# ヒトゲノムの機能解明に向けたENCODEの試み
## ゲノム機能の百科事典作製を目指して

ヒトゲノムを構成する全塩基配列の解明により，生命現象を分子レベルで理解するための基盤が築かれた．今やさまざまな疾患の原因をゲノムレベルで議論できるようになり，細胞分化や免疫機構などの生命現象の分子基盤をも解き明かされつつある．一方で，ヒトゲノムそれ自体に対する理解は限定的であり，(1) 配列情報をもとに予測されたタンパク質コード遺伝子の総数は25,000個程度で，線虫のような単純な生物やほかの哺乳類のそれと大差がないこと，(2) ゲノムの半数近くの領域は特定の配列が繰り返し現れる反復配列であること，さらに (3) ゲノムの98%もの領域はタンパク質の構造情報を担っていないことなどを明らかにするにとどまった[1]．ヒトゲノムの大部分の領域が担っている役割については，その塩基配列だけからは解き明かすことができなかったのである．

ゲノム（DNA）はトランスクリプトーム（RNA）の設計図としての一面をもっている[2]．そこで，ゲノムに符号化（コード）された情報の全体像の理解に先立ち，トランスクリプトームの全貌を明らかにする試みが行われた．菅野らはヒトゲノムから転写されたmRNAを完全長cDNAとして網羅的に収集し，およそ20,000種類の転写産物を特定した[3]．また理化学研究所が中心となったマウスcDNA収集と機能アノテーション（FANTOMプロジェクト）では，マウスゲノムからの転写産物を網羅的に収集し，その完全長cDNAをもとに転写開始点や終結点の解明に取り組んだ[4]．mRNAだけではなく，機能性RNAをも網羅することを目指したFANTOMプロジェクトによって，哺乳類ゲノムからは膨大な数の機能未知なタンパク質非コードRNAが転写されていることが明らかにされたのである．「RNA新大陸[5]」としてトランスクリプトームの機能探索を進める契機ともなった成果である[6~8]．特に (1) 多くのRNAには複数のスプライス・バリアントが存在し，細胞に応じて異なる転写開始点が使い分けられていること[9]，(2) これらRNAの発現制御にアンチセンス側の非コードRNAがかかわっていること[10, 11]，そして (3) 細胞の分化過程ではこれら非コードRNAの転写が，コードRNAと同様に転写因子の制御下にあるとしてモデル化できること[12]などが明らかとなるにつれて，生命現象の分子基盤を理解するにはこれらトランスクリプトームの制御にかかわるゲノム領域の特定が不可欠であると考えられるようになった[13]（図1）．

国際研究プロジェクト ENCODE（Encyclopedia of DNA Elements）は，ポストゲノム戦略としてヒトゲノムの機能要素とその全体像を明らかにすることを目的に組織された[14, 15]．ENCODEではゲノム上の機能をRNAの転写領域と転写制御にかかわる領域とに大別し，後者についてはさらにプロモーターや転写因子結合領域，エンハンサー，リプレッサーなどの転写制御領域と，クロマチン構造の変化やエピジェネティック修飾などを受ける領域とに分け，さまざまな細胞株においてヒトゲノムの各領域が担う役割の解明を目指した[14]．個々の細胞株におけるゲノム機能の調査は，(1) CAGE法やRACE法，RNA-Seq法などによるRNAの転写領域

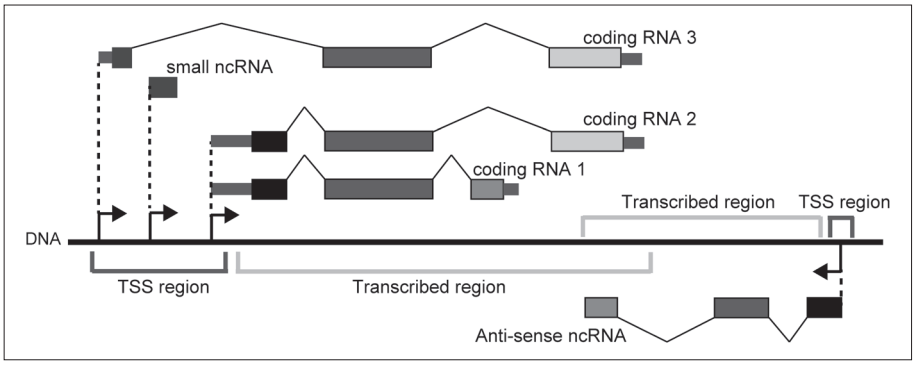



**図1 ■ 転写RNAの複雑性**
個々のゲノム領域からは異なる転写開始点をもつ複数のRNAが転写される．また，タンパク質コード領域の相補鎖（anti-sense）から転写される非コードRNAの一部は，コードRNAの発現制御にかかわっている．

表1■ヒトゲノムの機能要素

| ゲノムの機能要素 | 特徴 |
|---|---|
| CTCF-enriched element | 転写因子CTCFの相互作用領域．近接する転写制御因子の影響を絶縁するインシュレーターとして機能している可能性が高い． |
| Predicted enhancer | ヒストンのH3K4me1修飾が見られるユークロマチン領域．エンハンサーとして機能する転写因子の相互作用が見られ，近傍領域の転写効率を制御している可能性が高い． |
| Predicted promoter flanking region | RNAの転写開始点の近傍領域． |
| Predicted repressed or low-activity region | Polycombグループなどのタンパク質との相互作用により転写が抑制されている領域． |
| Predicted promoter region including TSS | 主にGencode遺伝子モデルの転写開始点近傍で，ヒストンのH3K4me3修飾が見られるユークロマチン領域．また，プロモーターとして機能する転写因子やRNA Pol2/3との相互作用から，RNAの転写開始点として機能している可能性が高い． |
| Predicted transcribed region | 転写伸張シグナル（H3K36me3）が見られ，RNAとして転写されていることが予想される領域． |
| Predicted weak enhancer or ope chromatin *cis*-regulatory lelement | プロモーターとしての機能が予想される領域． |

の特定，(2) ChIP-Seq法やFAIRE法による転写制御領域の特定，そして (3) RRBS法やMethyl-Seq法によるDNAの化学修飾領域の特定など，次世代シーケンシング技術を応用したさまざまな実験的手法を用いて行われた．

転写領域の網羅的な特定は，ゲノム機能を理解するうえで欠かせない行程である．ENCODEではヒトの細胞におけるトランスクリプトームを網羅的に解明するため，50以上ものヒト由来細胞種で発現しているRNAの解析を行った．ゲノムから転写されるRNAは，mRNAのように比較的長いRNA（長鎖RNA）と，遺伝子発現の制御などにかかわる200塩基以下の短いRNA（短鎖RNA）に大別できる．特にタンパク質の構造をもたない長鎖の非コードRNA（long non-coding RNA；lncRNA）の一部には (1) クロマチン制御タンパク質と結合し，特異的なゲノム領域のクロマチン構造の変化を誘導する，(2) 相補鎖側から転写されているRNAの分解を抑制することで発現の安定化に寄与するなど，ほかのRNAの転写制御にかかわる機能をもつことが知られていることから，ゲノムに備わった制御機構として重要な役割りを担っていると考えられる[8]．今回の網羅的な解析の結果，これまでにRNAへの転写が知られていない31%ものヒトゲノム領域から，膨大な数の新規lncRNAが転写されていることを明らかにした[16]．同様に短鎖RNAにも細胞機能の制御機構として重要な存在が含まれている．4つの主要なクラス（miRNA：mRNAの転写後調節に関与，snoRNA：rRNAの化学修飾などに関与，snRNA：RNAプロセシングなどに関与，tRNA：リボゾームへのアミノ酸運搬とコドン認識に関与）を中心に，10,000以上の機能性短鎖RNAが知られている[8]．さらに近年，遺伝子のプロモーターや転写終結点から転写されている新規の短鎖RNA，PASR（promoter-associated shortRNA）およびTASR（termini-associated shortRNA）が報告された[17]．今回の解析で特定された短鎖RNAのうち40%ものRNAが，これら新規のRNAと同様の特徴をもつことが示された．タンパク質コードRNAの数をしのぐこれら非コードRNAの機能解明は，ゲノムの制御機構の一つとして今後も注目されていくだろう．

そしてENCODEでは，ヒストンタンパク質が受ける12種類の化学修飾と，ゲノム上に結合している200種類以上の転写因子の分布状況をゲノムワイドに同定し，ゲノムが担う調節領域としての機能の特定に取り組んだ[18]．その結果，ヒトゲノムの8割以上の領域についての生化学的機能（biochemical function；RNAの転写や転写因子の結合，DNAのメチル化修飾など）を明らかにし，ゲノムが担う役割の全体像を初めて詳細に描き出したのである．ヒトゲノムの機能領域（functional element）は，その生化学的な状態に基づいて主に7つの機能要素へと分類され（**表1**），HeLa細胞をはじめとする代表的な細胞におけるゲノムの機能領域が決定された[19, 20]（**図2**）．

ヒトの身体は膨大な数の細胞で形作られており，その細胞は形態や機能から最低でも300種類に分類されている[21]．いずれの細胞でもゲノム配列が同じであることから，これら細胞の差異は発現しているRNAの違い—すなわちトランスクリプトームの違い—に起因していると考えられよう．特異性の高いこれらRNAを見落とさ

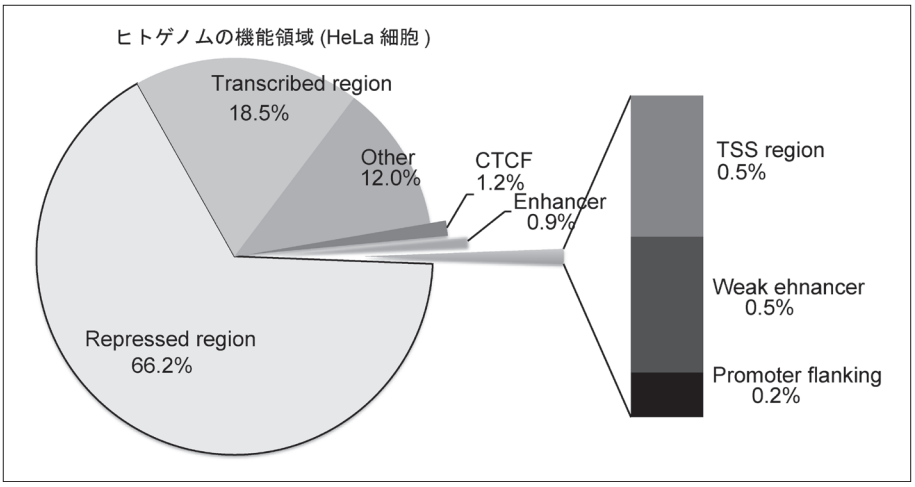



図2■HeLa細胞のゲノム機能領域

ずに特定するためには，ENCODEのように多彩な細胞を対象とした網羅的な解析が有効であると言える．現在，理化学研究所が主導で取り組んでいるFANTOM5プロジェクトでは，ヒトの身体を構成する200種類の細胞型に由来する初代培養細胞と，150種類のがん亜型に由来する細胞株，そして100種類以上の組織サンプルにおけるトランスクリプトームの網羅的な解析を行っている．生体内のトランスクリプトームの多様性と細胞の多様性を結びつけ，複雑な生命システムの全貌解明に向けた取り組みと言えよう（執筆中）．トランスクリプトームの発現パターンはさまざまな制御機構により逐次的に調節され，細胞の機能に必要なRNAを適切に発現することで生物が作り上げられることから，ENCODEによるゲノム機能の解明とFANTOMによるトランスクリプトームの解明により，生命システムの分子基盤の理解が大きく進むことが期待される．

これまでの生命科学研究は，特定の生命現象を司る未同定の因子の存在を仮定し，その存在を明らかにする仮説検証型のスタイルが一般的であった．しかし今日では，網羅的な分子プロファイリング技術と身近になった情報科学との融合により，膨大なデータの中から生命現象の本質を捉えることを目的としたデータ駆動型の研究が盛んになりつつある．膨大な配列により構成されるヒトゲノム機能の解明は，まさにデータ駆動型研究の典型例である．ENCODEでは次世代シーケンシング技術により得られた大規模なデータセットの解析を通して，ゲノムを構成するDNA配列の役割を整理し，ヒトゲノムの8割にも及ぶ領域に機能アノテーションを付けることでヒトゲノムのencyclopedia（百科事典）を作成した．この百科事典を有効に活用することで，今後の生命科学研究は大きく飛躍することだろう．

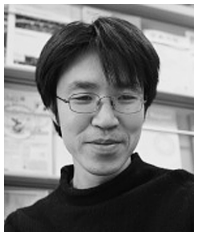

1) International Human Genome Sequencing Consortium : *Nature*, **431**, 931 (2004).
2) F. Crick : *Nature*, **227**, 561 (1970).
3) T. Imanishi *et al.* : *PLoS Biology*, **2**, e162 (2004).
4) P. Carninci *et al.* : *Science*, **309**, 1559 (2005).
5) Y. Hayashizaki : *Genes & Genetic Systems*, **86**, 221 (2011).
6) J. S. Mattick : *The Journal of Experimental Biology*, **210**, 1526 (2007).
7) M. U. Kaikkonen *et al.* : *Cardiovascular Research*, **90**, 430 (2011).
8) M. Esteller : *Nature Reviews Genetics*, **12**, 861 (2011).
9) S. Gustincich *et al.* : *The Journal of Physiology*, **575**, 321 (2006).
10) M. A. Faghihi *et al.* : *Nature Reviews Molecular Cell Biology*, **10**, 637 (2009).
11) C. Carrieri *et al.* : *Nature*, **491**, 454 (2012).
12) H. Suzuki *et al.* : *Nature Genetics*, **41**, 553 (2009).
13) P. J. Farnham : *Nature Reviews Genetics*, **10**, 605 (2009).
14) E. P. Consortium : *Science*, **306**, 636 (2004).
15) E. Birney *et al.* : *Nature*, **447**, 799 (2007).
16) S. Djebali *et al.* : *Nature*, **489**, 101 (2012).
17) P. Kapranov *et al.* : *Science*, **316**, 1484 (2007).
18) R. E. Thurman *et al.* : *Nature*, **489**, 75 (2012).
19) I. Dunham *et al.* : *Nature*, **489**, 57 (2012).
20) M. M. Hoffman *et al.* : *Nucleic Acids Research*, **41**, 827 (2013).
21) B. Alberts : "Molecular Biology of the Cell," 5th ed. Garland Science, New York, 2008.


（梶山和浩[*1,2]，林崎良英[*3]，川路英哉[*1,2,3]，[*1]理化学研究所ライフサイエンス技術基盤研究センター，[*2]横浜市立大学生命ナノシステム科学研究科，[*3]理化学研究所予防医療・診断技術開発プログラム）

プロフィル

**梶山　和浩**（Kazuhiro KAJIYAMA）
＜略歴＞2009年東京農工大学工学部生命工学科卒業／横浜市立大学大学院生命ナノシステム科学研究科に在籍時，理化学研究所オミックス基盤研究領域においてトランスクリプトーム研究に携わる＜研究テーマと抱負＞CAGE技術を次世代シーケンサーに応用したdeepCAGE法を用いて，抗がん剤に対する細胞応答を分子レベルで定量化することで，薬剤応答（特に副作用発現）にかかわる分子基盤の解明を目指している＜趣味＞トランペット吹奏

**林崎　良英**（Yoshihide HAYASHIZAKI）
＜略歴＞1982年大阪大学医学部医学科卒業／1986年同大学医学部大学院博士課程内科系卒業医学博士取得／1992年理化学研究所ライフサイエンス筑波研究センタージーンバンク室研究員／1995年同研究所ゲノム機能解析研究グループプロジェクトリーダー／1998年同研究所ゲノム科総合研究センター遺伝子構造・機能研究グループプロジェクトディレクター／2008年同研究所オミックス基盤研究領域領域長／2013年同研究所社会知創成事業予防医療・診断技術開発プログラムプログラムディレクター，現在に至る＜研究テーマと抱負＞オミックス科学の医療転換＜趣味＞ジョギング，建築デザイン，ワイン（飲むこと）



**川路　英哉**（Hideya KAWAJI）
＜略歴＞2003年大阪大学大学院基礎工学研究科情報数理系修了，博士（工学）／NTTソフトウェア(株)を経て，理化学研究所オミックス基盤研究領域研究員，同領域ユニットリーダーとしてゲノミクス，特にトランスクリプトームの情報処理・解析に従事．現在は同研究所予防医療・診断技術開発プログラム コーディネーター，横浜市立大学大学院生命医科学研究科客員准教授＜研究テーマと抱負＞私たちの生活に，ゲノミクスの成果が活用されはじめてきています．情報技術・ゲノミクスデータを使って，これをより有効に押し進める研究・開発を行っています＜趣味＞音楽