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 １７ 日現在

機関番号：８２６０６
研究種目：基盤研究(B)（一般）
研究期間：2018～2020
課題番号：１８Ｈ０３３２７

研究課題名（和文）転写異常に繋がるゲノム変異の網羅的探索法の開発とゲノム医療への応用

研究課題名（英文）Development of methods for comprehensive characterization of somatic variants causing transcription aberrations and their application to genomic medicine

研究代表者

白石　友一（Yuichi, Shiraishi）

国立研究開発法人国立がん研究センター・研究所・ユニット長

研究者番号：７０５１６８８０

交付決定額（研究期間全体）：（直接経費）　　10,700,000円

研究成果の概要（和文）：申請者が開発を続けていたSAVNetについての研究成果をまとめ、さらに国際がんゲノムコンソーシアムが主導するがん種横断的な全ゲノム解析プロジェクトに参画し、研究グループへの貢献を行なった。またSAVNetを希少疾患のデータに適用することで、いくつかの新規遺伝子変異の発見に成功した。公共データベースに登録されている大規模トランスクリプトームデータを活用するべく、トランスクリプトームデータのみからスプライシング関連変異を推定する方法論の開発を進めた。また、コンテナ仮想化技術、クラウド計算技術などの技術を統合的に組み合わせることで、大規模トランスクリプトームの情報解析基盤の開発を進めた。

研究成果の学術的意義や社会的意義
ゲノムシークエンス技術の革新により、様々な疾患の新規原因遺伝子変異が同定され、さらに、個人のゲノムのシークエンスを行い、診断、治療に役立てる「ゲノム医療」の試みが進んでいる。一方で、特に、非エキソン領域における変異の機能的意義付けが困難であり、未だにゲノムデータが有するポテンシャルを人類は十分に活かしきれていない。本研究の目的は、ゲノム変異における最も重要なクラスの一つである、スプライシングの異常を引き起こす変異に着目し、当該変異を同定するための方法論の開発を進め、大規模データ解析を通じたデータベース化を行い、ゲノム医療に役立つ基盤技術・リソースの開発を行うものである。

研究成果の概要（英文）：We have summarized the research results of SAVNet, which we had been developing, and contributed to The Pan-Cancer Analysis of Whole Genomes project led by the International Cancer Genome Consortium. In addition, by applying SAVNet to data on rare diseases, we have succeeded in discovering several novel genetic mutations. In order to utilize large-scale transcriptome data in public databases such as Sequence Read Archive, we have developed an algorithm and software to identify splicing-related variants from transcriptome data alone. In addition, we developed an information analysis infrastructure for large-scale transcriptomes by integrating technologies such as container virtualization and cloud computing.

研究分野：バイオインフォマティクス

キーワード：ゲノム　統計科学　がん

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

ゲノムシークエンス技術の革新により、人における生殖細胞系変異、体細胞変異の網羅的な同定が可能になり、様々な疾患の新規原因遺伝子変異、がんドライバー遺伝子が数多く同定されてきており、様々な疾患における遺伝学的知見の蓄積が急速に進んでいる。さらに、個人のゲノムのシークエンスを行い、診断、治療方針の策定に役立てる「ゲノム医療」の試みが進んできている。一方で、これまでの疾患・がんにおけるゲノム変異の探索においては、主にエキソン領域におけるものに限定されてきていた。これは、全ゲノムのシークエンスに比べて全エキソンのシークエンスが安価であったことに加え、非エキソン領域における変異の機能的意義付けが困難であるためであった。シークエンスコストの低下に伴い、全ゲノムシークエンスデータが急速に蓄積しつつあるが、膨大に検出されるゲノムワイドな変異リストから機能的なものを抽出する方法論は確立しておらず、未だに大量のゲノムデータが有するポテンシャルを人類は十分に活かしきれていない。非エキソン領域における重要な機能的変異を同定し、学術研究、臨床の現場で活用する体制を構築することは、今後のゲノム医療の革新のために急務である。
ゲノム変異における重要なクラスの一つに、スプライシングの異常を引き起こすものがある。その中で主要なものとしては、イントロンの両端２塩基（canonical splicing site, GT-AG）における変異であり、この部位における変異によりスプライシング因子の結合に影響が生じ、転写異常が起こることが知られている。また、研究代表者、分担者らは、ゲノム・トランスクリプトームシークエンスデータから、スプライシング異常に繋がるゲノム変異を網羅的に検出する手法、SAVNet を開発し、canonical splicing site 以外の変異が高頻度でスプライシング異常を起こすことを見出していた。

２．研究の目的

本研究の目的は、第１に、申請者の開発した手法を発展させて、さらに多種多様なスプライシング変異の同定を可能とすることである。深部イントロンにおける変異など、エキソン・イントロン境界以外の領域における変異や、遺伝子内の構造異常など、部分的にスプライシング異常との関連が指摘されている形式の変異は多数あるが、これらについての網羅的な解析はあまり進んでいない。多種多様な形式のスプライシング変異の検出を可能とすることで、スプライシング機能についての生物学的な見地からの理解、ゲノム医療への貢献に繋げることを目指す。第２に、開発した手法を大規模公共シークエンスデータに適用し、スプライシング異常を引き起こす変異のカタログ化を行い、データベースの構築を行うことである。

３．研究の方法

研究の当初は、深部イントロン領域においてスプライシング異常を引き起こす可能性のある変異の同定のために、深層学習の方法論などで判別機を構成することを検討していた。しかしながら、2019 年に spliceAI という方法論・データベースが発表ことから(Cell, Jaganathan et al., 2019)、研究の方法についての軌道修正を行った。現在では、深部イントロン領域におけるスプライシング変異の同定のために、トランスクリプトームデータを用いて、「データベースには登録のないスプライシング切断点が観測される」こと、「切断点付近において新しくスプライシング部位を活性化させる変異が観測される」という２つの条件を満たすということを条件として、深部イントロンを含むスプライシング部位生成変異を網羅的に同定する方法論の開発を遂行している。
既存手法の SAVNet においては、スプライシング変異の同定のためにゲノムデータとトランスクリプトームのペアが必要であった。しかしながら、大規模公共データベースにおいては、ゲノムデータとトランスクリプトームデータにおいてはゲノムデータとトランスクリプトームの片方しかないケースが一般的であり、そのことで、SAVNet を最大限に利活用できないという問題点があった。そこで、トランスクリプトームデータのみからスプライシング関連変異を推定するアルゴリズム・ソフトウェアの開発を進めた。

４．研究成果

SAVNet を応用して、スプライシング変異の全体像を解明した研究について、結果をまとめ、国際学術誌において受理・出版された（Shiraishi et al., Genome Research, 2018)。本論文にお

いては、TCGA の 31 がん種、8,976 検体のエキソーム・トランスクリプトームシークエンスデータのペアに応用し、検出された 14,438 のスプライシング変異の中で、約半数が正準スプライシング変異とは異なるタイプであったこと、検出されたスプライシング変異はがん抑制遺伝子に非常に強い集中しており、約４%の患者でがん遺伝子において非正準スプライシング変異が生じていたこと、喫煙による変異パターンとの関連から、肺がんにおいてスプライシング変異が増大することなどを示した。

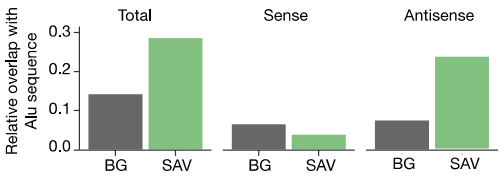

図１：深部イントロンにおけるスプライシング変異と Alu 配列とのオーバーラップについて。特に antisense に挿入された Alu 配列との共起が顕著であった。

また、国際がんゲノムコンソーシアムが主導するがん種横断的な全ゲノム解析プロジェクト（Pan-Cancer Analysis of Whole Genomes）における全ゲノムとトランスクリプトームの統合解析のワーキンググループにおいて、SAVNet を使った解析により、研究グループへの貢献を行なった（PCAWG Transcriptome Core Group et al., Nature, 2020）。1,900 個のスプライシング変異を同定し、その変異群と様々な機能（がん関連遺伝子への集積、Alu 配列とのオーバーラップ）についての解析を行った。特に、deep intron における偽エキソンを生成する変異と Alu 配列とは有意なオーバーラップが見られ（図１）、このことは種の進化で見られる Alu 配列の変異によるエキソンの進化ががんの進化でも生じていることを強く示唆するということを示した。

また、慶應大学医学部の小崎健次郎教授との共同研究において SAVNet を改良し、様々な希少疾患のデータに適用することで、いくつかの新規遺伝子変異の発見にも繋げることに成功した（Mol Genet Metab Rep, Yamada et al., 2019; Mol Genet Genomic Med, Yamada et al., 2020）。また、他のがんゲノム研究プロジェクトにおいても、SAVNet を使って新しいタイプの遺伝子異常の発見に繋がりつつある。

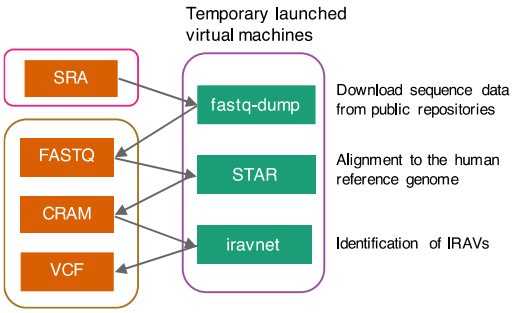

図２：イントロン残存を引き起こす変異を公共レポジトリに登録されているトランスクリプトームデータから抽出するためのワークフロー。

既存手法では、ゲノムデータとトランスクリプトームのペアが必要であったが、多くのプロジェクトにおいてはゲノムデータとトランスクリプトームの片方しかないケースも多く、そのことで、既存のデータを最大限に利活用できないという問題点があった。そこで、トランスクリプトームデータのみからスプライシング関連変異を推定する方法論の開発を進めた。例として、イントロン残存（イントロン残存の場合にはゲノム変異がトランスクリプトームリードに残る性質を利用）を引き起こすタイプのゲノム変異の探索を行う方法論（iravnet）を開発した。さらに、NCBI, EBI, DDBJ などの公共レポジトリに膨大な量のシークエンスデータの蓄積が進んでおり、これらの大規模データを最大限に有効活用するために、コンテナ仮想化技術、クラウド計算技術などの技術を統合的に組み合わせることで、大規模トランスクリプトームの情報解析基盤の開発を進めた（図２）。さらに、ClinVar などの疾患関連データベースに登録されている変異との位置関係から、疾患関連変異の同定を行うワークフローの開発し、総計で 1,000 以上の疾患関連変異の候補を同定した。本研究については現在学術論文に研究成果をまとめている最中である。

ゲノム変異によって新しくスプライシングモチーフが生成され、その場所で新たなスプライシングの切断点を生じさせる形式の変異がある（スプライスサイト生成変異）。こういった変異は、コーディング領域上の同義的置換変異や深部イントロンに潜んでおり同定が非常に難しい。そこで、我々はスプライスサイト生成変異をトランスクリプトームデータのみを用いて検出する方法論（juncmut）の開発を進めた。開発した方法論を The Cancer Genome Atlas の約 10,000 のデータに適用し、種々のがん遺伝子におけるスプライスサイト生成変異の同定を行なった。最後に、比較的高頻度にスプライシング変異が頻発する遺伝子・部位（TP53, CDKN2A など）において、転写異常の形態を特徴量とした機械学習の方法論の構築を行い、スプライシング変異の有無を予測するアルゴリズムの開発を行った。TCGA のデータに適用して、開発した方法論の検証を行なった。

５．主な発表論文等

〔雑誌論文〕　計4件（うち査読付論文　4件／うち国際共著　2件／うちオープンアクセス　1件）

１．著者名 PCAWG Transcriptome Core Group et al.	４．巻 578
２．論文標題 Genomic basis for RNA alterations in cancer	５．発行年 2020年
３．雑誌名 Nature	６．最初と最後の頁 129-136
掲載論文のDOI（デジタルオブジェクト識別子） 10.1038/s41586-020-1970-0	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 Mamiko Yamada, Hisato Suzuki, Yuichi Shiraishi, and Kenjiro Kosaki	４．巻 21
２．論文標題 Effectiveness of Integrated Interpretation of Exome and Corresponding Transcriptome Data for Detecting Splicing Variants of Genes Associated With Autosomal Recessive Disorders	５．発行年 2019年
３．雑誌名 Mol Genet Metab Rep.	６．最初と最後の頁 100531
掲載論文のDOI（デジタルオブジェクト識別子） 10.1016/j.ymgmr.2019.100531	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 －

１．著者名 Shiraishi Y, Kataoka K, Chiba K, Okada A, Kogure Y, Tanaka H, Ogawa S, Miyano S	４．巻 28
２．論文標題 A comprehensive characterization of cis-acting splicing-associated variants in human cancer	５．発行年 2018年
３．雑誌名 Genome Research	６．最初と最後の頁 1111-1125
掲載論文のDOI（デジタルオブジェクト識別子） 10.1101/gr.231951.117	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 該当する

１．著者名 Mamiko Yamada, Yuichi Shiraishi, Tomoko Uehara, Hisato Suzuki, Toshiki Takenouchi, Chihiro Abe-Hatano, Kenji Kurosawa, Kenjiro Kosaki	４．巻 8
２．論文標題 Diagnostic utility of integrated analysis of exome and transcriptome: Successful diagnosis of Au-Kline syndrome in a patient with submucous cleft palate, scaphocephaly, and intellectual disabilities	５．発行年 2020年
３．雑誌名 Mol Genet Genomic Med.	６．最初と最後の頁 e1364
掲載論文のDOI（デジタルオブジェクト識別子） 10.1002/mgg3.1364	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6件（うち招待講演　4件／うち国際学会　1件）

１．発表者名 白石 友一
２．発表標題 大規模トランスクリプトーム解析の情報基盤
３．学会等名 第92回日本生化学会大会（招待講演）
４．発表年 2019年

１．発表者名 白石 友一
２．発表標題 大規模トランスクリプトーム解析の情報基盤
３．学会等名 第2回日本メディカルAI学会学術集会（招待講演）
４．発表年 2020年

１．発表者名 白石友一
２．発表標題 がんゲノム・トランスクリプトームの統合解析
３．学会等名 2018年度統計関連学会連合大会
４．発表年 2018年

１．発表者名 白石友一
２．発表標題 がんゲノムにおけるスプライシング変異の網羅的な検出
３．学会等名 生命医薬情報学連合大会(IIBMP2018)
４．発表年 2018年

１．発表者名
白石友一

２．発表標題
がんゲノムにおけるスプライシング変異の網羅的な検出

３．学会等名
第91回日本生化学会大会（招待講演）

４．発表年
2018年

１．発表者名
白石友一

２．発表標題
Systematic Identification of Splicing Associated Variants toward Precision Medicine

３．学会等名
第49回高松宮妃癌研究基金国際シンポジウム（招待講演）（国際学会）

４．発表年
2018年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

Intron retention associated variant detection
https://github.com/friend1ws/iravnet
SAVNet github page
https://github.com/friend1ws/SAVNet
SAVNet documentation
https://savnet.readthedocs.io/en/latest/index.html

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
研究分担者	片岡　圭亮 (Keisuke Kataoka) (90631383)	国立研究開発法人国立がん研究センター・研究所・分野長 (82606)	

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			
米国	カリフォルニア大学 サンタクルーズ校			
スイス	チューリッヒ工科大学			