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　元 年　 ６ 月　 ４ 日現在

機関番号：１１３０１
研究種目：挑戦的萌芽研究
研究期間：2016～2018
課題番号：１６Ｋ１２５１９

研究課題名（和文）レアバリアントの遺伝子発現量に及ぼす影響の俯瞰的解析

研究課題名（英文）Feasibility study of an evaluation of the impact of rare variants on gene expression

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）　　2,600,000円

研究成果の概要（和文）：変異が発現量に及ぼす影響はまだ未解明な部分も多く、非常にチャレンジングである。まず、基礎となるデータの収集と本格的な解析の前処理として、A549とK562細胞に着目し、発現量データの大規模解析をすすめた。また、転写因子結合部位の解析のために、公共データベースに存在するChIP-seqデータを収集し解析を行った。これらの解析の実施に当たっては、MatatakiやChIP-seqのための正確なライブラリ長推定など、独自の解析手法を開発することができた。今後は、この研究課題で開発した要素手法を活用しながら、より多種類の細胞に展開し、非コード領域における変異のインパクト解析の手法開発へとつなげたい。

研究成果の学術的意義や社会的意義
変異が発現量に及ぼす影響に関して、その重要性はこれまでも繰り返し指摘されてきた問題である。これに対して、個別研究の蓄積はある一方で、どのような変異がどのように発現量に影響を与えるかの一般的な関係性は未だ明らかではない。一般的には、転写因子の結合部位に入る変異が重篤な影響を与える事が予想されるが、その一方で、ChIP-seqやHi-Cなどの研究データが明らかにしたように、ヌクレオソームの状態や遺伝子配列の空間的な近さなど大きなゲノム構造の影響など、未だ明らかになっていない部分も多い。本研究では、変異の発現量に与える影響を俯瞰的に見る事で、レアバリアントの評価に大きな影響を与えることが期待される。

研究成果の概要（英文）：The effect of mutations on expression levels is still largely unknown and very challenging. To begin with, the large-scale analysis of expression level data was carried out focusing on A 562 and K 549 cells as a pretreatment for the collection of basic data and the full-scale analysis. We also collected and analyzed ChIP-seq data in the public database for the analysis of transcription factor binding sites. In conducting these analyses, we were able to develop our own analysis methods, such as accurate library length estimation for Matataki and ChIP-seq. In the future, while utilizing the elemental methods developed in this research project, they will be deployed to a wider variety of cells and will not lead to the development of methods for the impact analysis of mutations in non-coding regions.

研究分野：バイオインフォマティクス

キーワード：ゲノム　レアバリアント　発現量

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９、ＣＫ－１９（共通）

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)、イギリスの１万人ゲノム（Nature, 2015）のような大規模ゲノム配列解析が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、2015 年にはついに、日本人 1000 人の高深度解析、イギリス人１万人の中深度の解析など、各国が大規模なゲノム解析を遂行している。

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、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。本年度から基盤研究 B のサポートで、タンパク質の立体構造情報を利用したレアバリアントの影響評価手法の開発を始めたが、タンパク質のコーディング領域の変異の影響評価にはタンパク質の立体構造情報という手がかりが有用である一方、非コード領域に関しては手が打たれていない状況である。そこで本研究提案では、COSMIC (catalog of somatic mutations in cancer)データベースに登録されているセルラインの変異情報から遺伝子の制御領域に入る変異と、NCBI/GEO に登録されている対応するセルラインの発現量情報を統合することで、変異が発現量に与える影響を系統的に評価する。その結果をもとに、発現量という観点からレアバリアントの影響を評価する手法の開発を目指す。

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発現量情報の相関解析を行うことで、変異情報が発現量に及ぼす影響を評価する手法の開発を目指す。

３．研究の方法

最初に COSMIC データベースを利用してセルライン毎の変異情報を取得する。その際、併せて各セルラインの元となる組織の情報を調べることで、後述するように組織特異性を排除する。次に、変異情報のゲノム上の位置と転写開始点データベース dbTSS の転写開始点データベース(Suzuki et al, Nuc. Acid. Res, 2015)を突き合わせることで、転写制御領域として転写開始点前後（上流１０００塩基、下流２００塩基などいくつかの定義を検討予定）の変異のリストアップをセルライン毎に行う。ただし、dbTSS の転写開始点情報は信頼性が高い一方で、利用できるセルラインの種類が限られるので、利用できないセルラインに関しては、refseq の内、同じ領域から転写されている転写産物の内、最も長い転写産物の転写開始点を利用する.この場合は、着目する制御領域を dbTSS から転写開始点情報が取れる転写産物より広めにとることで、refseq の信頼性の低さをカバーする。また、コーディング領域を見る際には着目されていない同義置換に関しても発現量への影響が報告されている例（古い例ではドーパミンレセプターに関する Hum. Mol. Genet. 2003 等）が存在するので、制御領域とは別に次年度以降に平均発現量への影響を調べる。

次に、これらのセルライン毎の遺伝子の発現データを NCBI GEO から取得する。この際、マイクロアレイのデータはノイズが多く特に変異による影響が心配されるので、RNA-seq のデータに限定する。通常、RNA-seq のデータを多サンプルで行うことは計算量の観点から難しいが、これまで我々が開発を行ってきた遺伝子共発現データベース COXPRESdb(Okamura et al, Nuc. Acid. Res., 2015)で利用している RNA-seq 発現量の解析パイプラインでは、5000 サンプルを超える RNA-seq データを解析した実績があるので、数が多いことは我々としては問題にならない。むしろ、正確な平均的発現量を見るためには出来るだけ多くのサンプルを活用する必要が

あると考えており、懸念されるのが、セルライン毎に発現量をまとめると利用できる RNA-seq の数が減ることである。2015 年 10 月現在の COSMIC データベースには、1026 種のセルラインが登録されているが、すべてのセルラインで十分量（最低でも１００サンプルは必要だと想定している）の RNA-seq データを取得できないので、発現量情報が十分にあるセルラインに限定して解析を進める。

なお、発現量に関しては、一番大きな違いは組織の違いである。そこで、小さな発現量の変化を見るために、セルライン毎に収集した組織情報に基づいて、組織毎の平均発現量を見積もり、同じ組織由来のセルラインの平均発現量と比較することで、セルラインに入った変異による発現量の変動を見積もる。この際、セルラインが癌由来であること、及び COSMIC が癌化に伴う体細胞変異に着目して変異を収集していることを念頭に、比較対象となる組織の発現量データはできるだけ非癌のサンプルを収集することで行う。この際、十分量の RNA-seq データが得られない場合には、以前、申請者が植物のデータに関して適用した手法を利用して、大きな寄与をする成分を抜くことで、組織特異性に対応すると想定される大きな変動を除外する（Kinoshita et al, Bioinformatics, 2009）

2 年目は、前年度に準備が整ったデータを利用して、変異と発現量の相関解析を行う。この際、平均発現量や変異データはセルライン毎に準備をし組織特異性の補正を行ったが、相関の解析は遺伝子毎に行うことで、より俯瞰的な変異と平均発現量の関係を見いだすことを目指す。また、調整領域の変異とコード領域の同義変異のそれぞれで平均発現量の差に対する影響を見る。調整領域に関しては、既知の転写因子結合部位モチーフに対応する部分の変異であるか否かも平均発現量変動の要因として検討する。既知の転写因子結合部位モチーフとしては、JASPAR データベースの利用を想定している(Stormo, Bioinformatics 2000, DB の最終更新は 2014 年)。これに加えて、一定以上の変動が観察できた遺伝子の転写制御領域に関しては、転写開始点からの距離に共通の傾向の有無や、MEME 等の既存のモチーフ探索法を適用することで、新規のモチーフの有無を検討する。その他にも転写制御領域で知られている知見として、CpG island の有無が発現量に与える影響（Fatami et al, Nuc. Acids. Res. 2005, Saxonov et al, PNAS, 2006）や、ENCODE の ChIP-seq のデータを利用したヒストン結合領域との関係の有無等、多角的に検討を加えることで単なる観察研究に終わらず変異と平均発現量の関係の一般的なルールの構築を目指す。

この際、一番懸念される問題として、変動幅の大きな遺伝子が少数しか見いだせないことであるが、既に個別の例では変異が発現量に影響を与える事がよく調べられている NRF2-Keap1 シグナル伝達系を具体例として、初年度の組織特異性の補正法を見直す。また、制御領域に見いだせる変異で影響が強い変異が見つからない場合には、検討する領域を広げたり、複数の転写開始点を持つ遺伝子に関しては別の転写開始点の検討も行うとともに、信頼性の高い dbTSS のデータのあるセルラインを重点的に解析する。なお、最初は代表転写開始点のみを用いて解析の効率化を図るが、代表転写開始点の定義は菅野らのグループの研究を参考にする(例えば、Kumura et al, Genome Research 2006)。

同義置換に関しては、コドン利用頻度との相関を主に検討するが、Yeast ではあるがコドンの連続性が発現量に大きな影響を与える（近くの同種のアミノ酸では同じコドンが好んで使われる）ことが示唆されている（Cannarozzi et al. Cell, 2010)。そこで、ヒトゲノム配列に関しても同様の解析を行い、平行して同義変異がコドン連続性に与える影響と発現量の解析を系統的に評価する。同義置換の疾患に及ぼす影響も議論されていることも踏まえて（Zuben&Kimchi-Sarfaty, Nat. Rev. Genetics, 2011)、平均発現量に影響を及ぼす同義置換に関しては、dbSNP にある OMIM の情報を参照し、疾患との関連の有無の関連も検討する。

RNA の 2 次構造が mRNA の安定性に関係することが古くから知られており、不安定な mRNA は分解が促進され結果として観測される発現量が低下する。そこで、RNA の 2 次構造予測を変異前後で行い結果を比較することで、変異の 2 次構造への影響も評価し、2 次構造が変わる場合に平均発現量への影響の有無も検討する。なお、RNA の 2 次構造予測に関しては歴史も長く、利用できる予測ツールが多数存在するので、既存の方法をいくつか併用する形で利用する。

2 年目の後半から 3 年目に掛けては、変異と発現量の関係のルールを整理し、定量的な予測の可能性を検討する。発現量そのものを定量的に予測することは非常に困難であると考えられるので、変異が平均発現量の変化に及ぼす影響の大きさを定量的に見積もる手法として実装を検討する。つまり、絶対値の予測ではなく、変化量の予測を行う。その結果を利用して、現在世界中で行われているゲノム解析の結果として公開データが利用できる制御領域も含む 6500 人のエクソームデータ（Tabor et al, AGHG, 2015）や日本人の 1070 人の全ゲノムデータ（Nagasaki et al, Nature Comm, 2015）の公開部分のデータに適用することで、平均発現量に大きな影響を与える変異のリストアップとデータベース化を行い、変異と平均発現量の関係のルールの一般性の妥当性を検討する。

４．研究成果

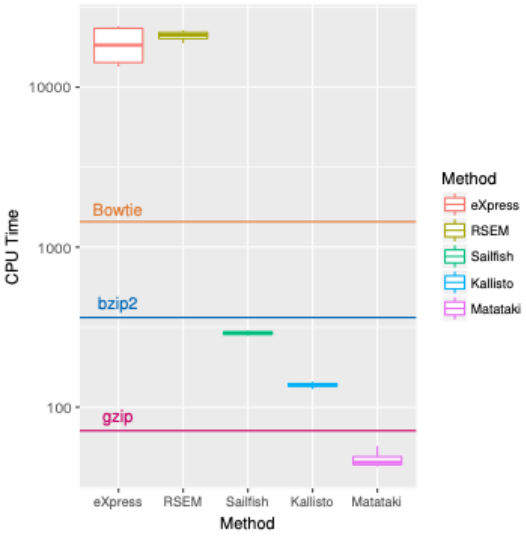
本計画は、これまで個別研究としてはなされていた変異と発現量の関係の解析を俯瞰的に見る初めての研究となる。変異がコーディング領域にアミノ酸の変異を伴う形で入る場合には、自身が基盤 B で展開しているように、タンパク質の立体構造情報を活用して影響を推定する手法の開発の実現性は高い。一方で、変異が発現量に及ぼす影響はまだ未解明な部分も多く、非常にチャレンジングである。チャレンジングではあるが、これでまで我々が展開してきた遺伝

子の発現量解析に関する研究とヒトを対象とした変異解析の基板を活用し、最低限、変異の発現量に及ぼす影響の統計的な描像を明らかにすることを目指す。さらに、定量的に個別の変異の発現量に及ぼす影響まで推定手法にまでつながるかどうかは、実際にどれぐらいの遺伝子とどのような変異が発現変動に寄与しているかを見いだすことが出来るかにかかっている。また、制御領域に入る変異の効果は非加算的であることも予想されるため、変異の発現量に及ぼす影響を定量的に見積もれるかどうかも含めた検討が必要となる。

変異の発現量への影響の網羅的な解析としては、背景で述べたような GWAS 解析が一般的である。また、個別研究としては ENCODE をはじめとして様々な転写因子に関して、転写因子結合部位の変異と転写因子の結合能の低下と発現量の変化の解析が、セルラインや生化学的な実験、あるいはノック会うとマウスの系で個別に検証されてきた。また、癌サンプルにおいては様々な癌腫において原因変異と発現量の解析がなされてきた。これに対して本研究では、変異情報として蓄積されてきたデータ（COSMIC データベース）と、発現量情報として蓄積されてきたデータ（NCBI・GEO）を会わせることで、変異の発現量に与える影響を俯瞰的に見る事に独自性がある。一方、直接的に転写因子を見ないという問題や、特定の条件下での発現量の変化に関して多くの事象を見る事が出来ないという弱点があるが、これまで特定の条件下での変化ばかりを追い、俯瞰的な描像を見ようとしていなかったアプローチに対する問題提起として、あえて変異と平均的な発現量の関係に着目する。その結果を受けて、定量的に変異の発現量への影響を見積もる方法の開発までを目指すが、変異の発現量への影響によって、手法を検討する必要があり、発現量制御のメカニズム（影響が相加的か相乗的か等）の一端にも何らかの知見を得たいと考えている。これらは、チャレンジングではあるが、これでまで我々が展開してきた遺伝子の発現量解析に関する研究とヒトを対象とした変異解析の基盤を活用し、変異の発現量に及ぼす影響の統計的な描像を明らかにすることを目指す。初年度は、ヒト培養細胞のうち比較的データが多い 337 個の ChIP-seq のデータを主に ENCODE から収集し独自に開発を行ったパイプラインで再解析を行った。解析を行うに当たっては、実験データのアノテーションが十分でない部分も多かったので、データから欠けているメタデータを補完するアルゴリズムの開発も行うことができた。発現量に関しては培養細胞系での整理を行い、比較的データが多い A549 細胞をターゲットとして変異と発現量の関係の解析を進めていくこととした。

2 年目は、前年度に準備が整ったデータを利用して、変異と発現量の相関解析を行う。この際、平均発現量や変異データはセルライン毎に準備をし組織特異性の補正を行ったが、相関の解析は遺伝子毎に行うことで、より俯瞰的な変異と平均発現量の関係を見いだすことを目指す。また、調整領域の変異とコード領域の同義変異のそれぞれで平均発現量の差に対する影響を見る。調整領域に関しては、ENCODE の ChIP-seq のデータを利用したヒストンや転写因子結合領域との関係の有無等、多角的に検討を加える。この際、一番懸念される問題として、変動幅の大きな遺伝子が少数しか見いだせないことであるが、既に個別の例では変異が発現量に影響を与える事がよく調べられている NRF2-Keap1 シグナル伝達系を具体例として、初年度の組織特異性の補正法を見直す。また、制御領域に見いだせる変異で影響が強い変異が見つからない場合には、検討する領域を広げたり、複数の転写開始点を持つ遺伝子に関しては別の転写開始点の検討も行うとともに、信頼性の高い dbTSS のデータのあるセルラインを重点的に解析する。なお、最初は代表転写開始点のみを用いて解析の効率化を図るが、代表転写開始点の定義は菅野らのグループの研究を参考にする。

まず、基礎となるデータの収集と本格的な解析の前処理として、A549 と K562 細胞に着目し、発現量データの大規模解析をすすめた。また、転写因子結合部位の解析のために、公共データベースに存在する ChIP-seq データを収集し解析を行った。当初計画では、Cosmic に入っているより多くの細胞種での検討を行いたかったが、公共データベースのアノテーションの揺らぎなど、大規模な解析を阻む要因が多く、本計画では 2 種の細胞に限った解析になってしまった点は残念である。一方、これらの解析の実施に当たっては、Matataki (Okamura and Kinoshita, 2019)や ChIP-seq のための正確なライブラリ長推定（Anzawa et al, in prep）など、独自の解析手法を開発することができた。今後は、この研究課題で開発した要素手法を活用しながら、より多種類の細胞に展開し、非コード領域における変異のインパクト解析の手法開発へとつなげたい。



５．主な発表論文等

〔雑誌論文〕（計　１件）

1. Okamura Y and Kinoshita K. Matataki: an ultrafast mRNA quantification method for

large-scale reanalysis of RNA-Seq data. BMC Bioinformatics 19(1), 266, 2018 doi:10.1186/s12859-018- 2279-y,査読有

〔学会発表〕(計　3件)

1. 木下賢吾,　東北における大規模コホート構築とゲノム・オミックス解析戦略, シンポジウム「理論生物物理学の現在と未来」,2019 年 2 月 14 日, 京都大学
2. 安澤隼人, 木下賢吾, ChIP-Seq データのクラスタリングによる実験条件・解析手法に起因するバイアスの可視化,　NGS 現場の会第五回研究会, 2017 年 5 月 22 日, 仙台国際センター
3. Hayato Anzawa, Kengo Kinoshita,　Model based discrimination method of ChIPed data from control data in ChIP-seq experiment dataset, 第 5 回 生命医薬情報学連合大会, 2016 年 9 月 29 日～10 月 1 日, 東京国際交流館プラザ平成

〔図書〕(計　　0件)

〔産業財産権〕

○出願状況（計　0件)

６．研究組織

(1) 研究分担者

無し

(2) 研究協力者

無し