トピックス

今までのトランスクリプトーム観を覆した「RNA 新大陸」の発見

独立行政法人　理化学研究所
ゲノム科学総合研究センター　遺伝子構造・機能研究グループ　高橋　真介，林崎　良英

はじめに

1953 年のワトソン・クリックによる DNA の構造と遺伝情報の流れであるセントラルドグマの発見が，20 世紀後半から現在に至る社会の様々な分野に与えた影響は計り知れない。遺伝子の物質的本体が DNA であったことの意義は，遺伝現象の解明のみならず，生化学分野に発展性を秘めた方向性が開けたことであった。その後，分子生物学として発展した分野において，長年の夢であった生命の本質に迫る発見が続いた。1970 年代の遺伝子組換え技術，DNA 塩基配列決定法の確立に続き 1980 ～ 90 年代の自動化技術の開発によってゲノムの全体像をつかむ方向へ研究は進んで，ついに国際ヒトゲノムコンソーシアム※1 は 2003 年 4 月 14 日にヒトゲノムの全塩基配列の決定を宣言した。しかし，ゲノムの塩基配列の中にいったい何が書かれているのかという情報に関して，いまだに詳細には知られていない。今回，２報の論文として米国の科学雑誌『Science』（９月２日号）に掲載された我々の研究グループの成果は，トランスクリプトーム※2 の実像の一端ですら我々の予想を超えた可能性が存在することを示したことである。またセントラルドグマに内包されない RNA の流れを発見したことである。

全世界 11 ヶ国／ 45 ヶ所の研究機関などでマウスゲノムの研究を展開している国際コンソーシアム「FANTOM」と，国家プロジェクトである「ゲノムネットワークプロジェクト」の両コンソーシアムは，哺乳動物の細胞が生産する RNA の中のトランスクリプトームの総合的解析を，今までにない大規模スケールで行った。

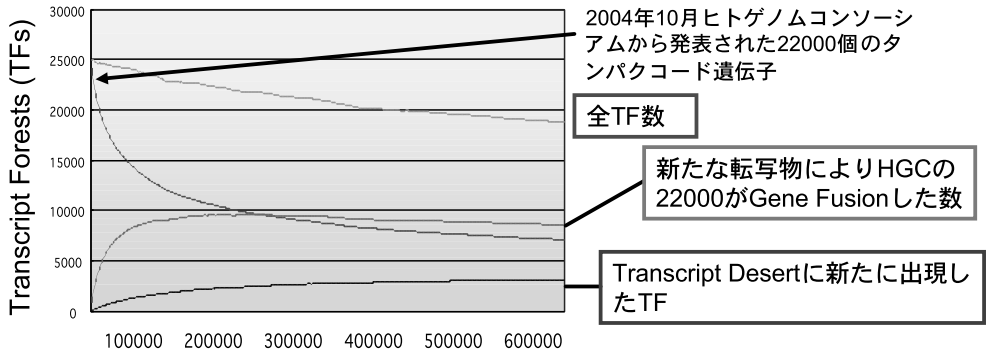
その結果，従来 100 個ぐらいしか知られていなかった「非タンパクコード RNA（Non-coding RNA; ncRNA※3）」が，当研究グループの以前からの報告数をはるかに超えて 23,000 個以上存在することを突きとめた。また，ゲノムに存在するプロモーター※4（転写開始点）が 18 万個以上であることなどを突き止め，さらに，二重鎖 DNA の双方の鎖が読まれるセンス／アンチセンス※5 の RNA ペアが従来考えられてきた数をしのぐ 36,000 ペアもあることを発見した。

この成果は，タンパク質がゲノムにコードされている最終機能物質であるという常識を覆し，予想を凌ぐトランスクリプトームの複雑さを認識させるもので，「RNA 新大陸」を発見したと言え，哺乳動物ゲノムの情報内容に対するこれまでの理解（「遺伝子」という領域が散在しているゲノムのイメージ）を根幹から変えてしまうものである。

背　景

2004 年 10 月，国際ヒトゲノムコンソーシアムより，ゲノムのたった約２％の領域が，生物の体を作り上げている主要部品である約 22,000 のタンパク質をコードしていると発表されたが，これは，一部既存の実験データを入れたものの，コンピューター予測によるところが多く，ゲノムにコードされている情報は本当にそれだけなのかが，依然として不明であった。また，これらの２％が，各生体内組織のどの発生ステージで選ばれていき，それらは如何にして制御されているのかという問題も残っていた。つまり，長いゲノムの中にオアシスのようにポツポツと散在している遺伝子がどう働くのかが課題であった。

トランスクリプトーム（転写物集団）とは，多くの個別 RNA 分子（転写物）によりなりたち，



それらはもともとゲノム DNA から転写されたものであり，多くの場合，RNA 分子はタンパク質へと翻訳され，最終生理活性物質となるとされていた。トランスクリプトーム解析がゲノム塩基配列決定よりもさらに労力を必要とする理由は，膨大な種類数の RNA 分子がゲノム DNA から生産され，これらは RNA を鋳型として合成する完全長 cDNA※6として個別に単離解析されなければならず，完全長 cDNA の合成に高度な技術を要するからである。

研究手法と成果

この研究では，トランスクリプトーム解析のために理化学研究所（理研）独自に開発した 4 種類の新技術を使った。研究グループが最初に開発した技術は完全長 cDNA クローニング法であり，それは完全な mRNA 配列を cDNA の形で写し取る技術で，この方法によって，257 種以上の組織から単離した総数 200 万個以上の完全長 cDNA を，その末端配列から分類しつつ，10 万 3 千個のマウス完全長 cDNA の配列を決定した。他の 3 つの技術全ては，RNA 配列の先頭である 5′ 末端と末尾である 3′ 末端の収集およびマッピングを高速かつ大量に行うものである。

マッピングの結果では，遺伝子の途中など色々なところから転写が開始されており，しかもこれは，ランダムではなく規則性をもって読まれていることがわかった。また，短いものや長いものが収集されたことから，ゲノム上の色々な部分が読まれていることもわかった。同一の遺伝子から，複数の転写を制御するプロモーター（転写開始点近傍配列），選択的スプライシング※7，複数の PolyA 付加サイト※8（3′ 末端：RNA 末尾）など，多様な RNA が生産されることが判明したのである。

この研究がなされるまでは，同一ゲノム領域から転写されても選択的スプライシングを起こして異なった転写物も異なった遺伝子産物として数えていたが，今回，これらを同一ゲノム上の 1 つの遺伝子産物として数え直したところ，22,000 個と推定されていた遺伝子の数が 2,300 個まで減少した。（図：遺伝子数の融合減少のグラフ）

以前から大量の転写物の塩基配列決定では，遺伝子産物の逆側の配列・アンチセンスがわずかに得られることが知られていたが，それらは mRNA から正しく合成された cDNA ではなく，逆転写酵素の忠実度の低さによる合成間違いか，ゲノム DNA 断片の混入による間違いだとして実験データから排除され続けていた。しかし，我々が執拗に確認したところ，ほとんどのアンチセンスが人工的なものでなく自然界に存在することがわかったので，ゲノムの両方の配列を差別することなく解析した。ゲノム DNA 上で同一鎖上にあり，エキソンに 1bp（ベースペア）以上の重なりがある転写物（transcript）をグループ化した際のエキソン領域の集合を Transcriptional Unit（TU）と定義し，約 200 万個の完全長 cDNA を詳しく分類したところ，41,147 種類の遺伝子 TU を発見した。これらの TU の半分以上が，タンパク質をコードしていない RNA（ncRNA）であることが明らかとなった。

同一ストランドにある転写産物をエキソンの重なりによりグループ化したゲノム上の連続領域を

Framework Cluster（FC）と呼び，DNA 二重鎖のどちらか一方が転写される場合，そのゲノム領域を Transcript Forest（TF），両鎖とも転写の対象にならないゲノム領域を Transcript Desert（TD）と定義して調べたところ，TF の割合は 70％を超えた。全ゲノムの 7 割以上の領域が RNA に転写され，そのうちの 5 割を超える RNA が ncRNA であったことは未踏の「RNA 新大陸」の存在を学術的に立証したわけである。

それらのエキソン領域は種間（ヒト－マウス）で保存されていないにもかかわらず，プロモーターの配列が保存されていたことは特筆するべき事実である。このことは，ncRNA では，センス／アンチセンス（S/AS）による 2 重鎖 RNA を介したメカニズムが機能しているのではないかと推察され，エキソンの配列よりも，いつどこで発現するのかということが重要であることを示唆している。

これらのデータは，哺乳類の分化や発生での転写制御の比較分析のための網羅的基盤となる。今回，新規完全長 cDNA 配列のうち，16,247 個のマウスの新しいタンパクコード転写産物を同定したが，そのうち 5,154 個の転写産物は，既知のタンパク質とは全く異なる新規タンパク質をコードしていた。

ゲノム上で双方の鎖が RNA に転写されているような DNA の領域，つまり，双方の RNA がペアを作ることが非常に多く見られ，36,000 個のセンス／アンチセンス（S/AS）のペアを発見した。この S/AS の RNA ペアはゲノムのほとんど全領域で普遍的に起こりうるということを示唆しており，細胞周期，タンパク質輸送，細胞死，細胞構造と接着，細胞分化，リン酸化酵素，インターロイキン，Ras タンパク質，ユビキチン化などの機能を持っている遺伝子によく見られた。この中には，重要なヒトの疾患原因遺伝子も含まれており，新たな薬剤の標的になりえることが考えられ，さらに，これらをノックアウトや強発現による手法を活用し，より詳細に解析すると S/AS による制御は通常の RNAi※9 現象で単純に説明できるものではないことが判明した。この研究で，アンチセンス RNA により，センス RNA の発現がコントロールされていることも明らかになり，アンチセンス転写は哺乳動物の転写制御に大きな役割を担って，それらのメカニズムに ncRNA が一役かっていることがわかった。

研究の意義

これらのデータは生物・医学研究領域において，高等動物のあらゆる生命現象を理解する手段となろう。ゲノム配列は，哺乳動物の部品（タンパク質）を作るための暗号であるのみならず，いつ，どの組織で発現するかという情報も含んでいる。今回作成された国際標準となるデータベースは，現在のところ世界で最も完全なトランスクリプトームの全体像を提供している。

哺乳動物ゲノム内には，「タンパクコード遺伝子」の種類はショウジョウバエのわずか 2 倍しかない（2004 年 10 月ヒトゲノムコンソーシアム発表数，約 22,000）。今回，FANTOM コンソーシアムの研究では，新たな配列を含む 56,722 種類の cDNA が見つかった。その中には，リボゾーム RNA，トランスファー RNA を除くと従来 100 個ぐらいしか知られていなかった ncRNA が，予想をはるかに超える 23,000 個以上存在することが突き止められた。さらに，これらが単なる間違って漏れ出てきた RNA ではなく，生体内で機能しているということを証明したことと合わせると，従来の「タンパク質がゲノムにコードされている最終機能物質である」という常識は覆り，人類未踏の領域である「RNA 新大陸」が発見されたことになる。

今回の研究で，最近まで我々の生物学領域で存在や機能を考慮されなかった大量の ncRNA によって遺伝子発現が制御されていることを示した。ほとんどのタンパク質が哺乳類では類似だが，生物種間に形質・形態の差異を生じさせる理由の多くは，タンパク質構成要素系より，さらに速く進化している RNA 調節制御系の違いに隠されているのだろう。もしこの考えが正しいなら，この発見は次に示すような生命科学，医学やバイオテク

ノロジーの将来にとっての重要な疑問に対する解答を劇的に変化させるだろう。

① いかにして遺伝情報が我々のゲノム中に蓄えられるのか？

② いかにしてこの遺伝情報が複雑な哺乳動物の発生過程を制御するために処理されるのか？

ここで示した研究は，一部ヒトのデータを含むもののマウスを中心とした解析である。現在我々の研究グループではヒトの大規模データも同様に準備中であり，この成果はヒトの疾患を理解する手助けになると期待される。

今後の展開

本研究を通じて，遺伝子とは何か，という基本的概念にパラダイムシフトが起きたと言えるであろう。ゲノムの中に遺伝子がオアシスのように散在するという旧来のゲノム観から，かつて「ジャンク DNA」と呼ばれていた領域は，実際には機能しており，ゲノムは総体として働いているという新しいゲノム観が生じたといっても過言ではないだろう。

さらに，本研究における「RNA 新大陸」の発見は，「タンパク質が最終生理活性物質であり，遺伝子とは，単にタンパク質をコードするもの」であるという既成概念を崩す結果となった。遺伝子から，表現形質を分子レベルで説明するネットワークの中に，新たに ncRNA が登場することとなったのである。

現在のトランスクリプトーム解析は，いまだに完成していない。完全長 cDNA とはまったく独立したアプローチであるタイリングアレイ※10 のデータと考えあわせると，PolyA RNA の約半分ぐらいが本研究の解析対象となったことが推察されている。さらに，Non-polyA RNA は，PolyA RNA とほぼ同数あることも予測されており，全体にとってわずかな部分しか手のつけられていないトランスクリプトーム解析は，まだまだ始まったばかりであるといえる。トランスクリプトームは，どのステージで，どの組織で発現しているのかという情報もあわせると，ゲノム解析と比べはるかに動的であり複雑である。将来には，さらなるトランスクリプトーム解析が必要となり，今回の成果は遺伝子機能を詳細に研究するための必須の知見をもたらしたといえよう。

補足説明

※1　国際ヒトゲノムコンソーシアム
ヒトの全ゲノム配列を解読することを目的とした研究機関の国際的な共同集団。

※2　トランスクリプトーム
RNA 合成酵素によってゲノム情報から写し取られた転写物集団。狭義な旧来のセントラルドグマの定義では，mRNA を主要なものとして考え，それ以外をジャンク（不要物）としていた。

※3　ncRNA
この RNA からはタンパク質は翻訳されない。

※4　プロモーター
転写開始を促す活性を持つ DNA 上の特定の領域・塩基配列。

※5　センス／アンチセンス
遺伝情報としてタンパク質に合成されうる配列の方向性をセンス，センス配列に対して相補的で逆の方向性をアンチセンスという。

※6　完全長 cDNA
成熟 mRNA を鋳型として合成された完全な cDNA のこと。cDNA とは相補 DNA のことで，分解し易い mRNA の情報を保存するため人為的に逆転写酵素を使って合成される。

※7　選択式スプライシング
真核生物の DNA から RNA が転写されるときに mRNA 前駆体のエキソン部分だけがつながり，イントロンが除かれて成熟 RNA となることをスプライシングという。順番どおりにつながらず異なったパターンで途中のエキソンを抜かしてつながり成熟 RNA となることは，選択的スプライシングと呼ばれる。

※8　PolyA 付加サイト
アデニル酸が 200 から 300 塩基重合する成熟 mRNA が 3' 末端末尾にもつ特異的配列部位。PolyA RNA は実際上 mRNA 識別の指標となり，Non（非）-polyA RNA は，この研究がなされるまでは，完全な mRNA が分解された無意味なものと考えられてきた。

※9　RNAi
RNA interference（RNA 干渉）の略で，二本鎖 RNA によるタンパク質翻訳の選択的阻害現象。

※10　タイリングアレイ tiling array
塩基配列を検出用プローブとしてシリコン基盤上に搭載した DNA チップで，ゲノムデータから等間隔に抜き出した配列を使えば，DNA の配列の違いを超高速で検出できる。