比較ゲノム解析に基づくヒト固有遺伝情報の同定と機能推定

●渡邉 日出海　◆小柳 香奈子
北海道大学大学院情報科学研究科

＜研究の目的と進め方＞

本研究課題では、ヒトに固有に備わる形質（ヒト固有形質）の発現分子機構を解明するための重要な手がかりを得ることを目指す。その重要な手がかりとして、ヒト固有形質に密接に関連すると推定されるヒト固有遺伝情報を同定する。この目的のために次の４点を実施する。(1) 平成 12 － 18 年度に獲得した科学研究費補助金を用いて構築してきたゲノム比較解析のための計算機環境を用いて、ヒトとチンパンジーが分かれた後にヒトゲノムに生じたヒト固有遺伝情報ならびにヒト特異的進化状態を示すゲノム領域を同定する。(2) その中からヒトの生物種としての個性（固有形質）に関連する可能性のあるものを選別する。(3) それらのヒト固有形質対応候補領域の有無をヒト集団と類人猿集団において実験的に確認する。(4) さらに基礎的分子生物学実験を行うことでそれらの表現型における特性を他の生物との間で比較する。ヒトの生物学的理解を進めるために、マウスなどのモデル生物を用いた実験を行う場合が一般的であるが、マウスのようにヒト固有形質を持たない生物を用いている限り、ヒト固有形質の発現分子機構を正確に解明することが出来ない。実際、ヒト固有形質の発現分子機構はまだほとんど何も解明されていない。本研究課題は、系統立ててヒト固有形質の発現分子機構を解明するための細胞レベルでの基礎的情報の蓄積を行うものである。遺伝情報解析を起点としたヒトゲノム機能解析は、表現型を起点とした機能解析とは異なり、ヒトの未知機能（表現型）の発見をもたらす可能性を持ち、ヒト（ゲノム）生物学の長足の発展に寄与することが期待できる。

＜ 2007 年度の研究の当初計画＞

これまでに作成してきたヒトーチンパンジーーマカク間ゲノムアラインメント（ヒトゲノムは完成配列データ、その他２種は概要配列データ）を基盤として、ヒト系統特異的ゲノム構造変化を様々な視点で検出する。特に、これまで解析を行ってきたヒト系統特異的塩基置換･アミノ酸置換の速度変化を高精度化し、また、CDS 領域内におけるヒト系統での構造変化については継続して実験による確認を進め、完了する。今年度に入ってからの解析により、74 個の遺伝子（予測遺伝子を含めると 110 個）の CDS 領域においてヒト系統において高い塩基置換速度を示すことが明らかになっている。

また、一方で、ゲノムアラインメントについても、曖昧な領域を実験による確認や分子進化学的解析を詳細に行うことによってその精度を再帰的に高めていく。これまでの解析の過程で、特にチンパンジーゲノム配列データの質がかなり低いことが明らかになってきているため、この手続きは本研究をまとめるために必須であると考えられる。

解析結果の確からしさを確認でき次第、班員等との共同研究を積極的に進め、ヒト特異的変化領域の生物学的意味を明らかにする。

＜ 2007 年度の成果＞

これまでに、新たにゲノム比較解析アルゴリズムとツールを開発し、それをヒトゲノム完成配列データとチンパンジー概要配列データに応用することでヒトーチンパンジー間高精度ゲノムアラインメントを作成した。両ゲノム間の相違に関する詳細なデータを作成した結果見出された相違箇所は膨大な数（計算の仕方によって値が変わるが塩基レベルでゲノムの５％を越える）に上ることが明らかになり、塩基置換（1.23%、2002 年に Science 誌に発表）よりも、挿入・欠失などによる違いが圧倒的に多いことが確認された。そこで、通常の分子進化学的解析以外の配列比較解析も実施する必要があると判断した。

次に、両ゲノム間の違いのうちどれがヒトゲノムで生じたものなのかを明らかにするために、米国で実施されているマカクゲノムのホールゲノムショットガン配列決定によるデータを用いた比較解析を実施した。使用したデータは、生リードデータとその精度データであり、リード数で 24,843,646 本、総塩基数 23,504,128,538 の冗長度約８のものである。この初期解析では、あえてアセンブルされたデータを用いなかった。その理由は、利用可能な情報を最大限に活用し、ミスアセンブリによる情報喪失を避けるためである。他に、オランウータン２種の概要配列データも利用可能であったが、リードの冗長度が２倍程度と低かったため解析に用いなかった。

ヒトーチンパンジー間で異なっている部分のみにおいて、マカクゲノムデータとの比較解析を行った。それは、ヒトーチンパンジー間最終共通祖先におけるゲノム構造の推定にはそれだけで十分だからである。その比較結果の全てに目を通して重要な情報を引き出すことは困難であったため、対象領域に優先順位をつけ、まず、最優先領域として遺伝子領域に対する詳細な解析を行った。その結果、Ensembl データベースでアノテーションが付されている遺伝子約 33,000 個（仮想遺伝子を含む）のうち、7,836 個の遺伝子でヒト固有領域が見出された。そのうち 443 個の遺伝子において、CDS 領域内挿入・欠失が見つかった。CDS 領域以外のイントロンや UTR 領域、遺伝子外領域においても多数のヒト固有領域が見つかった。

また、解析が困難であるために従来は対象外とされていたヒトとチンパンジーの種分岐後に遺伝子重複が起きた 125 遺伝子群についても、分子進化解析を行った。その結果、そのようなヒト系統特異的重複遺伝子間では進化速度にばらつきが認められ、そのうちヒト系統で特異的に進化速度の上昇の起きている遺伝子が 77 個見つかった。

以上の解析結果は、主として 2005 年に発表されたチンパンジーゲノム概要配列データに基づく解析によって得られたものであった。2006 年 9 月以降に、冗長度 6 のチンパンジー概要配列アセンブルデータ（PTR2.1）ならびにマカクゲノムのアセンブルデータ (MMUL1.1) が公共データベースから公開されたことを受けて、これらのデータを用いてゲノムアラインメントを作成しなおし、同様の解析を行ってみた。その結果、ヒト固有領域の推定に関して、2005 年版での結果との間に多くの矛盾が見つかった。つまり、チンパンジーゲノム配列データが 2005 年版のものから大幅に変更されたことを意味する。

この再解析の結果、ヒト全遺伝子領域のうち、非反復配列領域において登録データ上 50 塩基対以上のヒト固有領域が存在する可能性が示唆されたものは 142 個になった。これらの領域の多く

には、チンパンジーゲノム配列が未決定または曖昧な状態である部分が存在する。解析対象領域にこのような問題となる部分が多く見つかる理由として、ヒトゲノム全体の低 GC 率（40％程度）に最適化された条件で全ゲノムショットガン配列決定が実施されたため、高 GC 率領域を多数含むコード領域の配列は決定されずに残っている、という生物学的背景が考えられる。

この問題箇所は、本研究の推進と完了を妨げる重大な問題であり、どうしても解決しなくてはならない。そこで、それらチンパンジーゲノム中配列未決定領域を個別に配列決定することを進めている。この配列決定では、京都大学霊長類研究所で飼育されているメスのチンパンジー 1 個体から採取した 8ml の血液を処理し、増幅した核 DNA を用いている。BAC クローンではなくゲノム DNA を用いていることには理由がある。まず、いかなる対象領域でも同じ試料（ゲノム DNA）を用いて実験を行えることが挙げられる。配列決定対象領域は、比較的短く、また、ゲノム全体に散在している。したがって、BAC クローンを用いようとすると、その対象領域を含む BAC クローンを領域ごとに選び出し（この手続きには、前ゲノム特定領域研究において本研究代表者が開発したシステムを利用できる）、ライブラリー管理者からクローンを持つ大腸菌を取り寄せ、培養し、DNA を抽出し、その各々の DNA を分けて解析をする、という手順を踏む必要が生じる。一方、ゲノム DNA を直接用いると、これらの手順の全てが不要になる。もう一つの重要な理由は、クローニングの際に高 GC 率領域に変異が入りやすいという問題を避けることが挙げられる。我々は、塩基配列レベルで 1％程度の違いしかない配列同士を比較しているので、実験による人為的変異の導入を極力避ける必要がある。

上に示したように、配列決定対象領域のほぼ全てが高 GC 率（70％に達するものも少なくない）であったり、反復構造を持ったりするため、配列決定のための条件を工夫する必要があった。例えば、製品化されている複数の PCR 用ポリメラーゼを用いて、温度や時間、バッファに添加する物質の種類や濃度などに関すの様々な条件での PCR 産物の生成状態と配列決定状態を調べた。現在までに、いくつかの有用な条件を見出し、活用している。

以上に示したように、ある程度の長さを持つヒト固有ゲノム領域の同定を進める一方で、ヒト系統における特異的進化速度変化を示す領域の同定も進めている。具体的には、マカクゲノムを外群として、ヒトーチンパンジー間最終共通祖先（LCA）のゲノム配列を推定し、非反復配列領域における LCA- チンパンジー対 LCA- ヒトの進化速度比較を行い、有意差がある領域を同定している。この解析の結果、ヒト系統で有意に進化速度が大きくなっている遺伝子として、UGT2B10(UDP glucuronosyltransferase 2 family, polypeptide B10)、TRIP13(thyroid hormone receptor interactor 13)、FCGR2C(Fc fragment of IgG, low affinity IIc, receptor for (CD32))、ADCYAP1(adenylate cyclase activating polypeptide 1 (pituitary))、DLGAP2(discs, large (Drosophila) homolog-associated protein 2)、WNT5B(wingless-type MMTV integration site family, member 5B)、NAP1L4(nucleosome assembly protein 1-like 4)、ALDH1A3(aldehyde dehydrogenase 1 family, member A3) 等が検出された。

上記チンパンジーゲノム概要配列を用いた解析に並行して、今年度は国際塩基配列データベースに登録・更新され始めたチンパンジー BAC クローン 3261 個の配列を用いた解析も行った。2007 年 10 月の時点で利用可能であった約 570Mb の BAC クローン配列を、我々が開発した比較解析アルゴリズムを用いて、ヒトゲノム完成配列およびマカクゲノム概要配列とアラインした。その結果、進化速度の変動が少ないと考えられるヒト、チンパンジー、マカクゲノムにおいて重複の存在しない領域（約 77Mb）についてみると、ヒトゲノムーチンパンジー BAC クローン間の塩基相違度は 1.11424% であった。同手法、同領域におけるヒトゲノムーチンパンジーゲノム概要配列間の塩基相違度は 1.12781% であり、概要配列と比べて BAC クローン配列の精度が向上していることが示唆された。また、ヒト、チンパンジー、マカクゲノムにおいて重複の存在しない領域の中には、対応するチンパンジー概要配列が存在しない、チンパンジー BAC クローンデータ独自の領域も検出された。この BAC クローン領域の GC 含量（43.9%）は、対応する概要配列が存在する BAC クローン領域の GC 含量 (39.4％) と比べて高く、コード領域を多数含む高 GC 率領域が配列決定されずに残っているという推測を支持する結果であった。実際、BAC 内新規配列が 15 以上の遺伝子内に存在している。以上の結果から、BAC クローンデータを用いた解析の有用性が強く示された。今後も新規 BAC クローンデータが公開され次第随時追加解析を行い、ゲノムアラインメントの精度を高めていく。

以上の解析の過程では、新たな解析ツールの開発も行った。ゲノム再編成や重複などの複雑な進化過程を考慮したより高精度なゲノムアラインメントの作成や、ゲノムアラインメントに基づくヒト固有遺伝情報ならびにヒト特異的進化状態を示すゲノム領域の発見のためには、プログラムによる自動処理に加えて、人間の目によるデータの精査が重要である。我々が開発したツールは、このデータ精査の過程を支援するためのものであり、長大なゲノム間の対応領域を可視化し、それを自由に拡大縮小したり、対応領域の各ゲノム上における位置情報や塩基相違度等の必要な情報の取得を行うことが可能なものである。

＜国内外での成果の位置づけ＞

本研究課題は、系統立ててヒト固有形質の発現分子機構を解明するための細胞レベルでの基礎的情報の蓄積を行うものである。遺伝情報解析を起点としたヒトゲノム機能解析は、表現型を起点とした機能解析とは異なり、ヒトの未知機能（表現型）の発見をもたらす可能性を持ち、ヒト（ゲノム）生物学の長足の発展に寄与することが期待できる。

また、本研究において開発したツールは、ヒトーチンパンジーに限らず、他のあらゆる生物種のゲノム比較解析に有用なものである。

＜達成できなかったこと、予想外の困難、その理由＞

本研究で推定されたヒト固有形質対応候補領域について、基礎的分子生物学実験を行い、それらの表現型における特性を他の生物との間で比較する機能解析は達成することができなかった。その理由として、当初の予想以上にチンパンジー概要配列に多くの問題が存在したため、ヒト固有形質対応候補領域の推定および実験確認に時間を要したことがあげられる。

＜今後の課題＞

チンパンジーゲノム概要配列には未決定または曖昧な状態である部分が存在することが明らかとなったため、今後は、BAC クローンデータを用いた解析を行うことが重要である。ヒトーチンパンジーゲノム間で見出された相違箇所は、CDS 領域以外のイントロンや UTR 領域、遺伝子外領域などを含めると膨大であるが、そのような CDS 領域以外にも、ヒト固有の遺伝子発現量や発現パターンをもたらすものが存在する可能性があることから、ゲノムタイリングアレイを用いたヒト固有遺伝情報の網羅的な実験確認を行う。その際には、解析結果に基づいたゲノムタイリングアレイの設計を行う。また、ヒト固有の遺伝子発現量や発現パターンが予想される遺伝子については、リアルタイム PCR を用いて詳細な検証を行う。

＜成果公表リスト＞

投稿中のもののみ。