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　 ５ 月 ２１ 日現在

機関番号：１３９０１
研究種目：若手研究
研究期間：2018～2020
課題番号：１８Ｋ１４６８４

研究課題名（和文）一塩基置換（SNV）の統合的病因性予測システムの開発

研究課題名（英文）Development of integrated pathogenicity prediction system using single nucleotide variants (SNVs)

研究代表者

武田　淳一（Takeda, Junichi）

名古屋大学・医学系研究科・特任助教

研究者番号：６０６２５６７２

交付決定額（研究期間全体）：（直接経費）　　1,900,000円

研究成果の概要（和文）：ヒト遺伝子の塩基配列はタンパクに翻訳されるが、一つの塩基が変異することによってタンパクを構成するアミノ酸が変わり、その結果タンパク機能が正常でなくなり病気を引き起こすことがある。ただし、患者の全ゲノム塩基配列解析により大量に報告される一塩基変異の中には、病気に関係するのか不明なものも多い。本研究では、これら病的意義の不明な一塩基変異に対し、その病因性を予測するウェブツールであるInMeRF（https://www.med.nagoya-u.ac.jp/neurogenetics/InMeRF/）を開発・公開して論文による報告も行った。

研究成果の学術的意義や社会的意義
本研究では、遺伝子配列の中でアミノ酸の変化を引き起こす一塩基変異に対し、その病因性を予測するウェブツールであるInMeRF（https://www.med.nagoya-u.ac.jp/neurogenetics/InMeRF/）を構築した。InMeRFは比較可能な既存の9つのツールに対し、感度（既知の病因性一塩基変異が病因性だと予測される）・特異度（既知の非病因性一塩基変異が非病因性だと予測される）など7つの項目による性能評価が上回った。InMeRFを用いることにより、患者のゲノムから検出される未知の一塩基変異に対し、病因性かどうかを判断するのに役立つと考えている。

研究成果の概要（英文）：The nucleic acid sequence of a human gene is translated to the protein sequence. Some single nucleotide mutations change the corresponding amino acids, and the changes have the potential to damage the protein functions and may cause a sort of pathogenicity. In recently reported single nucleotide mutations detected from the whole genome sequences of patients, there are many mutations of undetermined significance. In this study, we developed the web service of pathogenicity prediction to the single nucleotide mutations of undetermined significance, and published the paper. The web service is called InMeRF (https://www.med.nagoya-u.ac.jp/neurogenetics/InMeRF/) .

研究分野：ゲノム生物学

キーワード：一塩基変異の病因性予測ツール

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

研究開始当初は、アミノ酸非同義置換（ミスセンス）を引き起こす一塩基変異（single nucleotide variant; SNV）の機能予測ツールが 23 種類報告されており、さらに、SNV の exonic splicing enhancers/silencers (ESE/ESS)、intronic splicing enhancers/silencers (ISE/ISS)への影響を予測するツールは 16 種類報告されていた。しかし、これらの評価指標の精度は満足できるレベルには達しておらず、ツール間での予測結果も異なっていた。また、SNV による ISE 生成・ISS 破断により、正常状態では発現されないエクソンが生成される偽エクソン活性化（pseudoexon activation）と呼ばれる現象が知られているが、偽エクソン活性化を予測するツールは報告されていなかった。次世代シークエンサー（NGS）を用いた大規模な疾患ゲノム解析から得られる SNV には意義不明の変異（variant of unknown significance; VUS）も数多く含まれており、VUS の病因性を明らかにするために SNV の統合的な機能予測システムを構築することは有用だと考えられた。

２．研究の目的

本研究の当初の目的は、(1)ミスセンス SNV の病因性予測ツール開発、(2) エクソン上の SNV のスプライシング効果（ESE/ESS）予測ツール開発、(3) SNV による偽エクソン活性化予測ツールの開発を行い、(4) SNV の機能を予測する統合環境を構築することである。

(1) ミスセンス SNV の病因性予測ツール開発：評価指標の異なる既存の 23 種類の機能予測ツールのスコアを用い、機械学習によりミスセンス SNV の病因性を予測するモデルを作成する。

(2) エクソン上の SNV のスプライシング効果（ESE/ESS）予測ツール開発：これまでに、エクソン上の SNV のスプライシングに対する影響を予測する複数のツールが報告されてきた。従来のツールは、人工的な in vitrro/in cellulo スクリーニング、もしくは文献データベースに基づいている。本研究では、ヒト肺がん培養細胞 26 種の同一細胞株の全ゲノム配列決定（Whole-genome sequencing; WGS）と RNA-seq (Suzuki et al. Nucleic Acids Res. 2014)の統合解析（共に in vivo）によって、エクソン上のスプライシング SNV を同定する。

(3) SNV による偽エクソン活性化予測ツール開発：SNV による偽エクソン活性化予測ツールは過去に報告されていない。SNV によって活性化する偽エクソンは、正常細胞・正常組織の RNA-seq 解析を精査すると偽エクソンが少数の転写物において含まれている（personal experiences and communications）。この知見に基づき、正常細胞・正常組織の RNA-seq 解析によりアノテーションされていない偽エクソン候補を同定する。

(4) SNV の機能を予測する統合環境構築：本研究で開発するミスセンス SNV の病因性予測ツールに加え、in vivo から同定したエクソン上のスプライシング SNV と、同じく in vivo から同定した偽エクソン活性化 SNV の予測モデルをそれぞれ作成してこれらを統合することにより、これまでにない情報量を有する網羅的な SNV の機能予測システムを構築する。

３．研究の方法

(1) ミスセンス SNV の病因性予測ツール開発：トレーニングデータ作成のために、病因性 SNV は HGMD Pro から、非病因性 SNV は dbSNP の MAF (minor allelic frequency)>0.001 から抽出した。特徴量は、dbNSFP v4.0a に含まれる 35 種類の既存のミスセンス SNV の機能予測ツールのスコアを用いた。モデルは 150 パターンのアミノ酸置換毎に、ランダムフォレストによって作成した。

(2) エクソン上の SNV のスプライシング効果（ESE/ESS）予測ツール開発：ヒト肺がん培養細胞 26 種の同一細胞株の RNA-seq から、3 エクソン以上で構成される遺伝子の最初と最後以外のエクソンの percent spliced-in (PSI)を計算した。PSI が計算されたエクソンに対し、ヒト肺がん培養細胞 26 種の同一細胞株の WGS によって検出された SNV を含む細胞と含まない細胞に分け（そのエクソンに近いイントロン、特にスプライスサイトに SNV がないことを確認し）、ウィルコクソンの符号順位検定を用いてそのエクソン上の SNV がスプライシングに関係しているかどうかを決めた（$P<0.05$）。

(3) SNV による偽エクソン活性化予測ツール開発：偽エクソン候補の抽出には、ENCODE プロジェクトで公開されたヒト胸部大動脈組織の RNA-seq を用いた。リファレンスのエクソンは AceView に登録されたエクソンを用いた。RNA-seq から検出したエクソンの中で、AceView のエクソンとして登録されておらず、両端をスプライスサイトに挟まれ、長さが 500 塩基以下で、PSI が 0.5 以下の領域を、未アノテーション（unannotated）エクソンとして抽出し、偽エクソン候補とした。

(4) SNV の機能を予測する統合環境構築：(1)のツールに加え、(2)と (3)のモデルを作成し、SNV の機能を予測する統合システムを構築した（予定であった）。

４．研究成果

(1) ミスセンス SNV の病因性予測ツール開発：SNV によりミスセンスを引き起こすアミノ酸置換は 150 パターンある。20 種類のアミノ酸はそれぞれ構造が異なり、性質が近いものもあればかなり違うものもある。ミスセンス SNV に対して、アミノ酸置換による性質の類似度である BLOSUM62 スコアと病因性の関連を調べたところ、置換によるアミノ酸性質の変化が大きいほど有意に病因性になることが明らかとなった（図１）。

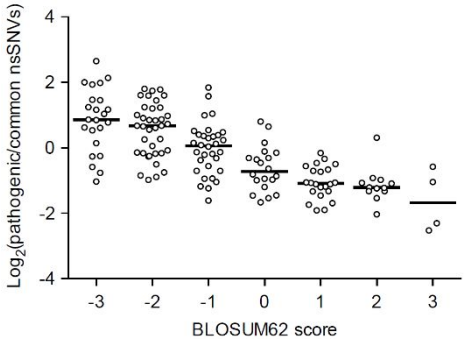

図１　縦軸は病因性のミスセンス SNV の割合（数字が高いほど高い）。横軸はミスセンス SNV によるアミノ酸置換の類似度（数字が低いほど異なる）。Jonckheere–Terpstra trend test (P=0.001)。

よって、アミノ酸置換の種類が予測に大きな影響を及ぼすと考え、150 パターンのアミノ酸置換毎にモデルを作成した。モデルの作成は、scikit-learn ライブラリに含まれるランダムフォレスト(RandomForestClassifier)を用いて Python3 で行った。本ツールの名称を Individual Meta Random Forest (InMeRF)とし、論文を報告して (Takeda et al. NAR Genom Bioinform. 2020)、ウェブサイトを公開した (https://www.med.nagoya-u.ac.jp/neurogenetics/InMeRF/)。InMeRF による、先天性筋無力症 (congenital myasthenic syndrome; CMS)の原因遺伝子の一つである MUSK と、二分脊椎症 (spina bifida)の原因遺伝子の一つである VANGL1 におけるミスセンス SNV の予測結果を図２に示す。

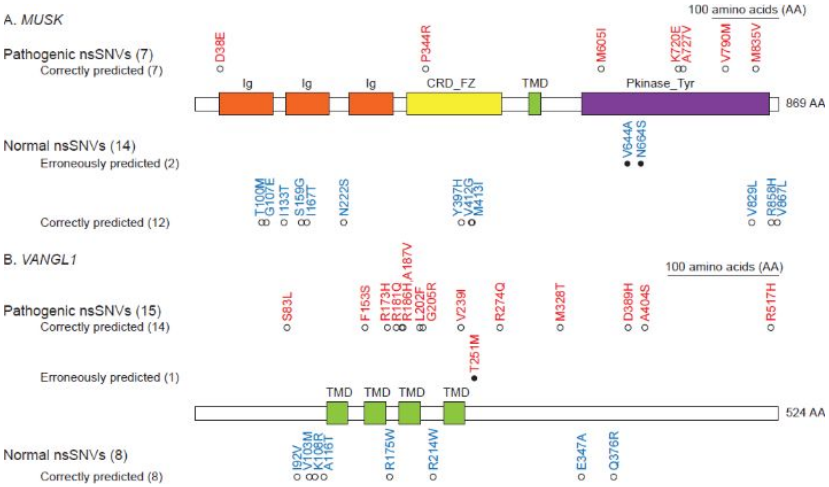

図２　A. MUSK に含まれるミスセンス SNV の予測結果。赤字は病因性として予測され、青字は非病因性として予測されたもの。白丸は正しく予測され、黒丸は間違って予測された印。感度：1.000 (7/7)、特異度：0.857 (12/14)。B. VANGL1 に含まれるミスセンス SNV の予測結果。感度：0.933 (14/15)、特異度：1.000 (8/8)。

(2) エクソン上の SNV のスプライシング効果 (ESE/ESS)予測ツール開発：ヒト肺がん培養細胞 26 種の同一細胞株の WGS と RNA-seq から、スプライシングを引き起こすエクソン上の SNV 候補を検出した。我々は、イントロンの 3’端-50～-3 塩基に存在する SNV のスプライシング効果を予測する IntSplice (Shibata et al. J Hum Genet. 2016)を報告しており、IntSplice と同じ 110 のスプライシングに関するシスエレメントを特徴として用いモデルを作成した。ただし、様々な機械学習手法とハイパーパラメーターを試したが、満足できるモデルを作成することができなかった。

(3) SNV による偽エクソン活性化予測ツール開発：ENCODE プロジェクトで公開されたヒト胸部大動脈組織の RNA-seq と、AceView のアノテーションデータから偽エクソン候補を検出した。 (2)と同様、IntSplice と同じ 110 のスプライシングに関するシスエレメントを特徴として用いモデルを作成した。ただし、これも (2)と同様、様々な機械学習手法とハイパーパラメーターを試したが、満足できるモデルを作成することができなかった。

(4) SNV の機能を予測する統合環境構築：本来は、 (1)で作成した InMeRF と、(2)と (3)で作成したツールを統合したシステムを作成する予定であった。 (2)と (3)のツール作成は上手く行かなかったが、IntSplice のアップデート版である IntSplice2 を作成した。モデルの作成は、勾配ブースティングのライブラリである LightGBM を用いて Python3 で行った。IntSplice2 のウェブサイトを公開し (https://www.med.nagoya-u.ac.jp/neurogenetics/IntSplice2/)、論文は投稿中である。 (2)と (3)については、解析する RNA-seq を増やしてデータ数を増やすとともに、特徴量が必要な機械学習から必要のないディープラーニングへの変更 (自然言語処理に用いられる Attention 機構など)も考慮して、ツールの完成を目指し、最終的に (4)のシステムを構築する。

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　0件／うちオープンアクセス　4件）

１．著者名 Jun-ichi Takeda, Kentaro Nanatsue, Ryosuke Yamagishi, Mikako Ito, Nobuhiko Haga, Hiromi Hirata, Tomoo Ogi, Kinji Ohno	４．巻 2
２．論文標題 InMeRF: prediction of pathogenicity of missense variants by individual modeling for each amino acid substitution	５．発行年 2020年
３．雑誌名 NAR Genomics and Bioinformatics	６．最初と最後の頁 lqaa038
掲載論文のDOI（デジタルオブジェクト識別子） 10.1093/nargab/lqaa038	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 Kun Huang, Jin Li, Mikako Ito, Jun-Ichi Takeda, Bisei Ohkawara, Tomoo Ogi, Akio Masuda, Kinji Ohno	４．巻 13
２．論文標題 Gene Expression Profile at the Motor Endplate of the Neuromuscular Junction of Fast-Twitch Muscle	５．発行年 2020年
３．雑誌名 Frontiers in Molecular Neuroscience	６．最初と最後の頁 154
掲載論文のDOI（デジタルオブジェクト識別子） 10.3389/fnmol.2020.00154	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 Akio Masuda, Toshihiko Kawachi, Jun-Ichi Takeda, Bisei Ohkawara, Mikako Ito, Kinji Ohno	４．巻 21
２．論文標題 tRIP-seq reveals repression of premature polyadenylation by co-transcriptional FUS-U1 snRNP assembly	５．発行年 2020年
３．雑誌名 EMBO Reports	６．最初と最後の頁 e49890
掲載論文のDOI（デジタルオブジェクト識別子） 10.15252/embr.201949890	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 TSUDA Takao、NONOME Tomoaki、GOTO Sae、TAKEDA Jun-ichi、TSUNODA Makoto、HIRAYAMA Masaaki、OHNO Kinji	４．巻 40
２．論文標題 Application of Skin Gas GC/MS Analysis for Prediction of the Severity Scale of Parkinson’s Disease	５．発行年 2019年
３．雑誌名 CHROMATOGRAPHY	６．最初と最後の頁 149～155
掲載論文のDOI（デジタルオブジェクト識別子） 10.15583/jpchrom.2019.014	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 －

〔学会発表〕　計3件（うち招待講演　0件／うち国際学会　1件）

1．発表者名 武田淳一、伊藤美佳子 、大野欽司
2．発表標題 ミスセンスSNVから表現型の病因性を予測する新規ツールの開発
3．学会等名 第8回生理研・名大医合同シンポジウム
4．発表年 2018年

1．発表者名 武田淳一、伊藤美佳子 、大野欽司
2．発表標題 ミスセンスSNVから表現型の病因性を予測する新規ツールの開発
3．学会等名 第41回日本分子生物学会
4．発表年 2018年

1．発表者名 J. Takeda, K. Nanatsue, R. Yamagishi, M. Ito, K. Ohno
2．発表標題 InMeRF: A set of random forest models for each amino acid substitution to predict pathogenicity of missense variants in the human genome
3．学会等名 ESHG2020.2（国際学会）
4．発表年 2020年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

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

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