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　５ 年　 ５ 月 ３０ 日現在

機関番号：１２６０１
研究種目：基盤研究(C)（一般）
研究期間：2019～2022
課題番号：１９Ｋ０６６２９

研究課題名（和文）複雑な形質の遺伝的背景の解明に向けた多因子的なゲノム情報処理技術の開発

研究課題名（英文）Development of multifactorial genome information processing technology for elucidating genetic the background of complex traits

研究代表者
中谷　明弘（Nakaya, Akihiro）

東京大学・大学院新領域創成科学研究科・特任教授

研究者番号：６０３０１１４９

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）：ゲノム情報の構造化のための変異データからの連鎖不平衡ブロックの抽出手法の開発を行った。検体群のゲノムDNA配列をスキャンしながら複雑に入り組んだ連鎖不平衡ブロックの抽出と可視化を行うソフトウェアとして実装した。ショートリード全ゲノムシーケンシング（WGS）情報解析パイプラインの整備も進めた。また、電子顕微鏡画像データからミトコンドリアや筋原繊維などの組織の空間的な形状や分布に関連する形質を定量化するソフトウェアの開発を行った。画像内の組織の分類情報に基づいた擬似カラー化処理の開発や分類情報を付加した画像内の矩形領域（ROI）のデータベースの整備を進めた。

研究成果の学術的意義や社会的意義
ゲノム配列データからの複雑な連鎖不平衡ブロックの抽出や電子顕微鏡画像データからの形質の定量を行うソフトウェアは十分には整備されていないため、独自に開発することによって解析対象の実データに合わせた開発や解析を手法自体の検討から行うことが可能となる。また、ソフトウェアとして実装することによってデータ解析に向けた手順や知見を汎用に共有することも可能となる。

研究成果の概要（英文）：We developed a method for extracting linkage disequilibrium blocks from mutation data for structuring genomic information. The method was implemented as software for extraction and visualization of intricate linkage disequilibrium blocks while scanning genomic DNA sequences of a sample set. We also developed a short-read whole genome sequencing (WGS) analysis pipeline. Software was developed to quantify traits related to spatial shape and distribution of tissues such as mitochondria and myofibrils from electron microscopy image data. We developed a pseudo-colorization method based on the classification information of tissues in images, and also developed a database of rectangular regions (ROIs) in images to which the classification information was added.

研究分野：ゲノム情報学

キーワード：ゲノム配列　配列変異　連鎖不平衡　解析パイプライン　ミトコンドリア　電子顕微鏡

１．研究開始当初の背景

ヒトをはじめとする血縁関係にない遺伝的背景が多様な検体群では、ゲノムワイドな DNA 配列の変異データの中に複数系統の連鎖不平衡領域が相互に分断されて入り組んでいる状況が見いだされる。ゲノムワイド関連解析（GWAS）では、染色体に沿って形質に対する有意さを表示するマンハッタンプロットが使用されるが、変異位置間の連鎖不平衡の程度をピラミッド状に描画した情報が併記されることはあっても、上記のような入り組んだ連鎖不平衡領域は十分には考慮されていない。そのため、変異パタンが類似した変異が冗長に使用されて、多重検定を加味した統計的な有意さの閾値の設定が必要以上に厳格になることも生じている。また、変異情報を入力とした数理モデルを用いて形質値を予測する際の説明変数の選択においても変異間の連鎖不平衡は説明変数間の相関（多重共線性）を制御する指標となるため、変異パタンに基づいて変異群をブロック化した高次元な情報を効率良く取得できる手法が必要となっている。

一方、ゲノム情報を用いた関連解析の対象となる形質情報に関しても高次元な情報が求められている。例えば、独自の DNA 配列をもつミトコンドリアは、外壁となる二重膜の内膜が陥入して形成するクリステと呼ばれる襞状の構造を有する細胞小器官の一つであるが、さまざまな生体機能や疾患（ミトコンドリア病や心筋症などの循環器疾患から認知症およびメタボリック症候群など多数）との関連が知られており、分裂や融合の程度および内部構造の形状は疾患に関連した重要な指標となっている。電子顕微鏡画像内でのミトコンドリアの形状が標準的か否かの目視による評価に止まらずに、画像全域にわたっての組織の判別や一粒ごとを切り出しての形状や相互の位置関係および空間分布などの高次元な形質の定量的な評価を行える手法が必要となっている。

２．研究の目的

複雑な形質の遺伝的背景の解明に向けた多因子的なゲノム情報処理技術を開発する。次世代シーケンシングで得られるゲノム DNA 配列の変異間の連鎖不平衡を抽出して構造化する手法の開発、および、全ゲノムシーケンシングの情報処理パイプラインの整備を進める。また、電子顕微鏡画像等の画像データとして得られる空間的な形状や分布に関する形質を抽出して定量化する手法を開発する。ヒト検体のミトコンドリア電子顕微鏡画像に適用した結果を蓄積したデータベースを構築し、画像データもクエリーとして指定できるビジュアルな検索を実現する手法の開発を進める。

３．研究の方法

(1) 多因子的なゲノム情報の構造化

ゲノム情報の構造化のための変異データからの連鎖不平衡（LD）ブロックの抽出手法の開発を行う。検体群のゲノム DNA 配列をスキャンしながら複雑に入り組んだ連鎖不平衡ブロックを抽出する手法の開発を行う。変異位置とそれらの間の連鎖不平衡の程度をそれぞれノードとエッジとしたネットワークを構成してグラフ理論に基づくアルゴリズムを開発し、VCF ファイルやタブ区切り形式データを入力とするソフトウェアとして実装する。

(2) 多次元的な形質情報の定量化

形質情報に関連した処理技術として、ミトコンドリアの形状の電子顕微鏡データの画像処理による定量化技術を開発する。ミトコンドリア電子顕微鏡画像からの形状特徴量の定量手法の開発を行う。グレースケール画像として得られる電子顕微鏡画像の全域をスキャンしてピクセル値に基づいて組織の種別の分類を支援するソフトウェアを開発する。組織の分類情報を付与した矩形領域（ROI）を学習用データセットとして集積したデータベースの整備を進める。

４．研究成果

(1) ゲノム情報の構造化のための連鎖不平衡（LD）ブロックの抽出ソフトウェアの開発

VCF形式ファイル内の変異を染色体に沿ってスキャンしながら連鎖不平衡（LD）ブロックを抽出するアルゴリズムを開発して LD ブロック抽出ソフトウェアとして実装した（特開 2020-202810）。２つの変異間の連鎖不平衡の評価はカイ二乗値に基づいた指標（引用文献①）などに基づいて行うことができる。ここでは、この指標が閾値以上の変異が連続する領域を LD ブロックと呼ぶことにする。実際のデータでは、図１に示すように DNA 配列に沿って集団内の特定の検体群に変異（赤●）が連続する領域（LD ブロック）の途中に別の変異（青●）がギャップとして挿入された状態が生じ得る。更に複雑な状況では、挿入される変異が別の LD ブロックとなって、複数の LD ブロックが入れ子な構造や相互に入り組んだ構

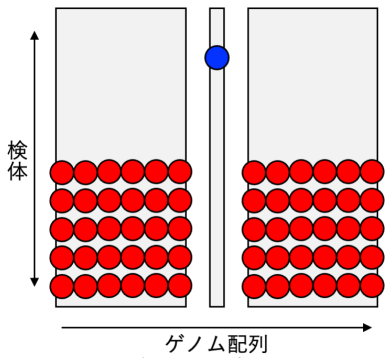

図１　LD ブロック内のギャップ

造となり得る。これらの構造は、単純なウィンドウ探索や特定の変異位置からの伸長探索では検出することが難しい。そこで、連鎖不平衡の断続状況の履歴を保持しながら DNA 配列に沿って変異をスキャンして、挿入されたギャップも考慮した LD ブロックを算出する手法を開発した。LD ブロックを構成する変異間の連鎖不平衡の程度の指標の閾値（下限値）や許容するギャップのサイズ（挿入される LD ブロック数や DNA 配列での物理距離）などの探索パラメータを指定できるようになっている。変異位置とそれらの間の連鎖不平衡の程度をそれぞれノードと重み付きエッジとしたネットワーク（LD ネットワーク）として構成できるようになっている。

図 2 は、1000 ゲノムプロジェクトによる日本人検体（N=10）の DNA 配列の変異データ（WES, 全エクソンシーケンシング）から LD ブロック抽出ソフトウエアを用いて抽出した LD ブロックの一部である。図 2 中央部では、DNA 配列中の位置（横軸方向）に沿って各検体（縦軸方向）の参照配列に対する変異情報（〇印および変異アリルを示す文字）を表示している。ここでは、二倍体を想定して〇印中の上下に 2 つのアリルを表示している。変異型アリルと参照型アリルのヘテロの場合は下半分を灰色で表示している。異なる 2 つの変異型アリルのヘテロの場合は灰色以外の 2 色で表示される（この例には含まれていない）。挿入と欠損はそれぞれ“i”と“d”で表示している。集団内の全ての検体が変異（この例では変異アリルを 1 つ以上）をもっている場合は検体群内では変異パタンが区別できない（連鎖不平衡が評価できない）ために読み飛ばしている（薄色で表示）。変異の有無が逆パタンとなって相関が高くなっている変異も考慮できるようになっている。

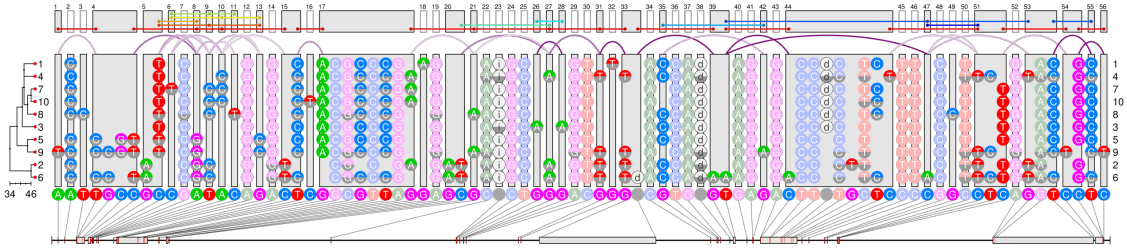

図 2　相互に入り組んだ連鎖不平衡（LD）ブロックの抽出と可視化

図 2 中央部では、灰色矩形はギャップ無しで連続している LD ブロックを表示している。矩形間を接続する曲線（紫色）でギャップによって分断されて不連続となっている LD ブロックを表示している。曲線の濃淡はギャップの前後での連鎖不平衡の程度の指標の高低を示している（濃色ほど連鎖不平衡の程度が高い）。図 2 上部には LD ブロックと LD ブロック間の接続の情報を表示している。前述の変異位置間の LD ネットワークで経路探索した結果に基づいて抽出した LD ブロック群がどのように入り組んでいるかを表示している。複数系統の LD ブロックが断続的な構造となって相互に入り組んでいることが分かる。図 2 左部には変異パタンの類似関係に基づいた検体のクラスタリング結果（デンドログラム）を表示している。図 2 下部は DNA 配列上でのエクソン領域の範囲（灰色矩形）と変異の位置（赤色線分）を表示している。

LD ブロック抽出ソフトウェアは C++言語で記述されている（インタプリタ言語での実装と比較して高速に実行可能）。520 検体の約 78 万塩基位置での変異を含む全エクソンシーケンシング（WES）データに適用した例では、染色体ごとの変異数と処理時間（ファイル入出力や描画処理も含む）の関係は図 3 に示す通りとなった。処理時間の評価には AMD Ryzen Threadripper 2990WX（32CPU コア/128GB メモリ）で動作する Windows 10 の Windows Subsystem for Linux（Ubuntu 18.04）上の GNU g++ compiler（version 7.4.0）を用いた。絶対的な処理時間はハードウェアの性能に依存するが、変異数と処理時間は概ね比例しているので、開発した LD ブロック抽出アルゴリズムおよびその実装の時間計算量は変異数に関して線形時間であることが示唆され、変異数が増加した場合でも計算時間の非実用的な増大に陥らないことを示している。また、染色体ごとの処理は互いに独立に実行できるので複数の異なった CPU コアに割り当てることによって並行処理による高速化も容易に実現できる。

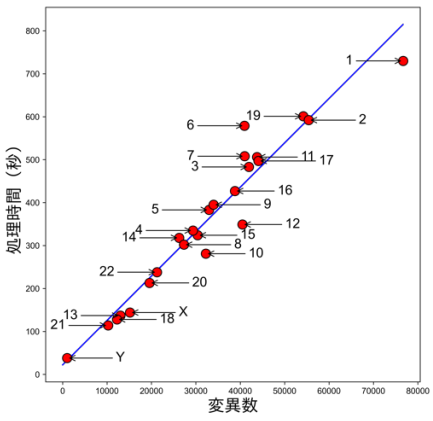

図 3　変異数と処理時間の関係

(2) ミトコンドリア電子顕微鏡画像からの形状特徴量の定量手法の開発

グレースケール画像として得られる電子顕微鏡画像の全域をスキャンして濃淡情報（ピクセル値）に基づいて画像内で組織の種別の分類を支援するソフトウェアを作成した。画像内のピクセル値（8 ビットグレースケールの場合は黒から白へのグラデーションに 0 から 255 のピクセル値が割り当てられる）と組織の種類との間に一定の相関があることを前提として、異なった組織が互いに強調されるようにグレースケール画像のピクセル値の分布（1 チャンネル）に色相情報

（RGB の 3 チャンネル）を割り当てること（カラーマッピング）によってモノクロの電子顕微鏡画像を擬似カラー化するソフトウェアを開発した。一般的に、カラー画像での赤から青までのグラデーションは、RGB 値を赤(255,0,0)→黄(255,255,0)→緑(0,255,0)→シアン(0,255,255)→青(0,0,255)と連続的に変化させることによって得られる（括弧内は R, G, B の 3 チャンネルの値）。これらの色の間をグレースケール画像のピクセル値（0 から 225）に対して均等の割り付けるのではなく、特定の組織が異なった色で表示されるようにカラー画像でのグラデーションを割り当てる。グレースケール画像でより多くの組織が含まれているピクセル値の範囲にはカラー画像でより多くの色を割り当てることになる。

例えば、図 4 は共同研究者らによる心筋の電子顕微鏡画像（引用文献②）を擬似カラー化した結果である。電子顕微鏡画像の原画像は図 4 上のようなグレースケール画像として得られる。図 4 中はグレースケール画像の全域にわたってピクセル値（0 から 255）の頻度を集計したヒストグラムである。この画像中ではピクセル値の範囲は概ね 90 から 240 の範囲となっていることが分かる。このピクセル値の範囲に図 4 中のヒストグラムに示す青から赤のグラデーションの割り当てを行い（カラーマッピング）、その割り当てを原画像に反映させると、図 4 下に示すように疑似カラー化した結果が得られる。ここで、上記のグラデーションの割り当ては独自に開発した GUI 付きのソフトウェアを用いて行った。グレースケール画像のピクセル値の範囲（0 から 255）上に配置された複数のカラー分岐点の調整を擬似カラー化画像への反映結果をリアルタイムで確認しながら行える機能が実装されている。

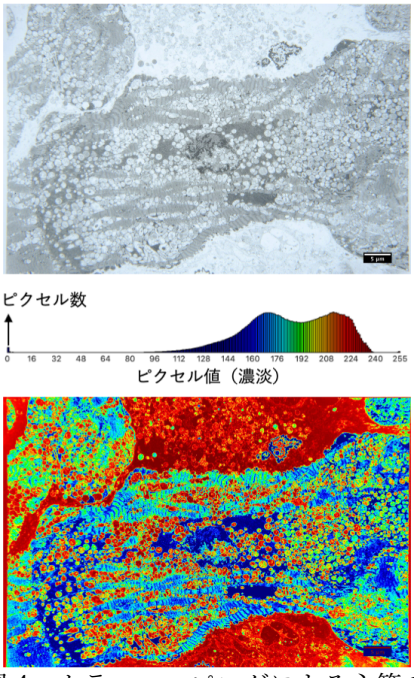

図 4　カラーマッピングによる心筋の電子顕微鏡画像の擬似カラー化

得られた疑似カラー化画像を参照しながら、ミトコンドリアと筋原繊維および周辺組織を対象として、クラスラベル（組織の分類情報）を付与した矩形領域（ROI）を多数切り抜いた学習用データセットの整備を進めた。色情報を援用して目視によってクラスラベルを指定して手動で切り抜いた矩形、あるいは、色情報をクラスラベルとして半自動的に切り抜いた矩形を画像内での座標情報と共に蓄積する仕組みを整備した。割り当てた色情報を用いて特定の組織を対象とした二値画像（マスク情報）を生成する際の閾値の決定も可能となっている。例えば、典型的なミトコンドリア粒子は赤色の範囲であるため、そのピクセル値が分布する領域に限定して輪郭抽出を行うことによって、画像内でのミトコンドリア粒子を抽出することが可能となる。また、筋原繊維は青色とシアン色の縞状となっていることに注目して抽出することが可能である。得られた組織の情報を初期解として、マスク情報の生成と形状の検出を繰り返すことによって矩形領域の選択とクラスラベルの改善を行い、畳み込みニューラルネットをはじめとする後続の解析の初期解として使用できるデータセットを集積したデータベースを整備した。

(3) ショートリード全ゲノムシーケンシング（WGS）情報解析パイプラインの整備

ショートリード全ゲノムシーケンシング（WGS）情報解析パイプラインの整備を進めた。Broad Institute によるワークフロー（“the GATK Best Practices”）に準拠した生殖細胞系列の一塩基変異（SNV）と短い挿入/欠失（Indel）の検出を行う情報解析パイプラインの整備を進めた。BWA と GATK による参照配列へのマッピングと変異検出を行い、シーケンサ出力の FASTQ 形式データから VCF 形式データを生成する。得られた塩基の品質スコアの再較正（BQSR）や検体群内の複数の検体を対象としたジョイントコールを行うように構成されている。ジョブ管理システム（AGE, Altair Grid Engine や Slurm Workload Manager）が使用できない環境（例えば Apple 社製 macOS）でも CPU コア数やメモリ量に合わせて効率良く処理を行うことができるようになっている。VCF 形式データとして得られる DNA 配列の変異情報は、項目(1)の連鎖不平衡（LD）ブロック抽出ソフトウェアを用いて処理することによって検体群内での変異パタンの類似関係を辿れるデータ構造として出力できるようになっている。

引用文献

① Yamazaki T. The effects of overdominance of linkage in a multilocus system. Genetics. 1977 May;86(1):227-36. doi: 10.1093/genetics/86.1.227. PMID: 885342; PMCID: PMC1213667.
② Takeda A, Murayama K, Okazaki Y, Imai-Okazaki A, Ohtake A, Takakuwa E, Yamazawa H, Izumi G, Abe J, Nagai A, Taniguchi K, Sasaki D, Tsujioka T, Basgen JM. Advanced pathological study for definite diagnosis of mitochondrial cardiomyopathy. J Clin Pathol. 2020 Aug 17:jclinpath-2020-206801. doi: 10.1136/jclinpath-2020-206801. Epub ahead of print. PMID: 32817174.

５．主な発表論文等

〔雑誌論文〕　計6件（うち査読付論文　6件／うち国際共著　1件／うちオープンアクセス　6件）

１．著者名 Imai-Okazaki A, Matsunaga A, Yatsuka Y, Nitta K R, Kishita Y, Sugiura A, Sugiyama Y, Fushimi T, Shimura M, Ichimoto K, Tajika M, Ogawa-Tominaga M, Ebihara T, Matsuhashi T, Tsuruoka T, Kohda M, Hirata T, Harashima H, Nojiri S, Takeda A, Nakaya A, Kogaki S, Sakata Y, Ohtake A, Murayama K, Okazaki Y	４．巻 341
２．論文標題 Long-term prognosis and genetic background of cardiomyopathy in 223 pediatric mitochondrial disease patients	５．発行年 2021年
３．雑誌名 International Journal of Cardiology	６．最初と最後の頁 48～55
掲載論文のDOI（デジタルオブジェクト識別子） 10.1016/j.ijcard.2021.06.042	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 －

１．著者名 Kikuchi M, Kobayashi K, Nishida N, Sawai H, Sugiyama M, Mizokami M, Tokunaga K, Nakaya A	４．巻 8
２．論文標題 Genome-wide copy number variation analysis of hepatitis B infection in a Japanese population	５．発行年 2021年
３．雑誌名 Human Genome Variation	６．最初と最後の頁 -
掲載論文のDOI（デジタルオブジェクト識別子） 10.1038/s41439-021-00154-w	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 －

１．著者名 Kaimori J, Mori T, Namba-Hamano T, Morimoto T, Takuwa A, Motooka D, Okazaki A, Kobayashi K, Asahina Y, Kajimoto S, Doi Y, Oka T, Sakaguchi Y, Nakaya A, Isaka Y	４．巻 145
２．論文標題 Cyclosporine A treatment of proteinuria in a new case of MAFB-associated glomerulopathy without extrarenal involvement: A case report	５．発行年 2021年
３．雑誌名 Nephron	６．最初と最後の頁 445～450
掲載論文のDOI（デジタルオブジェクト識別子） 10.1159/000516248	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 －

１．著者名 Ohnishi T, Kiyama Y, Arima-Yoshida F, Kadota M, Ichikawa T, Yamada K, Watanabe A, Ohba H, Tanaka K, Nakaya A, Horiuchi Y, Iwayama Y, Toyoshima M, Ogawa I, Shimamoto-Mitsuyama C, Maekawa M, ... , Kurokawa R, Suzuki K, Yoshikawa A, Toyota T, Hosoya T, Okuno H, Bito H, Itokawa M, Kuraku S, Manabe T, Yoshikawa T	４．巻 13
２．論文標題 Cooperation of LIM domain-binding 2 (LDB2) with EGR in the pathogenesis of schizophrenia	５．発行年 2021年
３．雑誌名 EMBO Molecular Medicine	６．最初と最後の頁 e12574
掲載論文のDOI（デジタルオブジェクト識別子） 10.15252/emmm.202012574	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 －

1．著者名 Shimamoto-Mitsuyama C, Nakaya A, Esaki K, Balan S, Iwayama Y, Ohnishi T, Maekawa M, Toyota T, Dean B, Yoshikawa T	4．巻 31
2．論文標題 Lipid pathology of the corpus callosum in schizophrenia and the potential role of abnormal gene regulatory networks with reduced microglial marker expression	5．発行年 2020年
3．雑誌名 Cerebral Cortex	6．最初と最後の頁 448～462
掲載論文のDOI（デジタルオブジェクト識別子） 10.1093/cercor/bhaa236	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 Shinohara M, Kikuchi M, Onishi-Takeya M, Tashiro Y, Suzuki K, Noda Y, Takeda S, Mukouzono M, Nagano S, Fukumori A, Morishita R, Nakaya A, Sato N	4．巻 3
2．論文標題 Upregulated expression of a subset of genes in APP; ob/ob mice: Evidence of an interaction between diabetes-linked obesity and Alzheimer's disease	5．発行年 2021年
3．雑誌名 FASEB BioAdvances	6．最初と最後の頁 323～333
掲載論文のDOI（デジタルオブジェクト識別子） 10.1096/fba.2020-00151	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1件（うち招待講演　1件／うち国際学会　0件）

1．発表者名 中谷明弘
2．発表標題 スギ花粉症舌下免疫療法の効果の評価に向けた網羅的計測データの多因子的解析
3．学会等名 第4回日本アレルギー学会関東地方会（招待講演）
4．発表年 2020年

〔図書〕　計1件

1．著者名 DSTEP教材作成委員会	4．発行年 2021年
2．出版社 羊土社	5．総ページ数 344
3．書名 東大式　生命データサイエンス即戦力講座（第1章-第4章）	

〔出願〕　計1件

産業財産権の名称 集積集合データの製造装置、製造方法、プログラム、および記録媒体	発明者 中谷明弘，小林香織，竹之内隆夫，上條憲一	権利者 日本電気株式会社，国立大学法人大阪大学
産業財産権の種類、番号 特許、特願2019-113456，特開2020-202810	出願年 2019年	国内・外国の別 国内

〔取得〕　計0件

〔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