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　３０　年　　５　月　２５　日現在

機関番号：　１１３０１
研究種目：　基盤研究(B)（一般）
研究期間：　2015～2017
課題番号：　１５Ｈ０２７７３

研究課題名（和文）ヒトゲノム低頻度変異の第一原理的理解に向けた基盤構築

研究課題名（英文）Construction of analyses basis for understanding rare variants using protein structural information

研究代表者

木下　賢吾（Kinoshita, Kengo）

東北大学・情報科学研究科・教授

研究者番号：６０３３２２９３

交付決定額（研究期間全体）：（直接経費）　　11,900,000円

研究成果の概要（和文）：最終的に、これまで開発を行ってきたシステムに関して、機能アノテーションの拡充、公開基盤の改良及び解析基盤の準備を進めながら、予定より遅れていた公開基盤としてwebサイトのリリースを行うことができた。ここでは、我々が開発を行った解析基盤を利用して変異をヒトモデル構造にマップすると共に、既知の疾患関連変異の可視化を行った。
これらのシステムでは計画にあるようにホモロジーモデリングと共に、今回新たに開発を行ったゲノム座標とタンパク質アミノ酸残基の対応付けに基づきマッピングを行った。これらの結果を解析することで、タミフルの代謝係わるタンパク質に関して、日本人特有の興味深い変異を見いだすことができた。

研究成果の概要（英文）：We could construct the proposed systems as planned, which can map genetic variations onto protein structures, and the results are available through the web site called snvmap. For the purpose, we generated model structures of human proteins, if they are not available in PDB.
By using the mapping results, we performed statistical analyses and found some interesting rare variations related with the metabolism of oseltamivir, an anti-virus drug, which specific for east Asian populations.

研究分野：　バイオインフォマティクス

キーワード：　低頻度変異　タンパク質立体構造　ホモロジーモデリング　ヒトタンパク質

１．研究開始当初の背景

2001 年に最初のヒトゲノム配列が明らかにされた。その後、HapMap プロジェクトで人類の多様性に関する多型マーカー探索が行われると共に、ゲノムワイドな変異関連解析（GWAS）が行われ、さまざまな疾患の原因多型が明らかにされてきた。また、次世代シークエンサと呼ばれる高速シークエンサ解析の低価格化により、低頻度の変異が次々に明らかにされ始めている。この試みの一つの到達点は、2012 年に多施設の共同研究による低深度での 1000 人ゲノム解読 (Nature, 2012)であり、これを皮切りにして、オランダ人ゲノム（Nature Genetics, 2014）のような大規模ゲノム配列解析がさまざまな民族を対象として行われるようになってきた。

ゲノム解析は当初、症例対照研究の枠組みでの GWAS が主として行われてきたが、従来行われていたコホート研究と結びつくことで、ゲノムコホートとして、疾患関連変異の同定へと方向転換しつつある。世界的に見てもこの潮流は明らかである。例えば、2013 年サルジニア人ゲノムプロジェクトでは 2500 人の全ゲノム解読が低深度ながら実行され（Cell, 2013）、2014 年にはオランダ人の 250 家系 769 人のゲノム解読が報告(Nature Genetics, 2014)されるなど、各国が大規模なゲノムコホートの立ち上げを行いつつある。このような情勢の中、国内では、100 万人ゲノムコホート計画の検討が進むのと平行して、多数のゲノムコホート研究が COI 事業に採択されるなど、日本としてもゲノムコホート研究を大規模に推進する方向に進みつつある。

大規模なゲノム解読が行われるようになって、GWAS 解析が対象としていた高頻度変異では多くの疾患の原因を十分に説明出来ないことが明らかとなった。その結果、配列解読によってのみ明らかになる低頻度変異（レアバリアント）を解析する必要性が認識されるようになり（下図）、ゲノム解読の低価格化と共に、レアバリアントが徐々に蓄積されてきている。例えば、米国で行われたエキソームプロジェクトでは、6503 人の非血縁者のエキソーム解析を行う事で 180 万以上の変異を同定しているが、その 89.3%は、マイナー対立遺伝子頻度（MAF）が 0.5%以下であるレアバリアントである。このようにレアバリアントのリストアップは着々と進む一方で、レアバリアントのもつ影響をどう評価するかという問題が出てきている。

２．研究の目的

近年、国内外で数多くのゲノムコホート研究が展開されつつある。ゲノムコホート研究では、従来型コホート研究に加えてゲノム解析を行い、遺伝型と環境要因の相互作用を解析し、疾患の原因を明らかにすることが試みられる。しかし、産出された変異データを解釈し、活用する基盤がまだ不十分なため、ゲノムデータが十分に活用されているとは言いがたい。特に、近年疾患との関連から注目されている「低頻度変異（レアバリアント）」の解析では、変異の観察頻度が低いため、従来のような統計的手法を適用するには、非常に大規模なコホートの形成を行う必要があるが、コスト的に現実的で無い。そこで本研究では、ヒトのゲノム情報とタンパク質の構造情報を統合することで、レアバリアントの解釈を行う基盤となる手法とデータベースの開発を行い、ゲノムコホートから生み出される多くの変異情報を最大限に活用する基盤を構築する必要がある。

これに対して我々は、これまでタンパク質科学の分野で培ってきた知見をゲノム科学の分野に持ち込むことで上記の問題の解決を試みる。

タンパク質科学の分野では変異実験と変異導入による機能変化に関する研究が数多くなされており、その結果は UniProt データベースという形で体系的に管理され専門家によるレビューを経てデータが共有されている。また、構造ゲノムプロジェクト（日本では、タンパク 3000、ターゲットタンパク質、創薬等基盤プラットフォーム事業）の進展により、タンパク質立体構造情報解明が急激に進んでおり、その結果は、wwPDB により体系的に管理がなされデータが共有されている。これらの情報を NCBI の RefSeq を仲立ちとし相互に参照することで、ゲノム上の変異をタンパク質立体構造にマッピングし、変異の影響をタンパク質立体構造上の位置、機能部位との空間的な距離、低分子結合部位やタンパク質相互作用に与える影響から評価することができるようになる。つまり、変異の影響を独立な部位のアミノ酸置換と見なすだけでなく、より分子的かつ立体構造に立脚した直接的な解釈が可能となることを目指す点が大きな特色と意義である。

３．研究の方法

木下がゲノムの変異の位置とタンパク質のアミノ酸の位置をつなぐことで、ゲノムの変異の解釈にタンパク質研究で培われた知見を活用する基盤を構築する。タンパク質研究で培われた知見としては、UniProt の情報だけでなく、木下と連携研究者の白井、太田が開発してきたデータベースの機能情報も活用する。構造情報が利用できないタンパク質に関しては、木下らが開発してきた遺伝子共発現データベース COXPRESdb を利用して変異の共発現への影響から、機能への影響を評価できるようにすることも検討する。ヒトゲノムデータは個人情報の観点から慎重に扱う必要があるため、公開可能なデータを取り扱う公開基盤の開発だけで無く、セキュリティの高い内部環境での利用を想定した解析基盤の構築も行う。

UCSC のゲノムとトランスクリプトームの

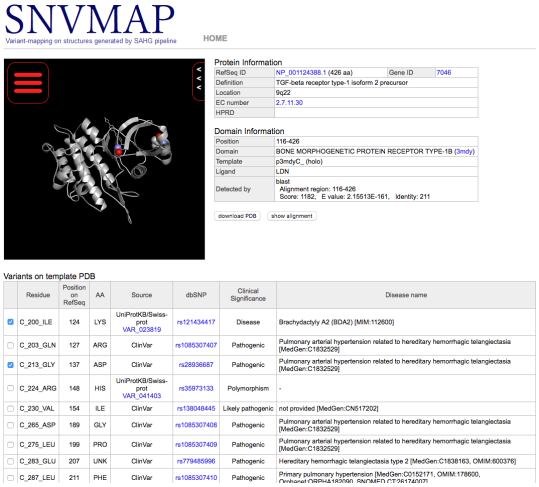
対応付けに応じて、ゲノム上の位置とトランスクリプト（RefSeq 配列）の位置の対応付けを行う。次に、太田らが開発を行った SAHG データベースでの解析パイプラインを利用して、RefSeq 配列と構造情報の対応付けを行う。この際、構造情報が無い場合は、ホモロジーモデリングによる構造構築も行う。対応付けの予備的な検証では、通常の BLAST や PSI-BLAST を用いた手法ではヒトの全トランスクリプトの 54%しかモデル構築を行えないが、SAHG 独自のアライメントツールを含む我々のパイプラインでは 70%の領域がモデル可能であり我々のグループに大きなアドバンテージがあることを確認している。さらに、機能情報として UniProt のアノテーション情報も統合するため、RefSeq と UniProt の対応付けを行う。これら対応付けは概念的には自明な作業に思えるが、予備的な検討の結果では、現在のヒトゲノム情報の不完全さや、タンパク質とゲノムのサンプルの違いによる配列の違い、タンパク質の立体構造解析がドメイン単位で行われることが多いことなどに起因する、いくつかの技術的な課題があることを確認している。例えば、タンパク質としては非常に良く研究されており UniProt にもタンパク質レベルで検証されているエントリーがゲノム上の位置すらわかっていないような場合も存在することを見いだしている。つまり、単純にマッピングするだけで無く、その対応付けを評価しながら注意深く進める必要があると同時に、現在のゲノムデータの限界をタンパク質の知見で補うことができると考えている。

４．研究成果

初年度は特にゲノム情報とタンパク質立体構造情報をつなぐ手法の開発を行った。具体的には、ゲノムとトランスクリプトームの対応付けに応じて、ゲノム上の位置とトランスクリプトの位置の対応付けを行い、平行して連携研究者の太田・白井らが開発を行ったモデリングパイプラインを利用して、RefSeq 配列と構造情報の対応付けを行う基盤を構築した。NHLBI が公開している６５００人の変異データを構造にマップし変異の登場頻度との関係の解析を行った。その結果、タンパク質相互作用部位に予想に反して、立体構造上は重篤に見えるが頻度が高くヒトには影響の無いと思われる興味深い変異を見つけることができた。関連する実験情報を集めることで、この変異が確かに相互作用を弱めることが確認出来ると同時に、相互作用が弱まっても、生体内でのタンパク質の存在量を考えると、複合体の形成自体は可能であることが明らかになった。これは、構造情報だけでは意味づけ困難な変異であるが、全体としては非常にまれなケースであることも確認することができた（Nishi et al, Protein Sci)。また、低分子結合部位周辺の変異も同様に解析を行い、概ね構造情報から予想される変異頻度であることが確認でき、我々のアプローチの妥当性が見えてきた（Yamada et al, BPPB, 2016)。

27 年度には前年度に開発をした公開基盤のプロトタイプとしての公開基盤の公開を行うと同時に、実験研究者に評価を依頼し、改良点の洗い出しと改良を行った。また、次年度に向けて解析基盤の構築を開始する。解析基盤は基本的には公開基盤の内部利用のためのパッケージという位置づけで構築を行うので、比較的短期間での構築が期待できる。Web インターフェースに関しては経験豊かな業者への外注を念頭に、専門家の協力を仰ぐことで実験研究者が直感的に使いやすいインターフェースのデザインを目指した。結果として、H28 年度にはゲノム情報とタンパク質立体構造情報をつなぐ内部用ツールの改良を行うことができた。ツールの基本的な部分は 27 年度に開発を終了していたので当初は 27 年度末頃にはヒト参照ゲノムの新しいバージョン（GRCh38, 2013 年 12 月リリース）に対応したアノテーションも充実してくると想定して、マッピングの更新を行うと共に、今後のゲノム情報、タンパク質立体構造情報も常に最新のバージョンに追随できるように開発を進める予定で、当初の開発はアノテーションが充実している hg19 をベースとして開発を行った。並行して、27 年度当初の国内外の状況を調査した上で GRCh38 の利用を検討したが、まだアノテーションが十分でなかったため、次年度以降に更新を行うこととした。一方、最新のデータが利用できるように、プログラム群の見直しを行い、PDB に収載されている最新のデータを利用できるようになった。

最終年度では、これまで開発を行ってきたシステムに関して、機能アノテーションの拡充、公開基盤の改良及び解析基盤の準備を進めながら、予定より遅れていた公開基盤として web サイトのリリースを行うことができた。（https://sahg.hgc.jp/snvmap)。ここでは、我々が開発を行った解析基盤を利用して変異をヒトモデル構造にマップすると共に、UniProt や CliVar にある疾患関連変異を MolMil で可視化を行った(下図参照)。



これらのシステムは計画にあるようにSAHG をベースとしたホモロジーモデリングと共に、今回新たに開発を行ったゲノム座標とタンパク質アミノ酸残基の対応付けに基づきマッピングを行った。この際、最新の立体構造情報を利用可能なように、パイプラインを見直し高速化を行うことができた。これらの結果を解析することで、タミフルの代謝係わるタンパク質に関して、日本人特有の興味深い変異を見いだすことができた。また、並行して可視化として VR を用いた可視化について検討も行った。VR に関しては特殊なデバイスが必要なため、DB への実装は行わなかったが、開発したプログラムに関しては、希望者には無償で配布することとした。29 年度末頃にはヒト参照ゲノムの新しいバージョン（GRCh38, 2013 年 12 月リリース）に対応したアノテーションも充実してくると予想していたが、予想に反して GRCh38 のアノテーション情報がそろわないため、今回のプロジェクトでは GRCh38 でのマッピングの更新は行わないこととしたが、解析基盤として開発をしたマッピング手法に関しては問題無く GRCh38 でも利用可能なことは確認できたので、今後アノテーションが充実した際には適用したいと考えている。

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 8 件）

(1) Kinjo AR, Bekker GJ, Wako H, Endo S, Tsuchiya Y, Sato H, Nishi H, Kinoshita K, Suzuki H, Kawabata T, Yokochi M, Iwata T, Kobayashi N, Fujiwara T, Kurisu G and Nakamura H. New tools and functions in data-out activities at Protein Data Bank Japan (PDBj). Protein Sci 27(1), 95-102, 2017. doi: 10.1002/pro.3273.（査読有り）

(2) Nishi H, Nakata J and Kinoshita K. Distribution of single nucleotide variants on protein-protein interaction sites and its relation to minor allele frequency. Protein Sci 25(2), 316-321,2016. doi: 10.1002/pro.2845.（査読有り）

(3) Kasahara K, Shirota M, Kinoshita K. Ion Concentration- and Voltage-dependent Push and Pull Mechanisms of Potassium Channel Ion Conduction. Plos One 11, e0150716, 2016. doi: 10.1371/journal.pone.0150716.（査読有り）

(4) Tadaka S and Kinoshita K. NCMine: Core-peripheral based functional module detection using near-clique mining. Bioinformatics 32(22), 3454-3460, 2016. doi: 10.1093/bioinformatics/btw488（査読有り）

(5) Shirota M and Kinoshita K. Discrepancies between human DNA, mRNA and protein reference sequences and their relation to single nucleotide variants in the human population. Database (Oxford) 2016, 1-15, 2016. doi: 10.1093/database/baw124（査読有り）

(6) Murakami Y, Omori S, and Kinoshita K. NLDB: a database for 3D protein-ligand interactions in enzymatic reactions. J Struct Funct Genomics 17, 101-110, 2016. doi: 10.1007/s10969-016-9206-0（査読有り）

(7) Kasahara K and Kinoshita K. IBiSA_Tools: A Computational Toolkit for Ion-Binding State Analysis in Molecular Dynamics Trajectories of Ion Channels. PLoS One 11(12), e0167524, 2016. doi: 10.1371/journal.pone.0167524.（査読有り）

(8) Yamada KD, Nishi H and Kinoshita K. Structural characterization of single nucleotide variants at ligand binding sites and enzyme active sites of human proteins. Biophysics and Physiology, 13, 157-163, 2016. doi: 10.2142/biophysico.13.0_157（査読有り）

〔学会発表〕（計 6 件）

(1) Kinoshita K. A Challenge to Understand Functional Impacts of Rare Variants Using Protein Strucral Information. 2nd Karolinska-Tohoku Joint Symposium on Medical Sciences, 2017/10/4,東北大学

(2) 栗本優美,木下賢吾.ゲノム変異に基づく味覚の人種差の分析. NGS 現場の会第 5 回研究会, 2017 年 5 月 22 日,仙台国際センター（仙台）

(3) 木下賢吾. ヒトゲノム変異の影響評価に対する生命情報科学的アプローチ. 日本分子生物学会シンポジウム, 2016 年 12 月 1 日, パシフィコ横浜（横浜）

(4) Nishi H, Yamada DK, Nakata J, Kinoshita K. Distribution of human single nucleotide at protein functional sites and its relation to minor allele frequency. Gordon Conference, 2016.USA

(5) 西羽美, 中田純一, 木下賢吾. ヒトゲノム一塩基変異のタンパク質構造からの理解：Exome 6500 と 1KJPN を例に. 第 16 回日本蛋白質科学会, 2016 年 6 月 8 日, 福岡国際会議場（福岡）

(6) Kinoshita K. Prediction of biological and biochemical functions of uncharacterized genes. Data Science in

Life Science and Engineering Collaboration and Symposium, 30th, Jul, 2015, Case Western Reserve University, USA

〔その他〕
ホームページ等
http://sahg.hgc.jp/snvmap

６．研究組織
(1)研究代表者
木下　賢吾（KINOSHITA, KENGO）
東北大学・大学院情報科学研究科・教授
研究者番号：60332293
(2)研究分担者
無し
(3)連携研究者
白井　剛（SHIRAI, TSUYOSHI）
長浜バイオ大学・バイオサイエンス学部・教授
研究者番号：00262890
太田　元規（MOTONORI, OHTA）
名古屋大学・大学院情報科学研究科・教授
研究者番号：40290895
(4)研究協力者
無し