様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　４ 年　 ５ 月 １３ 日現在

機関番号：１１３０１
研究種目：基盤研究(C)（一般）
研究期間：2019～2021
課題番号：１９Ｋ０６６２５

研究課題名（和文）深層学習による大規模ゲノムコホートの次世代シークエンス解析

研究課題名（英文）NGS analysis of a large genome cohort by deep learning

研究代表者
高山　順（Takayama, Jun）

東北大学・未来型医療創成センター・准教授

研究者番号：２０５７４１１４

交付決定額（研究期間全体）：（直接経費）　　3,300,000円

研究成果の概要（和文）：本研究は大規模ゲノムコホートや疾患ゲノム解析の主要な解析技術の一つであるNGSについて、その技術的限界を見極め、深層学習をはじめとした最新の技術でその限界を克服することを企図したものである。本研究のもと、NGSを用いたゲノム解析手法であるリシークエンシング法に潜むバイアスのうち民族集団の差によるバイアスに着目し、ヒトゲノム計画で構築された国際基準ゲノムに代わるものとして構築された日本人集団固有の基準ゲノム配列の性能評価を行った。結果は他の結果とともに論文化され、Nature Communications誌に発表された。また複数の招待講演を行なった。

研究成果の学術的意義や社会的意義
次世代シークエンシング法を用いるとヒトの全ゲノム情報を数日の内に解読可能であるが、完璧な方法ではない。本研究では次世代シークエンシング法の問題点を追求し、特に、民族集団の違いによるバイアスを克服することを試みた。より具体的には日本人のゲノム解析に最適化した日本人基準ゲノム配列JG1の性能評価を行った。日本人基準ゲノム配列は公開され、ゲノム解析サービスに利用されており、本研究はその有用性を示すものとして大きな意義を有すると考えられる。

研究成果の概要（英文）：This study is intended to identify the technical limitations of NGS, one of the major genome analysis techniques for large-scale genome cohort and disease analysis, and to overcome these limitations using state-of-the-art technologies such as deep learning. Under this study, we focused on the bias due to population differences among the biases latent in the resequencing method, a genome analysis method using NGS, and evaluated the performance of the reference genome sequence JG1, which is optimized for the Japanese population to replace the international reference genome constructed in the Human Genome Project. The results, together with other results, were published in Nature Communications. I also gave several invited lectures.

研究分野：ゲノム科学

キーワード：ヒトゲノム解析　次世代シークエンシング

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

近年、次世代シークエンシング(NGS)技術が発展したことで、ゲノム DNA 配列の解読は高速化したが、遺伝子型から表現型を予測することは困難なままである。例えば希少難病の原因特定に NGS 解析を用いても特定に至るのは 50%を下回るし、コモンディジーズの原因として GWAS 解析で同定された数千の遺伝的多型の効果を合算しても遺伝要因の大部分が説明されない。さらに同定したバリアントから表現型を予測する困難は一層深刻である。

NGS を用いたゲノム解析では一般に、検体のゲノム DNA から得られた大量のリード配列を基準ゲノム配列にマッピングしバリアントを検出して遺伝子型を推定する「リシークエンシング法」と呼ばれる方法が用いられる。しかしこのリシークエンシング法にはいくつかの問題が存在する。例えば基準ゲノム配列として用いられる国際ヒトゲノム計画由来の配列はレアバリアントを採用していたり、アフリカ系民族集団固有の配列が多数含まれていたりすることが知られている。これらの欠点を克服するための提案はいくつかあるもののいずれも効果が限定的であった。

２．研究の目的

そこで本研究は、NGS データが有する重要な情報を捨てることなく真に各検体の遺伝情報を抽出し、上述の問題を克服することを企図した。当初深層学習をはじめとした技術を用いることを念頭において克服することを想定していたが、その後の準備段階の予備調査において、まずは現状のリシークエンシング法に潜在するバイアス、特に検体と基準ゲノム配列の民族集団差に起因するバイアスの低減こそがまずとりくむべき課題と考えた。そのために現状のリシークエンシング法のバイアスを詳しく検討することを最初の目的と定めた。

３．研究の方法

申請者は同時期に、日本人基準ゲノム配列 JG1 を構築する研究も行っていた。これは日本人検体３名から長鎖リードデータやオプティカルマッピングデータ等を取得してゲノムアセンブリを行い、３アセンブリを統合し、多数派のバリアントを採用するようにした基準ゲノム配列である。単にアセンブリをしただけではなく、遺伝地図や放射線ハイブリッド地図を用いて染色体の単位にまとめ利便性を高めたことも特徴である。本研究を行う上で、この JG1 を用いることが有効であると考え、下記の解析を行った。解析結果は文献[1]として発表された。

（１）日本人基準ゲノム配列 JG1 の性状解析

（２）既存エクソーム解析データを用いた日本人基準ゲノム配列の性能評価

４．研究成果

（１）JG1 の性状解析 1:アセンブリ性能の統計量

日本人基準ゲノム配列 JG1 の FASTA ファイルを分析し、JG1 は 22 本の常染色体と 2 本の性染色体、1 つのミトコンドリアゲノムと 599 のその他配列からなり、全長はおよそ 3.1Gb で 473 の全長 251Mb のギャップ領域を含むことを明らかにした（図 1）。

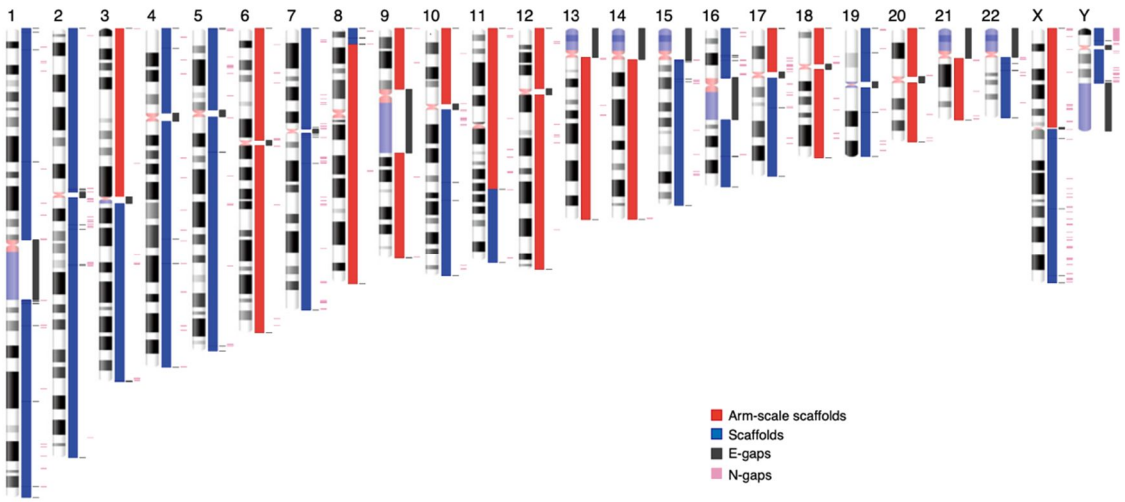

図１：日本人基準ゲノム配列 JG1 の全体像。イデオグラムの横にアセンブリ結果であるスキャフォールド（赤: 染色体腕全体をカバーするもの; 青: それ以外）とその間の未決定領域のギャップ（黒: E-gap, ピンク: N-gap）を図示した。

（２）JG1 の性状解析 2:民族集団の代表性

JG1が日本人集団を代表するものであるかを検証するため、JG1 に採用されたバリアントを用いて他民族集団とともに主成分分析(PCA)を行なった。その結果、確かに日本人集団のクラスタ内に存在し（図２）、また、どの日本人検体よりもヨーロッパ系・アフリカ系から遠いバリアント構成を有することが判明した[1]。また日本人集団のバリエーションパネルデータを活用することで、JG1 に採用されたバリアントが日本人集団の多数派をどの程度採用しているか検証した。その結果、日本人集団の 90%以上が有するバリアントサイトの 96%において確かに多数派バリアントを採用していることがわかった[1]。

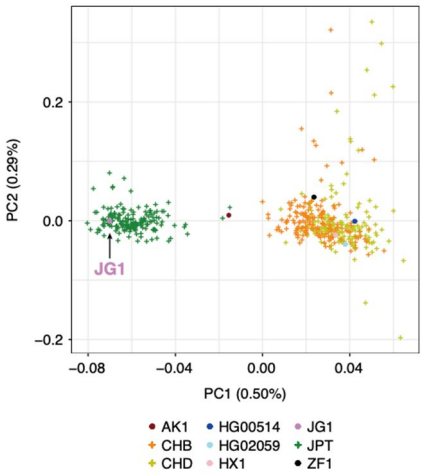

図２: PCA プロット。緑色の日本人クラスターに JG1 が存在する。

（３）JG1 の性能評価

JG1のクリニカルシークエンシングにおける基準ゲノム配列としての性能を評価するため、原因バリアントが既知の希少難病の日本人家系 7 家系のデータを用いてエクソーム解析を行なった。国際基準ゲノム配列と JG1 にそれぞれリードをマッピングし、GATK ベストプラクティス法に従ってバリアントコールを行なった。その結果、既知の原因バリアントは全て検出した一方で、原因以外のバリアントは JG1 で大幅に減っていることが判明した。特に、全検出バリアント数、疾患原因の候補となる機能へのインパクトの強いバリアント数も大幅に減っていることが判明した（図３）。さらに他の家系データも用いて、NGS で検出したバリアントが確かに存在することをサンガーシークエンシングでも確認できた。これらの結果は、民族固有の基準ゲノム配列を用いることが、現状の NGS 解析の高効率化に繋がることを示している。

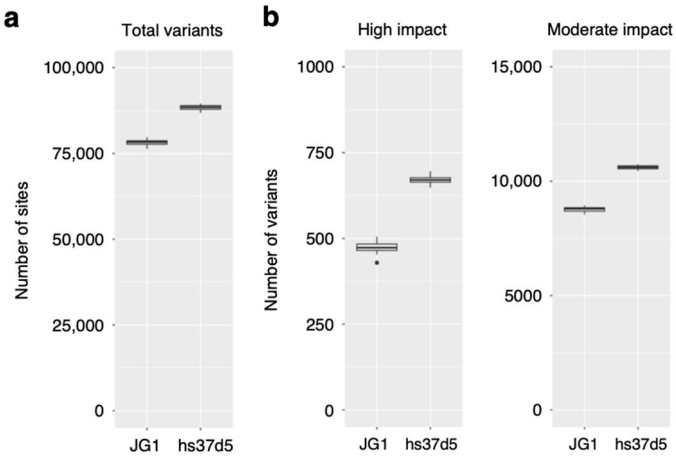

図３: JG1 の性能評価。日本人基準ゲノム配列は日本人検体のゲノム解析において、疾患原因以外の余計なバリアントコール数が減ることがわかる。hs37d5 は NGS 解析に最適化した国際基準ゲノム配列 GRCh37 である。

（４）研究成果の波及効果

本研究は、別の研究で構築された JG1 を用いて、現状の NGS 解析のバイアスの一つである、国際基準ゲノム配列と検体の民族集団バックグラウンドの差について詳しく解析し、集団固有の基準ゲノム配列が有用であることを示したものである。特にクリニカルシークエンシングのデータでその有用性を示したことは社会的にも重要な意義を有する。実際に JG1 は広く世界中に公開されており、これを利用した解析サービスも存在する。このことは本研究が基礎科学だけでなく実社会の応用にとっても有用であることを意味する。

参考文献

1. Takayama, J., Tadaka, S., Yano, K. et al. Construction and integration of three *de novo* Japanese human genome assemblies toward a population-specific reference. *Nat Commun* 12, 226 (2021). https://doi.org/10.1038/s41467-020-20146-8

５．主な発表論文等

〔雑誌論文〕　計1件（うち査読付論文　1件／うち国際共著　0件／うちオープンアクセス　1件）

１．著者名 Takayama Jun、Tadaka Shu、Yano Kenji、Katsuoka Fumiki、Gocho Chinatsu、Funayama Takamitsu、Makino Satoshi、Okamura Yasunobu、Kikuchi Atsuo、Sugimoto Sachiyo、Kawashima Junko、Otsuki Akihito、Sakurai-Yageta Mika、Yasuda Jun、Kure Shigeo、Kinoshita Kengo、Yamamoto Masayuki、Tamiya Gen	４．巻 12
２．論文標題 Construction and integration of three de novo Japanese human genome assemblies toward a population-specific reference	５．発行年 2021年
３．雑誌名 Nature Communications	６．最初と最後の頁 226
掲載論文のDOI（デジタルオブジェクト識別子） 10.1038/s41467-020-20146-8	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

〔学会発表〕　計4件（うち招待講演　2件／うち国際学会　2件）

１．発表者名 J. Takayama, S. Tadaka, K. Yano, F. Katsuoka, C. Gocho, T. Funayama, S. Makino, Y. Okamura, A. Kikuchi, J. Kawashima, A. Otsuki, J. Yasuda, S. Kure, K. Kinoshita, M. Yamamoto, G. Tamiya
２．発表標題 A population-specific reference genome built by integrating three de novo Japanese genome assemblies
３．学会等名 European Human Genetics Virtual Conference 2020（国際学会）
４．発表年 2020年

１．発表者名 高山順
２．発表標題 日本人基準ゲノムJG1の構築と小児希少疾患の全エクソーム解析への応用
３．学会等名 第27回日本遺伝子診療学会（招待講演）
４．発表年 2020年

１．発表者名 Jun Takayama
２．発表標題 Toward a Population-Specific Reference Genome:The Japanese Reference Genomes, JG1 and JG2
３．学会等名 T2T/HPRC Consortium 2020（国際学会）
４．発表年 2020年

１．発表者名 高山　順
２．発表標題 日本人基準ゲノムJG1の構築
３．学会等名 日本人類遺伝学会第64回大会（招待講演）
４．発表年 2019年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関