様式Ｆ－１９

科学研究費助成事業（学術研究助成基金助成金）研究成果報告書

平成２５年５月３１日現在

機関番号：８２４０１
研究種目：若手研究(B)
研究期間：2011～2012
課題番号：２３７９０３８９
研究課題名（和文）　全ゲノムシークエンスデータ解析による新規配列の網羅的同定と解析

研究課題名（英文）　Analysis of non-reference genome sequence in the human population

研究代表者
藤本　明洋（FUJIMOTO AKIHIRO）
独立行政法人理化学研究所・情報解析研究チーム・上級研究員
研究者番号：３０５２５８５３

研究成果の概要（和文）：
我々は、先行研究において日本人１個体のゲノムシークエンスより 3.4Mbp の非標準ゲノム配列を得た。非標準ゲノム配列の特性と多型性をさらに調査するため、複数個体の日本人の全ゲノムシークエンスより、標準ゲノム配列にマッピングされなかったリード配列を集めて *de novo* アセンブリを行った。その結果、約 4Mbp 程度の非標準ゲノム配列を得た。また、遺伝的多様性を検出する手法の開発を行った。

研究成果の概要（英文）：
We previously performed whole genome sequence of a Japanese individual and identified non-reference genome sequences from unmapped reads. To characterize non-reference genome sequences, we gathered unmapped read sequences from additional multiple Japanese individuals, and conducted *de novo* assembly. Approximately 4Mbp none-reference genome sequences were obtained.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
交付決定額	3,200,000	960,000	4,160,000

研究分野：医歯薬学
科研費の分科・細目：基礎医学・人類遺伝学
キーワード：全ゲノムシークエンス　非標準ゲノム配列 次世代シークエンサー

１．研究開始当初の背景

近年の著しいシークエンス技術の発展により、個人ゲノムシークエンスが可能となった。超並列シークエンサーの発展は著しく、現在では 600Gbp（ヒトゲノムの約 200 倍）の塩基配列データが約 2 週間で得られている。全ゲノムシークエンスは世界中で活発に行われており、全ゲノムシークエンスは疾患研究において、極めて重要な役割を担っていくと考えられる。

超並列シークエンサーは、シークエンスの対象となる DNA を細かく断片化し、同時に数千万ー数億配列を解読する。一般的に、次世代シークエンサーで決定できる長さは、100bp 程度であるため、そのままでは研究に用いることができない。シークエンスを行った後で、得られた塩基配列（リード配列）を標準ゲノム配列に計算機でマッピングしゲノム上の位置決めを行うか、*de novo* アセンブリを行うことでシークエンスされたリード配列から長い配列の再構成を行う。

我々は全ゲノムシークエンスデータ解析手法の確立を目的として、世界で初めて日本人の全ゲノムシークエンスを決定し、包括的な遺伝的多型解析を行った（Fujimoto et al ***Nature Genetics*** (2010)）。

国際ハップマッププロジェクトで解析さ

れた日本人男性1人の全ゲノムを、イルミナ社 Genome Analyzer II を使用してシークエンスを行い、全部で約 1200 億塩基対（ヒトゲノムの 40 倍）のデータを得た。得られたリード配列を Short-read mapping プログラムの BWA と blast を用いて標準ゲノム配列にマッピングした結果、99.1%のリード配列がマッピングされた。さらに、マッピングされたリード配列をもとに、高い精度で一塩基置換（SNV）やコピー数変化（CNV）を検出するアルゴリズムを開発し、実験的検証を行った。

我々は引き続き、マッピングされなかったリード配列の *de novo* アセンブルを行い、標準ゲノム配列に存在しない新規配列の検出を試みた。3 種類のアセンブラ（ABySS、SOAPdenovo、Velvet）を用いた結果、それぞれ 6,535 個、4,826 個、6,617 個の 100bp 以上のコンティグが得られ、長さの合計は 3-3.4Mbp であった。3 つのソフトウエアが出す結果は互い共通性が高く、特に SOAPdenovo から得られた配列は全体の 90.7%をカバーしていた。PCR 法で 186 コンティグの検証を行ったところ、181 コンティグが期待通りの増幅長を示した。また、サンガーシークエンス解析で 90%以上のコンティグが予測配列と高い相同性（>90%）を示すことが分かった。これらの結果より、我々のアプローチは高い精度で新規配列を検出していると考えられる。

これらのコンティグを NCBI のゲノムデータベースに対して相同性検索を行ったところ、約 90%が標準ゲノム配列以外のヒトゲノム配列にマッピングされた他、霊長類に相同性の高い配列や、Human Herpesvirus 4 (細胞株作成に用いられたウイルス)の配列も検出された。

我々が行った日本人ゲノムの解析は、一人当たり約 3Mbp 程度の新規配列を持つことを示唆している。それらの新規配列は、ヒトゲノム計画に用いられたサンプルには存在しなかった（欠失していた）可能性が高く、新規配列の有無は極めて個人差や人種差が大きいと考えられた。

２．研究の目的

全ゲノムシークエンスデータの解析から新規配列の同定を行い、新規配列の有無の個人差、新規配列上の配列の違い（SNP、挿入・欠失）の同定、新規配列上の未知遺伝子探索を行うことで、これまで全く明らかにされてこなかった人類集団の多様性を解明することを目的として本研究を行った。

３．研究の方法

複数個体の全ゲノムシークエンスより、標準ゲノム配列にマッピングされないリード配列を選出して *de novo* アセンブリを行った。また、一塩基多型、コピー数多型、構造多型の検出法を構築した。

４．研究成果

非標準ゲノム配列の検出と *de novo* アセンブリ

複数個体の日本人の全ゲノムシークエンスデータより、標準ゲノム配列にマッピングされないリード配列を検出した。先行研究では、short read 配列マッピング用のプログラム（BWA）を用いてマッピングした後、マッピングされなかった配列をさらに blast で標準ゲノム配列にマッピングした。しかし、blast は感度は高いものの、計算時間が長くメモリ使用量も多いため、複数サンプルの解析には不向きである。そこで、blast を BWA-SW プログラムに置き換えることにより解析を高速化した。

De novo アセンブリを、個体ごとに行う方法と、全個体のマッピングされない配列を合わせて行う方法で行い比較した。今回の解析の主たる目的は、日本人集団全体における新規配列の検出であるため、後者の方法で *de novo* アセンブリを行うこととした。

De novo アセンブリはパラメーターによって結果が大きく異なる。結果の評価を行うために、BWA を用いて *de novo* アセンブリにより構築された配列にリード配列の再マッピングを行いマッピングされた配列の総数、リードペア間の向きが異常でないペア数、N50、コンティグの全体の長さを総合的に用いて評価を行い、適切と思われる kmer サイズと、オプションを決定した。この解析を複数の日本人個体に対して行い、約 4Mbp の新規配列が検出した。

新規配列の特徴を調べるため、新規配列上のリピートを tandem repeat finder ソフトウエアで検出した。その結果、約 13%がリピート配列であることが示唆された。さらに、blast を用いて、標準ゲノム意外のヒトゲノム配列に対して、マッピングを行ったところ、（マッピングできたとする基準にも依存するが）、約 70%が Human alternative assembly、GRCh37.p9 に約 15%がマッピングされ、他のヒトゲノムにマッピングされない配列は約 20％程度であった。非標準ゲノム配列にマッピングされる配列も挿入や欠失を含んでいることが分かり、新規配列の有無に加えて、新規配列内の複数個体間差異の存在も示唆された。

構造異常の検出

ゲノムの構造異常（逆位、転座、挿入、欠失）は構造を大きく変え、周辺の遺伝子の発現量や構造に大きな影響を及ぼしうる。第２世代シークエンサーでは、読み取り長が短いため、マッピングエラー等により、偽陽性を生じる

可能性が高い。申請者らは、がんサンプルの全ゲノムシークエンスデータを用いて構造異常検出アルゴリズムの構築と改良を行った。

先行研究における構造異常の偽陽性の原因を調査したところ、ほとんどの偽陽性がマッピングエラーであることが判明した。また、偽陽性を生じたリード配列はblast等のより感度が高いアラインメントプログラムを用いて、マッピングを行うことで除きうることが分かった。これらの結果を基に、構造異常を偽陽性率 10%以下で検出する解析プログラムを開発した。

また、このプログラムを日本人の全ゲノムシークエンスデータに対して適用し、構造多型の検出を行い、日本人集団で頻度の高い構造多型を検出した。

一塩基多型の検出

一塩基多型は最も数が多く、疾患遺伝子検出のため一般的に用いられているマーカーであり、精度の良い検出法の開発は重要な課題である。我々は、先行研究において、複数の一塩基多型検出法（カウント法、頻度法、尤度比較法）の比較を行い、頻度法と尤度比較法の精度が高いことを見いだした。

そこで、尤度比較法をさらに改良し、感度と特異度を既存の SNP 検出ソフトウエア（samtools, GATK）と比較した。SNP genotyping アレイの結果を正解データとしたところ、感度は 99.98%、特異度は 99.99%であった。また、既存のソフトウエアと比べ、感度と特異度が優れていた。ソースコードを C 言語に置き換えることで高速化、省メモリ化を実現した。このプログラムは解析ツールとして公開している（投稿中）。

挿入・欠失の検出

挿入欠失の検出を行うため、頻度法を用いて挿入欠失の検出を行った。一塩基多型と異なり、挿入欠失は十分な量の正解データが存在しない。そのため、検出パラメーターを極めて緩く設定して挿入欠失を検出し、サンガー法で確認実験を行った。その結果より偽陽性を除き、真の挿入欠失を除かないようにパラメーターを決定した。その結果、偽陽性率 5%以内で挿入欠失を検出することができた。この手法を用いた結果、日本人 1 個体より約 50 万の挿入欠失（遺伝子領域内に約 300 個）を検出した。

コピー数多型の検出

コピー数多型は、シークエンス深度を解析して検出することができる。しかし、シークエンス深度は、GC 含量や配列のユニークさの影響を受け、絶対コピー数の検出は困難である。我々は、複数個体のシークエンス深度の比較と、全体の平均深度からの推定絶対コピー数を組み合わせることで絶対コピー数の正確な定量手法を開発した。この手法については、先行研究と比較を行い、精度の確認およびアルゴリズムの改良を行っている。

B 肝炎ウイルス関連肝がんへの適用

B 肝炎ウイルス（HBV）は、DNA ウイルスであり、炎症や宿主ゲノムへの挿入を介して発がんの原因となることが知られている。HBV ゲノムは肝臓がんのゲノムシークエンスより検出されることが期待されるため、HBV 関連肝がんの全ゲノムシークエンスデータを用いて、新規配列検出法の精度確認を行った。HBV 関連肝がんのゲノム配列より、標準ゲノム配列にマッピングされない配列を収集し、上記手法で解析を行った。その結果、HBV ゲノムに高い相同性を示す配列が確認された。

新規配列の多様性

以上の多型検出手法と検出された新規配列を総合的に評価することで、ゲノム上の新規配列の位置の特定、新規配列のコピー数、新規配列上の一塩基多型および挿入欠失の検出を行った。また、この研究を通して、日本人集団における非標準ゲノム配列の遺伝的多様性の一端を明らかにした。

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計　3 件）

①Fujimoto A, Totoki Y, Abe T, Boroevich KA, Hosoda F, Nguyen HH, Aoki M, Hosono N, Kubo M, Miya F, Arai Y, Takahashi H, Shirakihara T, Nagasaki M, Shibuya T, Nakano K, Watanabe-Makino K, Tanaka H, Nakamura H, Kusuda J, Ojima H, Shimada K, Okusaka T, Ueno M, Shigekawa Y, Kawakami Y, Arihiro K, Ohdan H, Gotoh K, Ishikawa O, Ariizumi S, Yamamoto M, Yamada T, Chayama K, Kosuge T, Yamaue H, Kamatani N, Miyano S, Nakagama H, Nakamura Y, Tsunoda T, Shibata T, and Nakagawa H. Whole genome sequencing of liver cancers identifies etiological influences on mutation patterns and recurrent mutations in chromatin regulators. ***Nat Genet*** 44: 760-764 (2012)（査読有）

②藤本明洋, 中川英刀, 角田達彦(2012) 次世代シークエンサーを用いた日本人ゲノム解読による遺伝的多様性の包括的解析. 最新医学 2012年 67巻1号 p.132 -136（査読無）

③藤本明洋, 中川英刀, 角田達彦(2011) 次世代シークエンサーを用いた日本人ゲノム解読

による遺伝的多様性の包括的解析. 生体の科学 Vol. 62 No.6（査読無）

〔学会発表〕（計８件）

①九州大学 ゲノミクスエピゲノミクス研究拠点セミナー
2013年1月21日　福岡
A Fujimoto, Y Totoki, S Miyano, T Tsunoda, T Shibata, and H Nakagawa
Comprehensive analysis of genetic variation by whole genome sequencing

②Radiation Research in the Post-genomic Era
2013年1月21日　広島
A Fujimoto
Whole genome sequence of Japanese individual and liver cancer

③ASHG 62nd Annual Meeting
2012年11月6日～10日　San Francisco, CA, USA
A Fujimoto, Y Totoki, T Abe, K A Boroevich, F Hosoda, H Hi Nguyen, M Aoki, N Hosono, M Kubo, F Miya, Y Arai, H Takahashi, T Shirakihara, M Nagasaki, T Shibuya, K Nakano, K Watanabe-Makino, H Tanaka, H Nakamura, J Kusuda, H Ojima, K Shimada, T Okusaka, M Ueno, Y Shigekawa, Y Kawakami, K Arihiro, H Ohdan, K Gotoh, O Ishikawa, S Ariizumi, M Yamamoto, T Yamada, K Chayama, T Kosuge, H Yamaue, N Kamatani, S Miyano, H Nakagama, Y Nakamura, T Tsunoda, T Shibata & H Nakagawa
Whole-genome sequencing of liver cancers identifies etiological influences on mutation patterns and recurrent mutations in chromatin regulators

④日本遺伝学会第84回大会
2012年9月24日～26日　福岡
藤本明洋, 十時泰, 細田文恵, Ha Hai Nguyen, 青木正幸, 久保充明, 宮野悟, 中釜 斉, 中村祐輔, 角田達彦, 柴田 龍弘, 中川英刀
肝臓がん27例の全ゲノム解析

⑤第71 回日本癌学会学術総会
2012年9月19日～21日 札幌
A Fujimoto, Y Totoki, S Miyano, T Tsunoda, T Shibata, and H Nakagawa
Whole genome sequence and comprehensive analysis of Hepatocellular carcinoma genome

⑥第11回東アジア人類遺伝学会
2011年11月9日～12日　千葉
A Fujimoto, H Nakagawa, N Hosono, K Nakano, T Abe, K Boroevich, M Nagasaki, R Yamaguchi, T Shibuya, M Kubo, S Miyano, N Kamatani, Y Nakamura & T Tsunoda[1,2]
Whole genome sequence of a Japanese individual with massively parallel sequencing technology

⑦人類学会第65回大会
2011年11月4日～6日　沖縄
藤本明洋, 中川英刀, 細野直哉, 中野かおる, 阿部哲雄, 長崎正朗, 山口類, 久保充明, 宮野悟, 中村祐輔, 角田達彦
次世代シークエンサーを用いた日本人一個体の全ゲノムシークエンスと遺伝的多様性の包括的解析

⑧進化学会第13回大会
2011年7月29日～31日　京都
藤本明洋, 中川英刀, 細野直哉, 中野かおる, 阿部哲雄, 長崎正朗, 山口類, 久保充明, 宮野悟, 中村祐輔, 角田達彦
次世代シークエンサーを用いた日本人一個体の全ゲノムシークエンスと遺伝的多様性の包括的解析

〔図書〕（計　0件）

〔産業財産権〕
○出願状況（計　0 件）

○取得状況（計　0 件）

〔その他〕
ホームページ等

６．研究組織
(1)研究代表者
藤本　明洋（FUJIMOTO AKIHIRO）
独立行政法人理化学研究所・情報解析研究チーム・上級研究員
研究者番号：３０５２５８５３

(2)研究分担者
　なし

(3)連携研究者
　なし