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　　６　月　１２　日現在

機関番号：　１２６０１
研究種目：　基盤研究(C)（一般）
研究期間：　2015～2017
課題番号：　１５Ｋ００３９８

研究課題名（和文）大規模がんゲノム変異データマイニングのための統計的手法の開発

研究課題名（英文）Development of statistical methods for large scale somatic mutation data mining

研究代表者
白石　友一（Yuichi, Shiraishi）

東京大学・医科学研究所・助教

研究者番号：７０５１６８８０

交付決定額（研究期間全体）：（直接経費）　　3,600,000円

研究成果の概要（和文）：がんの後天的変異データの集合から、特徴的なパターンを抽出する新たな統計的手法、pmsignatureを開発した（Shiraishi et al., PLoS Genetics, 2015）。この方法論は、「条件付き独立性を仮定したモデリングにより、変異パターンの因子数を増やしても首尾よく推定が可能である」、「提機械学習分野で文書分類に利用されるトピックモデルと類似したモデルとなっている」といった特徴を備えている。さらに、本手法を用いて、スプライシング異常を引き起こす変異において特徴的なパターンを見出した（Shiraishi et al., Biorxiv, 2017）。

研究成果の概要（英文）：We have developed a novel statistical method for extracting characteristic pattern from somatic mutation data (Shiraishi et al., 2015, https://github.com/friend1ws/pmsignature). Assuming the independence on each factor of mutation signatures and reducing the number of parameters, more robust and interpretable estimates can be obtained. Additionally, the proposed model has close relationships with the “mixed-membership models,” that have been intensively utilized in statistical machine learning and statistical genetics community. Furthermore, we have applied this approach to the set of splicing associated variants and identified several novel patterns (Shiraishi et al., BioRxiv, 2017).

研究分野：　統計科学

キーワード：　がんゲノム　機械学習

１．研究開始当初の背景

がんの変異には，がんの種類に応じて明らかな傾向があることが知られていた(喫煙歴のある肺がんについてはC>Aの変異，皮膚がんについてはC>T，CC>TTの変異)．さらに置換のパターンだけではなく，周辺の塩基情報も重要であることが知られていた(多くのがん種においてCpGサイトにおいてC>Tの変異が多く見られる)．近年のシークエンス技術の発展により，個々のがんゲノムからゲノムワイドに大量の変異の検出できるようになったことで，新規がん原因遺伝子の同定だけではなく，個々のがんゲノムごとに見られる変異のプロファイルの違いを見ることが可能になった.その一方で，がんゲノムシークエンス解析から得られる大量の変異データから，特徴的なパターンを抽出するために，新たな情報学的，統計学的手法の開発が求められるようになった．申請者自身は現在新しい手法，(pmsignature, probabilistic mutation signature, https://github.com/friend1ws/pmsignature)を開発中であった.

２．研究の目的

まず，申請者が開発している手法(pmsignature)についての開発，実際のデータを用いた評価を行い，既存手法と比べた優位性の検証を行うこと．コマンドラインに不慣れな生物・医学系研究者が簡単に利用できるように，ウェブアプリケーションの開発を行う．また,「一塩基置換変異だけではなく，挿入・欠失，構造変異などの種々の変異からのマイニングを可能にする」，「ゲノム機能情報を統合する」といった拡張を加え，大規模がんゲノム変異データをマイニングするための統合的プラットフォームの開発を行う.これにより，がんの分類や原因解明の研究を促進し，さらにはがんの治療戦略の方針決定に貢献することを目指した.

３．研究の方法

申請者自身の手法（pmsignature）については，ICGC, TCGAなどから収集した，大規模がんゲノム変異データベースに適用するなどして，既存手法と比較して新しい生物学的治験が得ることが可能かについて検証した.ウェブアプリケーションの開発にあたっては，Rパッケージをウェブ上にデプロイする上で一般的なフレームワークであるshiny(http://shiny.rstudio.com)を利用した.種々の変異の検出にあたっては，点変異以外の変異の特徴的なパターンを網羅的に検出する方法論の開発がさほど進んでいないことから，これらの変異を網羅的に同定する方法論・ソフトウェアの開発から着手した.

４．研究成果

まず，開発中であったがんの変異のデータマイニングのための統計的手法である，pmsignature をソフトウェアとして完成させた．提案手法は，「首尾よく条件付き独立性を仮定することにより，変異パターンの因子数を増やしても，頑健な推定が可能である」，「直管的にわかりやすいvisualizationを構成できる」,「提案モデルは機械学習分野で文書分類に利用されるトピックモデル(Blei et al., JMLR, 2003)と類似したモデルとなっており，過去にこれらの分野で蓄積されてきた膨大な知見を利用することが出来る」（図１）といった特徴を有している.さらにRpackageを開発し，コアとなる EM algorithm の計算部分はRcppというRでC++のコードを効率的に利用する為のパッケージを用いて実装した．これにより数百検体ほどのデータセットならば，デスクトップ PC で数十分程度で解析することが可能である．さらに，開発した手法を約７千検体のがんゲノムデータに適用することで，紫外線や APOBEC と呼ばれる酵素による変異パターンにおける新規パターンを見出した．また，その過程でウェブアプリケーションの開発も行い，現在まで稼働中である(Shiraishi et al., PLoS Genetics, 2015).

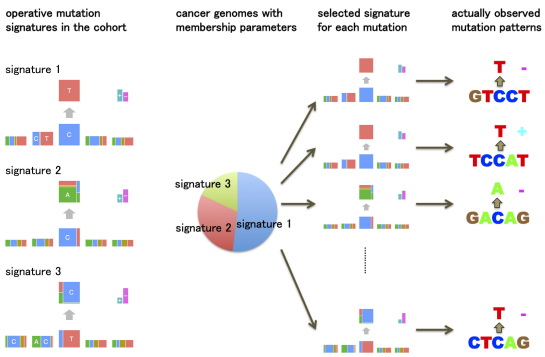

図１：pmsignature における後天的変異の生成モデル(Shiraishi et al., PLoS Genetics, 2015).

点変異だけではなく構造変異など様々な変異を統計モデルに組み込んだ手法の開発の準備として，構造変異を感度よく正確に検出する新たな方法論を開発したGenomonSV(https://github.com/Genomon-Project/GenomonSV)．本方法論は，染色体をまたぐ転座など，大域的な構造変異だけではなく,白血病でよく観察されるFLT3-ITDなど，数十から数百塩基の比較的小さな構造変異の検出も可能である．本方法論を用いて，医学系研究者との共同研究により，成人T細胞白血病における新しい分子異常メカニズムの解明（Kataoka, Nagata, Kitanaka,

Shiraishi et al., Nature Genetics, 2015) などを始め多数の疾患関連遺伝子の発見に貢献した．また，2016 年に，免疫チェックポイント遺伝子における新規ゲノム異常を見出すことに貢献した（Kataoka, Shiraishi et al., Nature, 2016）．
DNA シークエンスと RNA シークエンスデータを統合的に扱い，スプライシング異常を引き起こすゲノム変異を網羅的に抽出する新規統計手法，SAVNet(https://github.com/friend1ws/SAVNet)の開発を行った．さらに，開発した手法を TCGA の約 9000 検体のエキソーム，RNA 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ゲノム変異の全体像を明らかにした．また，スプライシング異常を引き起こす変異について，前年度までに開発した pmsignature を利用した変異パターンのマイニングを行い，スプライシング異常を引き起こす変異に特徴的な変異パターンを明らかにした．これらの研究成果は，現在論文投稿中である．

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 6 件）

1. Seki M, Shiraishi Y, et al. (他 42 名，14 番目)Recurrent SPI1 (PU.1) fusions in high-risk pediatric T cell acute lymphoblastic leukemia. Nat Genet. 2017 Aug;49(8):1274-1281. (査読有)
2. Makishima H, Shiraishi Y, et al. (他 35 名，18 番目)Dynamics of clonal evolution in myelodysplastic syndromes. Nat Genet. 2017 Feb;49(2):204-212. (査読有)
3. Kataoka K, Shiraishi Y, et al. (他 42 名，2 番目，第一著者と同等の寄与)Aberrant PD-L1 expression through 3'-UTR disruption in multiple cancers. Nature. 2016 Jun 16;534(7607):402-6. (査読有)
4. Fujimoto A, Shiraishi Y, et al. (他 49 名，6 番目)Whole-genome mutational landscape and characterization of noncoding and structural mutations in liver cancer. Nat Genet. 2016 May;48(5):500-9. (査読有)
5. Shiraishi Y, et al. (他 3 名) A Simple Model-Based Approach to Inferring and Visualizing Cancer Mutation Signatures. PLoS Genet. 2015 Dec 2;11(12):e1005657. (査読有)
6. Kataoka K, Shiraishi Y, et al. (他 57 名，4 番目，第一著者と同等の寄与)Integrated molecular analysis of adult T cell leukemia/lymphoma. Nat Genet. 2015 Nov;47(11):1304-15. (査読有)

〔学会発表〕（計１１件）

1. 大規模がんオミックス解析により明らかになったスプライシング変異の全体像，口頭，白石友一，生命科学系学会合同年次大会，2017/12/6，国内
2. The landscape of splicing associated variants revealed by large scale whole exome and transcriptome analysis，口頭，白石友一，日本癌学会学術総会，2017/9/29，国内
3. 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，口頭，白石友一，統計関連学会連合大会（招待講演），2017/9/5，国内
4. Latent Probabilistic Modeling for Mutational Signature in Cancer Genomes，口頭，白石友一，国際分類学会，2017/8/8，国内
5. Systematic identification of somatic variants causing cis-splicing alterations，口頭，白石友一，日本プロテオーム学会（招待講演），2017/7/26，国内
6. Systematic identification of somatic variants causing cis-splicing alterations，口頭，白石友一，International Workshop for Systems Genetics，2017/6/23，国内
7. Genomon-SV を使った大規模エキソーム解析で明らかになった構造変異の全体像，口頭，白石友一，日本癌学会学術総会，2016/10/6，国内
8. 大規模がんゲノム変異データマイニングのための統計学的手法，口頭，白石友一，生命医薬情報学連合大会，2016/9/30，国内
9. Large scale and reproducible cancer genome analysis using Genomon2 and Azure，口頭，白石友一，生命医薬情報学連合大会（招待講演），2016/9/30，国内
10. 大規模がんゲノム変異データマイニングのための統計学的手法，口頭，白石友一，統計関連学会連合大会，2016/9/6，国内
11. Extraction of Latent Probabilistic Mutational Signature in Cancer Genomes，口頭，白石友一，RECOMB-Seq，2015/4/10，ワルシャワ

〔図書〕（計 4 件）

1. 白石友一，「がんゲノムにおける後天的変異の変異シグナチャーのモデリング

と可視化について」, 遺伝子医学 MOOK 33 号, メディカルドゥ, 2018.

2. 白石友一,「低頻度体細胞変異を見つけるための情報解析技術」, 実験医学, 35(1) , 羊土社, 2017.
3. 白石友一, 新井田厚司, 宮野悟,「がん研究における RNA-Seq, がんのシステム異常を情報学的に俯瞰する」, 実験医学別冊 RNA-Seq 実験ハンドブック, 羊土社, 2016.
4. 白石友一, 「上級者の技術を見て学ぶ. 論文別コマンド解説」, 細胞工学別冊 次世代シークエンサーDry 解析超入門, 学研, 2015.

〔産業財産権〕

○出願状況（計 0 件）

○取得状況（計 0 件）

〔その他〕
ホームページ等
pmsignature:
https://github.com/friend1ws/pmsignature

pmsignature web application:
https://friend1ws.shinyapps.io/pmsignature_shiny/

GenomonSV:
https://github.com/Genomon-Project/GenomonSV

SAVNet:
https://github.com/friend1ws/SAVNet

６．研究組織
(1)研究代表者
白石　友一（Shitaishi, Yuichi）
東京大学・医科学研究所・助教
研究者番号：70516880