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　 ６ 月 １０ 日現在

機関番号：１４３０１
研究種目：基盤研究(C)（一般）
研究期間：2019～2021
課題番号：１９Ｋ０７３４９

研究課題名（和文）高精細化ゲノム情報による難治性疾患の原因遺伝子変異同定および高度解析技術の確立

研究課題名（英文）Elucidation of novel causative genes for rare diseases via high quality genomic information

研究代表者
川口　修治（Kawaguchi, Shuji）

京都大学・医学研究科・准教授

研究者番号：００５２５４０４

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）：本研究では、HLA遺伝子等とりわけ高多型で従来の解析手法の適用が難しい領域でも、高品質にゲノム情報を測定する技術を開発した。次に抽出情報を基に、希少難治性疾患の発症に関わる変異の同定技術を開発した。開発手法により、HTLV-1 関連脊髄症とIgG4関連疾患において発症と大きく関わるHLA-DRB1上のアミノ酸同定に成功した。HTLV-1 関連脊髄症ではウイルスゲノム情報を含めた発症リスク予測法も開発した。さらに、人工知能技術を用いて、ゲノム情報、臨床情報、データベース、文献情報等を統合解析することで、ゲノム情報のみの解析が困難であった低頻度の原因遺伝子の同定技術や疾患を分類する手法を開発した。

研究成果の学術的意義や社会的意義
希少難治性疾患は、患者数が少ないことから十分なデータが得られず、他の疾患と比べてゲノム解析の効率が低いという問題が存在した。本研究では、ゲノム情報の高品質化や様々な情報を人工知能技術を用いて統合的に解析するといった従来にない技術の確立を目指した。
開発した手法は、複数の難病においてその有効性を示すことができた。そのため、手法の汎用性は高く、本研究で解析した疾患のみならず様々な難病にも適用することが可能であり、多くの疾患解明研究への貢献が期待できる。

研究成果の概要（英文）：We established a technology to elucidate high-quality genomic information from high polymorphic region such as HLA locus at where conventional analysis method is difficult to adapt. The developed method succeeded to detect susceptible and protective amino acid residues related to a development of HTLV-1-associated myelopathy (HAM/TSP) and IgG4-related disease. In addition, we developed a prediction method for a development risk of HAM/TSP with mixing information of host and virus genome.
We also developed an estimation method of novel causative genes for rare diseases based on an artificial intelligence. The novel technique integrates genomic information with clinical information, database and literatures, and estimate candidates of novel causative genes that are difficult to elucidate by using only genome data.

研究分野：数理統計学

キーワード：HLA　人工知能　機械学習　希少難治性疾患　バイオインフォマティクス　次世代シークエンサー　HTLV-1 関連脊髄症　網膜色素変性

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

次世代シークエンサー（NGS）の登場により、高速・高精度にゲノムを解読できるようになり、希少難治性疾患（難病）等のゲノム全領域での解析が求められる疾患に対しての解析研究が発展してきた。一方、ヒト白血球抗原（*HLA*）やキラー細胞免疫グロブリン様受容体（*KIR*）といった免疫と深く関連する遺伝子は、非常に高多型な領域であるため、従来のゲノム解析手法が適用できず、疾患原因変異の検出が困難である。また、ウイルス性疾患やオリゴジェニックな疾患では、宿主とウイルスゲノムや環境因子が複雑に関連し合うため、解析のための高度なモデル構築と結果の解読技術が求められる。

一方、難病の一つである網膜色素変性は単一遺伝子疾患（メンデル遺伝病）ではあるが、全ゲノム解析（WGS）情報を駆使しても原因遺伝子の同定率は飛躍的に増加していない。網膜色素変性の原因遺伝子は既知のものでも百種類近く存在するが、既知の原因変異によって得られる同定率は半分にも満たず、多くの未知の原因遺伝子・変異が存在するとされる。これらの未知変異はそれぞれが極めて低頻度であると考えられるため、従来の家系情報を用いた原因遺伝子の探査研究で同定される事は稀であり、集団的な解析を用いる場合では統計的に十分な検体を集めるのに相当な労力が生じる。

以上のように、ゲノムシークエンス技術の向上が、必ずしも疾患研究の発展に直接繋がるわけではなく、ゲノム情報のみならず様々な情報を統合的に解析する必要がある。特に難病においては、患者数が少ない上に類似疾患が多く、疾患の定義が確固されていないものも多い。そのため、情報の分類があいまいで従来の解析モデルへの当てはめが難しい。しかしながら、近年の機械学習を始めとする人工知能技術の発展により、複雑かつ曖昧なデータに対しても適用できる技術が開発されつつある。そこで、これらの技術を応用することで、難病におけるゲノムと多種の情報を含めた統合解析手法の確立が期待できる。

２．研究の目的

本研究では、機械学習や人工知能技術を駆使してゲノム・オミックス情報および臨床情報の統合データから、疾患と関連する因子を導き出すための解析手法の確立を目指す。

本研究で目指す開発手法は以下の通りである。

・*HLA* 遺伝子の高精度配列解析技術と関連解析手法の確立

HLA は、自己・非自己を識別することで免疫応答の誘導に関わる重要な遺伝子であり、*HLA* のアレルが個人間で大きく異なる。この HLA アレルや遺伝子間ハプロタイプが様々な疾患と関連することが知られている。そのため、多くの疾患におけるゲノム解析では、コントロールとケース群との *HLA* アレル頻度を比較することが重要である。しかしながら、*HLA* は数万種のアレルが存在し、遺伝子間の相同性も極めて高いことから、通常のゲノム解析ではアレルを決定できない。また、国際データベースに登録されている HLA 遺伝子の完全長配列は一部のみで、多くの未知アレルが存在する。そこで、次世代シークエンサーを用いて日本人大規模検体に対する *HLA* 主要遺伝子のターゲットシークエンスをおこない、HLA 遺伝子の全長配列を決定する。得られた結果をデータベースとして整備した後広く公開し、*HLA* 遺伝子解析の高解像度リファレンスとして確立する。整備した *HLA* リファレンスはゲノム情報と統合することで、高精細なデータを構築する。

一方、難病データ解析に複雑なデータ構造を持つ *HLA* 情報を取り入れるには、従来とは異

なる解析モデルの構築が求められる。特に、ウイルス性疾患の多くは *HLA* との関連が示唆されるが、この解析には宿主のゲノム情報のみならず、ウイルスゲノム情報をモデルに含めて解析する必要がある。そこで、本研究では *HLA* 情報やその他の情報を含めた統合解析モデルの開発を行う。

・人工知能技術を用いた統合難病データ解析手法の確立

メンデル遺伝病である、網膜色素変性はこれまでのゲノム解析により、100 種類近くの原因遺伝子が報告されている。しかしながら、国内においては遺伝子診断による同定率は 3 割強であり、多くの未知原因遺伝子・変異が存在するとされ、ゲノム情報のみの解析には限界がある。そこで、ゲノム情報に加えて臨床情報や疾患データベースおよび膨大な文献情報を統合し、人工知能技術を用いて疾患解明に繋がる知見を導き出す技術の確立を目指す。本研究では、新規原因遺伝子同定技術と診断画像による疾患分類手法の開発を目標とする。開発した技術は、汎用化を進めることで様々な疾患に適用できるように改良する。

３．研究の方法

本研究では、開発項目毎に以下の方法で研究を進めた。

・*HLA* 遺伝子の高精度配列解析技術と関連解析手法の確立

日本人の HLA 遺伝子完全長配列を網羅するリファレンス作成に向けて、研究代表者らは、既に Long PCR 法と次世代シークエンサー（NGS）を用いた主要 *HLA* 6 遺伝子（*HLA-A,-C,-B,-DRB1,-DQB1,-DPB1*）の高効率ターゲットシークエンス法、データベースの全 *HLA* アレルを網羅しかつ高精度にタイピングする手法(HLA-HD)、NGS ショートリードからの遺伝子全長アセンブリ技術を開発している[1-3]。既に、開発手法によって 7,000 検体程の主要 *HLA* 6 遺伝子が完了している。本研究では新たに 2000 検体のシークエンスを行い、既存の結果と併せて、*HLA* の主要 6 遺伝子のシークエンス結果をアセンブリし、*HLA* 遺伝子全長の配列を決定する。得られた *HLA* 遺伝子の配列情報を整理し、日本全国における HLA リファレンスとしてデータベースを構築する。構築後、国際的 HLA データベースである IPD-IMGT/HLA への登録を行う。完成したリファレンスを用いて、全ゲノムデータと HLA タイピング結果を統合し、高精細ゲノムデータを作成する。

得られたゲノム情報を用いて、全ゲノムおよび HLA 領域情報から、疾患と関連する変異を同定する解析手法の開発を行う。自己免疫疾患、ウィルス感染に関連する稀少難治性疾患といった遺伝因子が複雑な疾患についての高度ゲノム解析を可能にすることを目標とする。開発は、希少難治性疾患の HTLV-1 関連脊髄症(HAM/TSP)を対象として進める。

・人工知能技術を用いた統合難病データ解析手法の確立

人工知能技術を用いて、ゲノムや臨床情報の統合データを入力として、疾患に関わる複合的な遺伝要因を出力可能なデータ解析技術を確立する。

新規原因遺伝子同定法は網膜色素変性を対象に以下の様に開発を進めた。はじめに、患者と健常者における全ゲノム解析データに対して、網膜/視神経疾患の国際的データベースである RetNet に登録されている原因遺伝子リストとこれまでに明らかになった既知の原因遺伝子とその変異情報を基に原因変異の特徴を学習し、新規原因遺伝子の候補を抽出した。次に PubMed の文献情報を収集し、既知遺伝子群と強く関連するワードとその遺伝子間の関係性を学習する。構築されたネットワークにおける、既知遺伝子の関連度合いをもとに、遺伝子候補をランキングする。

診断画像による原因遺伝子の分類手法では、網膜色素変性の患者の広角眼底写真を人工知

能技術により、色素変性パターンをクラスタリングし、原因遺伝子間のパターンの違いから疾患を分類する手法の開発を行なった。初めに原因遺伝子が診断されている網膜色素変性患者と健常者の広角眼底写真から深層学習を用いて、色素変性パターンの特徴を学習し識別器を作成する。作成した識別器を用いて、患者の画像から色素変性領域の画像を抽出する。色素変性領域画像を原因遺伝子の情報を考慮しながらクラスタリングすることで、原因遺伝子間での色素変性パターンの特徴の違いを検出し、遺伝子毎に疾患を分類する。
構築した手法に関しては、専門医の協力を得ながらその有効性を検証した。開発後は様々な難病に適用できる様に、手法を汎用化しクラウド環境下で動作できるシステム上に搭載する。

４．研究成果

HLA 全長配列予測プログラムの予測精度向上のために、次世代シークエンサーを用いた実験方法の改良を行い、より包括的に *HLA* アレルを増幅しかつ均質なカバレージが得られるようになった。得られたシークエンス結果を用いて全長配列予測プログラムを開発し、約 5,000 検体の主要 *HLA* 6 遺伝子（*HLA-A,-C,-B,-DRB1,-DQB1,-DPB1*)の全長配列予測を行った。得られた配列を検証したところ、特に HLA Class II 遺伝子において多くの未知全長配列が検出された（論文投稿準備中)。得られた配列については今後とりまとめて、日本人における頻度情報と共にデータベースとして公開する予定である。さらに、これまでの *HLA* アレルタイピング技術を *KIR* 遺伝子に応用し、全ゲノム解析データから、KIR 遺伝子のコピー数、アレル、ハプロタイプ推定手法および融合遺伝子検出法を開発した。開発結果を 1000 Genomes Project の WGS データに適用し、過去の研究結果との比較から開発手法の有効性を確認した（論文投稿準備中)。
得られた *HLA* 情報とその他のゲノム情報や臨床情報を統合して関連解析するためのモデル構築を進めた。開発したモデルを 651 人の HAM/TSP 患者と 804 人の無症候性 HTLV-1 キャリアに適用したところ、HLA-DRB1 の抗原提示部位である G-BETA ドメイン 7 番目のアミノ酸残基位置のロイシンが HAM/TSP 発症と感受性を、プロリンが抵抗性を示した。また、得られたアミノ酸は既知の発症リスクであるプロウイルス量とは異なる因子として発症に関連する事を見出した。これらの結果を基に、HLA 情報とプロウイルス量を組み合わせた発症リスク予測モデルを開発した[4]。同様のモデルを用いて、835 人の IgG4 関連疾患患者と 1,789 人の対照群の HLA アレルを解析し、HLA-DRB1 の G-BETA ドメイン 7 番目のアミノ酸残基位置のバリンを発症と関連するものとして同定した[5]。
次に、HTLV-1 プロウイルス全長配列のシークエンス結果から、HAM/TSP 発症との関連が知られている *Tax* 及び *HBZ* 遺伝子配列およびアミノ酸配列決定法を開発した。開発手法を HAM/TSP 患者(406 名)および HTLV-1 キャリア(1282 名)のシークエンス結果に適用し、配列の比較解析から HAM/TSP 発症リスクを有意に上昇させるアミノ酸残基を同定した。この情報を、同定した HLA-DRB1 遺伝子上のアミノ酸残基と統合して、より詳細な発症リスク予測モデルを開発した（論文投稿準備中)。
希少難治性疾患における新規原因遺伝子同定手法構築のため、網膜色素変性患者 522 例と健常者 2,142 例の全ゲノム解析結果と RetNet データベースの既知遺伝子情報および過去の研究で得られた既知の変異情報を用いてそれぞれの群の変異パターンを機械学習し、患者特異的な変異を持つ 994 遺伝子を新規原因遺伝子の候補遺伝子として選択した。次に、PubMed の膨大な文献を構文解析し、既知遺伝子と候補遺伝子の関係性のネットワークを

構築した。このネットワーク上の遺伝子間の距離関係から、候補遺伝子をランキングした。一部の既知遺伝子を候補遺伝子としてテストデータを作り開発手法を検証したところ、高い精度で既知遺伝子を検出できた(AUC=0.865)。

原因遺伝子を基準とした難病の分類手法開発では、はじめに深層学習を用いて、網膜色素変性患者 405 枚、健常者 104 枚の広角眼底画像から、色素変性パターンを学習し、患者の画像の中から色素変性箇所の抽出を行った。抽出した、色素変性画像を原因遺伝子の情報を用いて、クラスタリングした。クラスタリングされた画像の分布を解析した結果、原因遺伝子の違いや疾患の進行に伴い、色素変性パターンの違いが存在することが判明した。 今後、この結果を用いて、原因遺伝子を基にした疾患の再分類を進める。

以上の様に、本研究で開発された手法は汎用的で様々な疾患データへの有効性が示されため、今後様々な難病の解明研究への貢献が期待できる。

[1] Kawaguchi, S., Higasa, K., Shimizu, M., Yamada, R. and Matsuda, F., HLA-HD: An accurate HLA typing algorithm for next-generation sequencing data. *Hum. Mutat.*, 2017, **38**, 788-797.

[2] Kawaguchi, S., Higasa, K., Yamada, R. and Matsuda, F., Comprehensive HLA Typing from a Current Allele Database Using Next-Generation Sequencing Data. *Methods Mol Biol., HLA Typing*., 2018, 225-233.

[3] Kawaguchi, S. and Matsuda, F., High-Definition Genomic Analysis of HLA Genes Via Comprehensive HLA Allele Genotyping. *Methods Mol. Biol.* 2020, **2131**, 31-38.

[4] Penova M, Kawaguchi S, Yasunaga J, Kawaguchi T, Sato T, Takahashi M, Shimizu M, Saito M, Tsukasaki K, Nakagawa M, Takenouchi N, Hara H, Matsuura E, Nozuma S, Takashima H, Izumo S, Watanabe T, Uchimaru K, Iwanaga M, Utsunomiya A, Tabara Y, Paul R, Yamano Y, Matsuoka M, Matsuda F, Genome wide association study of HTLV-1-associated myelopathy/tropical spastic paraparesis in the Japanese population, *Proc Natl Acad Sci USA*. 2021, **118**, e2004199118.

[5] Terao, C., Ota, M., Iwasaki, T., Shiokawa, M., Kawaguchi, S., Kuriyama, K., Kawaguchi, T., Kodama, Y., Yamaguchi, I., Uchida, K., Higasa, K., Yamamoto, M., Kubota, K., Yazumi, S., Hirano, K., Masaki, Y., Maguchi, H., Origuchi, T., Matsui, S., Nakazawa, T., Shiomi, H., Kamisawa, T., Hasebe, O., Iwasaki, E., Inui, K., Tanaka, Y., Ohshima, K., Akamizu, T., Nakamura, S., Nakamura, S., Saeki, T., Umehara, H., Shimosegawa. T,, Mizuno, N., Kawano, M., Azumi, A., Takahashi, H., Mimori, T., Kamatani, Y., Okazaki, K., Chiba, T., Kawa, S. and Matsuda, F. on behalf of the Japanease IgG4-Related Disease Working Consortium., IgG4-related disease in the Japanese population: a genome-wide association study. *Lancet Rheumatol.*, 2019, **1**, e14-e22.

５．主な発表論文等

〔雑誌論文〕　計3件（うち査読付論文　2件／うち国際共著　1件／うちオープンアクセス　1件）

１．著者名 Penova Marina et al.	４．巻 118
２．論文標題 Genome wide association study of HTLV-1 associated myelopathy/tropical spastic paraparesis in the Japanese population	５．発行年 2021年
３．雑誌名 Proceedings of the National Academy of Sciences	６．最初と最後の頁 e2004199118
掲載論文のDOI（デジタルオブジェクト識別子） 10.1073/pnas.2004199118	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 該当する

１．著者名 Kawaguchi, S. and Matsuda, F.	４．巻 2131
２．論文標題 High-Definition Genomic Analysis of HLA Genes Via Comprehensive HLA Allele Genotyping	５．発行年 2020年
３．雑誌名 Methods Mol Biol.	６．最初と最後の頁 31～38
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/978-1-0716-0389-5_3	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Terao, C. et al.	４．巻 1
２．論文標題 IgG4-related disease in the Japanese population: a genome-wide association study	５．発行年 2019年
３．雑誌名 The Lancet Rheumatology	６．最初と最後の頁 e14～e22
掲載論文のDOI（デジタルオブジェクト識別子） 10.1016/S2665-9913(19)30006-2	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計4件（うち招待講演　0件／うち国際学会　0件）

１．発表者名 岩崎 毅，川口 修治，稲富 雄一，清水 正和，松田 文彦
２．発表標題 NGSショートリードによるHLA遺伝子全長配列の決定
３．学会等名 日本人類遺伝学会第66回大会 第28回日本遺伝子診療学会大会
４．発表年 2021年

１．発表者名
Shuji Kawaguchi, Hiroki Nakano, Shogo Numa, Yuichi Inadomi, Akio Oishi, Akitaka Tsujikawa, Atsushi Takano, Fumihiko Matsuda

２．発表標題
Classification of causative genes for retinitis pigmentosa by an image clustering based on deep learning

３．学会等名
日本人類遺伝学会第66回大会 第28回日本遺伝子診療学会大会

４．発表年
2021年

１．発表者名
北田 せり，川口 修治，清水 正和，安永 純一朗，佐藤 知雄，田 耕平，原田 瑛介，高橋 めい子，山野 嘉久，松岡 雅雄，松田 文彦

２．発表標題
HTLV-1 プロウイルスゲノム変異の大規模解析によるHAM/TSP発症リスク予測モデルの構築

３．学会等名
第7回日本HTLV-1学会学術集会

４．発表年
2021年

１．発表者名
川口 修治，清水 正和，安永 純一朗，高橋 めい子，岡山 昭彦，山野 嘉久，内丸 薫，JSPFAD，川上 純，松岡 雅雄，松田 文彦

２．発表標題
大規模検体における HLA・HTLV-1 プロウイルス量の統合解析による HAM/TSP 発症リスクの推定

３．学会等名
第6回日本HTLV-1学会学術集会

４．発表年
2019年

〔図書〕　計0件

〔出願〕　計1件

産業財産権の名称	発明者	権利者
HAM／TSP発症リスク判定方法	松田文彦、川口修治	同左
産業財産権の種類、番号	出願年	国内・外国の別
特許、特願2019-205747	2019年	国内

〔取得〕　計0件

〔その他〕

HLA-HD ホームページ
https://www.genome.med.kyoto-u.ac.jp/HLA-HD/

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