（別紙１）

# 総 括 研 究 報 告 書

課題番号：28－12
課 題 名：新規ヒトゲノム参照配列 GRCh38 および日本人基準配列を活用したゲノム変異診断

主任研究者名（所属施設） 国立成育医療研究センター
（所属・職名）システム発生・再生医学研究部 組織工学研究室 室長

（研究成果の要約） 小児疾患の多くはゲノムやエピゲノム変異が関わり、それらの診断に次世代シークエンサ(NGS)は今や不可欠の解析機器となっている。得られる配列長は従来のシークエンサと比べると短く、参照配列と呼ばれる既知のゲノム塩基配列と比較することで結果を得るため、参照配列の質が結果を大きく左右する。本センターでは 2009 年に公開された GRCh37 に基づくデータを一貫して使用しており、これまで特に問題は起きていないものの、新規参照配列 GRCh38 が発表されて 5 年以上が経過し、有用性が示され、かつ他機関との整合性を考慮すると移行する必要がある。このアップデートは NGS 登場後初めて、かつ複雑な作業である。センター全体にとって重要な事案ながら、どの研究チームも手をつけられない状態であった。本研究ではまずエクソーム解析を取り上げ、GRCh38 に対するマッピング、変異検出、アノテーションを行うパイプラインを再開発し、これまでと同等の変異検出ができることを確認した。特にミトコンドリアに多くの新規変異を検出し、実用性を実証することができた。この成果は全ゲノム解析をはじめとする他の NGS 解析にも活かすことができ、希望する参照配列を選択してデータ処理を行う環境を整えた。また、これらゲノム配列データのディープラーニングにより、自動的に NGS データベースを構築する方法を編み出した。

## １．研究目的

成人とは異なり小児に見られる異常のほとんどはゲノムやエピゲノム変異が関わる疾患で、次世代シークエンサ(NGS)が欠かせない研究や診断のための機器となっている。NGS が登場した当時、利用可能なヒトゲノム参照配列は NCBI Build 36 に基づく hg18 であったが、2009 年になって GRCh37 または hg19 と呼ばれる国際参照配列が発表され、次世代シークエンサやマイクロアレイなどの大規模なデータ解析に広く用いられてきた。現在でも特に断りがなければ、変異等の染色体上の位置は GRCh37 に基づいた数値で表され、国内外の研究者、医師、カウンセラー間で問題なく連絡を取ることができる状況になっている。しかしながらこれら参照配列は、今日見られる NGS の発展を視野に入れて用意されたものではなく、その後、マッピングと変異検出の正確さを上げるためにデコイ配列や alternate 配列が考案され、2013 年末に新規 GRCh38 が発表されるに至った。前バージョン GRCh37 の普及度合が高く、新規参照配列への切り換えはほとんど進んでこなかったが、発表から 5 年以上が経ってその有用性が認められつつある。さらに 1000 人ゲノムプロジェクトによって日本人を含む各ヒト集団が持つ配列多様性が明らかになり、国際的に統一された単一参照配列の利用では問題が残ることも指摘され、国内でも主任研究者を含めいく

つかの研究グループが日本人基準配列の決定に向け連絡を取り合っている。

本センターはこれまで一貫してGRCh37に基づいたデータ解析を行っており、新規参照配列への移行作業は全く行われていな買った。新しい配列の有用性は明らかで、かつ他機関との整合性の観点からも、新規参照配列に移行する必要があるが、この作業はデータベースを一つ入れ換えれば済むという単純なものではなく、解析パイプラインに組み込まれている個々のソフトウェアに対し設定と動作確認を行う必要があり、センター内の多くの部署で、また計算機の数だけいずれ必要になる作業でもあり、啓蒙も含め、イニシアチブを取って準備を進める必要がある。

まずは最も使用頻度の高いエクソーム解析を取り上げ、マッピング、変異検出、アノテーションを行う体制を整え、最終的には全ゲノム解析、DNA メチル化解析、クロマチン修飾解析、トランスクリプトーム解析、RNA メチル化解析等にも対応させたパイプラインソフトウェアの開発と整備を行う。また、近年、多くの状況で活用され始めたディープラーニングの技術を取り入れ、これら NGS データを自動的に振り分けてデータベースを構築する仕組みについても検討を行う。

２．研究組織

| 研究者 | 所属施設 |
|---|---|
| 岡村 浩司 | 国立成育医療研究センター |
| 片桐 沙紀 | お茶の水女子大学 理学部 |

３．研究成果

NGS から得られる配列データは従来のシークエンサから得られるものと比べると短く、解析方法は既知配列との比較が基本となる。通常は参照配列へのマッピングにより比較が行われるが、これはエクソーム解析、全ゲノム解析に限らず、DNA メチル化、クロマチン修飾のようなエピジェネティクス解析、さらにはトランスクリプトーム解析等においても同様である。小児や周産期疾患を中心とした研究を行っている本センターにおいてはエクソーム解析を行うことが最も多く、また所有する計算機クラスタがエクソーム解析を念頭に設計されたものであるため、初年度はエクソーム解析を最初のモデルケースとして取り上げることとした。

（１） GRCh38 は 2013 年 12 月に発表されて以来、13 のパッチがリリースされている。また、いくつかの研究機関が、研究者の便宜を計り、独自に改変した配列を公開しており、本研究課題においては欧州バイオインフォマティクス研究所(EBI)、カリフォルニア大学サンタクルーズ校(UCSC)、東北メディカル・メガバンク機構(ToMMo)から公開されているデータセットをダウンロード、比較検討することから研究を開始した。EBI に対し、UCSC では Y, R, K, M, W, S, B といった塩基が全て N で表記され、マッピングに利用しやすいことを確認した。また特に第 5, 14, 19, 21, 22 番染色体および Y 染色体において UCSC では多くの N が T, C, A, G の塩基で表記されていた。ToMMo の配列は UCSC の配列を基にしていると思われ、また国際参照配列に対し日本人ゲノムにおいて挿入により存在する配列を集約した配列をデコイ配列に加えた decoyJRG を含んでおり、特に日本人のゲノム配列解析に有用であると判断され、ToMMo GRCh38 を新規参照配列として利用することとした。

（２）GRCh38 とはいえ、ミトコンドリア DNA の配列は統一されたものではないこと

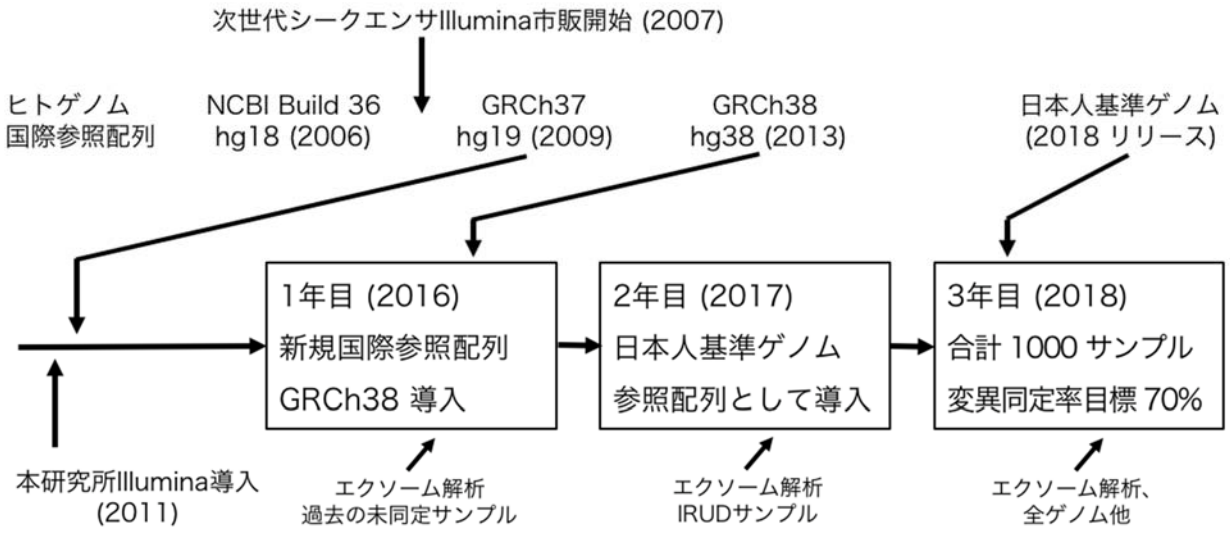


図 1. 参照配列 GRCh38 開発スキーム

が確認された。NC_012920 が広く参照配列として利用されているようであったが、3107 番目に N が入っていることが確認された。ToMMo が採用しているミトコンドリア DNA の配列は NC_012920 とは異なるが、主任研究者が別事業における「日本人由来細胞の同一性・ゲノム安定性評価法の確立」研究において決定した健常日本人 1 名 JOM005 のミトコンドリア DNA 配列 16,570 bp ときわめて類似していることが分かった。そこでミトコンドリア DNA の参照配列については、本研究所にて配列決定が行われた JOM005 を採用することとした。ToMMo の配列とは似ているものの、NC_012920 に対し 310 番目に C の挿入があるためそれ以降の座標がずれてしまうため、アノテーションデータを作り直す必要が生じた。そのため、dbSNP、ANNOVAR のデータベースを改変した。このツールを用いて得られたデータは、2017 年度に倫理審査申請を行った「ミトコンドリアの遺伝形式に関する研究」の基礎データとなった

（３） 全ゲノムに対し、エクソームにおいてはベイト設計領域周辺に限定した解析が必要となる。キャプチャーキットとしてはこれまで、Agilent SureSelect、Illumina TruSeq Exome Kit、Illumina Nextera Rapid Capture を採用したが、どれも GRCh38 に基づくデータが提供されていない。またそれぞれのキットに複数のバージョンが存在し、カバーしている領域もそれぞれに若干異なり、データ処理を複雑にしている。今後、これまでとは違ったキットが市販される可能性もある。そこで今回、これら既存のキャプチャーキットを全てカバーするような広い領域を解析対象として統一することとし、UCSC liftOver を利用して hg19 に基づくデータを GRCh38 に変換して作業を進めた。領域長は合わせて約 150 Mb となった。

（４） パイプラインは希少・未診断疾患イニシアチブ IRUD のために整備したスクリプトを基に、今回の作業においては、マッピング(BWA)、ヴァリアントコール(Picard および GATK)、アノテーション(ANNOVAR)等の各ツールはできる限り最新版を導入した。リードのトリミング、マッピング、リアラインメントによりサンプルごとの BAM 作成、さらにマルチサンプルコーリングによる VCF 作成までを行う 2 つのパイプラインを完成させた。

（５）インハウス健常者コントロールとし

て用いている 2 サンプル、および主任研究者が筆頭著者として *Sci. Rep.* 誌にて発表した色素性乾皮症患者由来の 1 サンプル(Okamura et al. 2016)、合計 3 サンプルをモデルとして従来の hg19 および GRCh38 の両方で解析を行った。マッピング率は 96.66%から 96.54%とわずかながら下がり、ヴァリアントコール数はそれぞれ 179,061 および 178,581 という結果が得られた。色素性乾皮症の原因である *XPA* 遺伝子のホモ変異は問題なく検出されており、診断に利用可能なパイプラインを完成することができた。マッピング率の低下は、GRCh38 に含まれる alternate 配列に起因することが分かった。

（6） アノテーションについては ANNOVAR および HGMD が GRCh38 に対応していることを確認したが、国内 5 拠点によるエクソーム解析から得られている日本人アレル頻度データ、ToMMo による健常日本人 2000 人全ゲノム解析から取得したアレル頻度データ等はどれも非対応で、位置情報を変換したデータベースを独自に作成する必要があり、作業を行った。この成果をエクソーム解析に限らず、全ゲノム解析、DNA メチル化解析、クロマチン修飾解析、トランスクリプトーム解析、RNA メチル化解析のパイプラインにも適用し、GRCh37 あるいは GRCh38 のいずれかを選択できる解析環境を整えた。次世代シークエンサから得られたデータが組み込まれている GRCh38 の優位性は明らかであり、日本人基準配列の利用も含め、本センター発の NGS 解析環境を活用し、小児未診断疾患への応用とともに、広く有用性を示したい。なお、本課題により作成したヒトゲノム参照配列は GRCh38s1 として以下のサイトより一般公開している。

http://epigenetics.nrichd.ncchd.go.jp/genomics/

（7）この数年間、ビッグデータと相まって社会のあらゆる分野でディープラーニングが取り入れられ、大きな成果を上げている。医学や生命科学の分野においてもこういった人工知能の活用が望まれており、本課題のテーマとなっている参照配列について検討を進めた。その結果、参照配列に加えてこれまでに取得されデータベースに登録されている塩基配列データを用いて、NGS から得られる FASTQ データを全ゲノム、全エクソーム、RNA-seq、ChIP-seq 等に分類し自動的にデータベースを構築するフレームワークを提案することに成功した。この結果は第 40 回日本分子生物学会年会の口頭発表に採択され(4P2T18-01)、研究協力者の学生が発表を行った。さらに本研究所の定例セミナーにおいて発表を行ったところ、病院の職員も含め、把握できる限りこれまでの定例セミナーの中で最も多い聴衆を集め、ディープラーニングに対する期待と関心の高さがうかがわれた。

その後、代表者は倫理審査申請「顔写真を判別するプログラム構築によるデータサイエンス研修」を経て、現在 24 名からなるデータサイエンス研修チームを率い、センター内における AI の啓蒙活動に携わっている。本成育医療研究開発費による成果が再び本センター全体で活用されている。

４．研究内容の倫理面への配慮

本研究課題においては、倫理審査委員会により承認を得て NGS によりシークエンシングされたサンプルのみを解析対象とした。