公共データを活用した転写因子結合ダイナミクスの解析

植 野 和 子

国立国際医療研究センター ゲノム医科学プロジェクト 戸山プロジェクト

１．はじめに

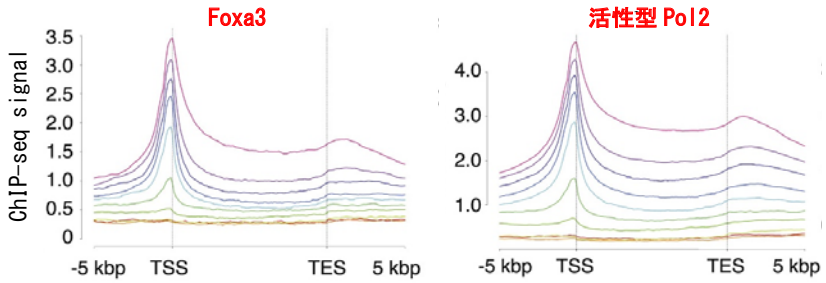
DNA に保存された生命の遺伝情報は、複製により子孫に受け継がれるのと同時に RNA への「転写」・タンパク質への「翻訳」の過程を経て表現型として表出する。一連の遺伝子発現メカニズムの制御において最も重要なステップであると考えられるのが、ゲノム DNA から RNA ポリメラーゼ II（Pol2）によりメッセンジャーRNA（mRNA）が写し取られる「転写」である。**転写因子**は DNA の特異的な塩基配列を認識し結合するタンパク質の一群であり、遺伝情報を RNA に転写する過程を促進または抑制することで、種々の細胞種毎の遺伝子発現状態をオーガナイズしている。ヒトゲノム上には、およそ 1,800 前後の転写因子がコードされており、それらの働きを網羅的に解明することが、究極的には生命の理解につながると考えられている。

一方近年では、転写因子の人為的操作によって細胞運命を工学的にリプログラミングすることが可能になりつつある。終末分化した体細胞を初期化し、あらゆる細胞への分化能を有する誘導多能性幹細胞（iPS 細胞）を誘導する技術がもっとも有名だが[1]、最近では様々な転写因子の組み合わせを導入することで、多種多様な細胞を直接誘導するダイレクトリプログラミング技術も盛んに開発が進められており[2]、いずれも再生医療に向けた応用が期待されている。今後は、それらのリプログラミング技術をより安全・安定的なものに改良し、臨床応用可能なレベルにまで発展させるために、導入転写因子セットが、ゲノム DNA の「どこ」に結合し「何の」遺伝子を「どのように」制御するのかを明らかにすることが必要とされている。

２．転写因子 Foxa3 は「動く」

我々の研究グループでは、マウスの皮膚や胎仔由来の線維芽細胞に Hnf4αおよび Foxa（ファミリータンパク質である Foxa1、Foxa2、Foxa3 のいずれか一つ）を強制発現させることにより作製される誘導肝細胞様細胞（iHep 細胞）[3]を研究対象として、転写因子の挙動と機能、およびリプログラミングにおける役割を解析してきた。特に**次世代シークエンサー（NGS）**を用いたゲノムワイドな転写因子の結合解析を通して、２つの転写因子が互いに影響しながら細胞の転写状態を生体の肝細胞に近い状態に誘導するメカニズムの一端を明らかにした[4]。

その解析データの中から我々は、リプログラミング誘導転写因子 Foxa ファミリーのうち**Foxa3** のみが、標的遺伝子のコード領域（genebody）上を Pol2 と共に移動し、なおかつその様な挙動をとること自体が iHep 細胞へのダイレクトリプログラミングの進行にとって必須であることを見出した（図１）。従来、転写因子は DNA 上の認識配列に強固に結合し、その部位に転写の活性化や抑制に関わる補因子をリクルートすることで転写を制御するという静的モデルが一般的に信じられてきたが、この発見は転写因子の既存イメージを大きく覆すものである。また、このような Foxa3 のダイナミックに「動く」分子挙動は、リプログラミング細胞だけでなく生体由来の肝細胞においても確認することができることから、生体の細胞内において何らかの生理的な役割を果たしているものと考えられる。



第 1 図: Foxa3 および Pol2 の genebody 上での結合分布。

左図に Foxa3 の、右図に活性型 Pol2 の遺伝子領域上での結合分布を示す metagene plot を記載した。それぞれの色別プロットは全遺伝子をそれらの転写量で１０分割したカテゴリーを示しており、Foxa3 は Pol2 と同様に転写量が減少すると genebody 上の結合が減少することを示している。このことは Foxa3 が Pol2 と結合を共にしている可能性を強く示唆している。また、転写開始点（TSS）に強いピーク、転写終結点（TES）下流に弱いピークが出現し、その間の genebody に裾野を引くような分布パターンとなることも、Foxa3 が genebody 上を移動するという「動く」分子モデルをサポートする結果である。

３．転写因子の大規模公共データベースとその再解析

Foxa3 と同様の分子挙動を示す転写因子の例はこれまでに報告が無く、上述したような「動く」分子モデルが一般的なものなのか、それとも Foxa3 特有のものなのかは全くわからない。転写因子のゲノムワイドな結合データは、論文投稿時にパブリックドメインにデポジットされることが求められることもあり、RAW データおよびその一次加工データが公共データベースに大量に格納されている（例：Gene Expression Omnibus[5]）。その登録数は日々増加しており、それらのデータを再解析することで、Foxa3 と似た「動く」分子挙動を示す転写因子をスクリーニングできるのではないかと考えた。

しかし、転写因子の結合データのほとんどは、**クロマチン免疫沈降シークエンシング法（ChIP-seq）**という手法で解析されたものであるが、ChIP-seq のプロトコールには様々なバリエーションがあり、かつ品質において玉石混合であることからも、無作為に収集したデータを機械的に解析するだけでは意味のある結果を得るのは難しいと考え、データの品質や実験手法が統一されている大規模プロジェクトのデータベースから情報を収集することにした。

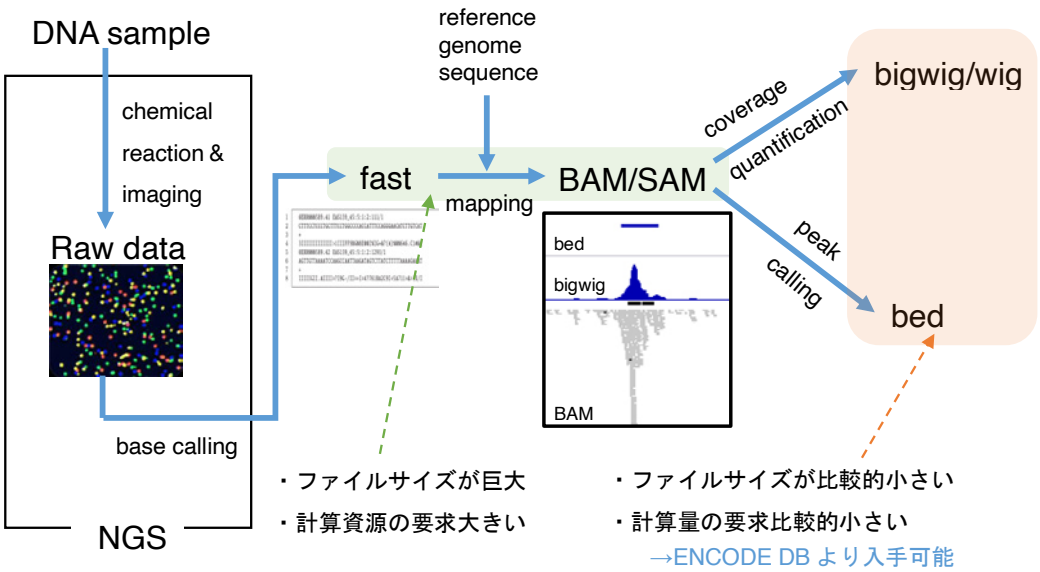
その意味で、現時点で最も信頼性が高く大量のデータを格納しているのが、アメリカ国立衛生研究所の主導で進められている **ENCODE（ENCyclopaedias Of DNA Elements）**計画のデータベースである[6]。このデータベースには 4,756 のヒトおよびマウスの細胞・組織から得られた ChIP-seq データが収められており（2022 年 6 月現在）、実験条件やサンプルなどの metadeta も整備されている。よって本プロジェクトでは ENCODE データベースから大量の ChIP-seq データを取得し、転写因子の結合パターンのスクリーニングを行うこととした。

４．転写因子結合の公共データの再解析

本プロジェクトでは、パイロットスタディとして 40 種類の転写因子データセットを取得し解析を行った。転写因子の選定基準は、①実験的に「動く」転写因子であることが示されている FOXA3（Foxa3 のヒトホモログ）、②実験により「動かない」転写因子であることが示されている FOXA1

および FOXA2（Foxa1 および Foxa2 のヒトホモログ）、③必ず遺伝子上を動く分子である POLRA2（Pol2 のサブユニット）、④TSS 近傍への結合が示唆されている MYC/MAX ファミリー転写因子、⑤他の FOX ファミリー分子、⑥AP-1 や CTCF など代表的な転写因子、についてそれぞれ種々の細胞をサンプルとして実験された ChIP-seq データを集めた。今回収集したデータは全てヒト由来の細胞を用いて実験したものである。

一般的に、ChIP-seq データなどの NGS データは short reads と呼ばれる 25〜150 bp ほどの大量の配列断片が格納された fastq と呼ばれるファイルを解析の起点として、マッピング、カバレッジの定量、ピークコールなどを続けて行う（図２）。ENCODE データベースでは、fastq に加えて一次解析後のファイル（BAM および bigwig）も配布しているため、それらをデータベースより取得して解析を行った。

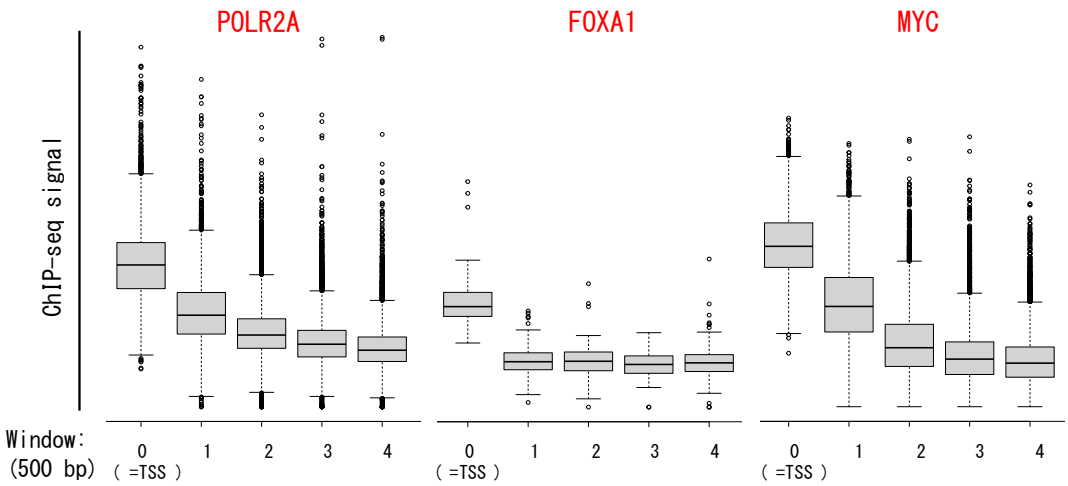


第 2 図：転写因子結合解析における NGS データの解析フロー。

ChIP-seq 実験により得られたサンプルから配列データ（fastq）の取得、さらにそこからマッピングデータ（BAM/SAM）、カバレッジデータ（bigwig/wig）、ピークデータ（bed）を取得する流れを示した。マッピング解析は非常に大きな計算資源とストレージを要求するため、今回は一次解析データであり、比較的ファイルサイズの小さい bed 及び bigwig を ENCODE データベースより取得して解析した。

まず取得したピーク情報（bed）を用いて、TSS 周辺のシグナルを確認した。TSS と重複しているピークが少ないデータは、十分な s/n 比が得られていない（または遺伝子近傍には結合しない転写因子）と判断し解析対象から除外した（2/40）。続いてカバレッジデータ（bigwig）を用いて TSS 周囲の領域でシグナルを 500 bp ウインドウで平均値を算出した。その分布が genebody において「裾野」を持つパターンになるか、それとも TSS のピーク以外はシグナルが低い平坦なパターンになるかどうかを、転写因子の「動く」「動かない」の判定基準とした（図３）。具体的に

は、500bp 幅のウインドウに分割した TSS 周辺の領域について各遺伝子座の ChIP-seq シグナルの平均値を算出し、１番目と２番目の分布に有意な差があるものを「動く」転写因子、差がなくフラットな低い値の分布を示すものを「動かない」転写因子と定義した。



第 3 図：転写因子の挙動を解析する計算方法。

転写因子の遺伝子近傍における挙動を解析した数値をウインドウ毎の boxplot で示している。「動かない」ことが既知の因子 Foxa1（左図）と、動くことが明らかな Pol2（中図）のデータを示す。いずれも TSS に強いピークがあるが、Foxa1 がその周囲の領域ではシグナルが低くフラットとなるパターンを示すのに対し、Pol2 ではシグナルが段階的に下降する裾野のようなパターンを示す。MYC は Pol2 と同じような結合パターンを示した(右図)。

解析の結果、Foxa1 および Foxa2 はこれまでの実験データの通り「動かない」転写因子であると判定された。一方 Foxa3 は、肝臓由来の細胞株による ChIP-seq データでは自分達の実験結果と同様に Pol2 と同じ「動く」パターンを示したが、血球由来の細胞株から得られたデータでは「動かない」結合パターンとなった。このことは、同じ転写因子でも細胞種の違いなどのコンテクストにより挙動を変化させる可能性があることを示唆している。他の転写因子の中で、**MYC/MAX ファミリー**の ChIP-seq データ全てが「動く」転写因子のパターンを示した。MYC/MAX は転写が活性化している遺伝子のプロモーターに積極的に結合する「invasion」と呼ばれる現象が知られており[7]、これまでの解析からは明らかにされてこなかったが、Foxa3 と同様に Pol2 と共に「動く」ことで機能を発揮する転写因子である可能性が示唆された。

５．まとめ

ENCODE データベースから取得した 40 個の転写因子結合データを再解析することで、Pol2 と類似した、遺伝子上を「動く」パターンを示す転写因子として、MYC/MAX ファミリーを同定した。今回の結果は、我々が構築した公共データの再解析法が転写因子の遺伝子上の動きのパターンを判別するうえで有効であることを示唆しており、今後は ENCODE データベースに格納される全 ChIP-seq データ等を用いて、より大規模な解析を実施したい。

参 考 文 献

[1] K. Takahashi and S. Yamanaka, Cell, 126, pp.663-676, (2006).

[2] K. Horisawa and A. Suzuki, Jpn Acad Ser B Phys Biol Sci., 96, pp. 131-158, (2020).

[3] S. Sekiya and A. Suzuki, Nature, 475, pp.390-393, (2011).

[4] K. Horisawa, M. Udono, K. Ueno, Y. Ohkawa, M. Nagasaki, S. Sekiya, A. Suzuki, Mol Cell, 79, pp.660-676, (2020).

[5] https://www.ncbi.nlm.nih.gov/geo/

[6] https://www.encodeproject.org/

[7] A. Sabò, T.R. Kress, M. Pelizzola, S. Pretis, M.M. Gorski, A. Tesi, M.J. Morelli, P. Bora, M. Doni, A. Verrecchia, C. Tonelli, G. Fagà, V. Bianchi, A. Ronchi, D. Low, H. Müller, E. Guccione, S. Campaner, B. Amati, Nature, 511, pp.488-492, (2014).