報道発表資料

2005 年 9 月 2 日
独立行政法人 理化学研究所

哺乳動物のトランスクリプトームの総合的解析による「RNA 新大陸」の発見

◇ポイント◇
- 哺乳動物のトランスクリプトームの総合的解析による「RNA 新大陸」の発見
- タンパクを作り出さない RNA（非タンパクコード RNA）について、予想をはるかに超える 23,000 個以上を発見した。（「RNA 新大陸」の発見）
- 二重鎖 DNA の両方の鎖が転写された RNA（センス／アンチセンスの RNA ペア）につい考えられてきた数をはるかにしのぐ 31,422 個を発見。

独立行政法人理化学研究所（野依良治理事長）を主体とした全世界 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※1」などは、細胞が生産する RNA を今までにない大規模スケールで、哺乳動物のトランスクリプトーム※2 の総合的解析をし、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ました。また、ゲノムに存在するプロモーター※4（転写※5 開始点）が 180,000 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※6／アンチセンス※7 の RNA ペア※8 が従来考えられてきた数をしのぐ 31,422 ペアもあることを発見しました。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示唆された。

これらの成果は、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※9」を発見したと言えます。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す。

本研究成果の詳細は、2 報の論文として米国の科学雑誌『Science』（9 月 2 日号）に掲載されます。また同時に、関連する基礎情報は、理研のサイト（http://fantom3.gsc.riken.jp/db/）と国立遺伝学研究所の日本 DNA データバンク（DNA Data Bank of Japan：DDBJ , http://www.ddbj.nig.ac.jp/）のデータベース上で公開されます。

論文のとりまとめにあたり、我が国においては、理研・ゲノム科学総合研究センター（榊佳之センター長）の遺伝子構造・機能研究グループの林崎良英プロジェクトディレクターらが中心となり今回の成果に大きく貢献しました。

なお、本研究成果には、ゲノムネットワークプロジェクトの一環から得られたものも一部含まれています。

1.背　景

過去 5 年間、数種類の哺乳動物のゲノムがつぎつぎと解読されてきましたが、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ません。2004 年 10 月、国際ヒトゲノムコンソーシアム※10により、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ましたが、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、本当にそれだけが、ゲノムにコードされている情報なのかは、依然として不明でした。また、各生体内組織のどの発生ステージで、これらの 2%が選ばれていき、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ます。

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）により成り立ち、それらはもともとゲノム DNA から転写されたものであります。そして多くの場合、RNA 分子はタンパク質へと翻訳※11 され、最終生理活性物質となります。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、これらは個別に RNA を鋳型として合成する完全長 cDNA※12 として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す。

理研・ゲノム科学総合研究センターの遺伝子構造・機能研究グループと FANTOM コンソーシアムは、マウスゲノムから生産された 2,000,000 個以上の cDNA 配列を解析し、**そこから 103,000 個の完全長 cDNA 配列を決定しました。さらに、遺伝**子構造・機能研究グループが開発した CAGE（Cap Analysis of Gene Expression）※13、GSC（Gene Signature Cloning）※14、シンガポールの Genome Institute of Singapore が開発した GIS（Gene Identification Signature）※15 と合わせ、転写の開始と終了の位置を表すタグ配列を作り出し、この解析に利用しました。

2. 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研と FANTOM コンソーシアムが独自に開発した 4 種類の新技術を使いました。研究グループが先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す。この方法によって、理研は 257 種以上の組織から単離した総数 2,000,000 個以上の完全長 cDNA を、その末端配列から分類わけしたのち、103,000 個のマウス完全長 cDNA の配列を決定しました。さらに 3 つの技術全ては、RNA 配列の先頭である 5'端※16（CAGE、GIS/GSC）と末尾である 3'端※16（GIS/GSC）の収集およびマッピングを高速かつ大量に行うものです。この技術を使ってマウスの 11,567,973 個、ヒトの 5,992,395 個の CAGE タグ情報と、マウスの 2,465,449 個の GIS/GSC タグ情報をそれぞれ得ました。

このような大規模データをゲノム上にマッピングし解析を行った結果、同一の遺伝子から、複数の転写を制御するプロモーター（転開始点）、選択的スプライシング※17、複数の PolyA 付加サイト※18（3'端:RNA 末尾）など、多様な RNA が生産されることが判明しました。また約 2,000,000 個のマウス完全長 cDNA を詳しく分類し、44,147 種類の遺伝子（Transcriptional Unit: TU※19）を発見しました。これは、ゲノムの 70%に相当する広大な領域が、一旦は RNA に読まれていることがわかりました。さらにこれらの TU の半分以上が、タンパク質をコードしていない RNA（ncRNA）が 23,218 個あることが明らかとなりました。それらのエクソン※20 領域は種

間（ヒトーマウス）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す。このことは、後述するように、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、エクソンの配列よりも、いつどこで発現※21 するのかということが重要であることを示唆しています。

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ります。今回、新規マウス完全長 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ましたが、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ました。

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、31,422 個のセンス／アンチセンス（S/AS）のペアを発見しました。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ます。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ます。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※22 現象で単純に説明できるものではないことが明らかになりました。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がわかりました。このことは、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おり、それらのメカニズムに ncRNA が一役かっている事実は、非常に面白い結果です。

3. 研究成果の意義

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ります。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ます。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ます。

哺乳動物ゲノム内には、「タンパクコード遺伝子※23」がショウジョウバエよりもわずかに 2 倍の種類数のしかありません（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 コンソーシアムは、マウス完全長 cDNA で新たな配列を含む 56,722 種類の cDNA を見つけました。その中には、リボゾーム RNA※24、トランスファーRNA※25 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を突き止めました。さらに、これらが単なる漏れ出てきた 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あわせると、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ります。

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ました。ほとんどのタンパク質が哺乳類では類似なので、生物種間に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ことを示唆しています。もしこの考えが正しいなら、この発見は下に示すような生物学研究、医学やバイオテクノロジーの将来にとって重要な疑問に対し、予測される解答を劇的に変化させます。

(1) 如何にして遺伝情報が我々のゲノム中に蓄えられるのか?

(2) 如何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?

これらの研究は、一部ヒトのデータを含むマウスを中心とした解析であり、主に理研を中心とするFANTOMコンソーシアムによって成し遂げられました。現在、先に挙げた新手法（完全長cDNA、CAGE、GSC、GIS）を用い、ヒトの大規模データも同様に準備中であり、ヒトの疾患を理解すると期待されます。

今回得た全ての情報は、日本時間2005年9月2日午前3時（2005年9月1日午後2時アメリカ東部沿岸時間）のサイエンス誌上発表と同時に、理研のサイト（http://fantom3.gsc.riken.jp/db/）と国立遺伝学研究所生命情報・DDBJ研究センター（五條掘孝センター所長）のデータベース;DDBJよりインターネット上で公開します。

4. 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考えています。ゲノムというもののなか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言い過ぎではありません。

さらに、本研究における「RNA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りました。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ります。これにより、RNAがいろいろなレベルで遺伝子の発現を調節する新たなメカニズムの研究がスタートすることになるでしょう。

現在のトランスクリプトーム解析は、いまだ完成していません。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※26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ます。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、トランスクリプトーム解析は、始まったばかりであるといえます。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す。将来には、さらなるトランスクリプトーム解析が必要性となり、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ことになりました。

（問い合わせ先）
独立行政法人理化学研究所　横浜研究所
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
プロジェクトディレクター　林崎　良英
Tel：045-503-9222 / Fax：045-503-9216

研究推進部　企画課　星野　美和子
Tel：045-503-9117 / Fax：045-503-9113

（報道担当）
独立行政法人理化学研究所 広報室
Tel：048-467-9272 / Fax：048-462-4715
Mail：koho@riken.jp

<補足説明>

※1 FANTOM

理研が中心となって結成された哺乳動物（マウス）の遺伝子を網羅的に機能注釈することを主眼とする国際的研究コンソーシアム共同集団の略称です。オーストラリア、シンガポール、スウェーデン、南アフリカ、イタリア、ドイツ、ギリシャ、スイス、英国、米国などを含む全世界の 11 ヶ国／45 ヶ所の研究機関等が参加しています。

※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）としていました。

※3 ncRNA

非タンパクコード RNA（Non-coding RNA）のことで、この RNA からはタンパク質は翻訳されません。

※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・塩基配列をいいます。

※5 転写

遺伝子 DNA から RNA が読み取られることです。

※6 センス

遺伝情報としてタンパク質に合成される配列の方向性です。

※7 アンチセンス
センス配列に対して相補的で逆の方向性です。

※8 RNA ペア
同一ゲノム上のセンスとアンチセンスの両方向の転写 RNA が、相補的に結合した複合物の状態です。

※9 RNA 新大陸
この研究で新たに提案された遺伝子の定義により、評価し直された ncRNA などの多様な細胞内 RNA 集団の莫大な可能性を示す比喩的表現です。

※10 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のことです。

※11 翻訳
転写された mRNA 情報をもとにリボゾームで行われるタンパク質合成を意味します。

※12 完全長 cDNA
cDNA は相補 DNA のこと。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ます。先頭のキャップ構造から末尾の polyA 付加まで備えた成熟 mRNA を鋳型として合成された完全な cDNA のことです。

※13 AGE （Cap Analysis of Gene Expression）
耐熱性逆転写酵素や cap-trapper 法を組み合わせて転写物の 5'末端から 20 塩基のタグ配列を切り出し、塩基配列を決定する実験技法です。

※14 GSC （Gene Signature Cloning）
次の GIS と同様に転写物の 5'末端と 3'末端の塩基配列同定する大量処理技術ですが、微量な mRNA からも検出できます。

※15 転写物の 5'末端と 3'末端の塩基配列同定する大量処理技術のことで、転写物の変動性を知ることが出来ます。

※16 5'端、3'端
核酸合成は、構成単位のヌクレオチド分子内の五単糖の炭素の位置で考えると 5'から 3'方向へ進むので、鎖の 5'端が先頭になり、3'端が末尾となります。

※17 選択的スプライシング
真核生物の DNA から転写された mRNA 前駆体が成熟 mRNA になるためにイントロン部分だけが選択に切り出される過程をスプライシングと呼びますが，イントロ

ンが複数存在するとき，異なったパターンのスプライシングが起こり，除去されるイントロンが異なる成熟 mRNA が産生されることを選択的スプライシングと呼びます。

※18 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'端末尾にもつ特異的配列部位のことです。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ました。

※19 Transcriptional Unit （TU）

ゲノム DNA 上で同一鎖上にあり、エクソン 1bp 以上 overlap がある transcript をグループ化した際のエクソン領域の集合を意味します。

※20 エクソン

mRNA の塩基配列をコードする DNA の構造配列。エクソン間に挟まれた非コード領域をイントロン呼びます。

※21 発現

遺伝子がその表現形質をあらわすこと。分子生物学の文脈では、遺伝子 DNA 情報が転写されること、またはさらに翻訳までいくことを示します。

※22 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をこう呼びます。

※23 タンパクコード遺伝子

開始コドンと終止コドンを両方持った mRNA を合成できる情報をもつ DNA 配列。実際にはプロモーターが必要になります。

※24 リボゾーム RNA

タンパク質と共に細胞内小器官（オルガネラ）であるリボゾームを構成する RNA の一種です。

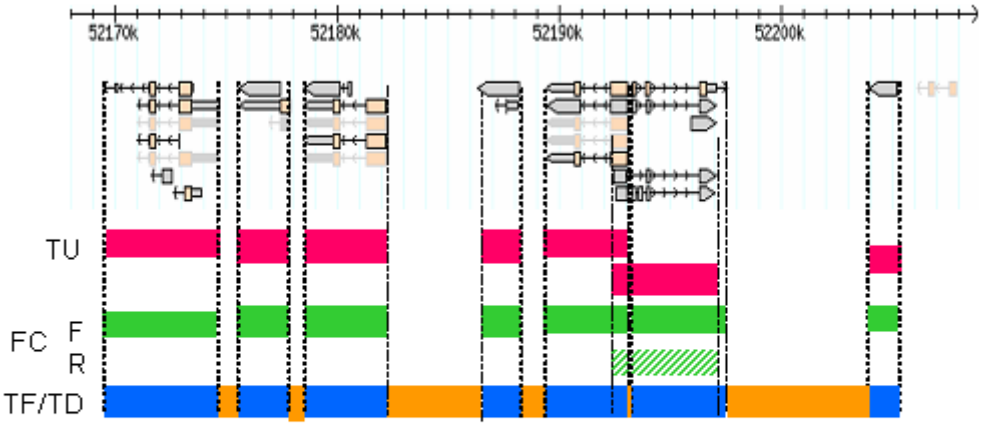
※25 トランスファーRNA

翻訳において、アミノ酸を運搬する機能をもつ RNA の一種です。

※26 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 チップです。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ます。

新たに提案する遺伝子の定義



Transcriptional Unit (TU): ゲノムDNA上で同一ストランドにあり、エクソンに1bp以上overlapがあるTranscriptをグループ化した際のゲノム領域。

Framework Cluster(FC): 同一ストランドにあるtranscript(イントロンも含む)をoverlapによりグループ化したゲノム上の連続領域。

Transcript Forest(TF): いずれかのストランドがmRNA前駆体として転写の対象となるゲノム領域。

Transcript Desert(TD): 両ストランド共に転写の対象とならないゲノム領域。