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　３ 年　 ６ 月　 ２ 日現在

機関番号：１１３０１
研究種目：基盤研究(C)（一般）
研究期間：2018～2020
課題番号：１８Ｋ０６１９４

研究課題名（和文）蛋白質立体構造を用いたゲノム塩基配列変化の機能解析のための情報科学的基盤の構築

研究課題名（英文）Development of information system for functional annotation of genomic sequence alterations by using protein 3D structures

研究代表者
城田　松之（Shirota, Matsuyuki）

東北大学・医学系研究科・講師

研究者番号：００５４９４６２

交付決定額（研究期間全体）：（直接経費）　　3,400,000円

研究成果の概要（和文）：本研究ではProtein Data Bankのタンパク質立体構造を用いて、タンパク質の疎水性相互作用や立体障害、タンパク質間相互作用、ジスルフィド結合や水素結合、DNAやRNA、イオン、低分子リガンドとの相互作用などの多様な評価指標を用いて、ヒトゲノムの多数の非同義バリアントの中からタンパク質機能に影響を与えるものをゲノムワイドに評価するための情報科学的基盤を構築した。これを用いて既知のバリアントについて評価を行い、またユーザが持つバリアントについてアノテーションを行うウェブシステムを構築した。

研究成果の学術的意義や社会的意義
ヒトが持つゲノム配列の違いは、個人の体質や病気へのかかりやすさの違いに関係していることが知られているが、どのような配列がどのような特徴に関係しているかはよくわかっていない。この研究ではゲノム配列がコードするタンパク質の立体構造に注目し、原子レベルの相互作用をよく解析することで、タンパク質の機能を大きく変えるようなゲノム配列変化を調べることを可能にした。この研究によってゲノム配列の違いによって病気になりやすくなる分子レベルのメカニズムの解明につながることが期待できる。

研究成果の概要（英文）：In this study, we developed an annotation system of nonsynonymous variants in the human genome to select those variants that can affect protein structure and function by evaluating various atomic interactions including hydrophobic interactions, protein-protein interactions, disulfide bonds, hydrogen bonds, and interactions with DNA, RNA, ions and other small ligands in the three-dimensional protein structures in Protein Data Bank. We applied this annotation system to known single nucleotide variants to facilitate interpretation of their functional roles and provided a web system that can annotate the variants of user's interest.

研究分野：バイオインフォマティクス

キーワード：タンパク質立体構造　ゲノムバリアント　データベース　機能欠失変異

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

ゲノム配列決定技術の発展に伴い、さまざまな一般集団や疾患患者において大規模ゲノム解析が行われ、多数のヒトゲノム配列変化（バリアント）が報告されており、これらのバリアントの中から個人の形質や疾患感受性といった表現型に影響するものを発見することが大きな課題となっている。タンパク質のアミノ酸配列を変える非同義バリアントは疾患原因の探索などでは優先的に検討される対象であるが、これら非同義バリアントに限っても多数が存在するために、特に分子機能への影響が大きいものを絞り込む手法が必要とされている。Protein Data Bank (PDB)のタンパク質立体構造情報はアミノ酸変化がタンパク質の安定性や機能に与える影響を推定する上で重要な情報である。非同義バリアントの重篤度を予測するために、これまで SIFT や Polyphen2 のようにアミノ酸配列の進化的保存度と、利用可能な立体構造上の特徴を用いた機械学習手法が広く利用されてきた。しかし、これらの手法は個々のアミノ酸変化がどのようなタンパク質の機能変化をもたらすかについての情報を与えてはくれるわけではない。また、これらの機械学習手法はある時点の PDB のデータで構築されるためそれより新しい立体構造情報の恩恵を受けられないという欠点がある。一方、ゲノムの非同義バリアントを与えると対応する PDB のアミノ酸残基情報を返す手法は PDB のほか、MuPIT、PhyreRisk、VarMap などが提供している。しかしこれらの手法ではアミノ酸置換の影響を解釈するための構造的特徴が十分に提供されておらず、ゲノム科学や医学の研究者が立体構造情報を利用する際の障害となっている。タンパク質の機能欠失やそれを原因とする疾患発症などの表現型変化を引き起こしやすいアミノ酸変化として、タンパク質内部の疎水性残基から極性残基への置換や立体障害を起こす置換、タンパク質間相互作用への障害、ジスルフィド結合や側鎖の水素結合の欠失、DNA や RNA、イオン、低分子リガンドとの相互作用部位の障害、活性部位の置換などが報告されている。立体構造を用いてアミノ酸置換を評価する場合、個別の置換についてこれらの原子間相互作用を詳細に検討することになり、また、PDB には同じタンパク質でも様々な条件や異なる分子との結合状態の構造が複数解析されていることがあり、利用可能な全ての状態を考慮する必要がある。このような多面的な立体構造を用いた分子機能への影響評価は、配列解析による進化的保存度の計算と比べてゲノムワイドで画一的に行うことが難しいという問題があった。そこで、ヒトの全ての非同義バリアントを最新の PDB 構造のアミノ酸に適切に対応づけ、その構造的特徴をあらかじめ計算して非同義バリアントの解釈に役立てられるように提供することが重要であるという発想を得た。これにより、構造生物学の専門家でないゲノム解析や医学の研究者が自分の興味のあるバリアントのセットからタンパク質機能に影響しうるものを検索し解析できることを目指し、本研究を行うこととなった。

２．研究の目的

本研究の目的は膨大なゲノムバリアントの中から立体構造をもとにタンパク質の構造と機能を損なう可能性が高いものを抽出することを可能とすることである。このためにヒトゲノム上の全ての非同義バリアントについて、変化が起こるタンパク質のアミノ酸残基、もしくはそれに相同なタンパク質の該当する残基を PDB の立体構造において同定するとともに、それらの残基についてタンパク質機能への影響を評価するために必要な情報をあらかじめ計算してデータベース化し、多数の非同義バリアントについてこれらの情報を検索・一覧表示できるような形で提供するような情報科学的基盤を構築する。この基盤を既知のバリアントに対して応用することで、立体構造に基づく情報をより手軽に利用できるようにする。これにより、タンパク質機能への効果がありそうなバリアントに研究者が注目し、その構造・機能変化を検証することで、遺伝型と表現型の大きなギャップを埋める動きを加速することが期待される。同時に、疾患患者等のゲノム解析を行った研究者が、解析によって得られた非同義バリアントのリストから、構造生物学や個々のタンパク質に関する専門的知識がなくても、PDB の立体構造を用いた非同義変異の評価を行い、タンパク質の分子機能に影響しうるバリアント候補を抽出することができ、また具体的にどのように分子機能に影響しそうかを理解できるようなシステムを構築する。これにより遺伝型（ゲノムバリアント）と疾患などの表現型の情報を持った研究者がこのような情報に容易にアクセスできるようになれば、疾患などの表現型を説明するゲノムバリアントの解明に大きく貢献すると期待される。

３．研究の方法

本研究の方法について図１に示す。大きく「ゲノム上の一塩基置換をタンパク質立体構造上のアミノ酸置換に変換する（マッピング）」ことと「タンパク質構造上のアミノ酸残基の構造的特徴を計算し表示する（構造アノテーション）」ことの二つに分けられ、それぞれについて毎週自動

更新できるようなパイプラインを作成した（図１）。

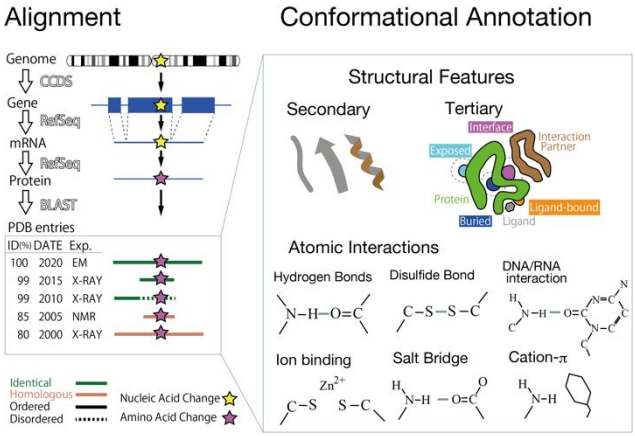

図 1　解析の流れ

マッピングパイプラインの構築

ゲノムバリアントは参照ゲノム上の染色体、塩基位置、塩基置換のパターンというセットの情報で与えられる。染色体と塩基位置の情報から Consensus CDS Project (CCDS)によるタンパク質コード領域（CDS）の情報を用いて、CDS に起こるバリアントを検出し、塩基置換のパターンからアミノ酸置換の有無を判定した。続いて、ヒトの全タンパク質配列と PDB の全タンパク質配列の間で配列相同性検索および配列アラインメントを行い、ゲノムバリアントの引き起こすアミノ酸変化を PDB の立体構造上の残基に対応づけた。同じタンパク質の複数の構造は異なる複合体状態や異なる薬剤との結合など、それぞれがタンパク質機能を評価する上で独自の情報を含んでいることがある。そのため、バリアントについて対応づけられるすべての PDB 立体構造の残基をリストアップして評価することにした。また、PDB には毎週新しいタンパク質構造が登録されるため、この作業を自動更新する機能を作成し、最新の情報にアクセスできるように実装した。

構造アノテーションパイプラインの構築

ゲノムの非同義バリアントについて「立体構造上重要な位置にある残基」に起こる変化を多数のバリアントの中から抽出することを可能にするために、PDB における全ての立体構造の全ての残基について、構造的特徴をあらかじめ計算し、データベース化した。構造的特徴としてはアミノ酸残基の二次構造、溶媒接触表面積（ASA）、複合体形成時の ASA の変化、水素結合、ジスルフィド結合、DNA や RNA との相互作用、低分子やイオンとの相互作用を考慮した。

これらの構造的特徴の情報をバリアントから PDB の残基への対応に基づいてバリアントへの注釈として行った。これにより多数のバリアントについて高速に構造情報をアノテーションできるだけでなく、「タンパク質内部に埋もれた残基」、「タンパク質相互作用面にある残基」、「低分子と結合している残基」などの絞り込みを可能としている。

構造サマリ情報の作成

バリアントごとに対応するアミノ酸残基の構造的特徴を表示する上で、該当する残基を含むタンパク質構造は複数ある場合がある。このような場合にそれらの特徴をまとめてサマリを表示することが必要である。まず、該当のアミノ酸残基のうち、座標が決定できない天然変性領域（disorder 領域）に存在するものの割合を計算した。その後、天然変性領域以外の構造をとる残基に絞って、ASA などの量的特徴では[中央値（最小値―最大値）構造数]のように表記し、二次構造のようなカテゴリ特徴ではαヘリックス、βシート、それ以外の割合を表記することでサマリを作成した。これにより、結果のバリアントリストを眺めることで、概ねどのような特徴をもつ残基に起こった変化なのかを俯瞰できるようにした。

日本人 8300 人の参照パネル等への適用

この手法の応用として東北メディカル・メガバンク機構で全ゲノム解析を行った日本人約 8300 人および Genome Aggregation Database (gnomAD)で全ゲノム解析を行ったヨーロッパ・アフリカ・東アジアなどの約 15000 人の全ゲノム解析の結果について構造アノテーションを行った。この情報は現在同機構のサイト jMorp (https://jmorp.megabank.tohoku.ac.jp/)にて公開されている。

Web サービスの構築

本研究ではこれらのバリアント解析を行えるウェブツールを作成した（https://wupsivs.sb.ecei.tohoku.ac.jp/）。これは、利用者がゲノムバリアントのリストをテキストボックスに入力・あるいはファイルとしてアップロードすることで、タンパク質構造に基づくアノテーションを取得できる。結果ページはバリアントごとに、対応する PDB のアミノ酸の構造的特徴のサマリを提供し、各バリアントページへのリンクを提供する。各バリアントのページはそれに対応する全ての PDB のアミノ酸残基の情報を表示し、また MolMil 構造ビューワによって残基の位置や構造を確認することができる。（二次構造、二面角、溶媒接触表面積、タンパク質間相互作用面にあるかどうかとその相互作用相手、空間的に近接するリガンドとその距離など）をあらかじめすべての PDB 構造について計算し、データベース化した。これらのアノテーションについても PDB に合わせて毎週の更新を行い、最新の構造情報を利用可能とする環境を整えた。これにより、既存の変異データにおいても構造上重要な変異を新しく発見することができると期待される。

４．研究成果

ヒトゲノムでのタンパク質構造情報の利用可能性について

本研究ではヒトの遺伝子がコードするタンパク質の全てのアミノ酸について PDB の構造にある残基との対応づけを行った。あるタンパク質やアミノ酸残基に複数の PDB 構造がヒットした場合、その構造の公開日を比較することで、初めて構造情報が利用可能となった日付を知ることができる。この情報をもとに過去に遡ってどのタンパク質やアミノ酸、バリアントについて立体構造が利用可能であったのか、あるいは新しく利用可能となったのかを網羅的に調査した。

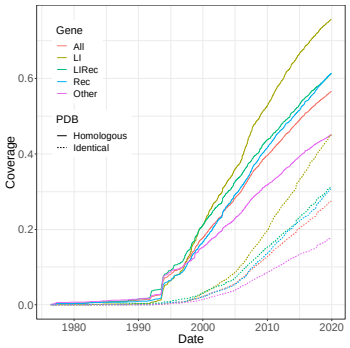

図 2　構造情報が利用可能な遺伝子の割合

図 2 に 2020 年までの PDB でタンパク質の構造情報を利用可能（50 残基以上のドメインがあるもの）な遺伝子の比率の推移を示す。それぞれの遺伝子に対して同一遺伝子のタンパク質構造のみを利用した場合（Identical、点線）と配列一致度 30%以上のホモログ（Homologous、実線）も含めた場合を表している。構造情報が利用可能な遺伝子の割合は全遺伝子で見るとホモログも含めて 49.8%（9678/19423）である一方、機能欠失が致死的な遺伝子（Loss-of-function Intorelant: LI）では 68.4%（2174/3177）に達していた。これは構造解析が機能的に重要なタンパク質を中心に行われていることを反映していると考えられる。疾患原因となるようなバリアントは機能欠失が致死的な遺伝子に多いと考えられるため、現時点でもタンパク質の構造情報は機能に影響するバリアントの評価において十分役立つことが考えられる。

次に、毎年どの程度のヒトタンパク質のアミノ酸残基について、初めて構造情報が利用可能となったかを、解析されたときの解析手法で分けて経時的に比較した（図 3）。この結果、新規のヒトタンパク質構造の決定は構造生物学の大規模プロジェクトが始まった 2000 年代に急増し、最近 5 年ほどは年間 10 万残基程度であった。特に電子顕微鏡（EM）の占める割合が近年急増しており、技術革新により EM のヒトのバリアントの解析に占める重要度が増加していることがわかった。

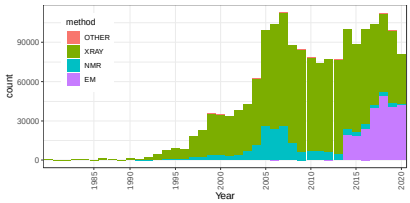

図 3 構造情報の利用可能な残基数の推移

次に、これらの毎年新規に構造解析された残基が疾患関連、また一般集団バリアントの機能推定にどの程度貢献するかを調べた（図 4）。ClinVar の疾患変異（Pathogenic）が数 100 程度、重要度不明な変異（VUS）が 1000 程度、良性（benign）が 100 程度である一方、gnomAD の集団変異のうち 1 アレルのみみられるシングルトン（singleton）やアレル頻度 1%以下のバリアント（lowfreq）、および COSMIC のがん関連変異が 10000 以上、gnomAD の高頻度（>1%、common）変異が 100 程度であった。この結果から新規のタンパク質立体構造がわかることで、実際に疾患原因の解明や表現型に影響するバリアントの解釈を加速することができることが示された。

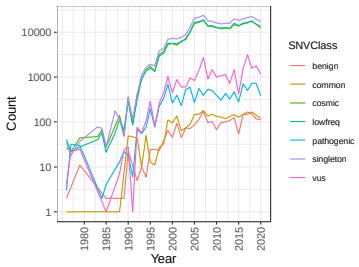

図 4 構造情報が利用できるようになったバリアント数の推移

既知のバリアントへの応用

タンパク質立体構造をもとにした情報の活用を促すため、既知の非同義バリアントについて構造情報を用いたアノテーションを行い、データベースとして公開した。既知のバリアントとしては東北メディカル・メガバンク機構で全ゲノム解析を行った日本人約 8300 人（8.3KJPN）および Genome Aggregation Database (gnomAD)で全ゲノム解析を行ったヨーロッパ・アフリカ・東アジアなどの約 15000 人の全ゲノム解析（gnomAD）の結果を用いている。この情報は現在同機構のサイト jMorp (https://jmorp.megabank.tohoku.ac.jp/)にて公開されている。同サイトで公開されているミスセンスバリアントの中から、PDB の立体構造情報を利用できるものについてはそれぞれのバリアントのページを用意し、変化する残基の立体構造上の位置と構造的特徴を簡単に見ることができる。図 5 にアルコールへの耐性に関与するアルコールデヒドロゲナーゼ ALDH2 遺伝子の K504E 変異について示す。PDB 5l13 の構造のチェイン A の 504 番目の残基をハイライトしており、この残基は二次構造がβシート（E）で ASA が 0.374 とサブユニットの表面にあり溶媒に露出しているが、

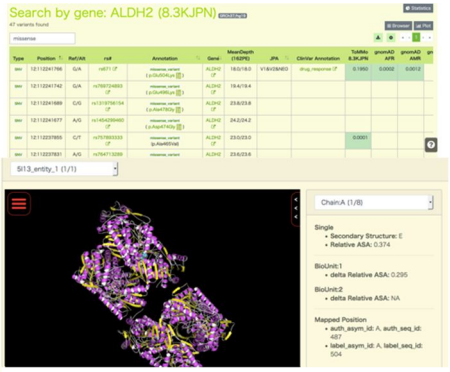

図 5 JMORP での構造の表示

複合体形成による ASA の変化が大きく（delta Relative ASA: 0.295）、複合体形成に関与していることがわかる。

構造アノテーション Web サーバの構築

ヒトのゲノムバリアントの機能解明を促進するため、バリアントのリストをアップロードすることで、非同義バリアントに対して、該当する PDB のアミノ酸とその構造特徴を出力する Web サーバを https://wupsivs.sb.ecei.tohoku.ac.jp/ で作成した。これは創薬等先端技術支援基盤プラットフォーム BINDS の課題「創薬等ライフサイエンス研究を促進する研究支援とデータサイエンス」と共同で開発したものである。

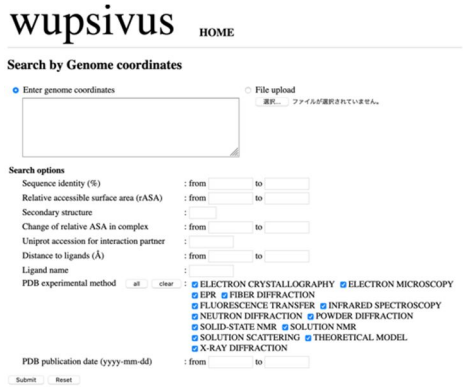

図 6 WUPSIVUS のトップページ

ユーザはゲノム座標を含むバリアントのリストをペーストするかファイルとしてアップロードすることで、その中の非同義バリアントについて PDB の立体構造情報を取得できる（図 6）。特に構造的に重要な残基に注目するために、検索オプションとして ASA の値や二次構造、複合体形成時の ASA の変化、相互作用相手の UniProt アクセッション、結合するリガンドの種類と相互作用の原子間距離などを設定することで、タンパク質内部に埋もれた残基や特定の二次構造をとる残基、タンパク質間相互作用面の残基、特定の相手と相互作用する部位にある残基、特定のリガンドと相互作用する残基などの絞り込みを行うことができる。また、実験手法（X 線結晶構造解析や電子顕微鏡など）およびデータ公開日で PDB の構造を絞ることもでき、特に新規に解析された構造に絞って結果を見ることができる。

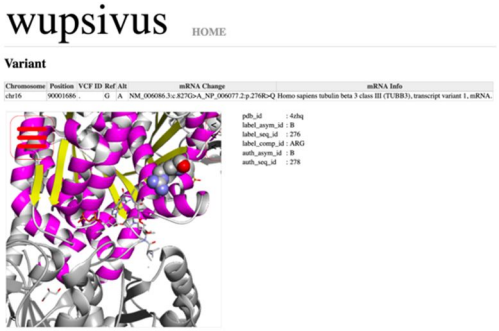

図 8 個々のバリアントページ

この検索結果は各バリアントについてヒットした PDB 構造の残基全ての構造特徴を要約した結果とともに表示される（図 7）。構造特徴として ASA や二次構造のほか、相互作用相手の UniProt アクセッションや天然変性残基の割合、ジスルフィド結合、水素結合、DNA/RNA との相互作用、リガンドとの相互作用の要約を表示している。これにより、とくに分子機能に影響が大きそうなバリアントを一覧の中から選択することができる。これらの各行にはバリアントページへのリンクがある。バリアントページでは該当のバリアントに対してヒットした PDB の残基が一覧表示され、構造ビューアを切り替えることで構造を見ながら残基の置換による相互作用を検討することができる（図 8）。

このような Web サイトを用いることで、膨大なバリアントの中から立体構造の特徴をもとに分子機能に影響が出る可能性の高いバリアントを選ぶことが可能となった。

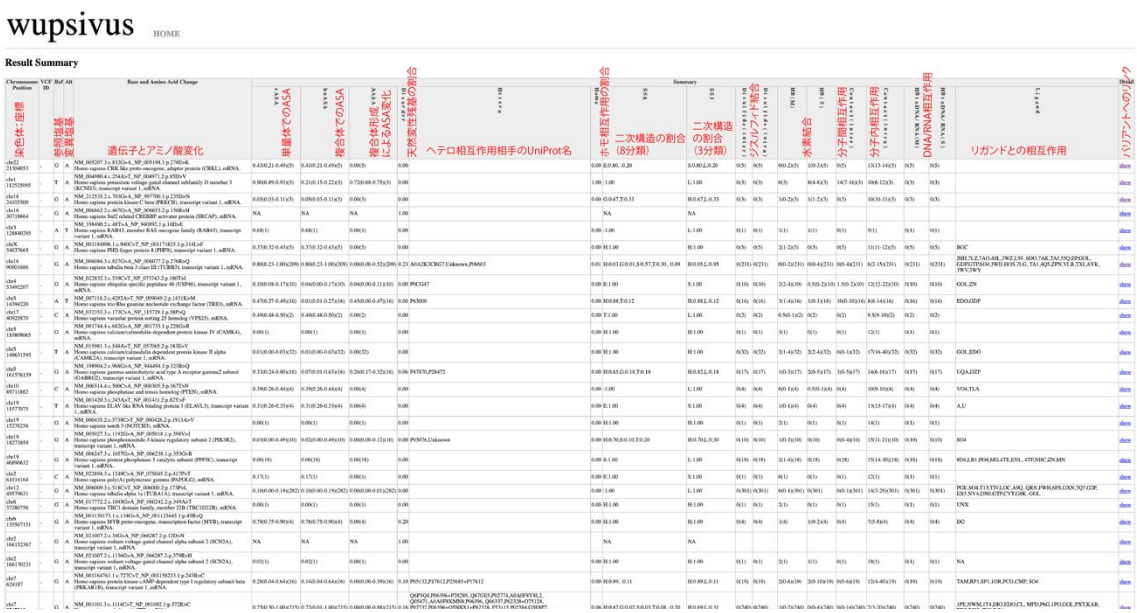

図 7 バリアント検索結果と構造サマリ

５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計6件（うち招待講演　0件／うち国際学会　2件）

１．発表者名 Matsuyuki Shirota
２．発表標題 Analysis on nonsynonymous variations with possible structural and functional impact on loss-of-function intolerant proteins
３．学会等名 ISMB 2019（国際学会）
４．発表年 2019年

１．発表者名 Matsuyuki Shirota
２．発表標題 Development of a genome annotation system based on protein three-dimensional structures
３．学会等名 ISMB2018（国際学会）
４．発表年 2018年

１．発表者名 城田松之
２．発表標題 蛋白質立体構造と機能に影響しうるゲノムバリアントの選択と解析
３．学会等名 日本生物物理学会第56回年会
４．発表年 2018年

１．発表者名 Matsuyuki Shirota
２．発表標題 Development of genome variant annotation system with protein 3D structures and applications to variants in Japanese population
３．学会等名 第 7 回生命医薬情報学連合大会
４．発表年 2018年

１．発表者名 城田松之
２．発表標題 蛋白質立体構造を用いたゲノムバリアントのアノテーション法の開発と日本人集団バリアントへの応用
３．学会等名 第41回日本分子生物学会年会
４．発表年 2018年

１．発表者名 Matsuyuki Shirota
２．発表標題 Comprehensive and Up-to-date Annotation of Nonsynonymous Single Nucleotide Variants of Human Genome with Protein Structure Information
３．学会等名 第20回日本蛋白質科学会
４．発表年 2020年

〔図書〕　計0件

〔産業財産権〕

〔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