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

平成　２９　年　　６　月　　７　日現在

機関番号：　１２６０１
研究種目：　若手研究(B)
研究期間：　2015～2016
課題番号：　１５Ｋ２０９４１

研究課題名（和文）超ロングリードを用いた包括的全ゲノム配列解析の確立と神経疾患解明への応用

研究課題名（英文）Comprehensive whole genome sequence analysis using a long read sequencer for delineating molecular mechanism of neurological diseases

研究代表者

石浦　浩之（Ishiura, Hiroyuki）

東京大学・医学部附属病院・助教

研究者番号：４０６３２８４９

交付決定額（研究期間全体）：（直接経費）　　3,200,000円

研究成果の概要（和文）：神経疾患の原因としては、一塩基多型や小さな挿入・欠失変異のみならず、構造変異やリピート伸長変異といった変異が原因となることがある。現在中心的に行われているshort readのシーケンスでは、構造変異などの変異を十分には検出することができず、またエクソーム解析では非翻訳領域やイントロンの変異を検出することができない。本研究では、短鎖シーケンサーによる全ゲノム配列解析に加えて、長鎖シーケンサーによる全ゲノム配列解析を施行した。短鎖シーケンサーでは解析しきれない変異が長鎖シーケンサーで検出できた例もあり、本方法は今後神経疾患の原因究明のために有用であると考える。

研究成果の概要（英文）：Several kinds of mutations including rearrangements and repeat expansions as well as single nucleotide variants and short insertion/deletions cause neurodegenerative and neuromuscular disorders. Usual exome sequencing, however, overlook some of the mutations. To overcome this, I performed whole genome sequencing using a short read sequencer and a long read sequencer. Some of the mutations can only be delineated by the long read sequencer, indicating the importance of long read sequencing. This strategy will help us to understand the molecular mechanism of neurodegenerative disorders.

研究分野：　神経遺伝学、神経内科学

キーワード：　ロングリードシーケンス　全ゲノム配列解析　神経変性疾患

１．研究開始当初の背景

神経筋疾患においては、通常の一塩基置換や、小さい挿入・欠失変異に加えて、リピート伸長変異が認められることがある。古くはハンチントン病、球脊髄性筋萎縮症、脊髄小脳変性症といった CAG リピート伸長によるポリグルタミン病の発見の例もあるが、近年では前頭側頭型認知症・筋萎縮性側索硬化症（C9orf72-linked frontotemporal dementia/amyotrophic lateral sclerosis）における GGGGCC リピート伸長変異、脊髄小脳変性症 31 型・36 型（spinocerebellar ataxias type 31 and 36）における TGGAA・GGCCTG のリピート伸長変異の発見などから、非翻訳領域のリピート伸長変異に脚光が当たっている。しかしながら、次世代シーケンサーによって飛躍的に塩基配列解析の効率が良くなったとしても、未だに高価であり、従来はエクソーム解析やターゲット解析と言った方法が行われることが多かった。例えば、エクソーム解析はエクソン領域を選択的に濃縮して解析を行う方法であるが、非翻訳領域を解析することができない。またそもそも、Illumina 社の次世代シーケンサーのように、短い配列を大量に解読する short read sequencer では、長いリピート伸長変異を検出することが非常に困難であった。こういった難点から、リピート伸長変異などを網羅的に解析できる方法はなかった。

以上のように、ヒトゲノム中には、リピート伸長変異をはじめとした難読領域が残されており、神経筋疾患の原因をさらに明らかにするためには、そのような難読領域をもきちんと解析できる方法の確立が必要と考えられる。

今回の研究は、難読領域を含めて解析を可能にする全ゲノム解析方法を確立することを目標としている。具体的には、PCR-free のライブラリを調整し、short read sequencer を用いて全ゲノム配列解析を行うと共に、long read sequencing を組み合わせて網羅的な全ゲノム配列解析を行うことを通して、神経筋疾患の病態解明を進めることを目指すものである。

２．研究の目的

Short read による全ゲノム配列解析と、long read による全ゲノム配列解析を組み合わせることで、包括的な全ゲノム配列解析を行う。

また、疾患の遺伝子座を同定するために、ハプロタイプ解析を精緻化することを目的とする。具体的には、本邦に多い疾患については、共通祖先由来の共通ハプロタイプを有することが多いので、全ゲノム SNP データから共通ハプロタイプを見つけ出すプログラムを作成する。

３．研究の方法

1 マイクログラムの genomic DNA から、TruSeq DNA PCR-Free Sample Prep Kit (Illumina)を用いて PCR を用いずサンプル調整を行う。HiSeq2500 (Illumina)を用いて、Rapid mode の 150 塩基ペアエンド法を用いて全ゲノム配列解析を行う。2 枚のフローセルによる解析を行うことによって、約 40X-50X のカバレッジとなることを想定した。

ショートリードデータについては、BWA (Li and Durbin. ***Bioinformatics*** 2009)を用いて参照配列 hg19 に alignment を行う。SAMtools を用いて一塩基置換、小欠失・挿入変異の同定を行う。リピート伸長変異についても TRhist (Doi et al. ***Bioinformatics*** 2014)等を用いて検出を行う。適宜自作スクリプトも作成して解析を行う。

一方、long read sequencer としては、PacBio RSII を用いた。約 100 マイクログラムの genomic DNA を準備した。

Pacific Biosciences 社の PacBio Template Prep Kit を用い、Guidelines for preparing 20 kb SMRTbell templates に従い、ライブラリを作成する。具体的には、g-TUBE for DNA Shearing (Covaris)を用いて genomic DNA の断片化を行った後、AMPure ビーズを用いて純化する。次に、ExoVII、DNA Damage Repair mix、End Repair Mix を用いて末端修復を行ったのち、Blunt end ligation で SMRTbell templates を作成する。Sequencing primer をアニリングさせた後、PacBio RS II で塩基配列解析を行う。得られた long read sequece については、blasr を用いて、参照配列（hg19）に alignment を行う。データについては、Integrative Genomic Viewer (http://software.broadinstitute.org/software/igv/)を用いて分析を行った。

本邦に多い疾患については、共通祖先由来の共通ハプロタイプを有することが多いため（Ishiura et al. ***Am J Hum Genet*** 2012, Ishiura et al. ***Arch Neurol*** 2012, Taira et al. ***Neurogenetics*** 2012, Mano KK et al. ***Neurol Genet*** 2016）、SNP タイピングデータを用いて、共通するハプロタイプを検出することができれば、それを元に疾患の遺伝子座を同定することができるのではないかと考えた。具体的には、Affymetrix Genome-Wide Human SNP array 6.0 を用いて、指定通りの方法で約 90 万個の SNP についてのデータを取得した。Genotype については、Genotyping Console を用いて解析・抽出を行った。得られた genotype について、Homozygosity Haplotype 法（Miyazawa et al. ***Am J Hum Genet*** 2007）を用いてハプロタイプを構築し、全ゲノムから共通ハプロタイプと考えられる領域を検出できるプログラムを作成した。既に共通ハプロタイプを持つということが判明している家系 (Mano KK et al. ***Neurol Genet*** 2016)の SNP タイピングデータを用いて検証を行った。

４．研究成果

神経変性疾患に罹患した 1 例（#9481）について、Illumina HiSeq2500 を用いて、Rapid mode 150bp ペアエンド法を用いて全ゲノム配列解析を行った。ライブラリは従来の PCR を用いる方法ではなく、PCR-free の方法を用いた。2 スライドで解析することにより、1,143,820,844 reads を得た。参照配列には 1,070,344,152 reads(93.6%)がマップされ、そのうちユニークにマップされたものが 975,343,052 read(85.3%)であった。ユニークにマップされたものだけを検討しても、平均カバレッジは 47.3X（図 1）となり、5X 以上のカバレッジとなる領域が 99.91%、10X 以上のカバレッジとなる領域が 99.79%となり、20X 以上と良好にカバーされる領域に限っても 98.78%（図 2）と、非常に良好なデータを得ることができた。

図 1：short read sequencer によるカバレッジの分布

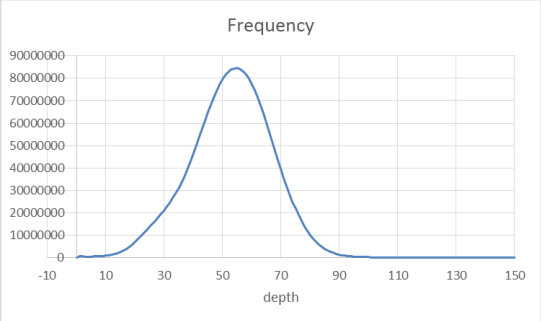
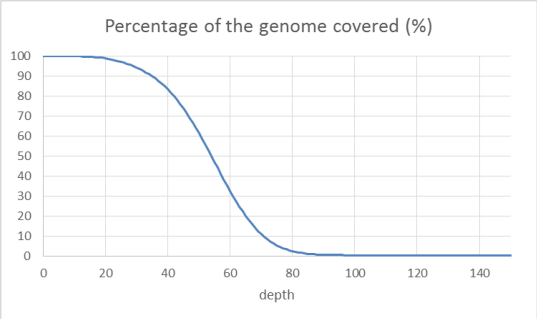


図 2：カバーされた参照配列上の領域の割合



今回は PCR-free のライブラリ調整をおこなったが、これにより、例えば GC content が 0%や 100%となるような read をも検出することができるようになった。Illumina 社のシーケンサーでは、元々GC content によりシーケンス効率が変化し、特に 0%や 100%となる極端な配列の解読効率が落ちることが指摘されており、またライブラリ調整の際に PCR を行うことによりこれが顕著になることが指摘されてきたが、PCR free でライブラリ調整を行うことで、より均一なデータをとることができることが判明した。

同一サンプルについては、PacBio RSII を用いて、全ゲノム配列解析を行った。サンプル調整を行った上で、P6C4 chemistry を用いて 6 回に分けてシーケンスを行った。結果、約 20X のカバレッジとなるデータを得ることができた。

PacBio RSII を用いることで、例えば short read sequence から構造変異が疑われながらもその構造が明らかではなかったような部分について、全長を読み通すことが可能になり、long read により解析精度を向上させることに成功した。

共通ハプロタイプを検出するプログラムについては、例として *PRNP* P105L 変異を持つ家系を用いて解析を行った（Mano KK et al. ***Neurol Genet*** 2016）。既に、*PRNP* 遺伝子周囲のハプロタイプが 3 家系で共通していることが示されている。作成したプログラムで、3 家系で共通したハプロタイプを持つ領域を検出したところ、*PRNP* 遺伝子座を含んでいた。このことより、本プログラムは、ある一定以上の長さを持つハプロタイプを検出するために有用であると考えられた。しかしながら、#9481 を含む類似疾患を用いたハプロタイプ解析では、共通ハプロタイプの検出をすることができなかった。類似の症状を呈する発症者 17 名と、対照者 198 名のゲノムワイド関連解析も施行したが、低い p 値を示す SNP を見いだすことができなかった。これらのことは、遺伝学的異質性（異なった原因遺伝子、もしくは同一遺伝子上の異なった遺伝子変異）を示唆する結果と考え、今後は家系調査を進め、候補領域を同定するためには、連鎖解析を行っていく方が有用と考えられた。候補領域を狭めることができれば、本研究で全ゲノム配列データを既に得られているため、速やかに原因遺伝子の同定につながるものと考える。

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 5 件）

①**Ishiura H**, Tsuji S. Epidemiology and molecular mechanism of frontotemporal lobar degeneration/amyotrophic lateral sclerosis with repeat expansion mutation in C9orf72. J Neurogenet 2015;29:85-94. doi: 10.3109/01677063.2015.1085980. 査読有り.

②**石浦浩之**. SCD・MSA 病型の新研究　遺伝型の最新研究進歩. 難病と在宅ケア 2016;22:23-26. 査読なし.

③**石浦浩之**. ゲノムビッグデータの解析. 神経内科 2016;84:585-589. 査読なし.

④辻省次、三井純、**石浦浩之**．神経遺伝医学研究の歴史的背景と今後の課題．遺伝子医学 MOOK　別冊　シリーズ：最新遺伝医学研究と遺伝カウンセリング．In press．査読なし．

⑤**石浦浩之．**次世代シーケンサー、次々世代シーケンサーとクリニカルシーケンシング．遺伝子医学 MOOK　別冊　シリーズ：最新遺伝医学研究と遺伝カウンセリング．In press．査読なし．

〔学会発表〕（計　0　件）

〔図書〕（計　0　件）

〔産業財産権〕

○出願状況（計　0　件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況（計　0　件）

名称：
発明者：
権利者：
種類：
番号：
取得年月日：
国内外の別：

〔その他〕
ホームページ等　なし

６．研究組織
(1)研究代表者
石浦浩之　（Hiroyuki, Ishiura）
東京大学・医学部附属病院・助教

研究者番号：40632849

(2)研究分担者
なし
（　　）

研究者番号：

(3)連携研究者
なし
（　　）

研究者番号：

(4)研究協力者
なし
（　　）