

科学研究費助成事業 研究成果報告書

平成 27 年 6 月 5 日現在

機関番号：12601

研究種目：新学術領域研究(研究領域提案型)

研究期間：2010～2014

課題番号：22129008

研究課題名(和文)脳疾患パーソナルゲノム多様性を分析する情報学の創成

研究課題名(英文)Exploring informatics for brain diseases based on personal genomics

研究代表者

森下 真一(Morishita, Shinichi)

東京大学・新領域創成科学研究科・教授

研究者番号：90292854

交付決定額(研究期間全体)：(直接経費) 79,500,000円

研究成果の概要(和文)：ヒトゲノム全域にわたって多様性情報を収集し分類することは人類遺伝学の大きな目標であるが、その情報量の大きさゆえに効率的に観測・分類を行うことは困難であった。本研究では、短い高精度なDNA解読リードおよび精度は若干劣るが1万塩基対を超える長いDNA解読リードという互いに補完的な特徴を持つ情報を組合せ、1塩基置換(SNP)/短い挿入削除等のミクロな多様性だけでなく、大規模な挿入削除領域/重複領域/染色体内逆位/染色体の融合転座等のマクロな多様性を検出するアルゴリズムを研究開発した。これらを利用して脳疾患に特徴的な構造多型を個人ゲノムデータから解読することを可能にした。

研究成果の概要(英文)：Disease-associated genetic variations in the human genome are divergent, including single nucleotide polymorphism (SNP) and short insertions/deletions as well as large-scale variations such as large deletions, long transposons, intra-chromosomal inversions, and chromosomal rearrangements. Comprehensive understanding of these variations has been challenging because an enormous volume of information had to be collected and classified properly. In this study, we utilize two types of high-throughput DNA sequencer of complementary characteristics; namely, one type outputs highly accurate but short DNA reads, while the other is able to generate extremely long reads of average length > 10,000 base pairs with moderate base accuracy. We developed a suite of algorithms that combined short and long reads to uncover large-scale structural variations. With these methods, we were able to detect a number of structural variations specific to brain diseases.

研究分野：脳疾患 人類遺伝学 ゲノムサイエンス バイオインフォマティクス アルゴリズム

キーワード：全ゲノム解析 エキソーム解析 脳疾患 家系解析 バイオインフォマティクス アルゴリズム 構造変異 異常リピート配列検出

1. 研究開始当初の背景

ヒトの個人間のゲノムの差異を描出するための究極の方法は、個人ゲノムの解読である。しかしながら、長大なヒトゲノムの最初の配列が解読されるには 1986 年の計画立案から 17 年の月日を費やし、2003 年までかかった。それまでの間は、ゲノム上の目印(マーカー)を見つけ、その間の距離を狭める手法によりヒトゲノムの多様性が収集されてきた。詳しくは RFLP、マイクロサテライト、そして SNP のマーカーが開発された。特に約 50 万個の SNP マーカーは、疾患感受性遺伝子の領域を狭めるのに近年大きな威力を発揮している。

それ以前の 2002 年から、我々はゲノム解読のソフトウェア技術の研究開発に取り組んだ。当時は、ヒトの個人ゲノムの解読は当面先という時代であり、発生物学のモデル生物であるメダカゲノムの解読を通じてソフトウェア技術を培った。メダカが利用された背景には、ゲノムサイズがヒトの約 1/4 であり、ハプロタイプの差が無い近交系が 10 系統以上存在し、ゲノム解読が容易である点があった。そのため、ゲノム解読の障害が少なく、2006 年には魚類では最も完成度が高い、染色体全体の約 90%を被覆するゲノムが完成した(Nature 2007)。この結果、染色体の重複/融合/転座/逆位等のマクロな変化が、脊椎動物間のゲノム進化過程でどのように起こってきたかを初めて詳細に分析できた(Genome Research 2007)。右図に魚類内のマクロな多様性と進化を示す(Nature 2007)。この技術はヒトゲノムのマクロな多様性を分析するのに役立つであろう。

加えてメダカゲノムプロジェクトでは、約 1800 万年前に分岐した 2 つの系統のゲノムを同時に解読した。これはヒトゲノムで例えるなら、2 人の個人ゲノムをフルシーケンシングしたことに相当する。メダカの 2 つの系統は交配可能な同一種でありながら、塩基置換率は 3.42%と、ヒト/チンパンジーの差 1.2% より遥かに高く、他に類のない遺伝学の研究資源となっている。さらに挿入削除の箇所もゲノム全体で網羅的に把握することにも成功した(挿入削除率 0.59%)。このデータを分析した結果、「転写開始点下流では、挿入削除率は +100, +300, +500 塩基対下流でピークをなし、挿入削除率は +200, +400, +600 塩基対下流でピークをなすという周期性がある。この周期性はヌクレオソームコア位置が +100, +300, +500 塩基対下流に存在する周期性と同期する」という発見が得られた(Science 2009)。フルシーケンシングした結果、新しいモチーフが得られる一つの例示になっている。

2. 研究の目的

ヒトゲノムの多様性には大きく分けて、1 塩

基置換(SNP)/短い挿入削除等のミクロな多様性、大規模な脆弱領域/染色体内逆位/染色体の融合転座等のマクロな多様性がある。ヒトゲノム全域にわたって多様性情報を収集し分類することは人類遺伝学の大きな目標であるが、その情報量の大きさゆえに効率的に観測・分類を行うことは困難であった。超高速 DNA 解読装置の登場は、疾患関連遺伝子座の状態を、1 塩基対レベルで観測することを可能にした。この結果、稀な変異が多様に起こり発症すると考える仮説の検証を可能にしつつある。本研究では、これまでの研究でゲノムと疾患の相関について先駆的研究が展開されてきた脳疾患を対象に、個人ゲノム情報を収集し、疾患と相関の高い DNA 変異を探し出し、生まれる機序を推定する新しい情報学を創成することを目指している。

3. 研究の方法

本研究では、互いに補完的な特徴を持つ 2 つの DNA 解読装置を用いた。

- 短いリード(100-250 塩基)を解読する DNA 装置(Illumina 社 HiSeq2500, MiSeq)。低コストでリードを収集でき、塩基精度も 95-98%と高い特長を備える。現在、1 塩基変異や短い挿入削除を検出する技術として普及している。一方、GC 率が極端に高い(もしくは極端に低い)DNA 断片は解読しにくいいため、見落としも多い。リード長が短いため 100 塩基を超える大規模な構造変異の検出にも弱い。
- 長いリード(1000~40,000 塩基)を解読する DNA 装置(Pacific Biosciences 社 PacBio RS II)。塩基あたりのコストは Illumina HiSeq 等に比べて約 1 桁高く、塩基精度も 85-90%と低い。しかし塩基エラーがランダムに入るため、リードの重複度を 50 倍ぐらいにすると塩基精度を 99.9%以上に上げることができる。しかも GC 率が極端に高い配列(たとえば 100%の GC 率の配列)も解読することが可能である。そのためヒトゲノム中の難読領域を解明するのに役立つ。また長いリードを解読可能なため、大規模な構造変異の検出も可能である。

これら 2 つの DNA 解読装置の特長を活かしながら、疾患と相関の高い変異を探し出すアルゴリズムやプログラムなどの情報分析方法を研究開発した。

4. 研究成果

1 塩基変異、挿入削除の検出: 本研究領域がはじまった当初は短いリードの DNA 解読装置を使って、全ゲノム解読および全エキソーム解読の情報解析パイプラインを作成した。本研究領域代表の辻研究室との共同研究を通じて、各変異の重篤性、体細胞変異の検出の精度をあげ、家系解析も充実した。本研究領域の研究者のために DNA 情報分析を支援し、

多数の論文を出版することに貢献した。

Short tandem repeat の検出方法： short tandem repeat と呼ばれる長さが 2-6 塩基のユニットが繰り返し異常な長さ(典型的には 1000 塩基以上)に伸長する現象が脳疾患と関連することが知られている。本研究領域代表の辻研究室と共同で、個人ゲノムの中に見出すための検出感度を上げるためのアルゴリズムを設計した。特に工夫したのは、長さ 100 塩基前後の短い配列を利用して、長い short tandem repeat の存在を予測することである。10-20 億本の短い配列を高速に処理するために様々な最適化技術を取り入れる必要があったが、1 日以下程度で処理することが可能になった(Bioinformatics, 2014)。この予測により、長い short tandem repeat が存在するゲノム上の位置までを推定できるようになったが、内部を完全解読するのは困難であった。幸い、位置情報が得られるため short tandem repeat 周辺の配列を使って PCR 増幅できる場合は増幅し、長い配列を解読可能な Pacific Biosciences 社の PacBio RS II シーケンサーにより解読ができるか否かを、SCA31 をサンプルとして検証した。その結果、いままでは中身を解読できていなかった short tandem repeat を完全解読できるようになった。この方法を、本研究領域の研究者のゲノムデータに適用してきており、最終年度以降もその努力を続けている。

構造変異の検出方法： short tandem repeat は DNA 構造を大規模に変化させる変異の一つである。ヒトゲノムには他にもトランスポゾン挿入・削除、逆位、染色体融合・転座、homologous recombination が生むゲノム重複など様々な大規模変異があると考えられている。しかし短いリードでは検出できないため、大規模構造変異が個人ゲノムにどの程度広がっているかは謎であった。そこで平成 24 年度から PacBio RS II を用いて構造変異を検出するアルゴリズムを研究開発した。コストを抑えるため、当初は低コストの短いリードを集め、長いリードのエラーを補正することに取り組み、ある程度の成果を得た。しかし最終年度の平成 26 年度には、長いリードの塩基長が改善され平均 2,000 塩基から 10,000 塩基へと伸び、コストも徐々に下がった。長いリードだけを組合せ、大規模構造変異を精度よく検出できる可能性が広がったため、新しいアルゴリズムも研究開発した。本研究領域代表の辻研究室と共同で複数の日本人ゲノムにおける大規模構造変異を解明しつつある。また世界的にもこの研究領域は注目されており、Pacific Biosciences 社および Howard Hughes Medical Institute 等の米国の研究機関の研究者とも連絡を取りながら、個人ゲノムにおける大規模構造変異の広がりを分析しつつある。論文は投稿準備

中である。

リピート配列上の DNA メチル化の検出方法：ヒトゲノムの半分近くは繰返し配列 (Alu, LINE 等のトランスポゾン、short tandem repeat expansion) により占められている。その DNA メチル化状態は従来技術 (マイクロアレー、パイサルファイトシーケンシング) では検出が難しく、理解が進んでいない。そこで、PacBio RS II が出力する微小な kinetic information を増幅し、DNA メチル化情報を抽出する高感度なアルゴリズムを平成 24 年度から研究開発し、平成 26 年度には実用的なレベルになった。現在論文を投稿中である。Pacific Biosciences 社および Howard Hughes Medical Institute 等の米国の研究機関の研究者とも連絡を取りながら、個人ゲノム間での変化、生殖細胞と癌のゲノムでの違いを分析している。

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 14 件)

1. K Ichikawa, S Morishita. "A linear time algorithm for detecting long genomic regions enriched with a specific combination of epigenetic states" BMC Genomics 16 (Suppl 2), S8 (2015) 査読有
2. Satoshi Yamashita, Akira Mori, Yasuto Nishida, Ryoichi Kurisaki, Nozomu Tawara, Tomo Nishikami, Yohei Misumi, Hidetsugu Ueyama, Shigehiro Imamura, Yujiro Higuchi, Akihiro Hashiguchi, Itsuro Higuchi, Shinichi Morishita, Jun Yoshimura, Makoto Uchino, Hiroshi Takashima, Shoji Tsuji, Yukio Ando. "Clinicopathological features of the first Asian family having vocal cord and pharyngeal weakness with distal myopathy due to a MATR3 mutation" Neuropathology and applied neurobiology (in press) 査読有
3. Doi K, Monjo T, Hoang PH, Yoshimura J, Yurino H, Mitsui J, Ishiura H, Takahashi Y, Ichikawa Y, Goto J, Tsuji S, Morishita S. "Rapid detection of expanded short tandem repeats in personal genomics using hybrid sequencing." Bioinformatics. 15;30(6):815-22.(2014) 査読有
4. Isojima T, Doi K, Mitsui J, Oda Y, Tokuhiko E, Yasoda A, Yorifuji T, Horikawa R, Yoshimura J, Ishiura H, Morishita S, Tsuji S, Kitanaka S. "A recurrent de novo FAM111A mutation

- causes Kenny-Caffey syndrome type 2." *J Bone Miner Res.* 29(4):992-8. (2014) 査読有
5. Yuji Takahashi, Yoko Fukuda, Jun Yoshimura, Atsushi Toyoda, Kari Kurppa, Hiroyoko Moritoyo, Veronique Belzil, Klaus Elenius, Guy Rouleau, Asao Fujiyama, Shinichi Morishita, Jun Goto, Shoji Tsuji. "Mutations in ERBB4 That Disrupt The NRG-ErbB4 Pathway Cause Autosomal Dominant Familial ALS Type 19" *Neurology* 82(10), 83 (2014) 査読有
 6. Kishin Koh, Hiroyuki Ishiura, Michiaki Miwa, Jun Yoshimura, Jun Mitsui, Jun Goto, Shinichi Morishita, Shoji Tsuji, Yoshihisa Takiyama. "Exome sequencing shows a novel de novo mutation in ATL1" *Neurology and Clinical Neuroscience* 2 (1), 1-4 (2014) 査読有
 7. Ishii A, Saito Y, Mitsui J, Ishiura H, Yoshimura J, Arai H, Yamashita S, Kimura S, Oguni H, Morishita S, Tsuji S, Sasaki M, Hirose S. "Identification of ATP1A3 mutations by exome sequencing as the cause of alternating hemiplegia of childhood in Japanese patients." *PLoS One.* 8(2):e56120 (2013) 査読有
 8. Takahashi Y, Fukuda Y, Yoshimura J, Toyoda A, Kurppa K, Moritoyo H, Belzil VV, Dion PA, Higasa K, Doi K, Ishiura H, Mitsui J, Date H, Ahsan B, Matsukawa T, Ichikawa Y, Moritoyo T, Ikoma M, Hashimoto T, Kimura F, Murayama S, Onodera O, Nishizawa M, Yoshida M, Atsuta N, Sobue G; JaCALs, Fifita JA, Williams KL, Blair IP, Nicholson GA, Gonzalez-Perez P, Brown RH Jr, Nomoto M, Elenius K, Rouleau GA, Fujiyama A, Morishita S, Goto J, Tsuji S. "ERBB4 mutations that disrupt the neuregulin-ErbB4 pathway cause amyotrophic lateral sclerosis type 19." *Am J Hum Genet.* 93(5):900-5. (2013) 査読有
 9. Ichikawa Y, Ishiura H, Mitsui J, Takahashi Y, Kobayashi S, Takuma H, Kanazawa I, Doi K, Yoshimura J, Morishita S, Goto J, Tsuji S. "Exome analysis reveals a Japanese family with spinocerebellar ataxia, autosomal recessive 1." *J Neurol Sci.* 15;331(1-2):158-60. (2013) 査読有
 10. Mitsui J, Matsukawa T, Ishiura H, Higasa K, Yoshimura J, Saito TL, Ahsan B, Takahashi Y, Goto J, Iwata A, Niimi Y, Riku Y, Goto Y, Mano K, Yoshida M, Morishita S, Tsuji S. "CSF1R mutations identified in three families with autosomal dominantly inherited leukoencephalopathy." *Am J Med Genet B Neuropsychiatr Genet.* 159B(8):951-7 (2012) 査読有
 11. Funakoshi M, Tsuda M, Muramatsu K, Hatsuda H, Morishita S, Aigaki T.: "A gain-of-function screen identifies wdb and Ikb1 as lifespan-extending genes in Drosophila." *Biochem Biophys Res Commun* 405(4). 667-72 (2011) 査読有
 12. Ogoshi K, Hashimoto S, Nakatani Y, Qu W, Oshima K, Tokunaga K, Sugano S, Hattori M, Morishita S, Matsushima K: "Genome-wide profiling of DNA methylation in human cancer cells" *Genomics* 98. 280-287 (2011) 査読有
 13. Hongyan Wu, Taro L Saito, Shinichi Morishita: "Accelerating Path-free XML Queries in RDBMS." *IPSI Online Transactions* 3. 206-217 (2010) 査読有
 14. Kuroshu RM, Watanabe J, Sugano S, Morishita S, Suzuki Y, Kasahara M.: "Cost-effective sequencing of full-length cDNA clones powered by a de novo-reference hybrid assembly." *PLoS One* 5(5). e10517 (2010) 査読有
 - 15.
- [学会発表](計7件)
1. Yuta Suzuki, Tatsuya Tsukahara, Hiroyuki Takeda, Shinichi Morishita. "Onserving heterozygotic DNA methylation patterns in diploid genomes using kinetic data from PacBio RS" *Advanced in Genome Biology and Technology*, (20150226). Florida, USA
 2. Koichiro Doi, S Morishita et al. "Detecting Significant Short Tandem Repeats in Personal Genomes" *Advanced in Genome Biology and Technology*, (20130221). Florida, USA
 3. Shoichiro Oishi, S Morishita et al. "Fast Error Correction of SMRT Sequencing Reads with Illumina Reads for Human Genome Resequencing" *Advanced in Genome Biology and Technology*, (20130221). Florida, USA
 4. Shinichi Morishita: "Searching Massive Epigenome Data for Evolutionarily Conserved Sequence Motifs" *First International IEEE Conference on Computational Advances in Bio and medical Sciences (ICCABS 2011)*. (20110203). Orlando, Florida, USA

5. Shinichi Morishita: "Genetic Variation Associated with Nucleosome Structure and DNA Methylation" Fish Genome Meeting. (20110311). Sanger Center, UK
6. Shinichi Morishita: "Genetic Variation Associated with Nucleosome Structure and DNA Methylation" BioSoft-2011. (20110323). Beijing, China
7. Masahiro Kasahara: "New Alignment Algorithm without Realignment for More Accurate Indel Detection" The 34th Annual Meeting of the Molecular Biology Society of Japan. (20111224). Pacifico Yokohama, Yokohama, JAPAN

〔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

<http://www.personal-genome.jp/>

6. 研究組織

(1)研究代表者

森下 真一 (MORISHITA, Shinichi)
東京大学・大学院新領域創成科学研究科・
教授
研究者番号：90292854

(2)研究分担者

笠原雅弘 (KASAHARA, Masahiro)
東京大学・大学院新領域創成科学研究科・
講師
研究者番号：60376605