxia-driven metabolic rewiring with KEAP1–NRF2 signaling and actin–endosomal remodeling (0.86)**

**Table 1. 核酸結合タンパク質の制御標的遺伝子の機能注釈と機能推定**

核酸結合タンパク質（NABP）の機能アノテーションは、UniProtデータベースから取得した。さらに、予測された制御標的遺伝子において有意にエンリッチされたアノテーションについて、PANTHERデータベースを用いて同定した。各NABPとその予測標的遺伝子で共通または類似の機能アノテーションを表に示した。

ChatGPTベースの手法により推定された機能的役割は黄色でハイライトされている。括弧内の数値は信頼スコア（confidence score）を示し、対応する生物学的プロセスに関与する標的遺伝子の割合を反映している。信頼スコアの範囲は0.00〜1.00であり、高（High）：0.87–1.00、中（Medium）：0.82–0.86、低（Low）：0.01–0.81、なし（None）：0として分類される。).

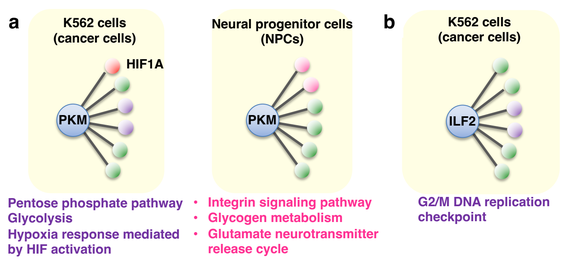



**Figure 4. 核酸結合タンパク質の細胞種に特異的な制御と機能**

がん細胞(K562)では、PKMが解糖系やペントースリン酸経路、HIF（Hypoxia-inducible factor: 低酸素誘導因子)の遺伝子発現制御に関わり、神経前駆細胞では有意に関連が見られなかった。

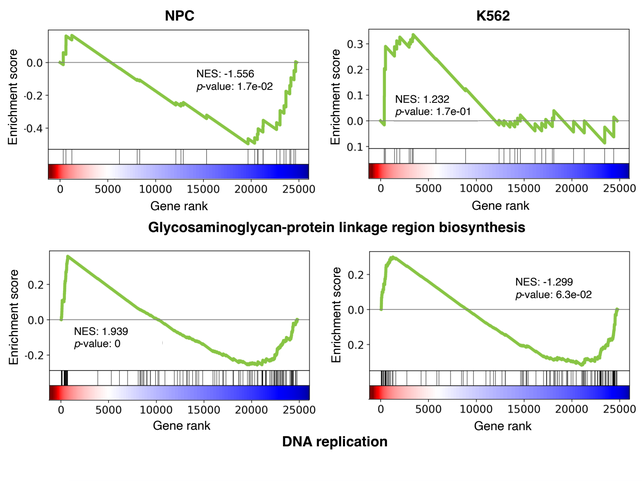



**Figure 5. GSEAを用いた核酸結合タンパク質の細胞種に特異的な制御と機能**

GSEA（Gene set enrichment analysis）は、DBPおよびRBPのDeepLIFT貢献度スコアを用いて行った。解析の結果、これらのスコアから推定された制御標的遺伝子が、各細胞種の機能やパスウェイに関連していることが示された。グラフはRNA結合タンパク質のPKMが発現制御に関わる遺伝子について、神経前駆細胞（NPC）とがん細胞（K562）で異なる機能・パスウェイを示した。

## 5. 結論

- 本研究は、遺伝子の共発現データから、核酸結合タンパク質（NABP）の制御標的および機能を予測したことを示した。ChIP-seqやeCLIPと比較して、RNA-seqは高速・低コストで、様々な細胞種や条件で利用可能である。
- 既知の遺伝子機能と整合し、文献的にも裏付けられる制御標的遺伝子の機能を新たに同定した。
- 機能エンリッチメント解析、ChatGPT、GSEAを用いた解析により、既知および新規のNABP機能が明らかとなった。
- Contribution scoreにより、直接的および間接的な相互作用を予測でき、NABPによる制御を包括的に理解できる。
- 細胞種特異的な予測を支援し、異なる生物学的コンテキストや複合体内で機能するNABPの発見につながると期待される。
- 本手法により、大規模なNABPの制御標的遺伝子と機能の予測ができ、新規の遺伝子発現制御メカニズムの発見につながる。

Osato N et al. *bioRxiv* 2025 doi: https://doi.org/10.1101/2025.03.03.641203