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　６ 年　 ６ 月 １２ 日現在

機関番号：８２６０６
研究種目：基盤研究(B)（一般）
研究期間：2021～2023
課題番号：２１Ｈ０３５４９

研究課題名（和文）大規模トランスクリプトームからの自律的知能獲得システム基盤の開発

研究課題名（英文）Development of autonomous intelligence acquisition system infrastructure from large scale transcriptome

研究代表者
白石　友一（Shiraishi, Yuichi）

国立研究開発法人国立がん研究センター・研究所・分野長

研究者番号：７０５１６８８０

交付決定額（研究期間全体）：（直接経費）　　13,200,000円

研究成果の概要（和文）：公共シークエンスレポジトリに蓄積されている大規模トランスクリプトームデータを利用し、各種スプライシング異常を引き起こすゲノム変異を同定するためのアルゴリズム・ソフトウェアを開発・実装した。これらをSequence Read Archiveの30万件以上のトランスクリプトームデータに適用し、数万以上のスプライシング変異を同定した。さらに、これらのスプライシング変異の生物学的意義について詳細な分析を行なった。例として、NOTCH1遺伝子における機能獲得型のスプライスサイト生成変異を同定し、実際の生物学実験を通じてこの変異が活性を引き起こし、核酸医薬によって抑制可能であることを実証した。

研究成果の学術的意義や社会的意義
今回開発したプラットフォームは、公共シークエンスレポジトリの蓄積されたデータを用いて実行することで、疾患に関連する変異および核酸医薬の創薬ターゲットとなる変異を自動的に検出可能となることを示した。これは、ゲノム医療における自律的な知識獲得の基盤となる実例となる。また、今後さらに蓄積が進むゲノムデータを効果的に活用するための実例となり、将来的にデータ駆動型の科学・医学を構築するための一つのモデルケースとなるだろう。

研究成果の概要（英文）：Using large-scale transcriptome data stored in public sequencing repositories, we developed and implemented algorithmic software to identify genomic mutations that cause various splicing abnormalities. These were applied to more than 300,000 transcriptome data in the Sequence Read Archive, and more than tens of thousands of splicing mutations were identified. In addition, we performed a detailed analysis of the biological significance of these splicing variants. As an example, we identified a gain-of-function splice-site generating mutation in the NOTCH1 gene and demonstrated through practical biological experiments that this mutation causes activity that can be suppressed by nucleic acid drugs.

研究分野：生命情報学

キーワード：大規模データ解析　ゲノム変異　クラウド　スプライシング

様　式　Ｃ－１９、Ｆ－１９－１（共通）

１．研究開始当初の背景

2000 年代後半からのシークエンス技術の革新、およびその後の研究により、シークエンス技術の有用性は生物学、遺伝学、医学、人類学など多岐にわたる分野で広く証明された。さらに、シークエンス技術は学術研究にとどまらず、患者のゲノムをシークエンスし、その結果を診断・治療法の決定に反映させるゲノム医療への実装が進展している。これに伴い、オミクスデータの蓄積速度は加速している。今後は、研究・医療で得られるオミクスデータをいかに活用するかが重要な課題となる。膨大なデータから新たな医学的知見を導出し、新規創薬やデータベースの精緻化を通じて医療現場へ還元し、さらなるデータ獲得の動機付けを行う「学習する医療システム」のサイクルを回すことが、我が国の医療の質を保ち、国際的なプレゼンスを示すために重要となる。

２．研究の目的

クラウド上に蓄積されている大規模オミクスデータから有用な知識を獲得するアルゴリズム・ソフトウェアの開発、並びにこれらのソフトウェアをクラウド上で効率的に実行する基盤の開発を通じて、自律的に知識を習得するプラットフォームの開発を目指す。本研究課題を通じて、現在我が国でも進められているゲノム医療、およびそれに伴うデータ収集の基盤についての検討を深化させ、将来的なデータ駆動型科学・医療の構築に寄与することを目指す。

３．研究の方法

ゲノム変異の中で重要なクラスの一つとして、スプライシング異常を引き起こす変異がある。典型的なケースとして、イントロンの両端の 2 塩基（GT-AT）に変異が生じると、スプライシングに必要な因子の結合が阻害され、異常なスプライシングが発生する。イントロンの両端以外にも、さまざまな形式のゲノム変異がスプライシング異常を引き起こすことが知られており、疾患を引き起こす病的変異のうち 15～60%がスプライシング異常に関連していると見積もられている（Park et al., Am. J. Hum. Genet., 2018 など）。本研究では、公共トランスクリプトームデータを利用し、転写異常を介して疾患に関与するゲノム変異のスクリーニングに焦点を当てる。公共トランスクリプトームデータの大規模解析を通じた疾患関連変異のスクリーニングは、世界でも類を見ない試みである。ゲノムデータとトランスクリプトームデータがペアで得られるケースはまだ少ないが、本研究の進展により既存データの有用性を大幅に高めることが期待される。

４．研究成果

(1) イントロン残存を引き起こす変異の網羅的同定

スプライシング異常の一形態であるイントロン残存を引き起こすゲノム変異に着目し、トランスクリプトームデータのみを用いてイントロン残存を引き起こすゲノム変異を同定する検出アルゴリズム（IRAVNet; https://github.com/friend1ws/iravnet）を開発した。また、1000 Genomes Project、The Cancer Genome Atlas など、ゲノム・トランスクリプトームデータがペアで与えられているデータセットを用いて、このアルゴリズムの感度・正答率を検証した。その後、SRA などに登録されている大規模なトランスクリプトームに対してこの方法論を適用するため、Amazon Web Service を利用したクラウドベースの解析プラットフォームとオンプレミスの計算クラスタを用いたプラットフォームを開発した（図１）。開発したプラットフォームを用いて、Sequence Read Archive(SRA)と The Cancer Genome Atlas(TCGA)に総計

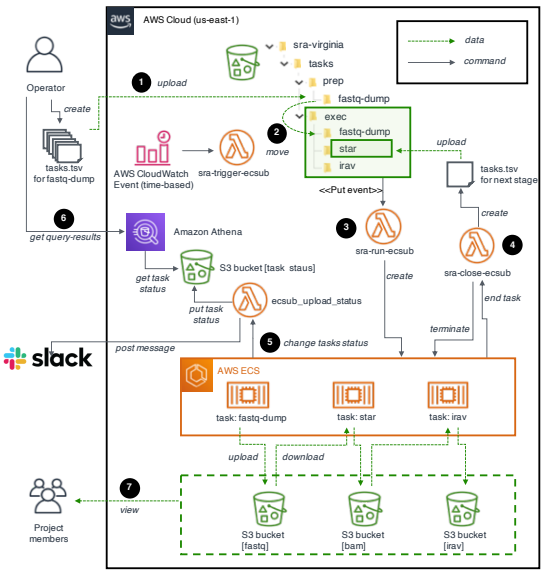

図１：Amazon Web Services を用いた大規模ゲノム解析インフラストラクチャー

230,988 件のトランスクリプトームデータを解析した。また、ClinVar に登録されている疾患関連変異との位置関係を比較し、3,000 の疾患関連変異候補を同定した。この中には、がんドライバー遺伝子や各種遺伝病に関連する遺伝子が多く含まれていた。検出した変異の一覧は、新たに開発したポータルサイト IRAV DB(https://iravdb.io/)で広く公開している。一連の研究成果は、国際学術誌に採択された（Shiraishi et al., Nature Communications, 2022)。

(2) スプライスサイト生成変異の網羅的同定

ゲノム変異によって新しくスプライシングモチーフが形成され、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「スプライスサイト生成変異」に焦点を当て、このクラスの変異をトランスクリプトームデータのみから検出する方法論（https://github.com/ncc-gap/juncmut）を開発した。また、1000 Genomes Project、The Cancer Genome Atlas など、ゲノム・トランスクリプトームデータがペアで与えられているデータセットを用いて、このアルゴリズムの感度・正答率を検証した。SRA と TCGA から取得される 30 万検体以上のトランスクリプトームから、約 30,000 のスプライスサイト生成変異を収集した。

獲得したスプライスサイト生成変異のリストについて、その意義をシステマティックに解明するための情報解析基盤を開発した。まず、転写に対する影響（partial exon loss、exon extension、cryptic exon inclusion など)、トランスクリプトが元の配列に対して in-frame か frameshift か、Premature Termination Codon(PTC)を生成するか、Nonsense Mediated Decay(NMD)を引き起こすかを予測するプログラムを開発し、スプライシング変異を分類した。さらに、in-frame のトランスクリプトを生成するスプライスサイト生成変異でも PTC を生成する割合を集計し、機能的影響を調べた。また、Alu 配列とスプライスサイト生成変異の関連を調査し、スプライスサイト生成変異が Alu exonization を頻繁に引き起こすこと、その際には Alu 配列の中で exon になる部分にホットスポットが存在することを高精度で示した（図 2)。がん関連遺伝子については、スプライスサイト生成変異の転写の結果、ドメインとの位置関係、COSMIC における変異頻度などを可視化するプログラムを開発した。その結果の一例として、CREBBP 遺伝子の KAT ドメインにスプライスサイト生成変異が集中して検出され、全てが in-frame の PTC を生成しないトランスクリプトを産生することが判明した（図 3)。これは、当該遺伝子においてスプライスサイト生成変異が単純な機能喪失型ではなく、なんらかの機能を有することの証左である。

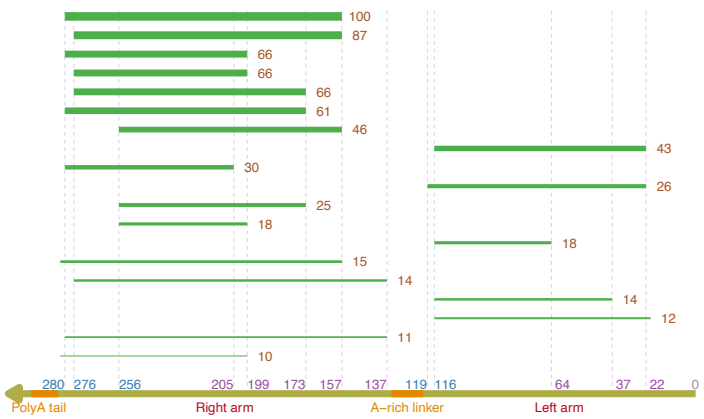

図 2: Alu exonization による偽エキソンのホットスポット

さらに、CRISPR-Cas9 を用いて肺がん細胞株 PC-9 から、機能活性が予測される NOTCH1 遺伝子のスプライスサイト生成変異（c.5048-132G>C, c.5048-132G>T）を持つ細胞モデルを作成した。ゲノム編集細胞から RNA を抽出し、RT-PCR により予想される標的エキソン・イントロン領域のスプライシングが実際に生じていることを確認した。検出した変異の一覧は、新たに開発したポータルサイト SSCV DB(https://sscvdb.io/)で広く公開している。一連の研究成果はプリプリントサーバーにアップロードされ、現在国際学術誌の査読中である。

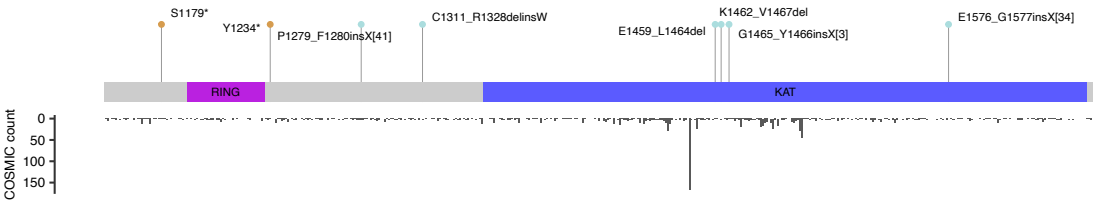

図 3: CREBBP におけるスプライスサイト生成変異の概要

(3) スプライシング異常から酸化ストレス応答系の活性予測アルゴリズムの開発

スプライシング異常のデータを用いて、NRF2, KEAP1 などの酸化ストレス応答系のパスウェイの活性を予測する方法論を開発した。この原理は、酸化ストレス応答系に異常がある場合、その下流の遺伝子の発現が亢進し、それに伴いスプライシングにも異常が生じることを利用したも

のである。The Cancer Genome Atlas においてスプライシングのデータと遺伝子変異プロファイルがペアで与えられているデータを用いて判別機を構築し、感度・特異度の検証を実施し、遺伝子変異の中でどのクラスがより活性を上昇させるかを評価した。また、活性が高いサンプルで一見関連する遺伝子が検出できない場合でも、詳細に元データを調査することにより構造異常などのゲノム異常を認めることができた。
また、Sequence Read Archive に登録されているトランスクリプトームデータを用いて酸化ストレス応答系の活性が高いサンプルをスクリーニングした。多くは肺がんを中心としたがんの検体であり、一部 LPS などにより酸化ストレスの活性が上昇したサンプルが見受けられた。

５．主な発表論文等

〔雑誌論文〕　計7件（うち査読付論文　5件／うち国際共著　1件／うちオープンアクセス　7件）

１．著者名 Iida Naoko、Okada Ai、Kobayashi Yoshihisa、Chiba Kenichi、Yatabe Yasushi、Shiraishi Yuichi	４．巻 NA
２．論文標題 Systematically developing a registry of splice-site creating variants utilizing massive publicly available transcriptome sequence data	５．発行年 2024年
３．雑誌名 bioRxiv	６．最初と最後の頁 NA
掲載論文のDOI（デジタルオブジェクト識別子） 10.1101/2024.02.21.581470	査読の有無 無
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 Nakamura Wataru、Hirata Makoto、19 authors、Shiraishi Yuichi	４．巻 9
２．論文標題 Assessing the efficacy of target adaptive sampling long-read sequencing through hereditary cancer patient genomes	５．発行年 2024年
３．雑誌名 npj Genomic Medicine	６．最初と最後の頁 11
掲載論文のDOI（デジタルオブジェクト識別子） 10.1038/s41525-024-00394-z	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 Shiraishi Yuichi、Koya Junji、Chiba Kenichi、Okada Ai、Arai Yasuhito、Saito Yuki、Shibata Tatsuhiro、Kataoka Keisuke	４．巻 51
２．論文標題 Precise characterization of somatic complex structural variations from tumor/control paired long-read sequencing data with nanomonsv	５．発行年 2023年
３．雑誌名 Nucleic Acids Research	６．最初と最後の頁 e74
掲載論文のDOI（デジタルオブジェクト識別子） 10.1093/nar/gkad526	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 Sakamoto Yoshitaka、Miyake Shuhei、Oka Miho、Kanai Akinori、Kawai Yosuke、Nagasawa Satoi、Shiraishi Yuichi、Tokunaga Katsushi、Kohno Takashi、Seki Masahide、Suzuki Yutaka、Suzuki Ayako	４．巻 13
２．論文標題 Phasing analysis of lung cancer genomes using a long read sequencer	５．発行年 2022年
３．雑誌名 Nature Communications	６．最初と最後の頁 3464
掲載論文のDOI（デジタルオブジェクト識別子） 10.1038/s41467-022-31133-6	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 Shiraishi Yuichi、Okada Ai、Chiba Kenichi、Kawachi Asuka、Omori Ikuko、Mateos Raul Nicolas、Iida Naoko、Yamauchi Hirofumi、Kosaki Kenjiro、Yoshimi Akihide	4．巻 13
2．論文標題 Systematic identification of intron retention associated variants from massive publicly available transcriptome sequencing data	5．発行年 2022年
3．雑誌名 Nature Communications	6．最初と最後の頁 5357
掲載論文のDOI（デジタルオブジェクト識別子） 10.1038/s41467-022-32887-9	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 Isobe Tomoya、Takagi Masatoshi、Sato-Otsubo Aiko、Nishimura Akira、Nagae Genta、Yamagishi Chika、Tamura Moe、Tanaka Yosuke、Asada Shuhei、Takeda Reina、Tsuchiya Akiho、Wang Xiaonan、Yoshida Kenichi、Nannya Yasuhito、Ueno Hiroo、Akazawa Ryo et al.	4．巻 13
2．論文標題 Multi-omics analysis defines highly refractory RAS burdened immature subgroup of infant acute lymphoblastic leukemia	5．発行年 2022年
3．雑誌名 Nature Communications	6．最初と最後の頁 4501
掲載論文のDOI（デジタルオブジェクト識別子） 10.1038/s41467-022-32266-4	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 Yuichi Shirishi et al.,	4．巻 NA
2．論文標題 Systematic identification of intron retention associated variants from massive publicly available transcriptome sequencing data	5．発行年 2021年
3．雑誌名 bioRxiv	6．最初と最後の頁 NA
掲載論文のDOI（デジタルオブジェクト識別子） 10.1101/2021.10.05.463278	査読の有無 無
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9件（うち招待講演　9件／うち国際学会　3件）

1．発表者名 白石友一
2．発表標題 完全がんゲノム配列の再構成に向けたロングリードシークエンス解析基盤の開発
3．学会等名 令和4年度国際がん研究シンポジウム「WGS, Long-read and Beyond」（招待講演）（国際学会）
4．発表年 2023年

１．発表者名
白石友一

２．発表標題
Integrated whole genome and transcriptome analysis platform applied to adolescent and young adult cancers

３．学会等名
第81回 日本癌学会学術総会（招待講演）

４．発表年
2022年

１．発表者名
白石友一

２．発表標題
高精度に構造異常を検出するための解析基盤

３．学会等名
日本メディカルAI学会 ToMMo 共催企画サテライトシンポジウム「コホート・バイオバンクとビッグデータ解析の可能性」（招待講演）

４．発表年
2022年

１．発表者名
白石友一

２．発表標題
知識発見を加速するゲノム解析プラットフォームについて

３．学会等名
和3年度国際がん研究シンポジウム「全ゲノム解析が変革するがん研究・がん医療」（招待講演）

４．発表年
2022年

１．発表者名
白石友一

２．発表標題
大規模公共トランスクリプトームレポジトリからの自律的知識獲得システム基盤

３．学会等名
2021年度国立遺伝学研究所研究会「ゲノム医科学とバイオインフォマティクスの接点と集学研究」（招待講演）

４．発表年
2022年

１．発表者名
白石友一

２．発表標題
知識発見を加速するオミクス解析基盤

３．学会等名
大阪大学医学系研究科バイオインフォマティクスセミナー（招待講演）

４．発表年
2022年

１．発表者名
白石友一

２．発表標題
クラウドを使ったゲノム解析基盤

３．学会等名
第11回マルチNGSオミクス解析研究会（招待講演）

４．発表年
2022年

１．発表者名
Yuichi Shiraishi

２．発表標題
Systematic identification of intron retention associated variants from massive publicly available transcriptome sequencing data

３．学会等名
International Conference on Genomics (IGC-16)（招待講演）（国際学会）

４．発表年
2021年

１．発表者名
Yuichi Shiraishi

２．発表標題
Massive in-silico screening of intron retention causing variants using publicly available transcriptome sequencing data

３．学会等名
nternational Caparica Conference in splicing 2021 (splicing2021)（招待講演）（国際学会）

４．発表年
2021年

〔図書〕　計3件

１．著者名 白石友一、岡田愛	４．発行年 2022年
２．出版社 医歯薬出版	５．総ページ数 8
３．書名 医学のあゆみ 283（9）	

１．著者名 白石友一、岡田愛、河野隆志	４．発行年 2023年
２．出版社 羊土社	５．総ページ数 6
３．書名 実験医学増刊 41（7）	

１．著者名 白石友一	４．発行年 2022年
２．出版社 クラウドを使ったゲノム解析環境の構築	５．総ページ数 5
３．書名 実験医学	

〔産業財産権〕

〔その他〕

IRAVNet software page
https://github.com/friend1ws/iravnet

IRAV DB
https://iravdb.io/

juncmut software page
https://github.com/ncc-gap/juncmut

SSCV DB
https://sscvdb.io/

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	飯田　直子 (Naoko Iida) (40360557)	国立研究開発法人国立がん研究センター・東病院・研究員 (82606)	
研究分担者	吉見　昭秀 (Akihide Yoshimi) (80609016)	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