プレスリリース

平成２３年　６月　３日

報道関係者各位

国立大学法人 奈良先端科学技術大学院大学

世界初！次世代ＤＮＡ解析の精度向上につながる改善点を解明
特定の塩基配列で読み取りエラーがあった
～解析装置だけで素早く完全なＤＮＡデータ取得へ～

【概要】

次世代シーケンサ（ＤＮＡ塩基配列解読装置）の急速な進歩により、ゲノムＤＮＡ配列の決定量が膨大になった。例えば、現在の最高性能のゲノムアナライザでは、約百塩基長の配列が一度に数千万本得られ、その大量な情報の処理が必要とされている。このようなシーケンサは主に欧米の３社で製造され、得られた大量情報にもとづき決定されたゲノムの塩基配列データの信頼性をいかに評価するかということが、バイオインフォマティクス（生物情報科学）分野の緊急課題になっている。中でもシーケンサの機種によって特定の配列が誤まって読み取られてしまうという装置の癖に注目し、決定された配列の信頼性を評価し、精度を上げるための技術開発はこれまでになかった。

こうしたなかで、奈良先端科学技術大学院大学（学長：磯貝彰）情報科学研究科計算システムズ生物学研究室の中村建介特任准教授と金谷重彦教授らは、世界の研究機関でもっともよく使われている次世代シーケンサ（米国イルミナ社製）について、この装置が読み取ったデータは特定の配列パターンに隣接した領域を読み取る際に極めて高い確率で誤った塩基と判定してしまうことを明らかにした。

次世代シーケンサーの重要な応用例としては、ヒトの個体差を識別して薬の副作用などを評価するいわゆるSNPs解析が挙げられるが、これまで、イルミナシーケンサーによるデータだけではSNP変異を断定できないと言われていた。今回我々が特定した情報にもとづき、ＤＮＡ塩基の変位箇所の詳しい解析を行うことで、他の実験技術による確認をしなくとも、高い信頼性で変位の存在を特定出来るようになることが期待される。さらには、ゲノム配列の知られていない生物種の解析においても、我々の特定した情報に基づいて塩基決定アルゴリズム（処理手順）の改善を進めることで、シーケンサの高い性能を最大限に引き出して、ゲノム決定が行える可能性もある。以上のように、医療を含めた生物学分野における遺伝子情報の解析研究の促進に大きく貢献することが期待される。

つきましては、関係資料を配付するとともに、下記のとおり記者発表を行いますので、是非ともご出席くださいますよう、お願い申し上げます。

記

＜日時＞　平成２３年６月８日（水）１５時００分～（1時間程度）
＜場所＞　奈良先端科学技術大学院大学　事務局棟２階　大会議室
奈良県生駒市高山町８９１６－５（けいはんな学研都市）
※アクセスについては、http://www.naist.jp/をご覧ください。

＜説明者＞
奈良先端科学技術大学院大学 情報科学研究科 計算システムズ生物学研究室 特任准教授　中村　建介

＜ご連絡事項＞
(1)本件につきましては、奈良先端科学技術大学院大学から、奈良県文化教育記者クラブをメインとし、学研都市記者クラブ、大阪科学・大学記者クラブ、文部科学記者会及び科学記者会に同時にご連絡しております。
(2)取材希望がございましたら、恐れ入りますが下記までご連絡願います。
(3)記者発表に関する問合せ先

奈良先端科学技術大学院大学　企画総務課　広報渉外係　瀬戸　克昭（せと　かつあき）
TEL: 0743-72-5026　FAX: 0743-72-5011　E-mail: s-kikaku@ad.naist.jp

世界初！次世代ＤＮＡ解析の精度向上につながる改善点を解明
特定の塩基配列で読み取りエラーがあった
～解析装置だけで素早く完全なＤＮＡデータ取得へ～

【概要】

次世代シーケンサと呼ばれる技術の進歩により、ゲノムのＤＮＡや、ＤＮＡを転写したＲＮＡなどの塩基配列の情報を従来にくらべて飛躍的に低いコスト、短時間で得ることができるようになりつつある。なかでも米国イルミナ社のゲノムアナライザは情報量に対するコストパーフォーマンスが高く、現時点でもっとも広く普及している機種となっている。奈良先端科学技術大学院大学（学長：磯貝彰）情報科学研究科計算システムズ生物学研究室の中村建介特任准教授と金谷重彦教授らは、現在のイルミナ社の解析データには読み取ることの難しい配列パターンが存在することを世界に先駆けて見いだした。この情報に基づいて、イルミナシーケンサの高い能力を最大限に引き出すことで、SNPs 変異（１つの塩基だけ異なる変異）の特定や、未知ゲノムの再構築を、これまでより高い精度で行うことが出来るようになる、と期待される。

【解説】

ゲノムＤＮＡなどの塩基配列の解読技術が近年急速に進んでいる。ほんの十年ほど前にはヒトゲノムプロジェクトに代表されるように国家プロジェクト規模の予算と人材（数年の時間・数十人・数十億円）を必要とした情報が、次世代シーケンサと呼ばれる装置を用いることにより研究室規模（数日・数名・数十万円）で得られるようになっている。その結果、たとえば個人のＤＮＡ配列のどの部分が標準的なヒトの塩基配列と異なっているかを簡便に調べることが出来る。こうした情報と、疾患や薬に対する副作用の感受性などの情報を組み合わせれば、効果が高く副作用の低い「テーラーメード医療」とよばれるような個々人の体質に合わせたきめ細かい医療の実現に大きく貢献できる。

また、数々のゲノムプロジェクトにより既にいくつかのモデル生物種についてゲノム配列が決定されているが、これらのゲノム配列情報の活用を進めてゆくポストゲノム研究においては、モデル生物にある程度類似していて、形質・機能がわずかに異なる近縁生物種のゲノム配列を特定することが、進化の研究などに有益な情報を与えてゆくと考えられる。このような基礎生物学におけるゲノム情報の機能解析においても次世代シーケンサから得られるデータは重要な役割を果たす。

次世代シーケンサと呼ばれるテクノロジーの中でイルミナ社のシーケンサは現在もっとも普及している機種であるが、得られる配列情報についていくつかの問題点が存在する。たとえば、イルミナ社自身の見解として、イルミナシーケンサにより特定された未知のＳＮＰｓ（一塩基変異）については、他の実験手法を用いて確認することが必要であるとアナウンスされている。結果、配列解析の結果として新しい変異を見いだしたとしても、他の実験手法による検証が必要とされるため価値が半減してしまう。

私たちは、このような検証が必要とされる理由がどこにあるのかを疑問に思いつつシーケンシングデータの解析を進めてゆく上で、イルミナシーケンサの配列データに共通する興味深い特徴が存在することを見いだした。

イルミナ社のシーケンサによる配列データの精度は実験の条件等にもよるが１％程度と言われている。既知のバクテリアゲノムに対してシーケンサにより得られたデータをマップした結果の一部分を図１に示す。それぞれの図の上部の数値は既知ゲノム配列上の塩基位置を表しており、その下に積み重なった短い線分がシーケンサから得られたリードと呼ばれる一つ一つの配列データ（各７５塩基長、ＤＮＡの断片）を適合するゲノム位置にマップしたものである。配列データの色はシーケンシング時の読み取り方向とゲノム方向が一致している場合には薄い灰色、逆向きの相補鎖として一致する場合には薄い青色で示されている。

また、それぞれのリードは標準データのレファレンスゲノムに対して最大 35 箇所のミスマッチまで許容して、レファレンス上で最もミスマッチがすくなくなる位置にマップされている。

従来のマッピングプログラムを用いた場合には、許されるミスマッチの数はリードあたり 2 個程度でギャップアラインメント(ギャップ調整)と呼ばれる短い塩基の挿入と欠損を考慮した手続きがとられるが、今回、私たちは多くのミスマッチを許しながら、ギャップアラインメントを行わないマッピングを行うこ

とで、図１に見られるように、レファレンスとのミスマッチが集中して発生する領域が存在することを見いだした。

さらにこの図で興味深いのは、赤い点で示されたミスマッチが右方向へ読み取っている灰色のリードに集中していて、逆方向の薄い青色のリードにはあまり見られないこと。またミスマッチが 1960 塩基位置付近を境界として右側、すなわち読み取り方向に特定の開始点が存在しているように見えることである。このことから配列読み取りにエラーを引きおこす要因が 1950 塩基位置付近に存在することが示唆される。私たちはこの配列特異的な読み取りエラーを ＳＳＥ（Sequence Specific Error）と名付け、その発生のメカニズムの推定とこれにより引きおこされるシーケンシング上の問題点を整理して今回の論文で報告した。

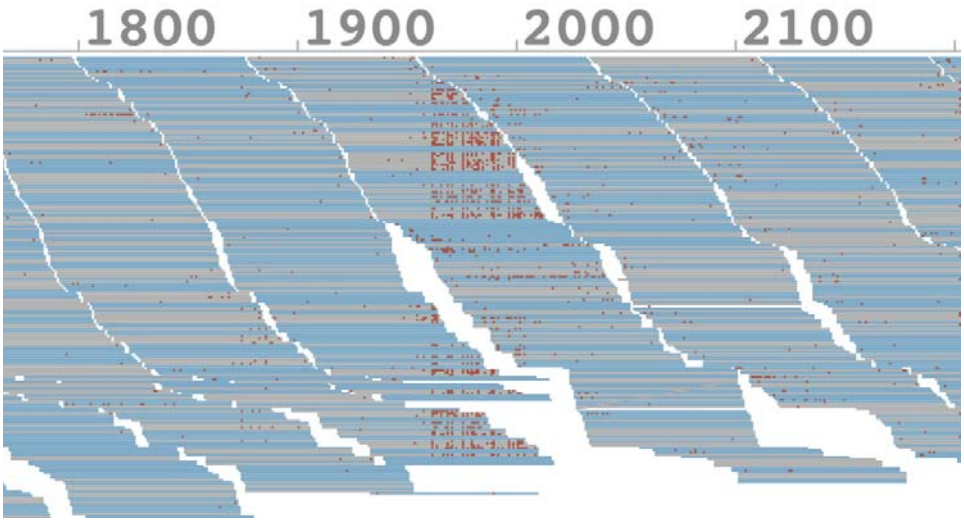

図１．配列特異的エラー（SSE）の発生位置、上部の数値はゲノム位置。赤い点はゲノム塩基とマップされたリードの塩基が異なる箇所を示す。１９６０塩基付近を境界として、右側（読み取り方向）に特定の開始点が存在しているとみられ､エラーの要因が１９５０塩基位置付近に存在すると示唆された。

図２．左上:サンプルとレファレンスの間に一塩基置換（ＳＮＰ）が起きている例　左下:SSE により SNP のようにミスマッチが起きている例　右　：複数の SSE により読み取りが困難になっている領域

今回特定されたエラーパターンによる主な問題として（１）一塩基置換（SNP ｓ）の誤認、（２）RNA-seq という方法による発現量を調べる解析などで、リード数から発現量を見積もる場合の定量性、（３）ゲノム配列の再構築（アセンブル）において読み取り困難な領域が存在することによる不連続領域―が挙げられる。

（１）については図２に示したように、真の一塩基置換が左上の図に示したように現れるのに対して、SSE による読み取りエラーが左下図の様に約半数のリード配列で同じ塩基部位に現れることがある。このようなエラーはたとえば真核生物の読み取りにおいてはハプロタイプ（片親由来のＤＮＡ塩基配列）の一方に一塩基置換が存在すると誤認される可能性が高い。図２右側では SSE によるエラーが狭い箇所に集中して起きているため、ミスマッチを多く許したマッピングでは上のように多くのミスマッチが現れ、従来のプログラムで通常のマッピング条件によりミスマッチを２つしか許さない場合にはリードがほとんど張り付かない状態（図２右下）を示している。

（２）では、細胞中に発現しているＲＮＡ配列の相対量をＲＮＡを鋳型として作成した DNA 配列のフラグメントの存在量から推定する手法において、図２の右側のように SSE に起因するエラーが多く含まれる配列領域を含むトランスクリプト（転写物）が存在する場合には、その発現量を過小評価してしまう可能性がある。また、このような領域の存在は（３）のゲノム配列の再構築をしようとする場合、イルミナシーケンサーデータからは配列を再現することが事実上不可能な不連続領域となってしまう。

【本研究の意義】

私たちの特定したイルミナシーケンサのエラープロファイルは、図２で示したように配列解析においてさまざまな問題の要因となっている。このエラープロファイルについて、基本的なメカニズムの推定ができていることから、今後、原因配列の特定を進めることで読み取りエラーが起きやすい領域を予測し、さらにはエラー発生のモデルを考慮して塩基決定（ベースコール）のアルゴリズムを設計することで、高性能なイルミナ社のシーケンサのデータを最大限に引き出すことが出来るようになる、と期待される。

【補足説明】

1. 次世代シーケンサー：従来の Sanger 法と呼ばれる塩基配列解析技術に対して、短い塩基配列を大量に読み出すことの出来る配列解析プラットフォームの総称。代表的なものとして(1) Illumina Genome Analyzer, (2) Life Technologies/ABI SOLiD System, (3) Roche/454 Genome Sequencer FLX の３つが広く使われている。
2. ＳＮＰｓ：一塩基多型、ある生物集団に一定の割合で存在する一塩基置換
3. マッピング：シーケンサから得られた短い塩基配列データ（リード）を既知のレファレンス塩基配列上で最も矛盾無く対応すると思われる場所にあてはめてゆく操作。アラインメントとも呼ばれる
4. アセンブル：シーケンサから得られた短い塩基配列データを、重複配列をのりしろとして重ね合わせながら伸張し、元の塩基配列を再構築しようとする操作。

【本研究内容についてコメント出来る方】

小笠原　直毅（おがさわら　なおたけ）先生
奈良先端科学技術大学院大学　バイオサイエンス研究科　教授
TEL:0743-72-5433　E-mail: nogasawa@bs.naist.jp

【本プレスリリースに関するお問い合わせ先】

奈良先端科学技術大学院大学　情報科学研究科　計算システムズ生物学研究室
特任准教授　中村 建介（なかむら　けんすけ）
TEL 0743-72-5396　　FAX 0743-72-5258
E-mail　kensuke-nm@is.naist.jp