(19)日本国特許庁(JP)　(12)公開特許公報(A)　(11)特許出願公開番号
特開2021-7035
(P2021-7035A)
(43)公開日　令和3年1月21日(2021.1.21)

(51)Int.Cl.		F I		テーマコード（参考）
G16B 40/30	***(2019.01)***	G16B	40/30	
G06N 3/02	***(2006.01)***	G06N	3/02	ZNA
G06N 3/04	***(2006.01)***	G06N	3/04	
C12N 15/09	*(2006.01)*	C12N	15/09	100

審査請求 有　請求項の数 17　ＯＬ　外国語出願　（全 141 頁）

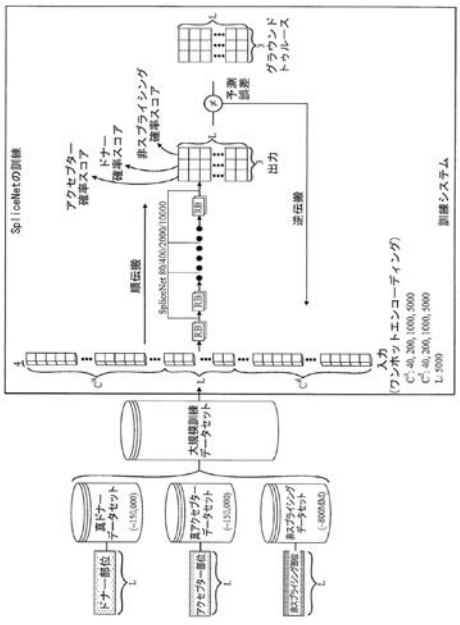
(21)出願番号　特願2020-163488 (P2020-163488)
(22)出願日　令和2年9月29日 (2020.9.29)
(62)分割の表示　特願2019-567663 (P2019-567663)の分割
原出願日　平成30年10月15日 (2018.10.15)
(31)優先権主張番号　62/573,125
(32)優先日　平成29年10月16日 (2017.10.16)
(33)優先権主張国・地域又は機関　米国 (US)
(31)優先権主張番号　62/573,131
(32)優先日　平成29年10月16日 (2017.10.16)
(33)優先権主張国・地域又は機関　米国 (US)

(71)出願人　500358711
イルミナ　インコーポレイテッド
アメリカ合衆国　カリフォルニア州　９２１２２　サンディエゴ　イルミナ　ウェイ　５２００
(74)代理人　100108453
弁理士　村山　靖彦
(74)代理人　100110364
弁理士　実広　信哉
(74)代理人　100133400
弁理士　阿部　達彦
(72)発明者　キショール・ジャガナタン
アメリカ合衆国・カリフォルニア・９２１２２・サン・ディエゴ・イルミナ・ウェイ・５２００

最終頁に続く

(54)【発明の名称】ディープラーニングベースのスプライス部位分類

(57)【要約】
【課題】バリアント分類のための畳み込みニューラルネットワークベースの分類器を構築する。
【解決手段】特に、開示される技術は、分類器の出力を対応するグラウンドトゥルース標識と徐々にマッチさせる逆伝搬ベースの勾配更新技術を使用して、訓練データで畳み込みニューラルネットワークベースの分類器を訓練することに関する。畳み込みニューラルネットワークベースの分類器は残差ブロックのグループを含み、残差ブロックの各グループは、残差ブロック内の畳み込みフィルタの数、残差ブロックの畳み込みウィンドウサイズ、および残差ブロックのAtrous畳み込みレートによってパラメータ化され、畳み込みウィンドウのサイズは、残差ブロックのグループの間で異なり、Atrous畳み込みレートは、残差ブロックのグループの間で異なる。訓練データは、良性バリアントおよび病原性バリアントから生成された翻訳済み配列対の良性訓練例および病原性訓練例を含む。
【選択図】図３０



【特許請求の範囲】
【請求項１】
pre-mRNAゲノム配列内のスプライス部位の可能性を予測するためのコンピュータ実施方法であって、
pre-mRNAヌクレオチド配列の訓練例でAtrous畳み込みニューラルネットワーク(ACNN)を訓練するステップであって、前記訓練例がドナースプライス部位の少なくとも50,000個の訓練例、アクセプタースプライス部位の少なくとも50,000個の訓練例、および非スプライシング部位の少なくとも100,000個の訓練例を含む、ステップを含み、
前記訓練するステップは、
前記ヌクレオチド配列のワンホットエンコードされた訓練例を入力するステップであって、各ヌクレオチド配列が少なくとも401個のヌクレオチドを含み、前記少なくとも401個のヌクレオチドが、少なくとも1つの標的ヌクレオチドと、前記標的ヌクレオチドの上流および下流の各側の少なくとも200個の隣接ヌクレオチドの構成とを含む、ステップと、
逆伝搬によって、前記ACNNのフィルタのパラメータを調整して、前記ヌクレオチド配列内の各標的ヌクレオチドがドナースプライス部位、アクセプタースプライス部位、または非スプライシング部位である可能性に対するスコアを予測するステップと
を含み、
それにより、訓練されたACNNは、ワンホットエンコードされた、少なくとも1つの標的ヌクレオチドおよび各側の少なくとも200個の隣接ヌクレオチドの構成を含む少なくとも401個のヌクレオチドからなるpre-mRNAヌクレオチド配列を入力として受け入れるとともに、前記標的ヌクレオチドがドナースプライス部位、アクセプタースプライス部位、または非スプライシング部位である可能性のスコアを決定するように構成される、コンピュータ実施方法。
【請求項２】
pre-mRNAヌクレオチド配列の前記訓練例および入力されるpre-mRNAヌクレオチド配列は、各々、前記標的ヌクレオチドの各側の2,500個の隣接ヌクレオチドを含み、それにより、前記訓練されたACNNは、少なくとも5,001個のヌクレオチドからなるpre-mRNAヌクレオチド配列を入力として受け入れるように構成される、請求項1に記載のコンピュータ実施方法。
【請求項３】
pre-mRNAヌクレオチド配列の前記訓練例および入力されるpre-mRNAヌクレオチド配列には、各々、前記標的ヌクレオチドの5,000個の上流構成ヌクレオチドおよび5,000個の下流構成ヌクレオチドが隣接し、それにより、前記訓練されたACNNは、少なくとも10,001個のヌクレオチドからなるpre-mRNAヌクレオチド配列を入力として受け入れるように構成される、請求項1に記載のコンピュータ実施方法。
【請求項４】
pre-mRNAヌクレオチド配列の前記訓練例および入力されるpre-mRNAヌクレオチド配列は、各々、各側に500個のヌクレオチドが隣接する前記標的ヌクレオチドを含む、請求項1に記載のコンピュータ実施方法。
【請求項５】
pre-mRNAヌクレオチド配列の前記訓練例および入力されるpre-mRNAヌクレオチド配列は、各々、1,000個の上流構成ヌクレオチドおよび1,000個の下流構成ヌクレオチドが隣接する前記標的ヌクレオチドを含む、請求項1に記載のコンピュータ実施方法。
【請求項６】
ドナースプライス部位の少なくとも150,000個の訓練例、アクセプタースプライス部位の少なくとも150,000個の訓練例、および非スプライシング部位の少なくとも800,000,000個の訓練例で前記ACNNを訓練するステップをさらに含む、請求項1から5のいずれか一項に記載のコンピュータ実施方法。
【請求項７】
前記ACNNは、残差ブロックのグループを含む、請求項1から6のいずれか一項に記載のコ

ンピュータ実施方法。
【請求項８】
残差ブロックの各グループは、残差ブロック内の畳み込みフィルタの数、残差ブロックの畳み込みウィンドウのサイズ、および残差ブロックの拡張係数によってパラメータ化される、請求項7に記載のコンピュータ実施方法。
【請求項９】
前記拡張係数は、前記残差ブロックのグループ間で非指数関数的に変化する、請求項8に記載のコンピュータ実施方法。
【請求項１０】
畳み込みウィンドウの前記サイズは、残差ブロックのグループ間で異なる、請求項8または9に記載のコンピュータ実施方法。
【請求項１１】
前記ACNNは、4つの残差ブロックおよび少なくとも1つのスキップコネクションからなる少なくとも1つのグループを含み、各残差ブロックは、32個の畳み込みフィルタ、畳み込みウィンドウサイズ11、および拡張係数1を有する、請求項1から9のいずれか一項に記載のコンピュータ実施方法。
【請求項１２】
前記ACNNは、500個の上流構成ヌクレオチドおよび500個の下流構成ヌクレオチドが隣接する前記標的ヌクレオチドを含む入力で訓練され、当該入力を評価するように構成され、4つの残差ブロックおよび少なくとも2つのスキップコネクションからなる少なくとも2つのグループをさらに含み、
前記2つのグループのうち、第1のグループ内の各残差ブロックは、32個の畳み込みフィルタ、畳み込みウィンドウサイズ11、および拡張係数1を有し、第2のグループ内の各残差ブロックは、32個の畳み込みフィルタ、畳み込みウィンドウサイズ11、および拡張係数4を有する、請求項1に記載のコンピュータ実施方法。
【請求項１３】
前記ACNNは、1,000個の上流構成ヌクレオチドおよび1,000個の下流構成ヌクレオチドが隣接する前記標的ヌクレオチドを含む入力で訓練され、当該入力を評価するように構成され、4つの残差ブロックおよび少なくとも3つのスキップコネクションからなる少なくとも3つのグループをさらに含み、
前記3つのグループのうち、第1のグループ内の各残差ブロックは、32個の畳み込みフィルタ、畳み込みウィンドウサイズ11、および拡張係数1を有し、第2のグループ内の各残差ブロックは、32個の畳み込みフィルタ、畳み込みウィンドウサイズ11、および拡張係数4を有し、第3のグループ内の各残差ブロックは、32個の畳み込みフィルタ、畳み込みウィンドウサイズ21、および拡張係数19を有する、請求項1に記載のコンピュータ実施方法。
【請求項１４】
前記ACNNは、5,000個の上流構成ヌクレオチドおよび5,000個の下流構成ヌクレオチドが隣接する前記標的ヌクレオチドを含む入力で訓練され、当該入力を評価するように構成され、4つの残差ブロックおよび少なくとも4つのスキップコネクションからなる少なくとも4つのグループをさらに含み、
前記4つのグループのうち、第1のグループ内の各残差ブロックは、32個の畳み込みフィルタ、畳み込みウィンドウサイズ11、および拡張係数1を有し、第2のグループ内の各残差ブロックは、32個の畳み込みフィルタ、畳み込みウィンドウサイズ11、および拡張係数4を有し、第3のグループ内の各残差ブロックは、32個の畳み込みフィルタ、畳み込みウィンドウサイズ21、および拡張係数19を有し、第4のグループ内の各残差ブロックは、32個の畳み込みフィルタ、畳み込みウィンドウサイズ41、および拡張係数25を有する、請求項1に記載のコンピュータ実施方法。
【請求項１５】
pre-mRNAゲノム配列内のスプライス部位の可能性を予測するための装置であって、
コンピュータ命令を格納したメモリと、

前記メモリに接続されたプロセッサと
を備え、
前記コンピュータ命令は、前記プロセッサによって実行されると、前記プロセッサに請求項1から14のいずれか一項に記載の方法の各ステップを実行させる、装置。
【請求項１６】
コンピュータに請求項1から14のいずれか一項に記載の方法の各ステップを実行させるためのコンピュータプログラムを記録したコンピュータ読み取り可能な記録媒体。
【請求項１７】
コンピュータに請求項1から14のいずれか一項に記載の方法の各ステップを実行させるためのコンピュータプログラム。
【発明の詳細な説明】
【技術分野】
【０００１】
付記
この付記は、本発明者らによる論文に列挙された場合によっては関連のある参考文献の書誌を含む。この論文の主題は、本出願が優先権/利益を主張する米国仮出願に記載されている。これらの参考文献は、要求に応じて代理人によって参照することができ、またはグローバルドシエを介して参照されてもよい。
【０００２】
優先出願
本出願は、2017年10月16日に出願された、Kishore Jaganathan、Kai-How Farh、Sofia Kyriazopoulou Panagiotopoulou、およびJeremy Francis McRaeによる「Deep Learning-Based Splice Site Classification」という名称の米国仮特許出願第62/573,125号(整理番号ILLM 1001-1/IP-1610-PRV)、2017年10月16日に出願された、Kishore Jaganathan、Kai-How Farh、Sofia Kyriazopoulou Panagiotopoulou、およびJeremy Francis McRaeによる「Deep Learning-Based Aberrant Splicing Detection」という名称の米国仮特許出願第62/573,131号(整理番号ILLM 1001-2/IP-1614-PRV)、2017年10月16日に出願された、Kishore Jaganathan、Kai-How Farh、Sofia Kyriazopoulou Panagiotopoulou、およびJeremy Francis McRaeによる「Aberrant Splicing Detection Using Convolutional Neural Networks (CNNs)」という名称の米国仮特許出願第62/573,135号(整理番号ILLM 1001-3/IP-1615-PRV)、ならびに2018年8月31日に出願された、Kishore Jaganathan、Kai-How Farh、Sofia Kyriazopoulou Panagiotopoulou、およびJeremy Francis McRaeによる「Predicting Splicing from Primary Sequence with Deep Learning」という名称の米国仮特許出願第62/726,158号(整理番号ILLM 1001-10/IP-1749-PRV)の優先権または利益を主張する。仮出願は、すべての目的に関して参照により本明細書に組み込まれる。
【０００３】
組込み
以下の文献は、あたかも全体が本明細書に記載されているかのように、すべての目的に関して参照により組み込まれている。
【０００４】
2018年10月15日に同時出願され(整理番号ILLM 1001-8/IP-1614-PCT)、その後PCT国際公開第WO2019/79200号として公開された、Kishore Jaganathan、Kai-How Farh、Sofia Kyriazopoulou Panagiotopoulou、およびJeremy Francis McRaeによる「Deep Learning-Based Aberrant Splicing Detection」という名称のPCT特許出願第PCT/US18/55919号。
【０００５】
2018年10月15日に同時出願され(整理番号ILLM 1001-9/IP-1615-PCT)、その後PCT国際公開第WO2019/79202号として公開された、Kishore Jaganathan、Kai-How Farh、Sofia Kyriazopoulou Panagiotopoulou、およびJeremy Francis McRaeによる「Aberrant Splicing Detection Using Convolutional Neural Networks (CNNs)」という名称のPCT特許出願第PCT/US18/55923号。

【０００６】
同時出願された、Kishore Jaganathan、Kai-How Farh、Sofia Kyriazopoulou Panagiotopoulou、およびJeremy Francis McRaeによる「Deep Learning-Based Splice Site Classification」という名称の米国非仮特許出願(整理番号ILLM 1001-4/IP-1610-US)。
【０００７】
同時出願された、Kishore Jaganathan、Kai-How Farh、Sofia Kyriazopoulou Panagiotopoulou、およびJeremy Francis McRaeによる「Deep Learning-Based Aberrant Splicing Detection」という名称の米国非仮特許出願(整理番号ILLM 1001-5/IP-1614-US)。
【０００８】
同時出願された、Kishore Jaganathan、Kai-How Farh、Sofia Kyriazopoulou Panagiotopoulou、およびJeremy Francis McRaeによる「Aberrant Splicing Detection Using Convolutional Neural Networks (CNNs)」という名称の米国非仮特許出願(整理番号ILLM 1001-6/IP-1615-US)。
【０００９】
文献1-S. Dieleman、H. Zen、K. Simonyan、O. Vinyals、A. Graves、N. Kalchbrenner、A. Senior、およびK. Kavukcuoglu、「WAVENET: A GENERATIVE MODEL FOR RAW AUDIO」、arXiv:1609.03499、2016、
【００１０】
文献2-S. O. Arik、M. Chrzanowski、A. Coates、G. Diamos、A. Gibiansky、Y. Kang、X. Li, J. Miller、A. Ng、J. Raiman、S. Sengupta、およびM. Shoeybi、「DEEP VOICE: REAL-TIME NEURAL TEXT-TO-SPEECH」、arXiv:1702.07825、2017、
【００１１】
文献3-F. YuおよびV. Koltun、「MULTI-SCALE CONTEXT AGGREGATION BY DILATED CONVOLUTIONS」、arXiv:1511.07122、2016、
【００１２】
文献4-K. He, X. Zhang、S. Ren、およびJ. Sun、「DEEP RESIDUAL LEARNING FOR IMAGE RECOGNITION」、arXiv:1512.03385、2015、
【００１３】
文献5-R. K. Srivastava、K. Greff、およびJ. Schmidhuber、「HIGHWAY NETWORKS」、arXiv: 1505.00387、2015、
【００１４】
文献6-G. Huang、Z. Liu、L. van der Maaten、およびK. Q. Weinberger、「DENSELY CONNECTED CONVOLUTIONAL NETWORKS」、arXiv:1608.06993、2017、
【００１５】
文献7-C. Szegedy、W. Liu、Y. Jia、P. Sermanet、S. Reed、D. Anguelov、D. Erhan、V. Vanhoucke、およびA. Rabinovich、「GOING DEEPER WITH CONVOLUTIONS」、arXiv:1409.4842、2014、
【００１６】
文献8-S. IoffeおよびC. Szegedy、「BATCH NORMALIZATION: ACCELERATING DEEP NETWORK TRAINING BY REDUCING INTERNAL COVARIATE SHIFT」、arXiv: 1502.03167、2015、
【００１７】
文献9-J. M. Wolterink、T. Leiner、M. A. Viergever、およびI. Isgum、「DILATED CONVOLUTIONAL NEURAL NETWORKS FOR CARDIOVASCULAR MR SEGMENTATION IN CONGENITAL HEART DISEASE」、arXiv:1704.03669、2017、
【００１８】
文献10-L. C. Piqueras、「AUTOREGRESSIVE MODEL BASED ON A DEEP CONVOLUTIONAL NEURAL NETWORK FOR AUDIO GENERATION」、Tampere University of Technology、2016、
【００１９】
文献11-J. Wu、「Introduction to Convolutional Neural Networks」、Nanjing University、2017、

【００２０】
文献12-I. J. Goodfellow、D. Warde-Farley、M. Mirza、A. Courville、およびY. Bengio、「CONVOLUTIONAL NETWORKS」、Deep Learning、MIT Press、2016、ならびに
【００２１】
文献13-J. Gu、Z. Wang、J. Kuen、L. Ma、A. Shahroudy、B. Shuai、T. Liu、X. Wang、およびG. Wang、「RECENT ADVANCES IN CONVOLUTIONAL NEURAL NETWORKS」、arXiv:1512.07108、2017。
【００２２】
文献1は、同じ畳み込みウィンドウサイズを有する畳み込みフィルタと、バッチ正規化層と、正規化線形ユニット(略語ReLU)層と、次元変換層と、指数関数的に高まるAtrous畳み込みレートを有するAtrous畳み込み層と、スキップコネクションと、ソフトマックス分類層とを含む残差ブロックのグループを使用して、入力配列を受け入れ、入力配列内のエントリをスコア化する出力配列を生成する深層畳み込みニューラルネットワークアーキテクチャを記載している。開示された技術は、文献1に記載されているニューラルネットワークコンポーネントおよびパラメータを使用する。一実装形態では、開示された技術は、文献1に記載されたニューラルネットワークコンポーネントのパラメータを修正する。たとえば、文献1とは異なり、開示された技術におけるAtrous畳み込みレートは、下位の残差ブロックグループから上位の残差ブロックグループへと非指数関数的に高まる。別の例では、文献1とは異なり、開示された技術における畳み込みウィンドウサイズは残差ブロックのグループ間で異なる。
【００２３】
文献2は、文献1に記載された深層畳み込みニューラルネットワークアーキテクチャについて詳細に説明している。
【００２４】
文献3は、開示された技術によって使用されるAtrous畳み込みについて説明している。本明細書では、Atrous畳み込みを「Dilated畳み込み」とも呼ぶ。Atrous/Dilated畳み込みは、訓練可能なパラメータが少ししかない大きい受容野を実現可能にする。Atrous/Dilated畳み込みは、Atrous畳み込みレートまたは拡張係数とも呼ばれるあるステップを用いて入力値をスキップすることによって、カーネルがその長さよりも大きい領域にわたって適用される畳み込みである。Atrous/Dilated畳み込みは、畳み込みフィルタ/カーネルの要素間の間隔を加え、それによって、畳み込み演算が実行されるときにより大きい間隔における近傍の入力エントリ(たとえば、ヌクレオチド、アミノ酸)が考慮される。これは長距離構成依存性を入力に組み込むことを可能にする。Atrous畳み込みは、部分畳み込み計算を隣接するヌクレオチドが処理されるときに再使用できるように保存する。
【００２５】
文献4は、開示された技術によって使用される残差ブロックおよび残差コネクションについて説明している。
【００２６】
文献5は、開示された技術によって使用されるスキップコネクションについて説明している。本明細書では、スキップコネクションを「ハイウェイネットワーク」とも呼ぶ。
【００２７】
文献6は、開示された技術によって使用される密結合畳み込みネットワークアーキテクチャについて説明している。
【００２８】
文献7は、開示された技術によって使用される次元変換畳み込み層およびモジュールベース処理パイプラインについて説明している。次元変換畳み込みの一例は1×1畳み込みである。
【００２９】
文献8は、開示された技術によって使用されるバッチ正規化層について説明している。
【００３０】

文献9も、開示された技術によって使用されるAtrous/Dilated畳み込みについて説明している。

【００３１】

文献10は、畳み込みニューラルネットワーク、深層畳み込みニューラルネットワーク、およびAtrous/Dilated畳み込みを伴う深層畳み込みニューラルネットワークを含む、開示された技術によって使用することのできる深層ニューラルネットワークの様々なアーキテクチャについて説明している。

【００３２】

文献11は、サブサンプリング層(たとえば、プーリング)および全結合層を含む畳み込みニューラルネットワークを訓練するためのアルゴリズムを含む、開示された技術によって使用することのできる畳み込みニューラルネットワークについて詳細に説明している。

【００３３】

文献12は、開示された技術によって使用することのできる様々な畳み込み演算について詳細に説明している。

【００３４】

文献13は、開示された技術によって使用することのできる畳み込みニューラルネットワークの様々なアーキテクチャについて説明している。

【００３５】

出願とともに電子的に提出された参照テーブルによる組込み

ASCIIテキストフォーマットの3つのテーブルファイルが本出願とともに提出されており、参照によって組み込まれている。ファイルの名称、作成日、およびサイズは以下の通りである。

【００３６】

table_S4_mutation_rates.txt	2018年8月31日	2,452KB

【００３７】

table_S5_gene_enrichment.txt	2018年8月31日	362KB

【００３８】

table_S6_validation.txt	2018年8月31日	362KB

【００３９】

開示された技術は、知能をエミュレートするための人工知能タイプのコンピュータおよびデジタルデータ処理システムならびに対応するデータ処理方法および製品(すなわち、知識ベースシステム、推論システム、および知識獲得システム)に関し、不確かさによって推論するためのシステム(たとえば、ファジー論理システム)、適応システム、機械学習システム、および人工ニューラルネットワークを含む。特に、開示された技術は、ディープラーニングベースの技法を使用して深層畳み込みニューラルネットワークを訓練することに関する。

【背景技術】

【００４０】

本節で説明されている主題は、本節において従来技術について言及した結果として単に従来技術であると仮定されるべきではない。同様に、本節で言及されている、または背景技術として提示される主題に関連付けられている問題は、従来技術においてすでに認識されている仮定されるべきではない。本節における主題は、単に異なるアプローチを表しているだけであり、それら自体で、請求される技術の実装形態に対応していてもよい。

【００４１】

機械学習

機械学習では、入力変数が出力変数を予測するために使用される。入力変数は、特徴と呼ばれることが多く、$X=(X_1, X_2, ..., X_k)$によって示され、この場合、各X_i、$i \in 1$、…、kが特徴である。出力変数は、応答変数または従属変数と呼ばれることが多く、変数Y_iによって示される。Yと対応するXとの関係は次の一般式で表すことができる。

$Y=f(X)+\in$

【００４２】
上式において、fは特徴(X_1、X_2、...、X_k)の関数であり、∈はランダム誤差項である。誤差項はXから独立しており、平均値がゼロである。
【００４３】
実際には、特徴Xは、Yを有することもなしに、またはXとYとの間に厳密な関係を知ることもなしに利用可能である。誤差項は平均値がゼロであるので、目標はfを推定することである。
【００４４】
【数１】

$$\hat{Y}=\hat{f}=(X)$$

【００４５】
上記の式において、
【数２】

$$\hat{f}$$

は∈の推定値であり、ブラックボックスとみなされることが多く、すなわち、
【数３】

$$\hat{f}$$

の入力と出力との間の関係のみが知られており、この値が作用する理由は不明である。
【００４６】
関数
【数４】

$$\hat{f}$$

は学習を使用して見つけられる。教師あり学習および教師なし学習が、このタスクに対する機械学習において使用される2つの方法である。教師あり学習では、ラベル付きデータが訓練に使用される。入力および対応する出力(=ラベル)を示すことによって、関数
【数５】

$$\hat{f}$$

は、出力を近似するように最適化される。教師なし学習では、目標は、ラベルなしデータから隠れ構造を見つけることである。このアルゴリズムは、入力データに対する精度の尺度を有さず、それによって教師あり学習と区別される。

【００４７】
ニューラルネットワーク
単層パーセプトロン(SLP)は、ニューラルネットワークの最も単純なモデルである。単層パーセプトロンは、図1に示されているように1つの入力層と1つの活性化関数とを備える。入力は、重み付きグラフに通される。関数fは、入力の和を引数として使用し、これを閾値θと比較する。
【００４８】
図2は、複数の層を含む全結合されたニューラルネットワークの一実装形態を示す。ニューラルネットワークは、互いにメッセージを交換する相互接続された人工ニューロン(たとえば、a_1、a_2、a_3)のシステムである。例示されているニューラルネットワークは、3つの入力と、隠れ層内の2つのニューロンと、出力層内の2つのニューロンとを有する。隠れ層は、活性化関数f(・)を有し、出力層は活性化関数g(・)を有する。各接続線は、訓練プロセスの実行中に調整される数値重み(たとえば、w_{11}、w_{21}、w_{12}、w_{31}、w_{22}、w_{32}、v_{11}、v_{22})を有し、それによって、適切に訓練されたネットワークは、認識すべき画像が供給されたときに正しく応答する。入力層は、未加工入力を処理し、隠れ層は、入力層と隠れ層との間の接続線の重みに基づき入力層からの出力を処理する。出力層は、隠れ層から出力をとり、隠れ層と出力層との間の接続線の重みに基づき処理する。ネットワークは、特徴検出ニューロンの複数の層を含む。各層は、前の層からの入力の異なる組合せに応答する多数のニューロンを有する。これらの層は、第1の層が入力された画像データにおけるプリミティブパターンのセットを検出し、第2の層がパターンのパターンを検出し、第3の層がそれらのパターンのパターンを検出する。
【００４９】
ゲノミクスにおけるディープラーニングの応用についての調査は、以下の文献に記載されている。
T. Chingら, Opportunities And Obstacles For Deep Learning In Biology And Medicine, www.biorxiv.org:142760, 2017、
Angermueller C、Parnamaa T、Parts L、Stegle O. Deep Learning For Computational Biology. Mol Syst Biol. 2016;12:878、
Park Y、Kellis M. 2015 Deep Learning For Regulatory Genomics. Nat. Biotechnol. 33, 825-826頁. (doi:10.1038/nbt.3313)、
Min, S.、Lee, B.およびYoon, S. Deep Learning In Bioinformatics. Brief. Bioinform. bbw068 (2016)、
Leung MK、Delong A、Alipanahi Bら Machine Learning In Genomic Medicine: A Review of Computational Problems and Data Sets 2016、ならびに
Libbrecht MW、Noble WS. Machine Learning Applications In Genetics and Genomics. Nature Reviews Genetics 2015;16(6):321-32。
【先行技術文献】
【特許文献】
【００５０】
【特許文献１】国際公開第07/010252号
【特許文献２】国際出願第2007/003798号
【特許文献３】米国特許出願公開第2009/0088327号明細書
【特許文献４】米国特許出願公開第2016/0085910号明細書
【特許文献５】米国特許出願公開第2013/0296175号明細書
【特許文献６】国際公開第04/018497号
【特許文献７】米国特許第7057026号明細書
【特許文献８】国際公開第91/06678号
【特許文献９】国際公開第07/123744号
【特許文献１０】米国特許第7329492号明細書
【特許文献１１】米国特許第7211414号明細書

【特許文献１２】米国特許第7315019号明細書
【特許文献１３】米国特許第7405281号明細書
【特許文献１４】米国特許出願公開第2008/0108082号明細書
【特許文献１５】米国特許第5641658号明細書
【特許文献１６】米国特許出願公開第2002/0055100号明細書
【特許文献１７】米国特許第7115400号明細書
【特許文献１８】米国特許出願公開第2004/0096853号明細書
【特許文献１９】米国特許出願公開第2004/0002090号明細書
【特許文献２０】米国特許出願公開第2007/0128624号明細書
【特許文献２１】米国特許出願公開第2008/0009420号明細書
【特許文献２２】米国特許出願公開第2007/0099208号明細書
【特許文献２３】国際公開第04/018497号
【特許文献２４】米国特許出願公開第2007/0166705号明細書
【特許文献２５】米国特許第7057026号明細書
【特許文献２６】米国特許出願公開第2008/0280773号明細書
【特許文献２７】米国特許出願第13/018255号
【特許文献２８】国際公開第00/4018497号
【特許文献２９】米国特許出願公開第2007/0166705号明細書
【特許文献３０】米国特許第7057026号明細書
【特許文献３１】国際特許出願第2013/030867号
【特許文献３２】国際公開第2014/142831号
【非特許文献】
【００５１】
【非特許文献１】T. Chingら, Opportunities And Obstacles For Deep Learning In Biology And Medicine, www.biorxiv.org:142760, 2017
【非特許文献２】Angermueller C、Parnamaa T、Parts L、Stegle O. Deep Learning For Computational Biology. Mol Syst Biol. 2016;12:878
【非特許文献３】Park Y、Kellis M. 2015 Deep Learning For Regulatory Genomics. Nat. Biotechnol. 33, 825-826頁. (doi:10.1038/nbt.3313)
【非特許文献４】Min, S.、Lee, B.およびYoon, S. Deep Learning In Bioinformatics. Brief. Bioinform. bbw068 (2016)
【非特許文献５】Leung MK、Delong A、Alipanahi Bら Machine Learning In Genomic Medicine: A Review of Computational Problems and Data Sets 2016
【非特許文献６】Libbrecht MW、Noble WS. Machine Learning Applications In Genetics and Genomics. Nature Reviews Genetics 2015;16(6):321-32
【非特許文献７】Bentleyら、Nature 456:53-59 (2008年)
【非特許文献８】Lizardiら、Nat. Genet. 19:225-232(1998年)
【非特許文献９】Dunn、Tamsen & Berry、Gwenn & Emig-Agius、Dorothea & Jiang、Yu & Iyer、Anita & Udar、Nitin & Stromberg、Michael. (2017). Pisces: An Accurate and Versatile Single Sample Somatic and Germline Variant Caller. 595-595. 10.1145/3107411.3108203
【非特許文献１０】T Saunders、Christopher & Wong、Wendy & Swamy、Sajani & Becq、Jennifer & J Murray、Lisa & Cheetham, Keira. (2012). Strelka: Accurate somatic small-variant calling from sequenced tumor-normal sample pairs. Bioinformatics (Oxford, England). 28. 1811-7. 10.1093/bioinformatics/bts271
【非特許文献１１】Kim, S.、Scheffler, K.、Halpern, A.L.、Bekritsky, M.A.、Noh, E.、Kallberg, M.、Chen, X.、Beyter, D.、Krusche, P.、およびSaunders, C.T. (2017). Strelka2: Fast and accurate variant calling for clinical sequencing applications
【非特許文献１２】Stromberg、Michael & Roy、Rajat & Lajugie、Julien & Jiang、Yu

& Li、Haochen & Margulies、Elliott. (2017). Nirvana: Clinical Grade Variant Annotator. 596-596. 10.1145/3107411.3108204
【非特許文献１３】Iossifovら、Nature 2014年

【発明の概要】
【課題を解決するための手段】
【００５２】
図面において、類似の参照文字は概して、異なる図全体にわたって類似の部分を指す。さらに、図面は必ずしも一定の縮尺で描かれているとは限らず、一般に開示された技術の原則を示すときには強調される。以下の説明では、開示された技術の様々な実装形態について以下の図面を参照しながら説明する。

【図面の簡単な説明】
【００５３】
【図１】単層パーセプトロン(SLP)を示す図である。
【図２】複数の層を含むフィードフォワードニューラルネットワークの一実装形態を示す図である。
【図３】畳み込みニューラルネットワークの機能の一実装形態を示す図である。
【図４】開示された技術の一実装形態による畳み込みニューラルネットワークの訓練のブロック図である。
【図５】開示された技術の一実装形態によるReLU非線形層の一実装形態を示す図である。
【図６】Dilated畳み込みを示す図である。
【図７】開示された技術の一実装形態によるサブサンプリング層(平均/最大プーリング)の一実装形態を示す図である。
【図８】畳み込み層の2層畳み込みの一実装形態を示す図である。
【図９】特徴マップ追加を介して事前情報を下流側に再注入する残差コネクションを示す図である。
【図１０】残差ブロックおよびスキップコネクションの一実装形態を示す図である。
【図１１】Stacked Dilated畳み込みの一実装形態を示す図である。
【図１２】バッチ正規化フォワードパスを示す図である。
【図１３】テスト時におけるバッチ正規化変換を示す図である。
【図１４】バッチ正規化バックワードパスを示す図である。
【図１５】バッチ正規化層を畳み込み結合層または密結合層とともに使用した状態を示す図である。
【図１６】1D畳み込みの一実装形態を示す図である。
【図１７】グローバルアベレージプーリング(GAP)がどのように作用するかを示す図である。
【図１８】開示された技術を実装するのに使用することができる訓練サーバおよびプロダクションサーバを含むコンピューティング環境の一実装形態を示す図である。
【図１９】本明細書において「SpliceNet」と呼ばれるAtrous畳み込みニューラルネットワーク(略語ACNN)のアーキテクチャの一実装形態を示す図である。
【図２０】ACNNおよび畳み込みニューラルネットワーク(略語CNN)によって使用することのできる残差ブロックの一実装形態を示す図である。
【図２１】本明細書において「SpliceNet80」と呼ばれるACNNのアーキテクチャの別の実装形態を示す図である。
【図２２】本明細書において「SpliceNet400」と呼ばれるACNNのアーキテクチャのさらに別の実装形態を示す図である。
【図２３】本明細書において「SpliceNet2000」と呼ばれるACNNのアーキテクチャのさらに別の実装形態を示す図である。
【図２４】本明細書において「SpliceNet10000」と呼ばれるACNNのアーキテクチャのさらに別の実装形態を示す図である。
【図２５】ACNNおよびCNNによって処理される様々な種類の入力を示す図である。

【図２６】ACNNおよびCNNによって処理される様々な種類の入力を示す図である。
【図２７】ACNNおよびCNNによって処理される様々な種類の入力を示す図である。
【図２８】少なくとも800万個の非スプライス部位上で訓練することのできるACNNおよび少なくとも100万個の非スプライス部位上で訓練することのできるCNNを示す図である。
【図２９】ワンホットエンコーダを示す図である。
【図３０】ACNNの訓練を示す図である。
【図３１】CNNを示す図である。
【図３２】ACNNおよびCNNの訓練、バリデーション、およびテストを示す図である。
【図３３】参照配列および代替配列を示す図である。
【図３４】異常スプライシング検出を示す図である。
【図３５】スプライス部位分類のためのSpliceNet10000の処理ピラミッドを示す図である。
【図３６】異常スプライス部位検出のためのSpliceNet10000の処理ピラミッドを示す図である。
【図３７Ａ】深層学習によって一次配列からスプライシングを予測する一実装形態を示す図である。
【図３７Ｂ】深層学習によって一次配列からスプライシングを予測する一実装形態を示す図である。
【図３７Ｃ】深層学習によって一次配列からスプライシングを予測する一実装形態を示す図である。
【図３７Ｄ】深層学習によって一次配列からスプライシングを予測する一実装形態を示す図である。
【図３７Ｅ】深層学習によって一次配列からスプライシングを予測する一実装形態を示す図である。
【図３７Ｆ】深層学習によって一次配列からスプライシングを予測する一実装形態を示す図である。
【図３７Ｇ】深層学習によって一次配列からスプライシングを予測する一実装形態を示す図である。
【図３７Ｈ】深層学習によって一次配列からスプライシングを予測する一実装形態を示す図である。
【図３８Ａ】RNA配列データにおける稀少潜在的スプライス突然変異のバリデーションの一実装形態を示す図である。
【図３８Ｂ】RNA配列データにおける稀少潜在的スプライス突然変異のバリデーションの一実装形態を示す図である。
【図３８Ｃ】RNA配列データにおける稀少潜在的スプライス突然変異のバリデーションの一実装形態を示す図である。
【図３８Ｄ】RNA配列データにおける稀少潜在的スプライス突然変異のバリデーションの一実装形態を示す図である。
【図３８Ｅ】RNA配列データにおける稀少潜在的スプライス突然変異のバリデーションの一実装形態を示す図である。
【図３８Ｆ】RNA配列データにおける稀少潜在的スプライス突然変異のバリデーションの一実装形態を示す図である。
【図３８Ｇ】RNA配列データにおける稀少潜在的スプライス突然変異のバリデーションの一実装形態を示す図である。
【図３９Ａ】潜在的スプライスバリアントが頻繁に組織固有の代替的スプライシングを形成する一実装形態を示す図である。
【図３９Ｂ】潜在的スプライスバリアントが頻繁に組織固有の代替的スプライシングを形成する一実装形態を示す図である。
【図３９Ｃ】潜在的スプライスバリアントが頻繁に組織固有の代替的スプライシングを形成する一実装形態を示す図である。

【図４０Ａ】予測される潜在的スプライスバリアントがヒト母集団において強い悪影響を及ぼす一実装形態を示す図である。
【図４０Ｂ】予測される潜在的スプライスバリアントがヒト母集団において強い悪影響を及ぼす一実装形態を示す図である。
【図４０Ｃ】予測される潜在的スプライスバリアントがヒト母集団において強い悪影響を及ぼす一実装形態を示す図である。
【図４０Ｄ】予測される潜在的スプライスバリアントがヒト母集団において強い悪影響を及ぼす一実装形態を示す図である。
【図４０Ｅ】予測される潜在的スプライスバリアントがヒト母集団において強い悪影響を及ぼす一実装形態を示す図である。
【図４１Ａ】稀少遺伝病の患者におけるデノボ潜在的スプライス突然変異の一実装形態を示す図である。
【図４１Ｂ】稀少遺伝病の患者におけるデノボ潜在的スプライス突然変異の一実装形態を示す図である。
【図４１Ｃ】稀少遺伝病の患者におけるデノボ潜在的スプライス突然変異の一実装形態を示す図である。
【図４１Ｄ】稀少遺伝病の患者におけるデノボ潜在的スプライス突然変異の一実装形態を示す図である。
【図４１Ｅ】稀少遺伝病の患者におけるデノボ潜在的スプライス突然変異の一実装形態を示す図である。
【図４１Ｆ】稀少遺伝病の患者におけるデノボ潜在的スプライス突然変異の一実装形態を示す図である。
【図４２Ａ】lincRNAに対する様々なスプライシング予測アルゴリズムの評価を示す図である。
【図４２Ｂ】lincRNAに対する様々なスプライシング予測アルゴリズムの評価を示す図である。
【図４３Ａ】TACTAAC分岐点およびGAAGAAエクソン内スプライスエンハンサーモチーフの位置依存効果を示す図である。
【図４３Ｂ】TACTAAC分岐点およびGAAGAAエクソン内スプライスエンハンサーモチーフの位置依存効果を示す図である。
【図４４Ａ】スプライシングに位置決めするヌクレオソームの影響を示す図である。
【図４４Ｂ】スプライシングに位置決めするヌクレオソームの影響を示す図である。
【図４５】複合効果を有するスプライス破断バリアントについてのエフェクトサイズを計算する例を示す図である。
【図４６Ａ】シングルトンおよび共通バリアントに対するSpliceNet-10kモデルの評価を示す図である。
【図４６Ｂ】シングルトンおよび共通バリアントに対するSpliceNet-10kモデルの評価を示す図である。
【図４６Ｃ】シングルトンおよび共通バリアントに対するSpliceNet-10kモデルの評価を示す図である。
【図４７Ａ】バリアントの位置によって分割されたスプライス部位形成バリアントのバリデーション率およびエフェクトサイズを示す図である。
【図４７Ｂ】バリアントの位置によって分割されたスプライス部位形成バリアントのバリデーション率およびエフェクトサイズを示す図である。
【図４８Ａ】訓練およびテスト染色体に対するSpliceNet-10kモデルの評価を示す図である。
【図４８Ｂ】訓練およびテスト染色体に対するSpliceNet-10kモデルの評価を示す図である。
【図４８Ｃ】訓練およびテスト染色体に対するSpliceNet-10kモデルの評価を示す図である。

【図４８Ｄ】訓練およびテスト染色体に対するSpliceNet-10kモデルの評価を示す図である。
【図４９Ａ】同義領域部位、イントロン領域部位、または非翻訳領域部位のみからの、稀少遺伝病の患者におけるデノボ潜在的スプライスバリアントを示す図である。
【図４９Ｂ】同義領域部位、イントロン領域部位、または非翻訳領域部位のみからの、稀少遺伝病の患者におけるデノボ潜在的スプライスバリアントを示す図である。
【図４９Ｃ】同義領域部位、イントロン領域部位、または非翻訳領域部位のみからの、稀少遺伝病の患者におけるデノボ潜在的スプライスバリアントを示す図である。
【図５０Ａ】ASDにおける潜在的スプライスデノボ突然変異を病原性DNMの割合として示す図である。
【図５０Ｂ】ASDにおける潜在的スプライスデノボ突然変異を病原性DNMの割合として示す図である。
【図５１Ａ】ASD患者における予測された潜在的スプライスデノボ突然変異のRNA配列バリデーションを示す図である。
【図５１Ｂ】ASD患者における予測された潜在的スプライスデノボ突然変異のRNA配列バリデーションを示す図である。
【図５１Ｃ】ASD患者における予測された潜在的スプライスデノボ突然変異のRNA配列バリデーションを示す図である。
【図５１Ｄ】ASD患者における予測された潜在的スプライスデノボ突然変異のRNA配列バリデーションを示す図である。
【図５１Ｅ】ASD患者における予測された潜在的スプライスデノボ突然変異のRNA配列バリデーションを示す図である。
【図５１Ｆ】ASD患者における予測された潜在的スプライスデノボ突然変異のRNA配列バリデーションを示す図である。
【図５１Ｇ】ASD患者における予測された潜在的スプライスデノボ突然変異のRNA配列バリデーションを示す図である。
【図５１Ｈ】ASD患者における予測された潜在的スプライスデノボ突然変異のRNA配列バリデーションを示す図である。
【図５１Ｉ】ASD患者における予測された潜在的スプライスデノボ突然変異のRNA配列バリデーションを示す図である。
【図５１Ｊ】ASD患者における予測された潜在的スプライスデノボ突然変異のRNA配列バリデーションを示す図である。
【図５２Ａ】カノニカル転写産物のみで訓練されたモデルのRNA配列に対するバリデーション率および感度を示す図である。
【図５２Ｂ】カノニカル転写産物のみで訓練されたモデルのRNA配列に対するバリデーション率および感度を示す図である。
【図５３Ａ】アンサンブルモデリングがSpliceNet-10k性能を向上させることを示す図である。
【図５３Ｂ】アンサンブルモデリングがSpliceNet-10k性能を向上させることを示す図である。
【図５３Ｃ】アンサンブルモデリングがSpliceNet-10k性能を向上させることを示す図である。
【図５４Ａ】エクソン密度が変化する領域におけるSpliceNet-10kの評価を示す図である。
【図５４Ｂ】エクソン密度が変化する領域におけるSpliceNet-10kの評価を示す図である。
【図５５】エフェクトサイズ計算および組織固有スプライシングを実証するために使用されるGTExサンプルの一実装形態を示す表S1である。
【図５６】それぞれに異なるアルゴリズムのバリデーション率および感度を評価するために使用されるカットオフの一実装形態を示す表S2である。

【図５７】遺伝子当たりエンリッチメント解析の一実装形態を示す図である。
【図５８】ゲノムワイドエンリッチメント解析の一実装形態を示す図である。
【図５９】開示された技術を実施するために使用することができるコンピュータシステムの簡略ブロック図である。
【発明を実施するための形態】
【００５４】
次の説明は、当業者が開示された技術を製作し使用することができるように提示され、特定の応用およびその要件の文脈においてなされている。開示されている実装形態に対し様々な修正を加えられることは、当業者にとっては明白であろうし、また本明細書において定義されている一般原理は、開示された技術の精神または範囲から逸脱することなく他の実装形態および応用にも適用され得る。したがって、開示された技術は、図示されている実装形態に限定されることを意図されておらず、本明細書で開示された原理および特徴と一致する最も広い範囲を適用されることを意図されている。
【００５５】
はじめに
畳み込みニューラルネットワーク
畳み込みニューラルネットワークは、特殊の種類のニューラルネットワークである。密結合層と畳み込み層との間の基本的な違いは以下の通りである。密層はその入力特徴空間においてグローバルパターンを学習するが、畳み込み層は、ローカルパターン、すなわち、画像の場合は、入力の小さい2Dウィンドウに見られるパターンを学習する。この主要特性は、畳み込みニューラルネットワークに2つの興味深い特性、すなわち、(1)畳み込みニューラルネットワークが学習するパターンは翻訳不変であり、(2)畳み込みニューラルネットワークはパターンの空間階層を学習することができる、という特性をもたらす。
【００５６】
第1の点に関して、畳み込み層は、写真の右上隅におけるあるパターンを学習した後、このパターンを任意の場所、たとえば、左上隅において認識することができる。密結合ネットワークは、このパターンが新しい位置に出現した場合にこのパターンを新たに学習する必要がある。これによって、畳み込みニューラルネットワークは、表現を学習するために必要とする訓練サンプルが少なくなり、一般化能力を有するので、データ効率が高くなる。
【００５７】
第2の点に関して、第1の畳み込み層は、エッジなどの小さいローカルパターンを学習することができ、第2の畳み込み層は、第1の層の特徴で構成されたより大きいパターンを学習し、以後の層についても同様である。これは、畳み込みニューラルネットワークが徐々に複雑にかつ抽象的になる視覚的概念を効率的に学習することを可能にする。
【００５８】
畳み込みニューラルネットワークは、多数の異なる層に配置された人工ニューロンの層を、各層を依存させる活性化関数を用いて相互接続することによって高度に非線形的なマッピングを学習する。これは、1つまたは複数の畳み込み層を含み、畳み込み層には、1つまたは複数のサブマッピング層および非線形層が散在し、一般にそれらの層の後に1つまたは複数の全結合層が続く。畳み込みニューラルネットワークの各要素は、前の層における特徴のセットから入力を受け取る。各畳み込みニューラルネットワークは、同じ特徴マップにおけるニューロンが同一の重みを有するので同時に学習する。これらのローカル共有重みは、ネットワークの複雑さを低減させ、それによって、多次元入力データがネットワークに入ったときに、畳み込みニューラルネットワークは特徴抽出および回帰または分類プロセスにおいてデータ再構成の複雑さを回避する。
【００５９】
畳み込みは、3Dテンソル上で動作し、3Dテンソルは特徴マップと呼ばれ、2つの空間軸(高さおよび幅)ならびに深度軸(チャネル軸とも呼ばれる)を有する。RGB画像では、画像が3つのカラーチャネル、すなわち、赤、緑、および青を有するので、深度軸の次元は3であ

る。白黒写真では、深度は1である(階調レベル)。畳み込み演算は、その入力特徴マップからパッチを抽出し、これらのパッチのすべてに同じ変換を適用し、出力特徴マップを生成する。この出力特徴マップは依然として3Dテンソルであり、幅と高さを有する。この出力特徴マップの深度は、出力深度が層のパラメータであり、その深度軸におけるそれぞれに異なるチャネルはもはやRGB入力とは異なり特定の色を表さず、むしろフィルタを表すので、任意であるものとしてよい。フィルタは入力データの特定の局面を符号化し、高さレベルでは、単一のフィルタはたとえば「入力における面の存在」という概念を符号化することが可能である。

【００６０】

たとえば、第1の畳み込み層は、サイズ(28，28，1)の特徴マップをとり、サイズ(26，26，32)の特徴マップを出力し、その入力上で32個のフィルタを計算する。これら32個の出力チャネルの各々が値の26×26グリッドを含み、26×26グリッドは、入力上のフィルタの応答マップであり、入力における異なる位置のそのフィルタパターンの応答を示す。それが、特徴マップという語が意味することであり、深度軸におけるあらゆる次元が特徴(またはフィルタ)であり、2Dテンソル出力[:，:，n]は、入力上のこのフィルタの応答の2D空間マップである。

【００６１】

畳み込みは2つの主要パラメータ、(1)入力から抽出されるパッチのサイズ--これらは一般的には1×1、3×3、または5×5、および(2)出力特徴マップの深度--畳み込みによって計算されるフィルタの数--によって定義される。多くの場合、これらは深度32から始まり、深度64まで続き、深度128または256で終わる。

【００６２】

畳み込みは、3D入力特徴マップ上でサイズ3×3または5×5のこれらのウィンドウをスライドさせ、すべての位置で停止させ、周囲の特徴(形状(window_height，window_width，input_depth))の3Dパッチを抽出することによって作用する。そのような各3Dパッチは次いで、(畳み込みカーネルと呼ばれる同じ学習された重み行列を有するテンソル積を介して)形状(output_depth，)の1Dベクトルに変換される。これらのベクトルのすべてが次いで、形状(height，width，output_depth)の3D出力マップに空間的に再アセンブルされる。出力特徴マップにおけるすべての空間位置が入力特徴マップ内の同じ位置に対応する(たとえば、出力の右下隅は入力の右下隅に関する情報を得る)。たとえば、3×3ウィンドウでは、ベクトル出力[i，j，:]は3Dパッチ入力[i-1:i+1，j-1:J+1，:]に由来する。完全なプロセスは図3に詳細に示されている。

【００６３】

畳み込みニューラルネットワークは、入力値と、訓練中に勾配更新を何回も繰り返して学習される畳み込みフィルタ(重みの行列)との間で畳み込み演算を実行する畳み込み層を備える。(m，n)をフィルタサイズとし、Wを重みの行列とすると、畳み込み層は、ドット積W・x+bを計算することによって入力XによるWの畳み込みを実行し、この場合、xはXのインスタンスであり、bはバイアスである。畳み込みフィルタが入力全体にわたってスライドする場合のステップサイズは、ストライドと呼ばれ、フィルタ領域(m×n)は受容野と呼ばれる。同じ畳み込みフィルタが入力の様々な位置にわたって適用され、学習される重みの数を低減させる。これはまた、位置不変学習を可能にする、すなわち、入力中に重要なパターンが存在する場合、そのパターンが配列内のどこにあるかにかかわらず畳み込みフィルタはそのパターンを学習する。

【００６４】

畳み込みニューラルネットワークの訓練

図4は、開示された技術の一実装形態による畳み込みニューラルネットワークの訓練のブロック図を示す。畳み込みニューラルネットワークは、入力データが特定の出力推定値になるように調整または訓練される。畳み込みニューラルネットワークは、出力推定値がグラウンドトゥルースに徐々に一致するかまたは近づくまで、出力推定値とグラウンドトゥルースとの比較に基づく逆伝搬法を使用して調整される。

【００６５】
畳み込みニューラルネットワークは、グラウンドトゥルースと実際の出力との差に基づいてニューロン間の重みを調整することによって訓練される。これは、以下のように数学的に記述され、ここで、δ=(グラウンドトゥルース)-(実際の出力)である。
【００６６】
【数６】

$$\Delta w_i = x_i \delta$$

【００６７】
一実装形態では、訓練規則は
$W_{nm} \leftarrow W_{nm} + \alpha (t_m - \phi_m) a_n$
と定義される。
【００６８】
上式では、矢印は、値の更新を示し、t_mはニューロンmの目標値であり、ϕ_mは、ニューロンmの計算された現在の出力であり、a_nは入力nであり、αは学習速度である。
【００６９】
訓練における中間ステップは、畳み込み層を使用して入力データから特徴ベクトルを生成することを含む。出力から始まる、各層における重みに対する勾配が、計算される。このことはバックワードパスまたは逆行と呼ばれる。ネットワークにおける重みは、負の勾配と前の重みの組合せを使用して更新される。
【００７０】
一実装形態では、畳み込みニューラルネットワークは勾配降下法によって誤差の逆伝搬を実行する確率的勾配更新アルゴリズム(ADAMなど)を使用する。シグモイド関数ベースの逆伝搬アルゴリズムの一例について以下に説明する。
【００７１】
【数７】

$$\varphi = f(h) = \frac{1}{1+e^{-h}}$$

【００７２】
上記のシグモイド関数では、hはニューロンによって計算される加重和である。シグモイド関数は、次の導関数を有する。
【００７３】
【数８】

$$\frac{\partial \varphi}{\partial h} = \varphi(1-\varphi)$$

【００７４】
アルゴリズムは、ネットワークにおけるすべてのニューロンの活性化を計算し、フォワ

ードパスについての出力を生成することを含む。隠れ層の中のニューロンmの活性化は、以下のように記述される。

【００７５】

【数９】

$$\varphi_m = \frac{1}{1+e^{-hm}}$$

$$h_m = \sum_{n=1}^{N} a_n w_{nm}$$

【００７６】

これは、すべての隠れ層に対して行われ、以下のように記述される活性化が得られる。

【００７７】

【数１０】

$$\varphi_k = \frac{1}{1+e^{hk}}$$

$$h_k = \sum_{m=1}^{M} \varphi_m v_{mk}$$

【００７８】

次いで、誤差および正しい重みが層ごとに算出される。出力における誤差は、

$\delta_{ok}=(t_k-\varphi_k)\varphi_k(1-\varphi_k)$

のように計算される。

【００７９】

隠れ層における誤差は、以下のように計算される。

【００８０】

【数１１】

$$\delta_{hm} = \varphi_m(1-\varphi_m)\sum_{k=1}^{K} v_{mk}\delta_{ok}$$

【００８１】

出力層の重みは、

$v_{mk} \leftarrow v_{mk} + \alpha\ \delta_{ok}\varphi_m$

のように更新される。

【００８２】

隠れ層の重みは学習速度αを使用して

$v_{nm} \leftarrow w_{nm} + \alpha\ \delta_{hm}a_n$

のように更新される。

【００８３】

一実装形態では、畳み込みニューラルネットワークは勾配降下最適化を使用してすべて

の層にわたって誤差を計算する。そのような最適化において、入力特徴ベクトルxおよび予測出力
【数１２】

$$\hat{y}$$

について、損失関数は、目標がyであるときに
【数１３】

$$\hat{y}$$

を予測するコストについてlとして定義され、すなわち、
【数１４】

$$l(\hat{y}, y)$$

である。予測出力
【数１５】

$$\hat{y}$$

は、関数fを使用して入力特徴ベクトルxから変換される。関数fは、畳み込みニューラルネットワークの重みによってパラメータ化され、すなわち
【数１６】

$$\hat{y} = f_w(x)$$

である。損失関数は、
【数１７】

$$l(\hat{y}, y) = l(f_w(x), y) \text{ または } Q(z, w) = l(f_w(x), y)$$

として記述され、この場合、zは入力および出力データ対(x, y)である。勾配降下最適化は、以下の式に従って重みを更新することによって実行される。
【００８４】

【数１８】

$$v_{t+1} = \mu v_t - \alpha \frac{1}{n} \sum_{i=1}^{N} \nabla w_t Q(z_t, w_t)$$

$$w_{t+1} = w_t + v_{t+1}$$

【００８５】

上式において、αは学習速度である。さらに、損失は、n個のデータ対のセットの平均として計算される。この計算は、線形収束時に学習速度αが十分低くなったときに終了する。一実装形態では、勾配は、ネステロフ加速勾配および適応勾配に送られる選択されたデータ対のみを使用して計算効率を高めることによって計算される。

【００８６】

一実装形態では、畳み込みニューラルネットワークは、確率的勾配降下法(SGD)を使用して費用関数を計算する。SGDは、損失関数における重みに対する勾配を、1つのランダム化されたデータ対z_tのみから算出することによって近似し、このことは次式のように記述される。

$v_t+1= \mu v - \alpha \nabla_w q(z_t, w_t)$

$w_t+1=w_t+v_t+1$

【００８７】

上式では、αは学習速度であり、μはモメンタムであり、tは、更新前の現在の重み状態である。SGDの収束速度は、学習速度αが速さと遅さの両方で十分に下げられたときにほぼO(1/t)になる。他の実装形態では、畳み込みニューラルネットワークは、ユークリッド損失およびソフトマックス損失などの様々な損失関数を使用する。さらなる実装形態では、畳み込みニューラルネットワークによってAdam確率的オプティマイザが使用される。

【００８８】

畳み込み層

畳み込みニューラルネットワークの畳み込み層は、特徴抽出器として働く。畳み込み層は、入力データを学習して階層的特徴に分解することのできる適応特徴抽出器として働く。一実装形態では、畳み込み層は、2つの画像を入力としてとり、第3の画像を出力として生成する。そのような実装形態において、畳み込みは、2次元(2D)における2つの画像に作用し、一方の画像は入力画像であり、他方の画像は「カーネル」と呼ばれ、入力画像上のフィルタとして適用され、出力画像を生成する。したがって、長さnの入力ベクトルfおよび長さmのカーネルgの場合、fおよびgの畳み込みf*gは、以下のように定義される。

【００８９】

【数１９】

$$(f*g)(i) = \sum_{j=1}^{m} g(j) \cdot f(i - j + m/2)$$

【００９０】

畳み込み演算は、カーネルを入力画像上でスライドさせることを含む。カーネルの各位置について、カーネルと入力画像の重畳値が乗算され、結果が加算される。積の和は、カーネルがセンタリングされる入力画像内の点における出力画像の値である。多数のカーネルから得られるそれぞれに異なる出力を特徴マップと呼ぶ。

【００９１】

畳み込み層は、訓練された後、新しい推論データに対する認識タスクを実行するように適用される。畳み込み層は、訓練データから学習するので、明示的な特徴抽出を回避し、訓練データから暗黙的に学習する。畳み込み層は、畳み込みフィルタカーネル重みを使用し、これらの重みは訓練プロセスの一部として決定され更新される。畳み込み層は、入力の異なる特徴を抽出し、これらの特徴は上位層において組み合わされる。畳み込みニューラルネットワークは、様々な数の畳み込み層を使用し、各畳み込み層が、カーネルサイズ、ストライド、パディング、特徴マップの数、および重みなどの様々な畳み込みパラメータを有する。

【００９２】

非線形層

図5は、開示された技術の一実装形態による非線形層の一実装形態を示す。非線形層は、異なる非線形トリガ関数を使用して各隠れ層上における可能性の高い特徴の明確な識別を示す。非線形層は、様々な特定の関数を使用して、正規化線形ユニット(ReLU)、双曲線正接、双曲線正接の絶対値、シグモイドおよび連続トリガ(非線形)関数を含む、非線形トリガリングを実装する。一実装形態では、ReLU活性化は、関数y=max(x, 0)を実装し、層の入力サイズと出力サイズを同じに維持する。ReLUを使用する利点は、畳み込みニューラルネットワークが何倍も速く訓練されることである。ReLUは、入力値がゼロよりも大きい場合は入力に対して線形であり、それ以外の場合はゼロである非連続非飽和活性化関数である。数学的には、ReLU活性化関数は、以下のように記述される。

【００９３】

【数２０】

$$\varphi(h) = \max(h, 0)$$

$$\varphi(h) = \begin{cases} h \ \ if \ \ h > 0 \\ 0 \ \ if \ \ h \leq 0 \end{cases}$$

【００９４】

他の実装形態では、畳み込みニューラルネットワークは、パワーユニット活性化関数を使用し、これは、

$\Phi(h)=(a+bh)^c$

によって記述される連続非飽和関数である。

【００９５】

上式において、a、b、およびcはそれぞれ、シフト、スケール、およびパワーを制御するパラメータである。パワー活性化関数は、cが奇数である場合にはxおよびy-反対称活性化を生成し、cが偶数である場合にはy-対称活性化を生成することができる。いくつかの実装形態では、このユニットは非正規化線形活性化を生成する。

【００９６】

さらに他の実装形態では、畳み込みニューラルネットワークはシグモイドユニット活性化関数を使用し、これは、以下のロジスティック関数によって記述される連続飽和関数である。

【００９７】

【数２１】

$$\varphi(h)=\frac{1}{1+e^{-\beta h}}$$

【００９８】

上式において、β=1である。シグモイドユニット活性化関数は、負の活性化を生成せず、y-軸に対してのみ反対称である。

【００９９】

Dilated畳み込み

図6は、Dilated畳み込みを示す。Dilated畳み込みは、Atrous畳み込みと呼ばれることもあり、文字通り穴を有することを意味する。このフランス語名は、高速2項ウェーブレット変換を計算するalgorithme a trousに由来する。これらの種類の畳み込み層では、フィルタのそれぞれのフィールドに対応する入力は近傍の点ではない。これは、図6に例示されている。入力間の距離は、拡張係数に依存する。

【０１００】

サブサンプリング層

図7は、開示された技術の一実装形態によるサブサンプリング層の一実装形態である。サブサンプリング層は、畳み込み層によって抽出された特徴の解像度を下げ、抽出された特徴または特徴マップをノイズおよび歪みに対してロバストにする。一実装形態では、サブサンプリング層は2種類のプーリング演算、平均プーリングおよび最大プーリングを使用する。プーリング演算は、入力を重なり合わない2次元空間に分割する。平均プーリングの場合、領域における4つの値の平均が計算される。最大プーリングの場合、4つの値の最大値が選択される。

【０１０１】

一実装形態では、サブサンプリング層は、前の層の出力を最大プーリングにおける入力のうちの1つのみにマップすること、および平均プーリングにおける入力の平均にマップすることによる、前の層におけるニューロンのセットに対するプーリング演算を含む。最大プーリングでは、プーリングニューロンの出力は、

$\Phi_0=\max(\Phi_1, \Phi_2, \cdots, \Phi_N)$

によって記述されるような入力内に存在する最大値である。

【０１０２】

上式において、Nはニューロンセット内の要素の総数である。

【０１０３】

平均プーリングでは、プーリングニューロンの出力は、以下の式によって記述されるような、入力ニューロンセットと一緒にある入力値の平均値である。

【０１０４】

【数２２】

$$\varphi_o=\frac{1}{N}\sum_{n=1}^{N}\varphi_n$$

【０１０５】

上式において、Nは入力ニューロンセット内の要素の総数である。

【０１０６】
図7において、入力はサイズ4×4を有する。2×2サブサンプリングでは、4×4画像がサイズ2×2の重なり合わない行列に分割される。平均プーリングでは、4つの値の平均は整数出力である。最大プーリングでは、2×2行列内の4つの値の最大値は整数出力である。
【０１０７】
畳み込みの例
図8は、畳み込み層の2層畳み込みの一実装形態を示す。図8において、サイズが2048次元の入力が畳み込まれる。畳み込み1において、入力は、サイズ3×3の16個のカーネルの2つのチャネルを備える畳み込み層によって畳み込まれる。次いで、その結果得られた16個の特徴マップは、ReLU1においてReLU活性化関数によって正規化され、次にサイズが3×3のカーネルを伴う16チャネルプーリング層を使用する平均プーリングによってプール1においてプーリングされる。次いで、畳み込み2において、プール1の出力は、サイズが3×3の30個のカーネルの16個のチャネルからなる別の畳み込み層によって畳み込まれる。この後に、さらに別のReLU2およびカーネルサイズが2×2のプール2における平均プーリングが続く。畳み込み層は、様々な個数の、たとえば、ゼロ個、1個、2個、および3個のストライドおよびパディングを使用する。その結果得られる特徴ベクトルは、一実装形態によれば512個の次元を有する。
【０１０８】
他の実装形態では、畳み込みニューラルネットワークは、異なる数の畳み込み層、サブサンプリング層、非線形層、および全結合層を使用する。一実装形態では、畳み込みニューラルネットワークは層の数がより少なく、層当たりのニューロンがより多い浅層ネットワークであり、たとえば、1つ、2つ、または3つの全結合層を有し、層当たり100個～200個のニューロンを有する。別の実装形態では、畳み込みニューラルネットワークは層の数がより多く、層当たりのニューロンがより少ない深層ネットワークであり、たとえば、5つ、6つ、または8つの全結合層を有し、層当たり30個～50個のニューロンを有する。
【０１０９】
フォワードパス
特徴マップにおけるf個の畳み込みコアについての第lの畳み込み層および第kの特徴マップにおける行x、列yのニューロンの出力は、以下の式によって決定される。
【０１１０】
【数２３】

$$O_{x,y}^{(l,k)} = \tanh\left(\sum_{t=0}^{f-1}\sum_{r=0}^{k_h}\sum_{c=0}^{k_w} W_{(r,c)}^{(k,t)} O_{(x+r,x+c)}^{(l-1,t)} + Bias^{(l,k)}\right)$$

【０１１１】
第lのサブサンプリング層および第kの特徴マップにおける行x、列yのニューロンの出力は、以下の式によって決定される。
【０１１２】
【数２４】

$$O_{x,y}^{(l,k)} = \tanh\left(W^{(k)}\sum_{r=0}^{S_h}\sum_{c=0}^{S_w} O_{(x\times S_h+r,y\times S_w+c)}^{(l-1,k)} + Bias^{(l,k)}\right)$$

【０１１３】
第lの出力層の第iのニューロンの出力は、以下の式によって決定される。
【０１１４】
【数２５】

$$O_{(l,i)} = \tanh\left(\sum_{j=0}^{H} O_{(l-1,j)} W^{l}_{(i,j)} + Bias^{(l,i)}\right)$$

【０１１５】
逆伝搬
出力層における第kのニューロンの出力偏差は、以下の式によって決定される。
【０１１６】
【数２６】

$$d(O^{o}_{k}) = y_k - t_k$$

【０１１７】
出力層における第kのニューロンの入力偏差は、以下の式によって決定される。
【０１１８】
【数２７】

$$d(I^{o}_{k}) = (y_k - t_k)\varphi'(v_k) = \varphi'(v_k) d(O^{o}_{k})$$

【０１１９】
出力層における第kのニューロンの重みおよびバイアスばらつきは、以下の式によって決定される。
【０１２０】
【数２８】

$$\Delta W^{o}_{k,x}) = d(I^{o}_{k}) y_{k,x}$$

$$\Delta Bias^{o}_{k}) = d(I^{o}_{k})$$

【０１２１】
隠れ層における第kのニューロンの出力バイアスは、以下の式によって決定される。
【０１２２】

【数２９】

$$d(O_k^H) = \sum_{i=0}^{i<84} d(I_i^o) W_{i,k}$$

【０１２３】
隠れ層における第kのニューロンの入力バイアスは、以下の式によって決定される。
【０１２４】
【数３０】

$$d(I_k^H) = \varphi'(v_k) d(O_k^H)$$

【０１２５】
隠れ層におけるk個のニューロンから入力を受け取る前の層の第mの特徴マップにおける行x、列yにおける重みおよびバイアスばらつきは、以下の式によって決定される。
【０１２６】
【数３１】

$$\Delta W_{m,x,y}^{H,k}) = d(I_k^H) y_{x,y}^m$$

$$\Delta Bias_k^H) = d(I_k^H)$$

【０１２７】
サブサンプル層Sの第mの特徴マップにおける行x、列yの出力バイアスは、以下の式によって決定される。
【０１２８】
【数３２】

$$d(O_{x,y}^{S,m}) = \sum_{k}^{170} d(I_{m,x,y}^H) W_{m,x,y}^{H,k}$$

【０１２９】
サブサンプル層Sの第mの特徴マップにおける行x、列yの入力バイアスは、以下の式によって決定される。
【０１３０】

【数３３】

$$d(I_{x,y}^{S,m}) = \varphi'(v_k) d(O_{x,y}^{S,m})$$

【０１３１】

サブサンプル層Sおよび畳み込み層Cの第mの特徴マップにおける行x、列yにおける重みおよびバイアスばらつきは、以下の式によって決定される。

【０１３２】

【数３４】

$$\Delta W^{S,m} = \sum_{x=0}^{fh} \sum_{y=0}^{fw} d(I_{[x/2],[y/2]}^{S,m}) O_{x,y}^{C,m}$$

$$\Delta Bias^{S,m}) = \sum_{x=0}^{fh} \sum_{y=0}^{fw} d(O_{x,y}^{S,m})$$

【０１３３】

畳み込み層Cの第kの特徴マップにおける行x、列yの出力バイアスは、以下の式によって決定される。

【０１３４】

【数３５】

$$d(O_{x,y}^{C,k}) = d(I_{[x/2],[y/2]}^{S,k}) W^k$$

【０１３５】

畳み込み層Cの第kの特徴マップにおける行x、列yの入力バイアスは、以下の式によって決定される。

【０１３６】

【数３６】

$$d(I_{x,y}^{C,k}) = \varphi'(v_k) d(O_{x,y}^{C,k})$$

【０１３７】

第lの畳み込み層Cの第kの特徴マップの第mの畳み込みコアにおける行r、列cにおける重みおよびバイアスばらつきは、以下の通りである。

【０１３８】

【数３７】

$$\Delta W_{r,c}^{k,m} = \sum_{x=0}^{fh} \sum_{y=0}^{fw} d(I_{x,y}^{C,k}) O_{x+r,y+c}^{l-1,m}$$

$$\Delta Bias^{C,k}) = \sum_{x=0}^{fh} \sum_{y=0}^{fw} d(I_{x,y}^{C,k})$$

【０１３９】

残差コネクション

図9は、特徴マップ追加を介して事前情報を下流側に再注入する残差コネクションを示す。残差コネクションは、過去の出力テンソルを後の出力テンソルに付加することによって前の表現をデータの下流に再注入することを含み、このことは、データ処理フローに沿った情報損失を防止する助けになる。残差コネクションは、あらゆる大規模ディープラーニングモデルに生じる2つの一般的な問題、すなわち、勾配消失および表現ボトルネックに対処する。概して、10個よりも多くの層を有する任意のモデルに残差コネクションを付加することは有益である可能性が高い。上述のように、残差コネクションは、前の層の出力を後の層の入力として利用可能にすることを含み、事実上シーケンシャルネットワークにおけるショートカットを形成する。後で活性化するために連結されるのではなく、前の出力が後の活性化に加算されるが、これはどちらの活性化も同じサイズであると仮定する。活性化が異なるサイズを有する場合、前の活性化を目標形状に再整形するための線形変換を使用することができる。

【０１４０】

残差学習およびスキップコネクション

図10は、残差ブロックおよびスキップコネクションの一実装形態を示す。残差学習の主要な考えは、残差マッピングが元のマッピングよりも学習するのがずっと容易であることである。残差ネットワークは、多数の残差ユニットを積み重ねて訓練精度の劣化を軽減する。残差ブロックは、特別な加法スキップコネクションを利用して深層ニューラルネットワークにおける勾配消失に対処する。残差ブロックの開始位置において、データフローは2つの流れに分離され、第1の流れがブロックの未変更入力を保持し、一方、第2の流れは重みおよび非線形性を適用する。ブロックの終了位置において、2つのストリームは要素ごとの和を使用してマージされる。そのような構成の主要な利点は、勾配がネットワーク内を容易に流れるようになることである。

【０１４１】

深層畳み込みニューラルネットワーク(CNN)は、残差ネットワークから利益を受け、容易に訓練することができ、画像分類および物体検出に関して精度の向上が実現されている。畳み込み層フィードフォワードネットワークは、第l層の出力を第(l+1)層への入力として接続し、それによって以下の層遷移$x_l=H_l(x_{l-1})$を生じさせる。残差ブロックは、識別関数$x_l=H_l(x_{l-1})+x_{l-1}$による非線形変換をバイパスするスキップコネクションを付加する。残差ブロックの利点は、勾配が識別関数を介して後の層から前の層へ直接流れることができることである。しかし、識別関数およびH_lの出力は、加算によって組み合わされるが、これはネットワーク内の情報フローを妨げ得る。

【０１４２】

WaveNet

WaveNetは、生オーディオ波形を生成するための深層ニューラルネットワークである。WaveNetは、比較的大きい「視野」を低コストで得ることができるので他の畳み込みニューラルネットワークから区別される。さらに、信号の調節をローカルおよびグローバルに追加することができ、それによってWaveNetを複数の音声を有する音声合成(TTS)エンジンへ

のテキストとして使用することが可能になり、すなわち、TTSがローカル調節を行い、特定の音声がグローバル調節を行う。

【０１４３】

WaveNetの主要ビルディングブロックは、因果的Dilated畳み込みである。因果的Dilated畳み込みに対する拡張機能として、WaveNetは、図11に示されているようにこれらの畳み込みのスタッキングも可能にする。この図におけるDilated畳み込みを有する同じ受容野を取得するには、別の拡張層が必要である。このスタックは、Dilated畳み込みの反復であり、Dilated畳み込み層の出力を単一の出力に接続する。これは、WaveNetが、比較的低い計算コストで1つの出力ノードの大きい「視野」を得ることを可能にする。比較として、512個の入力の視野を得るために、完全畳み込みネットワーク(FCN)は511個の層を必要とする。Dilated畳み込みネットワークの場合、8つの層が必要である。積み重ねられたDilated畳み込みでは、2つのスタックを有する7つの層または4つのスタックを有する6つの層のみが必要である。同じ視野をカバーするのに必要な計算力の差を把握するために、次の表は層ごとのフィルタが1つであり、フィルタ幅が2であるネットワークにおいて必要な重みの数を示す。さらに、ネットワークが8ビットの2進符号化を使用していると仮定される。

【０１４４】

【表１】

ネットワークタイプ	スタック数	チャネル当たり重み数	重み総数
FCN	1	$2.6 . 10^5$	$2.6 . 10^6$
WN	1	1022	8176
WN	2	1022	8176
WN	4	508	4064

【０１４５】

WaveNetは、残差コネクションが確立される前にスキップコネクションを追加し、それによって、すべての後続の残差ブロックをバイパスする。これらのスキップコネクションの各々は、それらを一連の活性化関数および畳み込みを通過させる前に加算される。直観的には、これは各層で抽出された情報の和である。

【０１４６】

バッチ正規化

バッチ正規化は、データ標準化をネットワークアーキテクチャの必須部分とすることによって深層ネットワーク訓練を加速するための方法である。バッチ正規化は、訓練中、時間の経過とともに平均および分散が変化するときでもデータを適応的に正規化することができる。バッチ正規化は、訓練中に見られるデータのバッチ式平均および分散の指数移動平均を内部に維持することによって作用する。バッチ正規化の主要な効果は、勾配伝搬を--残差コネクションと同様に--助け、したがって、深層ネットワークを可能にすることである。いくつかの非常に深いネットワークは、複数のバッチ正規化層を含む場合にのみ訓練することができる。

【０１４７】

バッチ正規化は、全結合層または畳み込み層と同様にモデルアーキテクチャに挿入することのできるさらに別の層とみなすことができる。バッチ正規化層は一般に、畳み込み層または密結合層の後に使用される。バッチ正規化層は、畳み込み層または密結合層の前に使用することもできる。両方の実装形態が。開示された技術によって使用することができ、図15に示されている。バッチ正規化層は軸引数をとり、軸引数は正規化すべき特徴軸を

指定する。この引数は、既定で-1であり、入力テンソルにおける最後の軸である。これは、data_formatが「channels_last」に設定されている密層、Conv1D層、RNN層、およびConv2D層を使用する際に正しい値である。しかし、data_formatが「channels_first」に設定されているConv2D層のニッチユースケースでは、特徴軸は軸1であり、バッチ正規化における軸引数を1に設定することができる。

【０１４８】

バッチ正規化は、入力をフィードフォワードし、バックワードパスを介してパラメータおよびそれら自体入力に関する勾配を計算するための定義を提供する。実際には、バッチ正規化層は、畳み込み層または全結合層の後で、ただし出力が活性化関数に送り込まれる前に挿入される。畳み込み層では、異なる位置における同じ特徴マップの異なる要素--すなわち活性化--が、畳み込み特性に従うように同じ方法で正規化される。したがって、ミニバッチにおけるすべての活性化が、活性化ごとではなくすべての位置において正規化される。

【０１４９】

内部共変量シフトは、深層アーキテクチャの訓練に悪名高くも時間がかかる理由である。これは、深層ネットワークが各層において新しい表現を学習するだけでなく、ネットワークの分布における変化を考慮する必要もあるからである。

【０１５０】

共変量シフトは、概して、ディープラーニング分野における既知の問題であり、実世界の問題において頻繁に生じる。一般的な共変量シフト問題は、訓練セットとテストセットとの分布の差であり、それによって一般化性能は最適なものではなくなる。この問題は、通常、標準化またはホワイトニング前処理ステップによって対処される。しかし、特にホワイトニング演算は計算コストがかかり、したがって、特に共変量シフトが異なる層全体にわたって生じる場合、オンライン設定では実用的ではない。

【０１５１】

内部共変量シフトは、訓練中のネットワークパラメータの変化に起因してネットワーク活性化の分布が各層にわたって変化する現象である。理想的には、各層は、それらが同じ分布を有するが機能関係は同じままである空間に変換されるべきである。共変量行列のコストのかかる計算を回避してあらゆる層およびステップにおいてデータを脱相関しホワイトニングするために、各ミニバッチ全体にわたって各層における各入力フィーチャの分布を正規化してゼロ平均および標準偏差1を有する。

【０１５２】

フォワードパス

フォワードパスの間、ミニバッチ平均および分散が計算される。これらのミニバッチ統計では、データは、平均を減算し、標準偏差で除算することによって正規化される。最後に、データは、学習されたスケールパラメータおよびシフトパラメータによってスケーリングされシフトされる。バッチ正規化フォワードパスf_{BN}は、図12に示されている。

【０１５３】

図12では、それぞれ、μ_βはバッチ平均であり、

【数３８】

$$\sigma_\beta^2$$

はバッチ分散である。学習されたスケールパラメータおよびシフトパラメータは、γおよびβによってそれぞれ示されている。わかりやすくするため、バッチ正規化手順は、本明細書では活性化ごとに説明され、対応するインデックスを省略する。

【０１５４】

正規化は、微分可能変換であるので、誤差はこれらの学習されたパラメータ内に伝搬さ

れ、したがって、識別変換を学習することによってネットワークの表現力を回復することができる。逆に、対応するバッチ統計と同一であるスケールパラメータおよびシフトパラメータを学習することによって、バッチ正規化変換は、実行すべき最適な動作であった場合、ネットワークに対して作用しない。テスト時に、入力がミニバッチからの他のサンプルに依存しないので、バッチ平均および分散は、それぞれの母集団統計によって置き換えられる。別の方法は、訓練中のバッチ統計の移動平均を維持し、これらを使用してテスト時のネットワーク出力を計算することである。テスト時に、バッチ正規化変換は、図13に例示されているように表され得る。図13では、μ_Dおよび

【数３９】

$$\sigma_D^2$$

は、バッチ統計ではなく、それぞれ母平均および母分散を示す。

【０１５５】

バックワードパス

正規化は、微分可能演算であるので、図14に示すようにバックワードパスが計算できる。

【０１５６】

1D畳み込み

1D畳み込みは、図16に示されているように、配列からローカル1Dパッチまたはサブ配列を抽出する。1D畳み込みは、入力配列内の時間的パッチから各出力時間ステップを取得する。1D畳み込み層は、配列内のローカルパターンを認識する。すべてのパッチに対して同じ入力変換が実行されるので、入力配列内のある位置において学習されたパターンを後で異なる位置において認識することができ、1D畳み込み層翻訳が時間的翻訳に対して不変になる。たとえば、サイズ5の畳み込みウィンドウを使用する塩基の1D畳み込み層処理配列は、長さ5以下の塩基または塩基配列を学習することができるべきであり、入力配列における任意の構成における塩基モチーフを認識することができるべきである。塩基レベル1D畳み込みは、塩基形態に関して学習できる。

【０１５７】

グローバル平均プーリング

図17は、グローバルアベレージプーリング(GAP)がどのように作用するかを示す。直前の層における特徴の空間的平均をとり記録することによって、グローバル平均プーリングを使用して分類のために全結合(FC)層を置き換えることができる。これは訓練負荷を低減し、過剰適合問題をバイパスする。グローバル平均プーリングは、モデルよりも前に構造を適用し、事前に定義された重みによる線形変換と同等である。グローバルアベレージプーリングは、パラメータの数を減らし、全結合層をなくす。全結合層は、典型的には、パラメータおよび結合に最も大きく依存する層であり、グローバルアベレージプーリングは、同様の結果を実現するうえでずっとコストが低い手法を構成する。グローバルアベレージプーリングの主要な考えは、直前の各層の特徴マップから平均値を記録される信頼係数として生成し、直接ソフトマックス層に送り込むことである。

【０１５８】

グローバルアベレージプーリングは3つの利点を有する。すなわち、(1)グローバルアベレージプーリング層には余分なパラメータがなく、したがって、グローバルアベレージプーリング層において過剰適合が回避される。(2)グローバルアベレージプーリングの出力は特徴マップ全体の平均であるので、グローバルアベレージプーリングは空間的変換に対してよりロバストである。(3)全結合層内のパラメータは非常に数が多く、通常、ネットワーク全体のすべてのパラメータにおける50%を占め、全結合層をグローバルアベレージプーリングで置き換えるとモデルのサイズを著しく縮小することができ、このため、グロ

ーバルアベレージプーリングはモデル圧縮において非常に有用である。

【０１５９】

グローバルアベレージプーリングは、直前の層におけるより強力な特徴がより高い平均値を有することが予期されるので有意である。いくつかの実装形態において、グローバルアベレージプーリングは、分類スコア用のプロキシとして使用することができる。グローバルアベレージプーリングの下での特徴マップは、信頼性マップ、および特徴マップとカテゴリとの間の力対応と解釈することができる。グローバルアベレージプーリングは、直前の層の特徴が直接分類に関して十分な抽象度を有する場合に特に効果的であるが、マルチレベル特徴を部品モデルのようなグループとして組み合わせるべきである場合、グローバルアベレージプーリングのみでは十分ではなく、この組合せは単純な全結合層またはその他の分類器をグローバルアベレージプーリングの後に付加することによって最もうまく実行される。

【０１６０】

用語

本出願で引用されるすべての文献および同様の題材は、限定はしないが、そのような文献および同様の題材のフォーマットとは無関係に、特許、特許出願、論文、書籍、専門書、およびウェブページを含み、全体が参照により明示的に組み込まれる。組み込まれる文献および同様の題材のうちの1つまたは複数が、限定はしないが定義された用語、用語の使用法、説明された技法などを含む本出願と異なるかまたは本出願と矛盾する場合、本出願が優先される。

【０１６１】

本明細書において使用されているように、次の用語は指示されている意味を有する。

【０１６２】

塩基は、ヌクレオチド塩基またはヌクレオチド、A(アデニン)、C(シトシン)、T(チミン)、またはG(グアニン)を指す。

【０１６３】

本出願は、「タンパク質」および「翻訳配列」という用語を交換可能に使用する。

【０１６４】

本出願は、「コドン」および「塩基トリプレット」という用語を交換可能に使用する。

【０１６５】

本出願は、「アミノ酸」および「翻訳ユニット」という用語を交換可能に使用する。

【０１６６】

本出願は、「バリアント病原性分類器」、「バリアント分類のための畳み込みニューラルネットワークベース分類器」、および「バリアント分類のための深層畳み込みニューラルネットワークベース分類器」という句を交換可能に使用する。

【０１６７】

「染色体」という用語は、生体細胞の遺伝子キャリアを指し、DNAおよびタンパク質成分(特にヒストン)を含むクロマチン鎖から得られる。本明細書では、伝統的な国際的に認知された個別ヒトゲノム染色体番号付与体系が使用されている。

【０１６８】

「部位」という用語は、基準ゲノム上の一意の位置(たとえば、染色体ID、染色体位置および配向)を指す。いくつかの実装形態では、部位は残差、配列タグ、または配列上のセグメントの位置であってもよい。「軌跡」という用語は、基準染色体上の核酸配列または形態の特定の位置を指す。

【０１６９】

本明細書では「サンプル」という用語は、一般に生体流体、細胞、組織、器官、あるいは核酸、または配列および/もしくは相を決定すべき少なくとも1つの核酸配列を含む核酸の混合物を含む有機体由来のサンプルを指す。そのようなサンプルは、限定はしないが、唾液/口腔液、羊水、血液、血液分画、微細針生検サンプル(たとえば、外科生検、微細針生検など)、尿、腹水、胸水、組織外植片、器官培養液および他の任意の組織または細胞

標品、あるいはそれらの割合もしくは誘導体またはそれらから分離された割合もしくは誘導体を含む。サンプルは、多くの場合に、人間被検体(たとえば、患者)から採取されるけれども、サンプルは、限定はしないが、犬、猫、馬、山羊、羊、牛、豚などを含む、染色体を有する任意の有機体から採取できる。サンプルは、生体源から得られた状態で直接使用されてもよく、またはサンプルの性質を修正する前処理の後で使用されてもよい。たとえば、そのような前処理は、血液から血漿を準備すること、粘性の流体を希釈することなどを含んでもよい。前処理の方法は、限定はしないが、濾過、沈殿、希釈、蒸留、混合、遠心分離、冷凍、凍結乾燥、濃縮、増幅、核酸断片化、干渉成分の不活性化、試薬の添加、溶解などを含んでもよい。

【０１７０】

「配列」という用語は、互いに結合されたヌクレオチドの鎖を含むかまたは表す。ヌクレオチドは、DNAまたはRNAに基づくものとしてよい。1つの配列が複数の部分配列を含んでもよいことを理解されたい。たとえば、単一の配列(たとえば、PCRアンプリコン)は350個のヌクレオチドを有し得る。読み取られるサンプルは、これらの350個のヌクレオチド内に複数の部分配列を含んでもよい。たとえば、読み取られるサンプルは、たとえば20個～50個のヌクレオチドを有する第1および第2の隣接部分配列を含んでもよい。第1および第2の隣接部分配列は、対応する部分配列(たとえば、40個～100個のヌクレオチド)を有する反復セグメントのいずれかの側に位置してもよい。隣接部分配列の各々は、プライマー部分配列(たとえば、10～30個のヌクレオチド)を含んでもよい(またはその一部を含んでもよい)。読みやすくするため、「部分配列」という用語は、「配列」と呼ばれるが、2つの配列は、必ずしも、共通鎖上で互いに分離してはいないことに留意されたい。本明細書で説明されている様々な配列を区別するために、配列には異なるラベルを付けられてよい(たとえば、標的配列、プライマー配列、隣接配列、参照配列、および同様のもの)。「対立遺伝子」などの他の用語には、類似の対象を区別できるように異なるラベルを付けてよい。

【０１７１】

「ペアエンドシーケンシング」という用語は、標的断片の両端の配列を決定するシーケンシング方法を指す。ペアエンドシーケンシングは、ゲノム再編成および反復断片、ならびに遺伝子融合および新規転写産物の検出を容易にする場合がある。ペアエンドシーケンシングのための方法は、PCT公開第WO07010252号、PCT出願第PCTGB2007/003798号、および米国特許出願公開第US2009/0088327号に記載されており、これらの文献の各々は参照により本明細書に組み込まれている。一例では、一連の動作は、(a)核酸のクラスタを生成し、(b)核酸を線形化し、(c)第1のシーケンシングプライマーのハイブリダイゼーションを行い、上述のエクステンション、スキャニング、およびデブロッキングの繰り返しサイクルを実行し、(d)相補的コピーを合成することによってフローセル表面上の目標核酸を「反転」させ、(e)再合成された鎖を線形化し、(f)第2のシーケンシングプライマーのハイブリダイゼーションを行い、上述のエクステンション、スキャニング、およびデブロッキングの繰り返しサイクルを実行する、というようにして実行され得る。反転動作は、ブリッジ増幅の単一サイクルについて上で述べたように試薬を送達することで実行され得る。

【０１７２】

「参照ゲノム」または「参照配列」という用語は、被検体からの識別された配列を参照するために使用され得る有機体の、部分であろうと全体であろうと、特定の任意の既知のゲノム配列を指す。たとえば、人間被検体さらには他の多くの有機体に使用される参照ゲノムは、全米バイオテクノロジー情報センターのncbi.nlm.nih.govにある。「ゲノム」は、核酸配列で表される、有機体またはウイルスの完全な遺伝情報を指す。ゲノムは、遺伝子およびDNAのノンコーディング配列の両方を含む。参照配列は、それに合わせて整列させられるリードより大きいものとしてよい。たとえば、これは、少なくとも約100倍大きい、または少なくとも約1000倍大きい、または少なくとも約10,000倍大きい、または少なくとも約10^5倍大きい、または少なくとも約10^6倍大きい、または少なくとも約10^7倍大きいものとしてよい。一例において、参照ゲノム配列は、完全長ヒトゲノムの配列である。

別の例において、参照ゲノム配列は、13番染色体などの特定ヒト染色体に限定される。いくつかの実装形態において、参照染色体は、ヒトゲノムversion hg19からの染色体配列である。そのような配列は、染色体参照配列と呼ばれてよいが、参照ゲノムという用語は、そのような配列を対象とすることを意図されている。参照配列の他の例は、他の種のゲノム、さらには任意の種の染色体、部分染色体領域(たとえば、ストランド)などを含む。様々な実装形態において、参照配列は、多数の個体に由来するコンセンサス配列であるか、または他の組合せである。しかしながら、特定の適用事例において、参照配列は、特定の個体から得られてもよい。

【０１７３】

「リード」という用語は、ヌクレオチドサンプルまたは参照の断片を記述する配列データの集合体を指す。「リード」という用語は、サンプルリードおよび/または参照リードを指すものとしてよい。典型的には、必ずというわけではないが、リードは、サンプルまたは参照中の連続する塩基対の短い配列を表す。リードは、サンプルまたは参照断片の塩基対配列によって(ATCGで)記号的に表され得る。これは、メモリデバイスに記憶され、適切に処理されて、リードが参照配列に一致するか、または他の基準を満たすかどうかを決定し得る。リードは、シーケンシング装置から直接的に、またはサンプルに関する記憶されている配列情報から間接的に、取得され得る。いくつかの場合において、リードはより大きい配列または領域を識別するために使用され得る、たとえば、染色体またはゲノム領域または遺伝子に対して整列され、特異的に割り当てられ得る、十分な長さ(たとえば、少なくとも約25bp)のDNA配列である。

【０１７４】

次世代シーケンシング法は、たとえば、合成によるシーケンシング技術(Illumina)、パイロシーケンシング法(454)、イオン半導体技術(Ion Torrentシーケンシング)、一分子リアルタイムシーケンシング(Pacific Biosciences)およびライゲーションによるシーケンシング(SOLiDシーケンシング)を含む。シーケンシング方法に応じて、各リードの長さは約30bpから10,000bp超まで変化し得る。たとえば、SOLiDシーケンサを使用するIlluminaのシーケンシング方法は、約50bpの核酸リードを生成する。別の例では、Ion Torrentシーケンシングは、最大400bpまでの核酸リードを生成し、454パイロシーケンシングは、約700bpの核酸リードを生成する。さらに別の例では、一分子リアルタイムシーケンシング法は、10,000bpから15,000bpまでのリードを生成し得る。したがって、いくつかの実装形態において、核酸配列リードは、30～100bp、50～200bp、または50～400bpの長さを有する。

【０１７５】

「サンプルリード」、「サンプル配列」、または「サンプル断片」という用語は、サンプルからの注目するゲノム配列に対する配列データを指す。たとえば、サンプルリードは、フォワードおよびリバースプライマー配列を有するPCRアンプリコンからの配列データを含む。配列データは、任意の選択配列方法から取得され得る。サンプルリードは、たとえば、合成によるシーケンシング(SBS)反応、ライゲーションによるシーケンシング反応、または反復要素の長さおよび/または素性を決定することが望ましい他の任意の好適なシーケンシング方法であってよい。サンプルリードは、複数のサンプルリードに由来するコンセンサス(たとえば、平均化または加重)配列であってよい。いくつかの実装形態において、参照配列を提供することは、PCRアンプリコンのプライマー配列に基づき注目する軌跡(locus-of-interest)を識別することを含む。

【０１７６】

「未処理断片」という用語は、サンプルリードまたはサンプル断片内の指定された位置または注目する二次的位置と少なくとも部分的に重なり合う注目するゲノム配列の一部に対する配列データを指す。未処理断片の非限定的な例は、二重ステッチ断片(duplex stitched fragment)、一重ステッチ断片(simplex stitched fragment)、二重アンステッチ断片(duplex un-stitched fragment)、および一重アンステッチ断片(simplex un-stitched fragment)を含む。「未処理」という語は、未処理断片がサンプルリード内の潜在的バリ

アントに対応し、認証するか、または確認する支持バリアントを示すかどうかに関係なく、サンプルリード内の配列データとの何らかの関係を有する配列データを含むことを示すのに使用される。「未処理断片」という用語は、断片がサンプルリード内のバリアントコールに対してバリデーションを行う支持バリアントを必ず含むということを示していない。たとえば、サンプルリードが第1のバリアントを示すようにバリアントコールアプリケーションによって決定されたときに、バリアントコールアプリケーションは、1つまたは複数の未処理断片がサンプルリード中のバリアントが与えられた場合に他の何らかの形で生じることが予想され得る対応するタイプの「支持」を欠いていることを決定し得る。

【０１７７】

「マッピングする」、「整列される」、「アライメント」、または「整列する」という語は、リードまたはタグを参照配列と比較し、それによって参照配列がリード配列を含むかどうかを決定するプロセスを指す。参照配列がリードを含む場合、リードは、参照配列にマッピングされ得るか、またはいくつかの実装形態では、参照配列内の特定の位置にマッピングされ得る。いくつかの場合において、アライメントは、単に、リードが特定の参照配列のメンバーかそうでないか(すなわち、リードが参照配列内に存在しているか、存在していないか)を伝えるだけである。たとえば、ヒトの13番染色体に対する参照配列へのリードのアライメントは、リードが13番染色体について参照配列内に存在しているかどうかを伝える。この情報を提供するツールは、セットメンバーシップテスター(set membership tester)と呼ばれるものとしてよい。いくつかの場合において、アライメントは、それに加えて、リードまたはタグがマッピングされる参照配列内の位置を示す。たとえば、参照配列がヒトゲノム配列全体である場合に、アライメントは、リードが13番染色体上に存在していることを示すものとしてよく、さらに、リードが13番染色体の特定のストランドおよび/または部位上にあることを示すものとしてよい。

【０１７８】

「インデル」という用語は、有機体のDNA内の塩基の挿入および/または欠失を指す。ミクロインデルは、1から50個のヌクレオチドの純変化を結果として引き起こすインデルを表す。ゲノムのコーディング領域では、インデルの長さが3の倍数でない限り、フレームシフト突然変異を引き起こす。インデルは、点突然変異と対比され得る。インデルは、配列からヌクレオチドを挿入し、削減するが、点突然変異は、DNA内の総数を変化させることなくヌクレオチドの1つを置き換える置換の一形態である。インデルは、また、Tandem Base Mutation(TBM)と対比されるものとしてよく、これは隣接するヌクレオチドにおける置換として定義され得る(もっぱら2つの隣接するヌクレオチドにおける置換であるが、3つの隣接するヌクレオチドにおける置換も観察されている)。

【０１７９】

「バリアント」という用語は、核酸参照と異なる核酸配列を指す。典型的な核酸配列バリアントは、限定することなく、単一ヌクレオチド多形(SNP)、短い挿入欠失多形(インデル)、コピー数バリエーション(CNV)、マイクロサテライトマーカーまたは縦列型反復配列および構造バリエーションを含む。体細胞バリアントコーリングは、DNAサンプル中に低頻度で存在するバリアントを識別する活動である。体細胞バリアントコーリングは、癌治療の状況において注目すべきものである。癌は、DNA内の突然変異の蓄積によって引き起こされる。腫瘍からのDNAサンプルは、一般的にヘテロ不均一性を有し、いくつかの正常細胞と、癌進行の初期段階にあるいくつかの細胞(突然変異は比較的少ない)と、いくつかの後期細胞(突然変異が比較的多い)を含む。このようにヘテロ不均一性があるため、腫瘍をシーケンシングするときに(たとえば、FFPEサンプルから)、体細胞突然変異は、低頻度で出現することが多い。たとえば、SNVは、所与の塩基をカバーするリードの10%にしか見られない可能性がある。バリアント分類器によって体細胞または生殖細胞として分類されるべきバリアントは、本明細書では「テスト対象のバリアント」とも呼ばれる。

【０１８０】

「ノイズ」という用語は、シーケンシングプロセスおよび/またはバリアントコールアプリケーションにおける1つまたは複数の誤差から結果として生じる間違ったバリアント

コールを指す。
【０１８１】
「バリアント頻度」という用語は、割合またはパーセンテージとして表される、母集団内の特定の軌跡における対立遺伝子(遺伝子のバリアント)の相対頻度を表す。たとえば、割合またはパーセンテージは、その対立遺伝子を持つ母集団内のすべての染色体の割合であるものとしてよい。たとえば、サンプルバリアント頻度は、個体から注目する遺伝子配列について取得されたリードおよび/またはサンプルの数に対応する「母集団」にわたる注目する遺伝子配列に沿った特定の軌跡/位置における対立遺伝子/バリアントの相対頻度を表す。別の例として、ベースラインバリアント頻度は、正常な個体の母集団からの1つまたは複数のベースライン遺伝子配列について取得されたリードおよび/またはサンプルの数に対応する「母集団」における1つまたは複数のベースライン遺伝子配列に沿った特定の軌跡/位置における対立遺伝子/バリアントの相対頻度を表す。
【０１８２】
「バリアント対立遺伝子頻度(VAF)」は、標的位置における全カバレッジによって除算されたバリアントと一致する観察されたシーケンシングされたリードのパーセンテージを指す。VAFは、バリアントを持つシーケンシングされたリードの割合の尺度である。
【０１８３】
「位置」、「指定された位置」、および「軌跡」という用語は、ヌクレオチドの配列内の1つまたは複数のヌクレオチドの配置または座標を指す。「位置」、「指定された位置」、および「軌跡」という用語は、ヌクレオチドの配列内の1つまたは複数の塩基対の配置または座標も指す。
【０１８４】
「ハプロタイプ」という用語は、一緒に受け継ぐ染色体上の隣接部位における対立遺伝子の組合せを指す。ハプロタイプは、もし生じた場合、軌跡の所与のセットの間で生じた組換え事象の数に応じて、1つの軌跡、複数の軌跡、または染色体全体であってよい。
【０１８５】
本明細書における「閾値」という用語は、サンプル、核酸、またはその一部(たとえば、リード)を特徴付けるためにカットオフとして使用される数値または数値でない値を指す。閾値は、実証的分析に基づき変えられ得る。閾値は、測定された、または計算された値と比較され、それによって、そのような値をもたらしたソースが特定の方式で分類されるべきかどうかを決定し得る。閾値は、経験的にまたは分析的に識別され得る。閾値の選択は、ユーザが分類を行わなければならないことを望む信頼度に依存する。閾値は、特定の目的のために選択されてよい(たとえば、感度と選択度とのバランスをとるため)。本明細書において使用されているように、「閾値」という用語は、分析のコースが変更され得る点および/または動作がトリガされ得る点を示す。閾値は、所定の数である必要はない。その代わりに、閾値は、たとえば、複数の係数に基づく関数であってもよい。閾値は、状況に対して適応的であり得る。さらに、閾値は、上限、下限、または上下限の間の範囲を示し得る。
【０１８６】
いくつかの実装形態において、シーケンシングデータに基づくメトリックまたはスコアは、閾値と比較され得る。本明細書において使用されているように、「メトリック」または「スコア」という用語は、シーケンシングデータから決定された値または結果を含み得るか、またはシーケンシングデータから決定された値または結果に基づく関数を含み得る。閾値と同様には、メトリックまたはスコアは、状況に対して適応的であり得る。たとえば、メトリックまたはスコアは、正規化された値であってよい。スコアまたはメトリックの一例として、1つまたは複数の実装形態が、データを分析するときにカウントスコアを使用し得る。カウントスコアは、サンプルリードの数に基づくものとしてよい。サンプルリードは、サンプルリードが少なくとも1つの共通特性または品質を有するように1つまたは複数のフィルタ処理段階に通されている可能性がある。たとえば、カウントスコアを決定するために使用されるサンプルリードの各々は、参照配列とすでに整列されているか、

または潜在的対立遺伝子として割り当てられ得る。共通特性を有するサンプルリードの数は、リードカウントを決定するためにカウントされ得る。カウントスコアは、リードカウントに基づくものとしてよい。いくつかの実装形態において、カウントスコアは、リードカウントに等しい値であってよい。他の実装形態では、カウントスコアは、リードカウントおよび他の情報に基づき得る。たとえば、カウントスコアは、遺伝子軌跡の特定の対立遺伝子に対するリードカウントおよび遺伝子軌跡に対するリードの総数に基づくものとしてよい。いくつかの実装形態において、カウントスコアは、リードカウントおよび遺伝子軌跡に対する以前に取得されたデータに基づき得る。いくつかの実装形態において、カウントスコアは、所定の値の間の正規化されたスコアであってよい。カウントスコアは、また、サンプルの他の軌跡からのリードカウントの関数または注目するサンプルと同時に行われた他のサンプルからのリードカウントの関数であってもよい。たとえば、カウントスコアは、特定の対立遺伝子のリードカウントおよびサンプル内の他の軌跡のリードカウントおよび/または他のサンプルからのリードカウントの関数であってよい。一例として、他の軌跡からのリードカウントおよび/または他のサンプルからのリードカウントは、特定の対立遺伝子に対するカウントスコアを正規化するために使用され得る。

【０１８７】

「カバレッジ」または「断片カバレッジ」という用語は、配列の同じ断片に対するサンプルリードの数のカウントまたは他の尺度を指す。リードカウントは、対応する断片をカバーするリードの数のカウントを表し得る。代替的に、カバレッジは、履歴的知識、サンプルの知識、軌跡の知識などに基づく指定された係数をリードカウントに乗算することによって決定され得る。

【０１８８】

「リード深度」(従来、数の後に「×」が続く)は、標的位置において重なるアライメントを有するシーケンシングされたリードの数を指す。これは、多くの場合に、間隔(エクソン、遺伝子、またはパネルなど)のセットにわたるカットオフを超える平均またはパーセンテージとして表される。たとえば、臨床報告書には、パネル平均カバレッジが1,105×で、標的塩基の98%が>100×をカバーしていると記載される可能性もある。

【０１８９】

「塩基コールクオリティスコア」または「Qスコア」は、0～20の範囲内のPHREDスケール確率を指し、これは単一のシーケンシングされた塩基が正しい確率に反比例する。たとえば、Qが20であるT塩基コールは、信頼P値が0.01の場合に正しい可能性が高いとみなされる。Q<20である塩基コールはどれも低品質と考えるべきであり、バリアントを支持するシーケンシングされたリードの実質的割合が低品質である識別されたバリアントはどれも潜在的に偽陽性であると考えるべきである。

【０１９０】

「バリアントリード」または「バリアントリード数」という用語は、バリアントの存在を支持するシーケンシングされたリードの数を指す。

【０１９１】

シーケンシングプロセス

本明細書において述べられている実装形態は、配列バリエーションを識別するために核酸配列を解析することに適用可能であるものとしてよい。遺伝子位置/軌跡の潜在的バリアント/対立遺伝子を解析し、遺伝子軌跡の遺伝子型を決定するか、または言い換えると、その軌跡に対する遺伝子型コールを提供するために実装形態が使用され得る。たとえば、核酸配列は、完全な主題が全体として本明細書に参照により明確に組み込まれている、米国特許出願公開第2016/0085910号および米国特許出願公開第2013/0296175号において説明されている方法およびシステムに従って解析され得る。

【０１９２】

一実装形態において、シーケンシングプロセスは、DNAなどの、核酸を含む、または含むことが疑われるサンプルを受け取ることを含む。サンプルは、動物(たとえば、人間)、植物、細菌、または菌類などの、既知の、または既知でないソースからのものであってよ

い。サンプルは、そのソースから直接採取され得る。たとえば、血液または唾液は、個体から直接採取されてよい。代替的に、サンプルは、そのソースから直接取得され得ない。次いで、1つまたは複数のプロセッサが、シーケンシングのためのサンプルを調製するようシステムに指令する。この調製は、異物を取り除き、および/または特定の物質(たとえば、DNA)を分離することを含み得る。生体サンプルは、特定のアッセイに対する特徴を含むように調製され得る。たとえば、生体サンプルは、合成によるシーケンシング(SBS)用に調製され得る。いくつかの実装形態において、調製は、ゲノムのいくつかの領域の増幅を含み得る。たとえば、調製は、STRおよび/またはSNPを含むことが知られている所定の遺伝子軌跡を増幅することを含み得る。遺伝子軌跡は、所定のプライマー配列を使用して増幅され得る。

【０１９３】

次に、1つまたは複数のプロセッサは、サンプルをシーケンシングするようシステムに指令する。シーケンシングは、多種多様の既知のシーケンシングプロトコルを通じて実行され得る。特定の実装形態において、シーケンシングはSBSを含む。SBSにおいて、光学的基板の表面(たとえば、フローセル内に少なくとも部分的にチャネルを画成する表面)上に存在する増幅されたDNAの複数のクラスタ(場合によっては数百万個のクラスタ)をシーケンシングするために複数の蛍光標識されたヌクレオチドが使用される。フローセルは、シーケンシングのための核酸サンプルを含むものとしてよく、フローセルは、適切なフローセルホルダー内に置かれる。

【０１９４】

核酸は、未知の標的配列に隣接する、既知のプライマー配列を含むように調製できる。第1のSBSシーケンシングサイクルを開始するために、1つまたは複数の異なる標識を付けられているヌクレオチド、およびDNAポリメラーゼ、などは、流体流れサブシステムによってフローセル内に流れ込む/フローセルを通って流れることができる。単一の種類のヌクレオチドが一度に追加され得るか、またはシーケンシング手順において使用されるヌクレオチドは、可逆終止特性を持つように特に設計され、それにより、シーケンシング反応の各サイクルがいくつかの種類の標識ヌクレオチド(たとえば、A、C、T、G)の存在下で同時に生じることを可能にすることができる。ヌクレオチドは、発蛍光団などの、検出可能な標識部分を含むことができる。4つのヌクレオチドが混ぜ合わされた場合、ポリメラーゼは、組み込むべき正しい塩基を選択することができ、各配列は、一塩基によって伸長される。非組込みヌクレオチドは、洗浄液をフローセルに流し込むことによって洗い流され得る。1つまたは複数のレーザが核酸を励起し、蛍光を誘起するものとしてよい。核酸から放射される蛍光発光は、組み込まれた塩基の発蛍光団に基づいており、異なる発蛍光団は、異なる波長の放射光を放射し得る。デブロッキング試薬がフローセルに追加され、それにより、伸長され、検出されたDNAストランドから可逆ターミネーター基を取り除くことができる。次いで、デブロッキング試薬は、洗浄液をフローセルに流し込むことによって洗い流され得る。次いで、フローセルは、上で述べたように標識ヌクレオチドの導入から始まるシーケンシングのさらなるサイクルの準備が整っている。流体および検出操作は、数回繰り返されて、シーケンシングランを完了することができる。例示的なシーケンシング方法は、たとえば、参照により本明細書に組み込まれている、Bentleyら、Nature 456:53-59 (2008年)、国際公開第WO04/018497号、米国特許第7,057,026号、国際公開第WO91/06678号、国際公開第WO07/123744号、米国特許第7,329,492号、米国特許第7,211,414号、米国特許第7,315,019号、米国特許第7,405,281号、および米国特許出願公開第2008/0108082号において説明されている。

【０１９５】

いくつかの実装形態において、核酸は表面に付着され、シーケンシングの前またはシーケンシング中に増幅され得る。たとえば、増幅はブリッジ増幅を使用して実行され、表面上に核酸クラスタを形成することができる。有用なブリッジ増幅方法は、たとえば、参照により本明細書に組み込まれている、米国特許第5,641,658号、米国特許出願公開第2002/0055100号、米国特許第7,115,400号、米国特許出願公開第2004/0096853号、米国特許出願

公開第2004/0002090号、米国特許出願公開第2007/0128624号、および米国特許出願公開第2008/0009420号において説明されている。表面上の核酸を増幅する他の有用な方法としては、ローリングサークル増幅(RCA)があり、たとえば、参照に本明細書に組み込まれている、Lizardiら、Nat. Genet. 19:225-232 (1998年)および米国特許出願公開第2007/0099208A1号において説明されている。

【０１９６】

例示的なSBSプロトコルの1つは、たとえば、参照により本明細書に組み込まれている、国際公開第WO04/018497号、米国特許出願公開第2007/0166705A1号、および米国特許第7,057,026号において説明されているような、脱離可能3'ブロックを有する修飾ヌクレオチドを利用する。たとえば、SBS試薬の反復サイクルは、たとえば、ブリッジ増幅プロトコルの結果として、標的核酸が付着されているフローセルに送達され得る。核酸クラスタは、線形化溶液を使用して単一のストランド化形式に変換され得る。線形化溶液は、たとえば、各クラスタの1つのストランドを開裂することができる制限エンドヌクレアーゼを含むことができる。開裂の他の方法は、就中、化学分解(たとえば、過ヨウ素酸塩とのジオール連鎖の開裂)、熱またはアルカリに曝すことによる、エンドヌクレアーゼを用いた開裂による脱塩基部位の開裂(たとえば、NEB、Ipswich、Mass.、USA、part number M5505Sによって供給されるような「USER」)、他の場合にはデオキシリボヌクレオチドからなる増幅産物に組み込まれるリボヌクレオチドの開裂、光化学開裂、またはペプチドリンカーの開裂を含む、制限酵素または切断酵素の代替として使用できる。線形化操作の後に、シーケンシングプライマーが、シーケンシングされるべき標的核酸へのシーケンシングプライマーのハイブリダイゼーションのための条件の下でフローセルに送達され得る。

【０１９７】

次いで、フローセルは、単一ヌクレオチド添加により各標的核酸にハイブリダイズされたプライマーを伸長する条件の下で脱離可能3'ブロックおよび蛍光標識とともに修飾ヌクレオチドを有するSBS伸長試薬と接触させることができる。単一ヌクレオチドのみが各プライマーに付加されるが、それは、修飾ヌクレオチドがシーケンシングされるテンプレートの領域に相補的な成長するポリヌクレオチド鎖内に組み込まれた後、さらなる配列伸長を導くのに利用可能な遊離3'-OH基がなく、したがって、ポリメラーゼはさらなるヌクレオチドを付加することができないからである。SBS伸長試薬は、除去され、放射による励起下においてサンプルを保護する成分を含む走査試薬で置き換えられ得る。走査試薬に対する例示的な成分は、参照によって本明細書に組み込まれている、米国特許出願公開第2008/0280773A1号および米国特許出願第13/018,255号において説明されている。次いで、伸長された核酸は、走査試薬の存在の下で蛍光検出することができる。蛍光発光が検出されると、3'ブロックは、使用されるブロッキング基に適切なデブロック試薬を使用して脱離されてもよい。それぞれのブロッキング基に有用な例示的なデブロック試薬は、参照により本明細書に組み込まれている、国際公開第WO004018497号、米国特許出願公開第2007/0166705A1号、および米国特許第7,057,026号において説明されている。デブロック試薬は、現時点において、さらなるヌクレオチドの付加に適切な3'OH基を有する伸長されたプライマーにハイブリダイズされた標的核酸を残して洗い流すことができる。したがって、伸長試薬、走査試薬、およびデブロック試薬を添加するサイクルは、これらの操作のうちの1つまたは複数の間の任意選択の洗浄とともに、所望の配列が得られるまで、繰り返すことができる。上記のサイクルは、修飾ヌクレオチドの各々が、特定の塩基に対応することが知られている、それに付着される異なる標識を有するときに、サイクル毎に単一の伸長試薬送達操作を使用して実施することができる。異なる標識は、各組込み操作中に付加されるヌクレオチド同士の区別を円滑にする。代替的に、各サイクルは、伸長試薬送達の別個の操作と、その後に続く、走査試薬送達および検出の別個の操作とを含むことができ、その場合、ヌクレオチドのうちの2つまたはそれ以上は、同一標識を有することができ、既知の送達順序に基づき区別することができる。

【０１９８】

シーケンシング操作は、特定のSBSプロトコルに関して上で説明されたけれども、多種

多様の他の分子解析のうちのどれかをシーケンシングするための他のプロトコルは望み通りに実行され得ることは理解されるであろう。

【０１９９】

次いで、システムの1つまたは複数のプロセッサは、その後の解析のためのシーケンシングデータを受け取る。シーケンシングデータは、.BAMファイルなどの、様々な方式でフォーマットされ得る。シーケンシングデータは、たとえば、多数のサンプルリードを含み得る。シーケンシングデータは、ヌクレオチドの対応するサンプル配列を有する複数のサンプルリードを含み得る。1つのサンプルリードしか説明されていないけれども、シーケンシングデータは、たとえば、数百個、数千個、数十万個、または数百万個のサンプルリードを含み得ることは理解されるべきである。異なるサンプルリードは、異なる数のヌクレオチドを有し得る。たとえば、サンプルリードは、10から約500またはそれ以上の個数のヌクレオチドを有し得る。サンプルリードは、ソースのゲノム全体に及ぶものとしてよい。一例として、サンプルリードは、疑わしいSTRまたは疑わしいSNPを有するそれらの遺伝子軌跡などの、所定の遺伝子軌跡の方へ向けられる。

【０２００】

各サンプルリードはヌクレオチドの配列を含むものとしてよく、これはサンプル配列、サンプル断片、または標的配列と称されてよい。サンプル配列は、たとえば、プライマー配列、隣接配列、および標的配列を含み得る。サンプル配列内のヌクレオチドの数は、30、40、50、60、70、80、90、100、またはそれ以上を含み得る。いくつかの実装形態において、サンプルリード(またはサンプル配列)のうちの1つまたは複数は、少なくとも150個のヌクレオチド、200個のヌクレオチド、300個のヌクレオチド、400個のヌクレオチド、500個のヌクレオチド、またはそれ以上を含む。いくつかの実装形態において、サンプルリードは、1000個を超えるヌクレオチド、2000個のヌクレオチド、またはそれ以上を含み得る。サンプルリード(またはサンプル配列)は、一端または両端にプライマー配列を含み得る。

【０２０１】

次に、1つまたは複数のプロセッサは、潜在的バリアントコールを取得するためのシーケンシングデータと、サンプルバリアントコールのサンプルバリアント頻度とを解析する。この操作は、バリアントコールアプリケーションまたはバリアントコーラーとも称され得る。したがって、バリアントコーラーはバリアントを識別するか、または検出し、バリアント分類器は検出されたバリアントを体細胞または生殖細胞として分類する。代替的なバリアントコーラーは、本明細書における実装形態により利用されてよく、異なるバリアントコーラーは、注目しているサンプルの特徴および同様のものに基づき、実行されているシーケンシング操作のタイプに基づき使用され得る。バリアントコールアプリケーションの他の非限定的な例は、https://github.com/Illumina/Piscesでホストされ、その完全な主題は全体が参照により本明細書に明示的に組み込まれている、論文、Dunn、Tamsen & Berry、Gwenn & Emig-Agius、Dorothea & Jiang、Yu & Iyer、Anita & Udar、Nitin & Str_mberg、Michael. (2017). Pisces: An Accurate and Versatile Single Sample Somatic and Germline Variant Caller. 595-595. 10.1145/3107411.3108203において説明されている、Illumina Inc.(カリフォルニア州サンディエゴ所在)によるPisces(商標)アプリケーションなどである。

【０２０２】

そのようなバリアントコールアプリケーションは、次のような4つの逐次的に実行されるモジュールを含むことができる。

【０２０３】

(1)Pisces Read Stitcher:BAM内のペアドリード(同じ分子のリード1とリード2)をコンセンサスリードにステッチングすることによってノイズを低減する。出力は、ステッチングされたBAMである。

【０２０４】

(2)Pisces Variant Caller:小SNV、挿入、および欠失をコールする。Piscesは、リード

境界によってバラバラにされたバリアントを融合するためのバリアント折り畳みアルゴリズム、基本フィルタリングアルゴリズム、および単純ポアソンベースバリアント信頼スコアリングアルゴリズムを含む。出力は、VCFである。

【０２０５】

(3)Pisces Variant Quality Recalibrator (VQR):バリアントコールが熱損傷またはFFPE脱アミノ化に関連付けられるパターンに圧倒的に従う場合、VQRステップは、疑わしいバリアントコールのバリアントQスコアをダウングレードする。出力は調整済みVCFである。

【０２０６】

(4)Pisces Variant Phaser (Scylla):リードバックドグリーディクラスタリング(read-backed greedy clustering)法を使用して、小バリアントをクローン部分母集団からの複合対立遺伝子にアセンブルする。これは、下流ツールによる機能的結果のより正確な決定を可能にする。出力は調整済みVCFである。

【０２０７】

それに加えて、または代替的に、この操作は、https://github.com/Illumina/strelkaでホストされ、その完全な主題は全体が参照により本明細書に明示的に組み込まれている、論文、T Saunders、Christopher & Wong、Wendy & Swamy、Sajani & Becq、Jennifer & J Murray、Lisa & Cheetham、Keira. (2012). Strelka: Accurate somatic small-variant calling from sequenced tumor-normal sample pairs. Bioinformatics (英国、オックスフォード). 28. 1811-7. 10.1093/bioinformatics/bts271において説明されている、Illumina Inc.によるバリアントコールアプリケーションStrelka(商標)アプリケーションを利用し得る。さらに、それに加えて、または代替的に、この操作は、https://github.com/Illumina/strelkaでホストされ、その完全な主題は全体が参照により本明細書に明示的に組み込まれている、論文、Kim, S.、Scheffler, K.、Halpern, A.L.、Bekritsky, M.A.、Noh, E.、Kallberg, M.、Chen, X.、Beyter, D.、Krusche, P.、およびSaunders, C.T. (2017). Strelka2: Fast and accurate variant calling for clinical sequencing applicationsにおいて説明されている、Illumina Inc.によるバリアントコールアプリケーションStrelka2(商標)アプリケーションを利用し得る。さらに、それに加えて、または代替的に、この操作は、https://github.com/Illumina/Nirvana/wikiでホストされ、その完全な主題は全体が参照により本明細書に明示的に組み込まれている、論文、Stromberg、Michael & Roy、Rajat & Lajugie、Julien & Jiang、Yu & Li、Haochen & Margulies、Elliott. (2017). Nirvana: Clinical Grade Variant Annotator. 596-596. 10.1145/3107411.3108204において説明されている、Illumina Inc.によるNirvana(商標)などの、バリアントアノテーション/コールツールを利用し得る。

【０２０８】

そのようなバリアントアノテーション/コールツールは、Nirvanaにおいて開示されているような、異なるアルゴリズム技術を適用することができる。

【０２０９】

a. すべての重なり合う転写産物をインターバル配列で識別する。機能的アノテーションについては、われわれは、バリアントを重ね合わせるすべての転写産物を識別することができ、インターバルツリーが使用され得る。しかしながら、インターバルのセットは静的であり得るので、われわれは、インターバル配列に合わせてそれをさらに最適化することができた。インターバルツリーはすべての重なり合う転写産物をO(min(n, k lg n))時間以内に返し、nはツリー内のインターバルの数であり、kは重なり合うインターバルの数である。実際、kはほとんどのバリアントに対してnと比較して本当に小さいので、インターバルツリー上の有効実行時間はO(k lg n)となるであろう。われわれは、最初の重なり合うインターバルを見つけて、次いで残りの(k-1)を数え上げるだけでよいようにすべてのインターバルがソートされた配列内に記憶されるインターバル配列を作成することによってO(lg n + k)まで改善した。

【０２１０】

b. CNVs/SVs (Yu)。コピー数バリエーションおよび構造バリアントに対するアノテー

ションが提供され得る。小バリアントのアノテーションと同様に、SVおよび前にも報告されている構造バリアントと重なり合う転写産物は、オンラインデータベース内でアノテーションを付けられ得る。小バリアントと異なり、重なり合う転写産物がすべて、アノテーションを付けられる必要はないが、それは大SVと重なり合う転写産物が多すぎるからである。その代わりに、部分的な重なり合う遺伝子に属すすべての重なり合う転写産物がアノテーションを付けられるものとしてよい。特に、これらの転写産物については、影響を受けるイントロン、エクソン、および構造バリアントによって引き起こされる結果が報告され得る。すべての重なり合う転写産物の出力することを可能にするオプションが利用可能であるが、遺伝子記号、それがカノニカルオーバーラップであるかまたは転写産物と部分的に重なり合うかを示すフラグなどのこれらの転写産物に対する基本情報は報告され得る。各SV/CNVについて、これらのバリアントが研究されているかどうか、異なる母集団内のその頻度を知ることもまた重要である。したがって、われわれは、1000ゲノム、DGV、およびClinGenなどの、外部データベース内の重なり合うSVを報告した。任意のカットオフを使用してどのSVが重なり合っているかを決定することを回避するために、その代わりに、すべての重なり合う転写産物が使用され、相互重なり合いが計算され得る、すなわち、重なり合う長さがこれら2つのSVの長さの最小値によって除算され得る。

【０２１１】

ｃ．補足アノテーションの報告。補足アノテーションは、小バリアントと構造バリアント(SV)の2種類がある。SVは、インターバルとしてモデル化され、上で説明されているインターバル配列を使用して重なり合うSVを識別することができる。小バリアントは、点としてモデル化され、位置および(任意選択で)対立遺伝子によってマッチングされる。そのようなものとして、それらは二分探索法に似たアルゴリズムを使用して探索される。補足アノテーションデータベースは、極めて大きくなる可能性があるので、かなり小さいインデックスが作成され、染色体位置を補足アノテーションが置かれるファイル位置にマッピングする。インデックスは、位置を使用して二分探索され得るオブジェクトのソートされた配列(染色体位置およびファイル配置からなる)である。インデックスサイズを小さく保つために、複数の位置(特定の最大カウントまでの)が第1の位置に対する値とその後の位置に対するデルタのみを記憶する1つのオブジェクトに圧縮される。われわれは、二分探索法を使用するので、実行時間はO(lg n)であり、nはデータベース内の項目の数である。

【０２１２】

ｄ．VEPキャッシュファイル

【０２１３】

ｅ．転写産物データベース。転写産物キャッシュ(cache)および補足データベース(SAdb)ファイルは、転写産物および補足アノテーションなどのデータオブジェクトのシリアル化されたダンプである。われわれは、アンサンブルVEPキャッシュをキャッシュに対するわれわれのデータソースとして使用する。キャッシュを作成するために、すべての転写産物はインターバル配列内に挿入され、配列の最終状態はキャッシュファイル内に記憶される。したがって、アノテーション実行中に、われわれは、事前計算されたインターバル配列をロードし、それに対して探索を実行するだけでよい。キャッシュはメモリ内にロードされ、探索は非常に高速なので(上で説明されている)、重なり合う転写産物を見つけることは、Nirvanaでは極端に高速である(総実行時間の1%未満のプロファイルを有する?)。

【０２１４】

ｆ．補足データベース。SAdbに対するデータソースは、補足物質の下でリストされる。小バリアントに対するSAdbは、データベース内の各オブジェクト(参照名および位置によって識別される)がすべての関連する補足アノテーションを保持するように、すべてのデータソースのkウェイマージ(k-way merge)によって作成される。データソースファイルを解析している間に遭遇した問題は、Nirvanaのホームページに詳しく説明されている。メモリ使用量を制限するために、SAインデックスのみがメモリにロードされる。このインデックスは、補足アノテーションに対するファイル位置の高速検索を可能にする。しかしながら、データはディスクからフェッチされなければならないので、補足アノテーションを

追加することは、Nirvanaの最大のボトルネックとして識別されている(総実行時間の～30%でプロファイルされる)。

【０２１５】

g. 結果および配列オントロジー。Nirvanaの機能的アノテーション(提供されたときの)は、Sequence Ontology(SO)(http://www.sequenceontology.org/)の指針に従う。時々、われわれは、現在のSOにおける問題を識別し、アノテーションの状態を改善するためにSOチームと共同作業する機会を有していた。

【０２１６】

そのようなバリアントアノテーションツールは前処理を含むことができる。たとえば、Nirvanaは、ExAC、EVS、1000ゲノムプロジェクト、dbSNP、ClinVar、Cosmic、DGV、およびClinGenのような、外部データソースからの多数のアノテーションを含んでいた。これらのデータベースをフルに活用するため、われわれは、それらからの情報の不適切な部分を削除しなければならない。われわれは、異なるデータソースから存在する異なるコンフリクトを取り扱うために異なる戦略を実装した。たとえば、同じ位置および代替対立遺伝子に対する複数のdbSNPエントリの場合、われわれは、すべてのidsをidsのコンマ区切りリストに結合し、同じ対立遺伝子に対して異なるCAF値を有する複数のエントリがある場合に、われわれは第1のCAF値を使用する。コンフリクトしているExACおよびEVSエントリについて、われわれは、サンプルカウントの数を考慮し、より高いサンプルカウントを有するエントリが使用されている。1000ゲノムプロジェクトでは、われわれは、コンフリクトしている対立遺伝子の対立遺伝子頻度を取り除いた。別の問題は情報の不正確さである。われわれは、もっぱら、1000ゲノムプロジェクトから対立遺伝子頻度情報を抽出したが、われわれは、GRCh38について、infoフィールド内に報告された対立遺伝子頻度は、遺伝子型が利用可能でないサンプルを除外せず、そのため、すべてのサンプルについて利用可能でないバリアントに対して頻度が低下したことに気付いた。われわれのアノテーションの精度を保証するために、われわれは、個体レベルの遺伝子型のすべてを使用して真の対立遺伝子頻度を計算する。ご存じの通り、同じバリアントは、異なるアライメントに基づき異なる表現を有することができる。われわれがすでに識別されているバリアントに対する情報を正確に報告することができることを確実にするために、われわれは、一貫性のある表現にするために異なるリソースからのバリアントを前処理しなければならない。すべての外部データソースについて、われわれは、参照対立遺伝子および代替対立遺伝子の両方における重複ヌクレオチドを除去するために対立遺伝子をトリミングした。ClinVarでは、われわれは、xmlファイルを直接解析し、すべてのバリアントに対して5'アライメントを実行したが、これは多くの場合にvcfファイルにおいて使用される。異なるデータベースは、情報の同じセットを含むことができる。不必要な重複を回避するために、われわれは、いくらかの重複情報を取り除いた。たとえば、われわれは、データソースを1000ゲノムプロジェクトとして有するDGV内のバリアントを除去したが、それは、われわれはすでにより詳しい情報とともに1000個のゲノムにおけるこれらのバリアントを報告しているからである。

【０２１７】

少なくともいくつかの実装形態により、バリアントコールアプリケーションは、低頻度バリアントに対するコール、生殖細胞系列コーリング、および同様のものを提供する。非限定的な例として、バリアントコールアプリケーションは、腫瘍細胞のみのサンプルおよび/または腫瘍細胞-正常細胞ペアサンプルに対して実行し得る。バリアントコールアプリケーションは、単一ヌクレオチドバリエーション(SNV)、複数ヌクレオチドバリエーション(MNV)、インデル、および同様のものを探索するものとしてよい。バリアントコールアプリケーションは、シーケンシングまたはサンプル調製誤差による不整合をフィルタ処理している間に、バリアントを識別する。各バリアントについて、バリアントコーラーは、参照配列、バリアントの位置、および潜在的バリアント配列(たとえば、AからC SNV、またはAGからA欠失)を識別する。バリアントコールアプリケーションは、サンプル配列(またはサンプル断片)、参照配列/断片、およびバリアントコールをバリアントが存在してい

るという指示として識別する。バリアントコーラーアプリケーションは未処理断片を識別し、未処理断片の指定、潜在的バリアントコールを検証する未処理断片の数のカウント、支持バリアントが生じた未処理断片内の位置、および他の関連情報を識別し得る。未処理断片の非限定的な例は、二重ステッチ断片、一重ステッチ断片、二重アンステッチ断片、および一重アンステッチ断片を含む。

【０２１８】

バリアントコールアプリケーションは、.VCFまたは.GVCFファイルなどの、様々なフォーマットでコールを出力し得る。例にすぎないが、バリアントコールアプリケーションは、MiSeqReporterパイプラインに含まれていてもよい(たとえば、MiSeq(登録商標)シーケンサ計測器上に実装されたとき)。任意選択で、このアプリケーションは、様々なワークフローで実装されてよい。解析は、所望の情報を取得するために指定された方式でサンプルリードを解析する単一プロトコルまたはプロトコルの組合せを含み得る。

【０２１９】

次いで、1つまたは複数のプロセッサは、潜在的バリアントコールと関連してバリデーション操作を実行する。バリデーション操作は、以下で説明されているように、クオリティスコア、および/または段階的テストの階層に基づくものとしてよい。バリデーション操作が、潜在的バリアントコールを認証または検証するとき、バリデーション操作は、バリアントコール情報を(バリアントコーラーアプリケーションから)サンプル報告生成器に受け渡す。代替的に、バリデーション操作が潜在的バリアントコールを無効にするか、または不適であるとみなしたとき、バリデーション操作は、対応する指示(たとえば、負のインジケータ、コールインジケータなし、無効コールインジケータ)をサンプル報告生成器に受け渡す。バリデーション操作は、また、バリアントコールが正しいか、または無効コール指定が正しいかの信頼度に関係する信頼度スコアを受け渡すものとしてよい。

【０２２０】

次に、1つまたは複数のプロセッサは、サンプル報告を生成し、記憶する。サンプル報告は、たとえば、サンプルに関する複数の遺伝子軌跡に関する情報を含み得る。たとえば、遺伝子軌跡の所定のセットの各遺伝子軌跡について、サンプル報告は、遺伝子型コールを提供するか、遺伝子型コールが行うことができないことを指示するか、遺伝子型コールの確実さに関する信頼度スコアを提供するか、または1つまたは複数の遺伝子軌跡に関するアッセイの潜在的問題を指示する、のうちの少なくとも1つを行うものとしてよい。サンプルレポートは、また、サンプルを提供した個体の性別を指示し、および/またはサンプルが複数のソースを含むことを指示するものとしてよい。本明細書において使用されているように、「サンプル報告」は、遺伝子軌跡もしくは遺伝子軌跡の所定のセットのデジタルデータ(たとえば、データファイル)および/または遺伝子軌跡もしくは遺伝子軌跡のセットの印刷された報告書を含み得る。したがって、生成すること、または提供することは、データファイルを作成すること、および/またはサンプル報告を印刷すること、またはサンプル報告を表示することを含み得る。

【０２２１】

サンプル報告は、バリアントコールが決定されたが、バリデーションが行われていないことを指示し得る。バリアントコールが無効であると決定されたときに、サンプル報告は、バリアントコールのバリデーションを行わないという決定に対する基準に関する追加の情報を指示し得る。たとえば、報告内の追加の情報は、未処理断片の記述と、未処理断片がバリアントコールを支持するか、または否定する程度(たとえば、カウント)とを含み得る。それに加えて、または代替的に、報告内の追加の情報は、本明細書において説明されている実装形態により取得されるクオリティスコアを含み得る。

【０２２２】

バリアントコールアプリケーション

本明細書において開示されている実装形態は、潜在的バリアントコールを識別するためにシーケンシングデータを解析することを含む。バリアントコーリングは、すでに実行されたシーケンシング操作に対する記憶されているデータに基づき実行されてよい。これは

、それに加えて、または代替的に、シーケンシング操作が実行されている間にリアルタイムで実行されてよい。サンプルリードの各々は、対応する遺伝子軌跡に割り当てられる。サンプルリードは、サンプルリードのヌクレオチドの配列、または、言い換えると、サンプルリード内のヌクレオチドの順序(たとえば、A、C、G、T)に基づき対応する遺伝子軌跡に割り当てられ得る。この解析結果に基づき、サンプルリードは、特定の遺伝子軌跡の可能なバリアント/対立遺伝子を含むものとして指定され得る。サンプルリードは、遺伝子軌跡の可能なバリアント/対立遺伝子を含むものとして指定されている他のサンプルリードとともに収集され(または集められ、またはビンに入れられ)得る。この割り当て操作は、サンプルリードが特定の遺伝子位置/軌跡に場合によっては関連付けられているものとして識別されているコーリング操作とも呼ばれ得る。サンプルリードは、そのサンプルリードを他のサンプルリードから区別するヌクレオチドの1つまたは複数の識別配列(たとえば、プライマー配列)を特定するために解析され得る。より具体的には、識別配列は、他のサンプルリードからのサンプルリードを特定の遺伝子軌跡に関連付けられているものとして識別し得る。

【０２２３】

割り当て操作は、その識別配列のn個のヌクレオチドからなる一連のヌクレオチドが選択配列のうちの1つまたは複数と効果的にマッチするかどうかを決定するために識別配列のn個のヌクレオチドからなる一連のヌクレオチドを解析することを含み得る。特定の実装形態において、割り当て操作は、サンプル配列の最初のn個のヌクレオチドが選択配列のうちの1つまたは複数と効果的にマッチするかどうかを決定するためにサンプル配列の最初のn個のヌクレオチドを解析することを含み得る。数nは、プロトコルに埋め込むようにプログラムされるか、またはユーザによって入力され得る、多種多様の値を有し得る。たとえば、数nは、データベース内の最短の選択配列のヌクレオチドの数として定義され得る。この数は、所定の数であり得る。ヌクレオチドの所定の数は、たとえば、10、11、12、13、14、15、16、17、18、19、20、21、22、23、24、25、26、27、28、29、または30であるものとしてよい。しかしながら、他の実装形態において使用されるヌクレオチドはこれより少なくても多くてもよい。数nは、システムのユーザなどの、個人によって選択されてもよい。数nは、1つまたは複数の条件に基づくものであってよい。たとえば、数nは、データベース内の最短のプライマー配列のヌクレオチドの数、または指定された数のうちの、いずれか小さい方の数として定義され得る。いくつかの実装形態において、ヌクレオチドが15個未満であるプライマー配列が例外として指定され得るように15などの、nに対する最小値が使用されてよい。

【０２２４】

いくつかの場合において、識別配列のn個のヌクレオチドからなる一連のヌクレオチドは、選択配列のヌクレオチドと正確にマッチしないことがある。そうであっても、識別配列は、識別配列が選択配列とほぼ同一である場合に選択配列と効果的にマッチするものとしてよい。たとえば、サンプルリードは、識別配列のn個のヌクレオチドからなる一連のヌクレオチド(たとえば、最初のn個のヌクレオチド)が指定された数(たとえば、3)以下のミスマッチおよび/または指定された数(たとえば、2)以下のシフトで選択配列とマッチした場合に遺伝子軌跡についてコールされ得る。規則は、各ミスマッチまたはシフトがサンプルリードとプライマー配列との間の差としてカウントするように確立され得る。差の数が指定された数より小さい場合、サンプルリードは、対応する遺伝子軌跡についてコールされ(すなわち、対応する遺伝子軌跡に割り当てられ)得る。いくつかの実装形態において、サンプルリードの識別配列と遺伝子軌跡に関連付けられている選択配列との間の差の数に基づくマッチングスコアが決定され得る。マッチングスコアが指定されたマッチング閾値を通過した場合、選択配列に対応する遺伝子軌跡は、サンプルリードに対する潜在的軌跡として指定されてよい。いくつかの実装形態において、その後の解析は、サンプルリードが遺伝子軌跡に対してコールされるかどうかを決定するために実行され得る。

【０２２５】

サンプルリードがデータベース内の選択配列のうちの1つと効果的にマッチした(すなわ

ち、上で説明されているように正確にマッチするか、またはほぼマッチした)場合、サンプルリードは、選択配列に相関する遺伝子軌跡に割り当てられるか、または指定される。これは、軌跡コーリングまたは仮の軌跡コーリングと呼んでよく、サンプルリードは、選択配列に相関する遺伝子軌跡に対してコールされる。しかしながら、上で説明されているように、サンプルリードは、複数の遺伝子軌跡に対してコールされてよい。そのような実装形態において、潜在的遺伝子軌跡のうちの1つのみに対してサンプルリードをコールするか、または割り当てるためにさらなる解析が実行され得る。いくつかの実装形態において、参照配列のデータベースに比較されるサンプルリードは、ペアドエンドシーケンシングからの第1のリードである。ペアドエンドシーケンシングを実行するときに、サンプルリードに相関する第2のリード(未処理断片を表す)が取得される。割り当てた後、割り当てられたリードで実行されるその後の解析は、割り当てられたリードに対してコールされている遺伝子軌跡のタイプに基づくものとしてよい。

【０２２６】

次に、サンプルリードは、潜在的バリアントコールを識別するために解析される。他にもあるがとりわけ、解析の結果は、潜在的バリアントコール、サンプルバリアント頻度、参照配列、およびバリアントが生じた注目するゲノム配列内の位置を識別する。たとえば、遺伝子軌跡が、SNPを含むことについて知られている場合、遺伝子軌跡に対してコールされている割り当てられたリードは、割り当てられたリードのSNPを識別するために解析を受けるものとしてよい。遺伝子軌跡が、多形反復DNA要素を含むことについて知られている場合、割り当てられたリードは、サンプルリード内の多形反復DNA要素を識別するか、または特徴付けるために解析され得る。いくつかの実装形態において、割り当てられたリードがSTR軌跡およびSNP軌跡と効果的にマッチした場合、警告またはフラグがサンプルリードに割り当てられ得る。サンプルリードは、STR軌跡およびSNP軌跡の両方として指定され得る。解析することは、アライメントプロトコルに従って割り当てられたリードを整列させて割り当てられたリードの配列および/または長さを決定することを含み得る。アライメントプロトコルは、参照により全体が本明細書に組み込まれている、2013年3月15日に出願した国際特許出願第PCT/US2013/030867号(国際公開第WO2014/142831号)において説明されている方法を含み得る。

【０２２７】

次いで、1つまたは複数のプロセッサは、未処理断片を解析して、支持バリアントが未処理断片内の対応する位置に存在しているかどうかを決定する。様々な種類の未処理断片が識別され得る。たとえば、バリアントコーラーは、元のバリアントコールのバリデーションを行うバリアントを示す未処理断片の種類を識別し得る。たとえば、未処理断片の種類は、二重ステッチ断片、一重ステッチ断片、二重アンステッチ断片、または一重アンステッチ断片を表し得る。任意選択で、他の未処理断片は、前述の例の代わりに、またはそれに加えて、識別されてもよい。未処理断片の各種類を識別することに関連して、バリアントコーラーは、支持バリアントが生じた、未処理断片内の位置、さらには支持バリアントを示した未処理断片の数のカウントも識別する。たとえば、バリアントコーラーは、未処理断片の10個のリードが特定の位置Xにおいて支持バリアントを有する二重ステッチ断片を表すように識別されたことを示す指示を出力し得る。バリアントコーラーは、また、未処理断片の5個のリードが特定の位置Yにおいて支持バリアントを有する一重アンステッチ断片を表すように識別されたことを示す指示も出力し得る。バリアントコーラーは、参照配列に対応していた、したがってそうでなければ注目するゲノム配列における潜在的バリアントコールのバリデーションを行う証拠を提供するであろう支持バリアントを含まなかった多数の未処理断片も出力し得る。

【０２２８】

次に、支持バリアントを含む未処理断片のカウント、さらには支持バリアントが生じた位置が保持される。それに加えて、または代替的に、注目する位置(サンプルリードまたはサンプル断片内の潜在的バリアントコールの位置に相対的な)で支持バリアントを含んでいなかった未処理断片のカウントが保持され得る。それに加えて、または代替的に、参

照配列に対応し、潜在的バリアントコールを認証または確認することをしない未処理断片のカウントが保持され得る。決定された情報は、潜在的バリアントコールを支持する未処理断片のカウントおよび種類、未処理断片内の支持バリアントの位置、潜在的バリアントコールを支持しない未処理断片のカウント、および同様のものを含めて、バリアントコールバリデーションアプリケーションに出力される。

【０２２９】

潜在的バリアントコールが識別されたとき、プロセスは、潜在的バリアントコール、バリアント配列、バリアント位置、およびそれに関連付けられている参照配列の指示を出力する。バリアントコールは、誤差が、コールプロセスが誤ったバリアントを識別する原因となり得るので、「潜在的」バリアントを表すように指定される。本明細書の実装形態により、誤ったバリアントまたは偽陽性を低減し排除するために、潜在的バリアントコールが解析される。それに加えて、または代替的に、プロセスはサンプルリードに関連付けられている1つまたは複数の未処理断片を解析し、未処理断片に関連付けられている対応するバリアントコールを出力する。

【０２３０】

ゲノミクスにおけるディープラーニング

遺伝的バリエーションは、多くの疾病の説明の助けとなり得る。すべての人間は固有の遺伝子コードを有し、個体のグループ内に多くの遺伝的バリアントがある。悪影響のある遺伝的バリアントのほとんどは、自然選択によってゲノムから枯渇している。どの遺伝的バリエーションが病原性を有するか、または悪影響をもたらす可能性が高いかを識別することは重要である。これは、研究者が病原性を持つ可能性の高い遺伝的バリアントに集中し、多くの疾病の診断および治療のペースを加速するのに役立つ。

【０２３１】

バリアントの特性および機能的効果(たとえば、病原性)をモデル化することは、ゲノミクスの分野において重要であるが困難なやりがいのある課題である。機能的ゲノムシーケンシング技術が急速に進歩しているにもかかわらず、バリアントの機能的結果の解釈は、細胞型特有の転写調節系の複雑さ故に依然として大きな問題である。

【０２３２】

病原性分類器に関して、深層ニューラルネットワークは、複数の非線形の複雑な転換層を使用して高水準の特徴を次々にモデル化する人口ニューラルネットワークの一種である。深層ニューラルネットワークは、パラメータを調整するための観察された出力と予測された出力との間の差を伝えるフィードバックを逆伝搬を介して提供する。深層ニューラルネットワークが発達したのは、大きな訓練データセットが利用可能であること、並列および分散コンピューティングの力、および高度な訓練アルゴリズムがあったおかげである。深層ニューラルネットワークは、コンピュータビジョン、音声認識、および自然言語処理などの多数の分野において大きな進歩を促した。

【０２３３】

畳み込みニューラルネットワーク(CNN)および回帰型ニューラルネットワーク(RNN)は、深層ニューラルネットワークの構成要素である。畳み込みニューラルネットワークは、特に、畳み込み層、非線形層、およびプーリング層を備えるアーキテクチャを用いる画像認識において成功を収めている。回帰型ニューラルネットワークは、パーセプトロン、長・短期記憶ユニット、およびゲート付き回帰型ユニットのような構成単位の間で環状結合を有する入力データの順次的情報を利用するように設計されている。それに加えて、制限された構成に対して、深層時空間ニューラルネットワーク、多次元回帰型ニューラルネットワーク、および畳み込みオートエンコーダなどの、他の多くの新興の深層ニューラルネットワークが提案されている。

【０２３４】

深層ニューラルネットワークを訓練することの目標は、各層における重みパラメータの最適化であり、これはより単純な特徴を、最も適している階層表現がデータから学習されるように複雑な特徴に徐々に組み合わせる。最適化プロセスの単一のサイクルは次のよう

に編成される。最初に、訓練データセットが与えられると、フォワードパスは順次各層において出力を計算し、ネットワークを通して機能信号を前方に伝搬する。最終的な出力層では、目的損失関数が推論された出力と所与のラベルとの間の誤差を測定する。訓練誤差を最小化するため、バックワードパスは連鎖法則を使用して誤差信号を逆伝搬させ、ニューラルネットワーク全体を通してすべての重みに関して勾配を計算する。最後に、重みパラメータが、確率的勾配降下法に基づき最適化アルゴリズムを使用して更新される。バッチ勾配降下法は、各完全なデータセットに対してパラメータ更新を実行するが、確率的勾配降下法は、データ例の各小セットに対して更新を実行することによって確率的近似を提供する。いくつかの最適化アルゴリズムが確率的勾配降下法に由来している。たとえば、Adagrad訓練アルゴリズムおよびAdam訓練アルゴリズムは、それぞれ、各パラメータに対する勾配の更新頻度およびモーメントに基づき学習速度を適応的に修正しながら確率的勾配降下法を実行する。

【０２３５】

深層ニューラルネットワークの訓練における別のコア要素は正則化であり、これは過剰適合を回避し、したがって良好な一般化性能を達成することを意図されている戦略を指す。たとえば、重み減衰は、重みパラメータがより小さい絶対値に収束するように目的損失関数にペナルティ項を追加する。ドロップアウトは、訓練中に隠れユニットをニューラルネットワークから取り除き、可能なサブネットワークのアンサンブルとみなされてよい。ドロップアウトの能力を増強するために、新しい活性化関数、maxoutおよびrnnDropと呼ばれる回帰型ニューラルネットワークに対するドロップアウトのバリアントが提案されている。さらに、バッチ正規化は、ミニバッチ内で各活性化に対してスカラー特徴の正規化ならびに各平均および分散をパラメータとして学習することを通して新しい正則化方法を提供する。

【０２３６】

シーケンシングされたデータが多次元および高次元であるとした場合に、深層ニューラルネットワークは、広範な応用性および高い予測能力を有することでバイオインフォマティクス研究にとって非常に有望である。畳み込みニューラルネットワークは、モチーフ発見、病原性バリアント識別、および遺伝子発現推論などのゲノミクスにおける配列ベースの問題を解決するように適合されている。畳み込みニューラルネットワークは、DNAを研究する上で特に有用である重み共有戦略を使用するが、それは、配列モチーフを捕捉し、これらの配列モチーフは短く、有意な生物学的機能を有すると想定されているDNAにおけるローカルパターンを再帰させることができるからである。畳み込みニューラルネットワークのホールマークは、畳み込みフィルタの使用である。緻密に設計され、手作業で作られる特徴に基づく伝統的な分類アプローチとは異なり、畳み込みフィルタは、未処理入力データを知識の情報表現にマッピングするプロセスに類似する、特徴の適応的学習を実行する。この意味で、畳み込みフィルタは一連のモチーフスキャナとして働くが、それは、そのようなフィルタのセットが訓練手順実行中に入力の中の関連するパターンを認識し、それ自身を更新することができるからである。回帰型ニューラルネットワークは、タンパク質またはDNA配列などの、可変長の順次的データにおける長距離依存性を捕捉することができる。

【０２３７】

したがって、バリアントの病原性を予測するための強力な計算モデルが、基礎科学およびトランスレーショナルリサーチの両方に対して巨大な恩恵をもたらし得る。

【０２３８】

現在のところ、奇病患者の25～30%しかタンパク質コード配列の調査からの分子診断を受けておらず、このことは、残りの診断率は非コードのゲノムの99%内にあり得ることを示唆している。ここで、われわれは、非コードバリアントのスプライス変更効果の正確な予測を可能にする、任意のpre-mRNA転写産物配列からスプライス接合を正確に予測する新規性のあるディープラーニングネットワークを説明する。予測されたスプライス変更結果を有する同義語およびイントロン突然変異はRNA-seq上で高い率によりバリデーションし

、ヒト母集団内で大きな悪影響を及ぼす。予測されたスプライス変更結果を有するデノボ突然変異は、自閉症を患い、健康な対照と比較して知的障害のある患者において著しくエンリッチされ、これらの患者28人のうちの21人においてRNA-seqデータと突き合わせてバリデーションがなされる。われわれは、稀少遺伝性疾患を患っている患者における病原性突然変異の9～11%が疾病バリエーションのこの以前には正しく評価されていなかったクラスによって引き起こされると推定している。

【０２３９】

エクソームシーケンシングは、稀少遺伝性疾患を患っている患者および家族の臨床診断を転換しており、第一線のテストとして採用されたときに、診断を求める終わりなき旅の時間とコストとを著しく低減する(Monroeら、2016年、Starkら、2016、Tanら、2017年)。しかしながら、エクソームシーケンシングの診断率は、稀少遺伝病コホートにおいて約25～30%であり、患者の大半はエクソームおよびマイクロアレイ併合テストの後であっても診断のないままである(Leeら、2014年、Trujillanoら、2017年、Yangら、2014年)。非コード領域は遺伝子調節に著しい役割を果たし、ヒトの複雑な疾病の無バイアスのゲノムワイド関連研究において発見された因果疾病軌跡の90%を占めるが(Ernstら、2011年、Farhら、2015年、Mauranoら、2012年)、このことは、浸透非コードバリアントも稀少遺伝病の因果的突然変異の著しい負担の原因となり得ることを示唆している。実際、潜在的スプライスバリアントと称されることが多い、本質的なGTおよびAGスプライスジヌクレオチドの外側にあるにもかかわらずmRNAスプライシングの通常パターンを壊す浸透非コードバリアントは、稀少遺伝病に著しい役割を果たしていると長い間認識されてきた(Cooperら、2009年、Padgett、2012年、ScottiおよびSwanson、2016年、WangおよびCooper、2007年)。しかしながら、潜在的スプライス突然変異は、スプライシングコードをわれわれが不完全にしか理解していないこと、およびその結果本質的なGTおよびAGジヌクレオチドの外側のスプライス変更バリアントを正確に識別することが困難であることにより、臨床診療において見落とされることが多い(WangおよびBurge、2008年)。

【０２４０】

最近、RNA-seqがメンデル性疾患におけるスプライシング異常を検出するための有望なアッセイとして浮かび上がったが(Cummingsら、2017年、Kremerら、2017年)、今までのところは、臨床現場における有用性は関連する細胞型が知られており、バイオプシーにアクセス可能である症例のうちの少数に限られたままである。潜在的なスプライス変更バリアントのハイスループットスクリーニングアッセイ(Soemediら、2017年)は、スプライシングバリエーションのキャラクタリゼーションを拡大したが、スプライス変更突然変異が生じ得るゲノム空間は、極端に大きいので、遺伝病におけるランダムなデノボ突然変異を評価するためにはあまり実用的でない。任意のpre-mRNA配列からのスプライシングの一般的予測は、潜在的に、非コードバリアントのスプライス変更結果の正確な予測を可能にし、遺伝病を患っている患者の診断を実質的に改善するであろう。今まで、スプライセオソームの特異性に接近する未処理配列からのスプライシングの一般予測モデルは、コアスプライシングモチーフの配列特性をモデル化すること(YeoおよびBurge、2004年)、エクソンスプライスエンハンサーおよびサイレンサーを特徴付けること(Fairbrotherら、2002年、Wangら、2004年)、およびカセットエクソン包含を予測すること(Barashら、2010年、Jhaら、2017年、Xiongら、2015年)など、特定のアプリケーションにおいて進歩があったにもかかわらず、いまだ明確になっていない。

【０２４１】

長いpre-mRNAを成熟転写産物にスプライスすることはその精度、およびプライス変更突然変異の臨床的悪性度に関して顕著であり、その上、細胞機構がその特異性を決定する際の機構に対する理解は不完全なままである。ここで、われわれは、コンピュータでスプライセオソームの精度に近づくディープラーニングネットワークを訓練し、95%の精度でpre-mRNA配列からエクソン-イントロン境界を識別し、RNA-seq上で80%を超えるバリデーション率により機能的潜在的スプライス突然変異を予測する。スプライシングを変更することが予測される非コードバリアントは、ヒト母集団内で強い悪影響を有し、新しく生成され

た潜在的スプライス突然変異の80%が陰性選択を受け、これはタンパク質切り詰めバリエーションの他のクラスの影響に類似している。自閉症および知的障害のある患者のデノボ潜在的スプライス突然変異は、タンパク質切り詰め突然変異によって回帰的に突然変異する同じ遺伝子に当たり、追加の候補疾病遺伝子の発見を可能にする。われわれは、稀少遺伝性疾患を患っている患者の浸透因果的突然変異の最大24%までが疾病バリエーションのこの以前には正しく評価されていなかったクラスによるものであると推定し、臨床的シーケンシングアプリケーションに対して非コーディングであるゲノムの99%の解釈を改善する必要性を強調している。

【０２４２】

臨床的エクソームシーケンシングは、稀少遺伝性疾患を患っている患者および家族に対する診断に革命を起こし、第一線のテストとして採用されたときに、診断を求める終わりなき旅の時間とコストとを著しく低減する。しかしながら、エクソンシーケンシングに対する診断率は、稀少疾病患者およびその親の複数の大きなコホートにおいて25～30%と報告されており、患者の大半はエクソームおよびマイクロアレイ併合テストの後でも診断がないままである。非コードゲノムは、遺伝調節において高活性であり、非コードバリアントは共通疾病に対するGWASヒットの～90%を占め、非コードゲノムの稀少バリアントも稀少遺伝性疾患および腫瘍研究などの浸透性疾病における因果的突然変異の著しい割合を占める可能性のあることを示唆している。しかしながら、非コードゲノム内のバリアントを解釈することが困難であることは、大きな構造バリアントの外では、非コードゲノムは現在、臨床管理に対する最大の影響を有する稀少浸透性バリアントに関して診断上の追加のメリットをほとんどもたらさないことを意味する。

【０２４３】

カノニカルGTおよびAGスプライスジヌクレオチドの外側のスプライス変更突然変異の役割は、長い間稀少病において評価されてきた。実際、これらの潜在的スプライスバリアントは、グリコーゲン貯蔵症XI(ポンペ病)および骨髄性プロトポルフィリアなどの、いくつかの稀少遺伝性疾患に対する最も一般的な突然変異である。イントロンの5'および3'末端における伸長スプライスモチーフは大きく変性し、等しく良好なモチーフがゲノム内に頻繁に出現し、そのためどの非コードバリアントが潜在的スプライシングを引き起こす可能性があるかの正確な予測が既存の方法では非実用的なものとなる。

【０２４４】

スプライセオソームがその特異性をどのように達成するかをより理解するために、われわれは、転写産物配列のみをその入力として使用して、pre-mRNA内の各ヌクレオチドに対してそれがスプライスアクセプターであるか、スプライスドナーであるか、またはそのいずれでもないかを予測するようにディープラーニングニューラルネットワークを訓練した(図37A)。偶数染色体上のカノニカル転写産物を訓練セットとして、奇数染色体上の転写産物をテスト用に使用することで(パラログを除き)、ディープラーニングネットワークは、95%の精度でエクソン-イントロン境界をコールし、CFTRなどの100KBを超える転写産物であっても、ヌクレオチド精度で完全に再構成されることが多い(図37B)。

【０２４５】

われわれは、次に、そのような際立った精度でエクソン-イントロン境界を認識するためにネットワークによって使用される特異性決定因子を理解することに努めた。統計学的または人間工学的に設計された特徴に基づいて動作する前の分類器とは対照的に、ディープラーニングは、階層的方式で配列から特徴を直接的に学習し、追加の特異性が長距離配列構成から与えられることを可能にする。実際、われわれは、ネットワークの精度がネットワーク内に入力として提供される予測の下でヌクレオチドに隣接する配列構成の長さに大きく依存することを見いだし(Table 1(表1))、われわれが40-ntの配列のみを使用するディープラーニングモデルを訓練したときに、パフォーマンスは既存の統計的方法を控えめに超える程度でしかない。これは、ディープラーニングが個別の9～23ntスプライシングモチーフを認識するために既存の統計的方法を超えて付け加えるものはほとんどないが、より広い配列構成が等しく強いモチーフを持つ非機能的部位から機能的スプライス部位

を区別する鍵となることを示している。エクソン上で配列がどこで摂動されるかを予測するようネットワークに求めることで、ドナーモチーフを切断することもアクセプター信号が消失する(図37C)ことを引き起こすことが示され、これはインビボでエクソンスキッピング事象により頻繁に観察されることであり、著しい特異度が単純に許容可能な距離で強いアクセプターモチーフとドナーモチーフとの間のペアリングを要求することによって与えられることを示している。

【０２４６】

大きな一連の証拠が、エクソンの長さの実験的摂動がエクソン包含対エクソンスキッピングに対する強い効果を有することを示しているけれども、これは、ディープラーニングネットワークの精度がなぜ構成の1000-ntを超えて増加し続けるかを説明していない。ローカルスプライスモチーフ駆動特異性と長距離特異性決定因子とをより適切に区別するために、われわれは、入力として構成の100-ntのみを取るローカルネットワークを訓練した。ローカルネットワークを使用して既知の接合にスコアを与えることで、われわれは、エクソンおよびイントロンの両方がモチーフ強度が最小となる最適な長さ(エクソンに対して～115nt、イントロンに対して～1000nt)を有することを見いだしている(図37D)。この関係は、10000-ntのディープラーニングネットワークには存在せず(図37E)、これはイントロンおよびエクソンの長さの変動がすでに広い構成のディープラーニングネットワーク内に完全に組み入れられていることを示す。特に、イントロンおよびエクソンの境界は、広い構成のディープラーニングモデルに決して与えられておらず、これは、エクソンおよびイントロンの位置を配列だけから推論することによってこれらの距離を導出することができることを示していた。

【０２４７】

六量体空間の体系的探索も、ディープラーニングネットワークがエクソン-イントロン定義におけるモチーフ、特に位置-34から-14への分岐点モチーフTACTAAC、エクソンの末端近くの適切に特徴付けられたエクソンスプライスエンハンサーGAAGAA、および典型的にはポリピリミジントラクトの一部であるポリUモチーフを利用するが、エクソンスプライスサイレンサーとして働くようにも見える(図21、22、23、および24)。

【０２４８】

われわれは、参照転写産物配列およびバリアントを含む代替転写産物配列の両方でエクソン-イントロン境界を予測し、エクソン-イントロン境界の変化がないか探すことによって、ディープラーニングネットワークをスプライス変更機能に対する遺伝的バリアントの評価に拡張する。60,706人の人間からの集計エクソームデータが最近利用可能になったことで、われわれは、予測されたバリアントに対する陰性選択の影響を評価し、対立遺伝子頻度スペクトル内の分布を調べることによってスプライス機能を変更することが可能になっている。われわれは、予測された潜在的スプライスバリアントが強く陰性選択の下にあり(図38A)、これは予想されるカウントと比較した高い対立遺伝子頻度における相対的枯渇によって明らかであり、その枯渇の大きさはAGまたはGTスプライス切断バリアントおよびストップゲインバリアントに匹敵することを見いだしている。インフレーム変化を引き起こすバリアントを超えてフレームシフトを引き起こす潜在的スプライスバリアントを考えたときに陰性選択の影響はより大きくなる(図38B)。タンパク質切り詰めバリエーションの他のクラスと比較したフレームシフト潜在的スプライスバリアントの枯渇に基づき、われわれは、自信を持って予測された潜在的スプライス突然変異の88%が機能すると推定している。

【０２４９】

エクソームデータほどの集約全ゲノムデータは利用できないけれども、能力を深イントロン領域における自然選択の影響を検出するように制限することで、われわれは、エクソン領域から隔たる潜在的スプライス突然変異の観察されたカウント対予想カウントを計算することもできた。全体的に、われわれは、エクソン-イントロン境界から>50ntの距離で潜在的スプライス突然変異の60%の枯渇を観察している(図38C)。信号減衰は、全ゲノムデータがエクソームと比較されたより小さいサンプルサイズと、深イントロンバリアントの

影響を予測する困難がより大きいこととが組み合わさったことである可能性が高い。
【０２５０】
われわれは、また、観察された数対予想された数の潜在的スプライスバリアントを使用して、選択下の潜在的スプライスバリアントの数、およびこれとタンパク質切り詰めバリエーションの他のクラスとの比較方法を推定することができる。潜在的スプライスバリアントは、スプライス機能を部分的にしか無効にしないので、われわれは、また、より緩和された閾値における観察された潜在的スプライスバリアント対予想された潜在的スプライスバリアントの数を評価し、ExACデータセット内の稀少AGまたはGTスプライス切断バリアントと比較して約3倍多い悪影響のある稀少潜在的スプライスバリアントがあると推定した(図38D)。各個体は、タンパク質切り詰めバリアント(図38E)の数にほぼ等しい、約～20個の稀少潜在的スプライス突然変異を持つが、これらのバリアントのすべてがスプライス機能を完全に無効にするわけではない。
【０２５１】
全ゲノムシーケンシングおよび複数の組織部位からのRNA-seqの両方を有する148人の個体を含む、GTExデータの最近のリリースは、われわれがRNA-seqデータ内で直接的に稀少潜在的スプライスバリアントの効果を探すことを可能にする。稀少疾病シーケンシングにおいて遭遇するシナリオを近似するために、われわれは、稀少バリアント(GTExコホート中のシングルトン、および1000ゲノムにおける<1%の対立遺伝子頻度)のみを考察し、これらとバリアントを有する個体に固有であったスプライシング事象とをペアリングした。遺伝子発現および組織発現における差ならびにスプライス異常の複雑さは、ディープラーニング予測の感度および特異性を評価することを困難にするけれども、われわれは、厳しい特異性閾値において、90%を超える稀少潜在的スプライス突然変異がRNA-seq上で有効であるとバリデーションがなされることを見いだしている(図39A)。RNA-seq内に存在している多数の異常スプライシング事象が、ディープラーニング分類器により適度の効果を有すると予測されるバリアントに関連付けられているように見え、これはスプライス機能に部分的にしか影響を及ぼしていないことを示唆している。これらのより敏感な閾値において、新規の接合の約75%は、スプライシング機能において異常を引き起こすことが予測される(図38B)。
【０２５２】
母集団シーケンシングデータ上で強く悪影響を及ぼし、RNA-seq上で高い率で有効であるとバリデーションがなされる潜在的スプライスバリアントを予測することにディープラーニングネットワークが成功したことから、この方法は、稀少疾病シーケンシング研究における追加の診断を識別するために使用することが可能であることが示唆される。この仮説をテストするために、われわれは、自閉症および神経発生障害に対するエクソームシーケンシング研究におけるデノボバリアントを調査し、潜在的スプライス突然変異は影響を受ける個体対健康な兄弟姉妹たちにおいて著しくエンリッチされることを実証した(図40A)。さらに、潜在的スプライス突然変異のエンリッチメントはタンパク質切り詰めバリアントに対するものに比べてわずかに低く、これはわれわれの予測された潜在的スプライスバリアントの約90%が機能していることを示す。これらの値に基づき、疾病を引き起こすタンパク質切り詰めバリアントの約～20%がエクソンおよびエクソンのすぐそばに隣接するヌクレオチドにおける潜在的スプライス突然変異に帰因するものとしてよい(図40B)。イントロン配列全体をインテロゲートすることができる全ゲノム研究にこの数字を外挿することで、われわれは、稀少遺伝性疾患における因果的突然変異の24%が潜在的スプライス突然変異によるものであると推定している。
【０２５３】
われわれは、各個別遺伝子に対するデノボ潜在的スプライス突然変異をコールする確率を推定し、確率と比較して候補疾病遺伝子における潜在的スプライス突然変異のエンリッチメントの推定を可能にした。デノボ潜在的スプライス突然変異は、タンパク質切り詰めバリエーションによって以前にヒットしたが、ミスセンスバリエーションによってヒットしなかった遺伝子内で強くエンリッチされ(図40C)、このことはほとんどが他の作用機構

ではなくむしろハプロ不全を通じて疾病を引き起こし得ることを示していた。予測された潜在的スプライス突然変異をタンパク質切り詰めバリアントのリストに追加することで、われわれは、タンパク質切り詰めバリエーションのみを使用した場合と比較して自閉症における3個の追加の疾病遺伝子と、知的障害における11個の追加の疾病遺伝子とを識別することができる(図40D)。

【０２５４】

疾病の可能性の高い組織が利用できなかった(この場合、脳)患者における潜在的スプライス突然変異のバリデーションを行う実現可能性を評価するために、われわれは、Simon's Simplex Collectionからの予測されたデノボ潜在的スプライス突然変異を有する37人の個体に深RNA-seqを実行し、その個体に存在し、実験における他のすべての個体およびGTExコホートからの149人の個体には存在しない、異常スプライシング事象を探した。われわれは、37人の患者のうちのNNが、予測された潜在的スプライスバリアントによって説明されるRNA-seq(図40E)上で固有の異常スプライシングを示したことを見いだしている。

【０２５５】

要約すると、われわれは、稀少遺伝性疾患における因果的疾病突然変異を識別するのに役立つ十分な精度で潜在的スプライスバリアントを正確に予測するディープラーニングモデルを実証しているということである。われわれは、潜在的スプライシングによって引き起こされる稀少疾病診断の実質的部分が、タンパク質コード領域のみを考慮することによって現在見逃されていると推定し、非コードゲノムにおける浸透性稀少バリエーションの効果を解釈するための方法を開発する必要があることを強く主張する。

【０２５６】

結果

ディープラーニングを使用する一次配列からのスプライシングの正確な予測

われわれは、pre-mRNA転写産物のゲノム配列のみを入力として使用して、pre-mRNA転写産物中の各位置がスプライスドナー、スプライスアクセプター、またはそのいずれでもないか(図37A、図21、図22、図23、および図24)を予測するディープレシデュアルニューラルネットワーク(Heら、2016a)を構築した。スプライスドナーおよびスプライスアクセプターは数万個のヌクレオチドによって分離され得るので、われわれは、非常に大きなゲノム距離に及ぶ配列決定因子を認識することができる32個のDilated畳み込み層(YuおよびKoltun、2016年)からなる新規のネットワークアーキテクチャを採用した。エクソン-イントロン境界に隣接する短いヌクレオチドウィンドウのみを考慮するか(YeoおよびBurge、2004年)、または人間工学により設計された特徴(Xingら、2015年)、または表現もしくはスプライス因子結合(Jhaら、2017年)などの実験データに頼っている以前の方法とは対照的に、われわれのニューラルネットワークは、隣接構成配列の10,000個のヌクレオチドを評価してpre-mRNA転写産物中の各位置のスプライス機能を予測することによって一次配列から直接的にスプライシング決定因子を学習する。

【０２５７】

われわれは、ニューラルネットワークのパラメータを訓練するためにヒト染色体のサブセット上のGENCODEアノテーションされたpre-mRNA転写産物配列(Harrowら、2012年)を使用し、またパラログを除いて、残りの染色体上の転写産物を使用し、ネットワークの予測をテストした。テストデータセット内のpre-mRNA転写産物について、ネットワークは、閾値において正しく予測されたスプライス部位の割合である95%のTop-k精度でスプライス接合を予測し、予測された部位の数はテストデータセット内に存在するスプライス部位の実際の数に等しい(Boydら、2012年、YeoおよびBurge、2004年)。CFTRなどの100kbを超える偶数遺伝子は、ヌクレオチド精度で完全に再構成されることが多い(図37B)。ネットワークがエクソン配列バイアスに単純に依存していないことを確認するために、われわれは、長い非コードRNA上でネットワークをテストした。われわれの精度を低下させると予想される、非コード転写産物アノテーションが不完全であるにもかかわらず、ネットワークは、84% top-k精度でlincRNAs内の既知のスプライス接合を予測し(図42Aおよび図42B)、これはタンパク質コード選択圧のない任意の配列上のスプライセオソームの挙動を近似する

ことができることを示している。

【０２５８】

テストデータセット内の各GENCODEアノテーションされたエクソン(各遺伝子の最初のエクソンと最後のエクソンを除く)について、われわれは、また、Gene and Tissue Expression atlas(GTEx)(The GTEx Consortiumら、2015年)からのRNA-seqデータに基づき、ネットワークの予測スコアがエクソン包含対エクソンスキッピングを支持するリードの割合と相関するかどうかを調査した(図37C)。GTEx組織にわたって構造的にスプライスインまたはスプライスアウトされたエクソンは、それぞれ、1または0に近い予測スコアを有していたが、実質的な程度の代替スプライシング(サンプル上で平均化された10～90%の間のエクソン包含)を受けたエクソンは中間スコアに向かう傾向があった(ピアソン相関=0.78、P 0)。

【０２５９】

われわれは、次に、際立った精度を達成するためにネットワークによって利用される配列決定因子を理解することに努めた。われわれは、アノテーションされたエクソンの近くの各ヌクレオチドの体系だったコンピュータ内置換を実行し、隣接するスプライス部位におけるネットワークの予測スコアに対する効果を測定した(図37E)。われわれは、スプライスドナーモチーフの配列を切断すると、インビボでのエクソンスキッピング事象により観察されているように、上流のスプライスアクセプター部位も失われることをネットワークが予測することを頻繁に引き起こすことを見いだしたが、これは最適な距離でペアの上流アクセプターモチーフと下流のドナーモチーフセットとの間のエクソン定義によって著しい特異度が与えられることを示している(Berget、1995年)。スプライシング信号に寄与する追加のモチーフは、SRタンパク質ファミリーの適切に特徴付けられている結合モチーフおよび分岐点を含む(図43Aおよび図43B)(Fairbrotherら、2002年、ReedおよびManiatis、1988年)。これらのモチーフの効果は、エクソン内のそれらの位置に大きく依存し、これはその役割が競合するアクセプター部位とドナー部位とを区別することによってイントロン-エクソン境界の正確な位置決めを指定すること含むことを示唆している。

【０２６０】

可変入力配列構成でネットワークを訓練することは、スプライス予測の精度に際立った影響を及ぼし(図37E)、これはスプライス部位から最大10,000nt離れている長距離配列決定因子は最適に近いモチーフを有する多数の非機能部位から機能的スプライス接合を識別するうえで本質的であることを示している。長距離および短距離特異性決定因子を調べるために、われわれは、80ntの配列構成で訓練されたモデル(SpliceNet-80nt)によってアノテーションされた接合に割り当てられたスコアと、10,000ntの構成で訓練された完全なモデル(SpliceNet-10k)とを比較した。80ntの配列構成で訓練されたネットワークは、典型的な長さ(エクソンに対しては150nt、イントロンに対しては～1000nt)のエクソンまたはイントロンに隣接する接合により低いスコアを割り当てる(図37F)が、これはそのような部位が異常に長いまたは短いエクソンおよびイントロンのスプライス部位と比較して弱いスプライスモチーフを有する傾向があることに呼応する以前の観測結果(Amitら、2012年、Gelfmanら、2012年、Liら、2015年)と一致する。対照的に、10,000ntの配列構成上で訓練されたネットワークは、エクソンまたはイントロン長によって与えられる長距離特異性を説明できることから、スプライスモチーフが比較的弱いにもかかわらず、平均長のイントロンおよびエクソンに対する選好を示す。長い途切れのないイントロンにおけるより弱いモチーフのスキッピングは、エクソンポージングが存在しないときに実験的に観察されるより高速なRNAポリメラーゼII伸長と一致し、これはスプライセオソームが次善のモチーフを認識する時間を短縮することを可能にし得る(Closeら、2012年、Jonkersら、2014年、Velosoら、2014年)。われわれの発見から、平均スプライス接合は実質的な特異性を与える有利な長距離配列決定因子を保有し、ほとんどのスプライスモチーフにおいて許容される配列縮退の程度が大きいことを説明する。

【０２６１】

スプライシングは同時転写で生じるので(Cramerら、1997年、Tilgnerら、2012年)、ク

ロマチン状態と同時転写スプライシングとの間の相互作用も、エクソン定義をガイドし(Lucoら、2011年)、クロマチン状態が一次配列から予測可能である範囲でネットワークによって利用される潜在的可能性を有するものとしてよい。特に、ヌクレオソームの位置決めのゲノム全体の研究から、ヌクレオソーム占有率がエクソンではより高いことが示されている(Anderssonら、2009年、Schwartzら、2009年、Spiesら、2009年、Tilgnerら、2009年)。ネットワークがスプライス部位予測にヌクレオソーム位置決めの配列決定因子を使用しているかどうかをテストするために、われわれは、ゲノム上で150nt(おおよそ、平均エクソンのサイズ)だけ隔てられている最適なアクセプターモチーフとドナーチーフとの対を調べ、モチーフの対が結果としてその軌跡においてエクソン包含を引き起こすかどうかを予測することをネットワークに求めた(図37G)。われわれは、エクソン包含に対して有利であると予測された位置が、遺伝子間領域にあっても、高いヌクレオソーム占有率を有する位置と相関した(スピアマン相関=0.36、P 0)ことを見いだしており、この効果はGC内容に対する制御の後も持続する(図44A)。これらの結果は、ネットワークが一次配列からのヌクレオソーム位置決めを予測することを暗黙のうちに学習しており、それをエクソン定義における特異性決定因子として利用することを示唆している。平均長のエクソンおよびイントロンと同様に、ヌクレオソーム上に位置決めされたエクソンは弱いローカルスプライスモチーフを有し(図44B)、これは代償的因子が存在する場合の縮退モチーフに対するより高い許容性と一致する(Spiesら、2009年)。

【０２６２】

複数の研究において、エクソンとヌクレオソーム占有率との間の相関関係が報告されているけれども、エクソン定義中のヌクレオソーム位置決めに対する因果的役割は、まだ確固たるものではない。149人の個体からのデータをGenotype-Tissue Expression (GTEx)コホート(The GTEx Consortiumら、2015年)からの両方のRNA-seqおよび全ゲノムシーケンシングとともに使用して、われわれは、単一の個体にとってはプライベートであり、プライベートのスプライス部位生成遺伝子突然変異に対応する新規のエクソンを識別した。これらのプライベートのエクソン生成事象は、K562およびGM12878細胞(P=0.006、並べかえ検定による、図37H)における既存のヌクレオソーム位置決めに、これらの細胞系が対応するプライベートの遺伝子突然変異を欠いている可能性が最も高いとしても、有意に関連付けられた。われわれの結果は、遺伝子バリアントが、結果として生じる新規のエクソンが既存のヌクレオソーム占有領域に重なる場合に新規のエクソンの生成をトリガする可能性がより高いことを示しており、これはエクソン定義を促進する際にヌクレオソーム位置決めに対する因果的役割を支持する。

【０２６３】

RNA-seqデータにおける予測された潜在的スプライス突然変異の検証

われわれは、参照pre-mRNA転写産物配列およびバリアントを含む代替転写産物配列の両方でエクソン-イントロン境界を予測し、スコア間の差(Δスコア)を取ることによって、ディープラーニングネットワークをスプライス変更機能に対する遺伝的バリアントの評価に拡張した。重要なことは、ネットワークが参照転写産物配列およびスプライス接合アノテーション上で訓練されただけであり、訓練中にバリアントデータを決して見ることなく、バリアントの効果の予測をスプライシングの配列決定因子を正確にモデル化するネットワークの能力に関する困難なテストにしたことである。

【０２６４】

われわれは、全ゲノムシーケンシングおよび複数の組織からのRNA-seqの両方を有する149人の個体を含む、GTExコホート(The GTEx Consortiumら、2015年)におけるRNA-seqデータ内の潜在的スプライスバリアントの効果を探した。稀少疾病シーケンシングにおいて遭遇したシナリオを近似するために、われわれは、最初に、稀少なプライベート突然変異(GTExコホートのただ1人の個体に存在している)に集中した。われわれは、ニューラルネットワークにより機能的結果を有することが予測されたプライベート突然変異がプライベートの新規スプライス接合、およびプライベートのエクソンスキッピング事象におけるスキップオーバーされたエクソンの境界で強くエンリッチされたことを見いだしており、これ

はこれらの予測の大部分が機能的であることを示唆している。

【０２６５】

正常および異常なスプライスイソ型の相対的産生に対するスプライス部位生成バリアントの効果を定量化するために、われわれは、部位をカバーするリードの総数の割合として新規スプライス事象を支持するリードの数を測定した(図38C)(Cummingsら、2017年)。スプライス部位切断バリアントについて、われわれは、多数のエクソンがエクソンスキッピングの低いベースライン率を有し、バリアントの効果がエクソンスキッピングリードの割合を増やすということであることを観察した。したがって、われわれは、切断接合点でスプライスしたリードの割合の減少と、エクソンをスキップしたリードの割合の増大の両方を計算し、2つの効果のうちの大きい方を取った(図45およびSTAR法)。

【０２６６】

高い信頼度で予測された潜在的スプライスバリアント(ΔスコアΔ0.5)は、本質的なGTまたはAGスプライス切断のレートの3/4でRNA-seq上でバリデーションする(図38D)。潜在的スプライスバリアントのバリデーション率およびエフェクトサイズは両方ともそのΔスコアを密に追跡し(図38Dおよび図38E)、モデルの予測スコアがバリアントのスプライス変更潜在的可能性に対する良好なプロキシであることを実証する。バリデーションが行われたバリアント、特に低いスコア(Δスコア<0.5)を有するバリアントは、不完全な浸透性を有することが多く、その結果、代替スプライシングはRNA-seqデータにおける異常転写産物と正常転写産物の両方の混合を生成する。バリデーション率およびエフェクトサイズのわれわれの推定は控えめであり、説明が付かないスプライスイソ型の変化および、未成熟終止コドンを頻繁に導入するので異常スプライス転写産物を優先的に分解する、ナンセンス変異依存分解の両方により、真の値を過小評価していることもあり得る(図38Cおよび図45)。これは、本質的なGTおよびAGスプライスジヌクレオチドを切断するバリアントの平均エフェクトサイズが完全に浸透性のあるヘテロ接合バリアントについて予想される50%より小さいことが証拠である。

【０２６７】

mRNA転写産物の観察されたコピーの少なくとも3/10で異常スプライスイソ型を生成する潜在的スプライスバリアントでは、ネットワークは、バリアントがエクソンに近いときに71%、バリアントが深イントロン配列内にあるときに41%の感度を有する(ΔスコアΔ0.5、図38F)。これらの発見は、深イントロンバリアントは、場合によってエクソンの近くに存在するように選択されているより少ない特異性決定因子を深イントロン領域が含むので予測することがより難しいことを示している。

【０２６８】

既存の方法に対してわれわれのネットワークのパフォーマンスをベンチマーキングするため、われわれは、稀少遺伝病診断の文献で参照されている3つの人気がある分類器、すなわち、GeneSplicer(Perteaら、2001年)、MaxEntScan(YeoおよびBurge、2004年)、およびNNSplice(Reeseら、1997年)を選択し、可変閾値においてRNA-seqバリデーション率および感度をプロットした(図38G)。この分野の他の人々の経験と同様に(Cummingsら、2017年)、われわれは、おそらくローカルモチーフに集中し、大部分は長距離特異性決定因子に関わっていないので、場合によってはスプライシングに影響を及ぼし得るゲノム規模の非常に多くの非コードバリアントが与えられた場合に既存の分類器が不十分な特異性を有することを見いだしている。

【０２６９】

既存の方法と比較してパフォーマンスに大きなギャップがある場合、われわれは、追加の制御を実行して、RNA-seqデータのわれわれの結果が過剰適合によって混乱する可能性をなくすようにした。第1に、われわれは、プライベートバリアントおよびGTExコホート内の複数の個体に存在するバリアントについて別々にバリデーションおよび感度解析を繰り返した(図46A、図46B、および図46C)。スプライシング機構もディープラーニングモデルも対立遺伝子頻度情報へのアクセスを有していないので、ネットワークが対立遺伝子頻度スペクトルにわたって類似のパフォーマンスを有することを検証することが重要な制御

である。われわれは、同じΔスコア閾値において、プライベートおよび共通潜在的スプライスバリアントがRNA-seqにおいてバリデーション率の有意な差を示さず(P>0.05、フィッシャーの正確確率検定)、これはネットワークの予測が対立遺伝子頻度に対してロバストであることを示している、ことを見いだしている。

【０２７０】

第2に、新規スプライス接合を作成することができる異なる種類の潜在的スプライスバリアント上のモデルの予測のバリデーションを行うために、われわれは、新規GTまたはAGジヌクレオチドを生成するバリアント、伸長されたアクセプターまたはドナーモチーフに影響を及ぼすもの、およびより遠くにある領域内で生じるバリアントを別々に評価した。われわれは、潜在的スプライスバリアントが3つのグループの間におおよそ等しく分散されること、および同じΔスコア閾値において、グループ間にバリデーション率またはエフェクトサイズの有意な差がないことを見いだしている(それぞれ、P>0.3の一様性のχ^2検定およびP>0.3のマン・ホイットニーのU検定、図47Aおよび図47B)。

【０２７１】

第3に、われわれは、訓練に使用される染色体上のバリアントおよび染色体の残りの上のバリアントについて別々にRNA-seqバリデーションおよび感度解析を実行した(図48Aおよび図48B)。ネットワークは、参照ゲノム配列およびスプライスアノテーションでのみ訓練され、訓練中にバリアントデータに曝されなかったけれども、われわれは、バリアント予測におけるバイアスがネットワークが訓練している染色体における参照配列を見ていたという事実から生じる可能性を除外することを望んでいた。われわれは、ネットワークが訓練およびテスト染色体からのバリアントで等しくうまく機能し、バリデーション率または感度に有意な差がなく(P>0.05、フィッシャーの正確確率検定)、ネットワークのバリアント予測が訓練配列を過剰適合させることによって説明される可能性がないことを示していることを見いだしている。

【０２７２】

潜在的スプライスバリアントを予測することは、われわれのモデルさらには他のスプライス予測アルゴリズムの結果によって反映されているように、アノテーションされたスプライス接合予測することよりも難しい問題である(図37Eおよび図38Gと比較)。重要な理由は、2種類の解析の間のエクソン包含率の基礎となる分布に差があることである。膨大なGENCODEアノテーションされたエクソンは強い特異性決定因子を有し、その結果、構成スプライシングおよび予測スコアは1に近いものとなる(図37C)。対照的に、ほとんどの潜在的スプライスバリアントは部分的にしか浸透性を有さず(図38Dおよび図38E)、低から中までの予測スコアを有し、頻繁に、正常転写産物と異常転写産物の両方の混合が生成される代替スプライシングを引き起こす。このため、潜在的スプライスバリアントの効果を予測する後者の問題は、アノテーションされたスプライス部位を識別することに比べてより本質的に難しい問題になっている。ナンセンス変異依存分解、説明のつかないイソ型変化、およびRNA-seqアッセイの制限などの追加の因子が、さらに、RNA-seqバリデーション率を引き下げることに関わっている(図38Cおよび図45)。

【０２７３】

弱い潜在的スプライスバリアントから組織特有の代替スプライシングが頻繁に生じる

代替スプライシングは、異なる組織における転写産物の多様性および発生段階を増やすのに使われる遺伝子調節の主要様式であり、その調節不全は、疾病過程に関連付けられる(Blencoweら、2006年、Irimiaら、2014年、Kerenら、2010年、LicatalosiおよびDarnell、2006年、Wangら、2008年)。予想外に、われわれは、潜在的スプライス突然変異によって形成される新規のスプライス接合の相対的使用度は、組織間で実質的に異なり得ることを見いだしている(図39A)。さらに、スプライシングにおける組織特有の相違を引き起こすバリアントは複数の個体において再現性を有し(図39B)、確率論的効果ではなく組織特有の生物学がこれらの相違の基礎にあることを示している。われわれは、弱いおよび中間の予測されたスコア(Δスコア0.35～0.8)を有する潜在的スプライスバリアントの35%が組織間で生成される正常および異常の転写産物の割合の著しい差を示すことを見いだしてい

る(χ²検定に対してボンフェローニ相関P<0.01、図39C)。これは、高い予測スコア(Δスコア>0.8)を有するバリアントと対照的であり、これは組織特有の効果を生成する可能性は著しく低かった(P=0.015)。われわれの発見は、代替的にスプライスされたエクソンが、それぞれ1または0に近いスコアを有する、構造的にスプライスインまたはスプライスアウトされているエクソンと比較して、中間の予想スコア(図37C)を有する傾向があるという以前の観察結果と整合している。

【０２７４】

これらの結果は、RNA結合タンパク質のクロマチン構成および結合などの、組織特有の因子が好ましさにおいて近い2つのスプライス接合の間の競争を揺らす可能性があるモデルを裏付ける(Gelfmanら、2013年、Lucoら、2010年、Shuklaら、2011年、Uleら、2003年)。強い潜在的スプライスバリアントは、エピジェネティック構成に関係なく正常イソ型から異常イソ型へスプライシングを完全にシフトする可能性が高いが、より弱いバリアントはスプライス接合選択を決定境界に近づけ、その結果、異なる組織型および細胞構成において代替接合が使用されることになる。これは、新規の代替スプライシング多様性を生成する際に潜在的スプライス突然変異によって果たされる予想外の役割を際立たせ、次いで、自然選択は有用な組織特有の代替スプライシングを形成する突然変異を温存する機会を有するであろう。

【０２７５】

予測された潜在的スプライスバリアントはヒト母集団において強い悪影響を有する

予測された潜在的スプライスバリアントはRNA-seqにおいて高率でバリデーションを行うけれども、多くの場合において、それらの効果は完全な浸透性を有さず、正常および異常のスプライスイソ型の両方の混合が生成され、これらの潜在的スプライス変更バリアントの割合が機能的に有意であり得ない確率を高める。予測された潜在的スプライスバリアント上の自然選択のシグネチャを調べるために、われわれは、Exome Aggregation Consortium (ExAC)データベース(Lekら、2016年)からの60,706個のヒトエクソーム中に存在している各バリアントにスコアを付け、エクソン-イントロン境界を変更すると予測されたバリアントを識別した。

【０２７６】

予測されたスプライス変更バリアントに作用する陰性選択の程度を測定するために、われわれは、共通の対立遺伝子頻度に見られる予測されたスプライス変更バリアントの数をカウントし(ヒト母集団内で≧0.1%)、それをExACにおけるシングルトン対立遺伝子頻度での予測されたスプライス変更バリアントの数と比較した(すなわち、60,706人の個体のうちの1人で)。ヒト母集団サイズの最近の指数関数的拡大により、シングルトンバリアントは、純化淘汰によって最低限フィルタ処理された最近形成された突然変異を表す(Tennessenら、2012年)。対照的に、共通バリアントは、純化淘汰の篩いに通された中立突然変異のサブセットを表す。したがって、シングルトンバリアントに関する共通対立遺伝子頻度スペクトルにおける予測されたスプライス変更バリアントの枯渇は、悪影響を有する、したがって機能的である予測されたスプライス変更バリアントの割合の推定をもたらす。タンパク質コード配列に対する交絡効果を回避するために、われわれは、解析を本質的なGTおよびAGジヌクレオチドの外に置かれている同義バリアントおよびイントロンバリアントに制限し、スプライス変更効果を有することも予測されるミスセンス突然変異を除外した。

【０２７７】

共通対立遺伝子頻度において、自信を持って予測された潜在的スプライスバリアント(Δスコア≧0.8)は、予想と比較して相対的枯渇が証拠として示すように、強い陰性選択の下にある(図40A)。この閾値では、大半のバリアントがRNA-seqデータ内で完全浸透性に近いと予想される場合(図38D)、予測された同義およびイントロン潜在的スプライス突然変異は共通対立遺伝子頻度で78%だけ枯渇しており、これはフレームシフト、ストップゲイン、および本質的なGTまたはAGスプライス切断バリアントの82%の枯渇に匹敵する(図40B)。インフレーム変化を引き起こすバリアントを超えてフレームシフトを引き起こす潜在的

スプライスバリアントを考えたときに陰性選択の影響はより大きくなる(図40C)。フレームシフト結果を伴う潜在的スプライスバリアントの枯渇は、タンパク質切り詰めバリエーションの他のクラスのものとほぼ同一であり、これは、近イントロン領域内の自信を持って予測された潜在的スプライス突然変異の大部分(既知のエクソン-イントロン境界から≦50nt)は機能的であり、ヒト母集団内で強い悪影響を有することを示している。

【０２７８】

この解析結果を既知のエクソン-イントロン境界から>50ntだけ深イントロン領域に拡大するために、われわれは、Genome Aggregation Database(gnomAD)コホートからの15,496人から全ゲノムシーケンシングデータを集約し(Lekら、2016年)、共通対立遺伝子頻度における潜在的スプライス突然変異の観察されたカウントおよび予想されたカウントを計算した。全体的に、われわれは、エクソン-イントロン境界(図40D)から>50ntの距離で共通潜在的スプライス突然変異の56%の枯渇を観察した(ΔスコアAG0.8)が、これはRNA-seqデータで観察した通り、深イントロンバリアントの影響を予測することが非常に困難であることと一致している。

【０２７９】

われわれは、次に、gnomADコホート内の個体毎に稀少潜在的スプライス突然変異の数を測定することによって、潜在的スプライス突然変異がタンパク質コードバリエーションの他の種類に関して浸透性遺伝病に関わる潜在的可能性を推定することに努めた。陰性選択の下にある予測された潜在的スプライス突然変異の割合に基づくと(図40A)、平均的な人間は、～11個の稀少タンパク質切り詰めバリアント(図40E)と比較して、～5個の稀少機能的潜在的スプライス突然変異(対立遺伝子頻度<0.1%)を保持する。潜在的スプライスバリアントは、数で、本質的なGTまたはAGスプライス切断バリアントをおおよそ2:1で上回る。われわれは、インフレーム変更を形成すること、またはスプライシングを異常イソ型に完全にシフトさせないことから、これらの潜在的スプライスバリアントの有意な割合が遺伝子機能を完全には無効にし得ないことに注意している。

【０２８０】

デノボ潜在的スプライス突然変異は稀少遺伝性疾患の主原因である

自閉症スペクトラム障害および重大な知的障害を患っている患者の大規模シーケンシング研究では、神経発達経路内の遺伝子を切断するデノボタンパク質コード突然変異(ミスセンス、ナンセンス、フレームシフト、および本質的スプライスジヌクレオチド)の中心的役割を実証した(Fitzgeraldら、2015年、Iossifovら、2014年、McRaeら、2017年、Nealeら、2012年、De Rubeisら、2014年、Sandersら、2012年)。変更されたスプライシングを通じて作用する非コード突然変異の臨床的影響を評価するために、われわれは、ニューラルネットワークを適用してDeciphering Developmental Disordersコホート(DDD)(McRaeら、2017年)からの知的障害を患っている4,293人の個体、Simons Simplex Collection(De Rubeisら、2014年、Sandersら、2012年、Turnerら、2016年)およびAutism Sequencing Consortiumからの自閉症スペクトラム障害(ASD)を患っている3,953人の個体、ならびにSimons Simplex Collectionからの2,073人の影響を受けていない兄弟姉妹の対照におけるデノボ突然変異の効果を予測した。研究におけるデノボバリアント確認の差を制御するために、われわれは、個体毎の同義突然変異の数がコホート全体にわたって同じになるようにデノボバリアントの予想される数を正規化した。

【０２８１】

スプライシングを切断することを予測されているデノボ突然変異は、健康な対照と比較して、知的障害では1.51倍(P=0.000416)、および自閉症スペクトラム障害では1.30倍(P=0.0203)にエンリッチされる(ΔスコアAG0.1、図41A、図43A、および図43B)。スプライス切断突然変異は、また、エンリッチメントが二重タンパク質コーディングおよびスプライシング効果を有する突然変異によってのみ説明される可能性のあることを除き、同義およびイントロン突然変異のみを考慮したときに患者対対照において著しくエンリッチされる(図49A、図49B、図49C)。影響のある個体対影響のない個体におけるデノボ突然変異の過剰に基づき、潜在的スプライス突然変異は、各研究におけるシーケンシングカバレッジまた

はバリアント確認を欠いた領域内の突然変異の予想される割合を調整した後に、自閉症スペクトラム障害では病原性突然変異の約11%、および知的障害では9%(図41B)を含むと推定される。影響を受ける個体のほとんどのデノボ予測潜在的スプライス突然変異は、ΔスコアA<0.5を有しており(図41C、図50A、および図50B)、GTEx RNA-seqデータセット内の類似のスコアを持つバリアントに基づき正常転写産物と異常転写産物との混合物を生成すると予想されるであろう。

【０２８２】

可能性と比較して候補疾病遺伝子内の潜在的スプライス突然変異のエンリッチメントを推定するために、われわれは、トリヌクレオチド構成を使用して突然変異率を調整することで各個別遺伝子に対するデノボ潜在的スプライス突然変異をコールする確率を計算した(Samochaら、2014年)(Table S4)。新規遺伝子発見において潜在的スプライス突然変異とタンパク質コード突然変異の両方を組み合わせることで、知的障害に関連付けられている5個の追加の候補遺伝子、およびタンパク質コード突然変異のみを考慮したときに発見閾値(FDR<0.01)を下回っている自閉症スペクトラム障害(図41Dおよび図45)に関連付けられている2個の追加の遺伝子が得られる(Kosmickiら、2017年、Sandersら、2015年)。

【０２８３】

自閉症患者におけるデノボ潜在的スプライス突然変異の実験的バリデーション

われわれは、LCL表現の少なくとも最小レベルを有する遺伝子における予測されたデノボ潜在的スプライス突然変異を持っている、Simons Simplex Collectionからの36人の個体から末梢血液由来リンパ芽球様細胞株(LCL)を取得し(De Rubeisら、2014年、Sandersら、2012年)、各個体はその近親内の自閉症の患者のみを代表していた。ほとんどの稀少遺伝病の場合と同様に、関連する組織および細胞型(たぶん発生中の脳)はアクセス可能でなかった。したがって、われわれは、LCLにおけるこれらの転写産物の多くの弱い表現を補償するために高深度mRNAシーケンシング(サンプル毎に～3億5000万×150bpの単一リード、GTExのカバレッジのおおよそ10倍)を実行した。われわれが、単純に最上位予測ではなく、予測された潜在的スプライスバリアントの代表的セットのバリデーションを行っていることを確実にするために、われわれは、比較的許容性のある閾値(スプライス損失バリアントに対してΔスコア>0.1およびスプライス利得バリアントに対してΔスコア>0.5、STAR法)を適用し、これらの基準を満たすすべてのデノボバリアント上で実験的バリデーションを実行した。

【０２８４】

注目する遺伝子で不十分なRNS-seqカバレッジを有していた8人を除いた後、われわれは、患者28人中21人で予測されたデノボ潜在的スプライス突然変異に関連付けられている固有の異常スプライシング事象を識別した(図41Eおよび図51A、図51B、図51C、図51D、図51E、図51F、図51G、図51H、図51I、および図51J)。これらの異常スプライシング事象は、深LCL RNA-seqが取得された他の35人の個体、さらにはGTExコホートからの149人の個体にはなかった。われわれは、21の確認されたデノボ潜在的スプライス突然変異のうち、新規の接合生成の9症例、エクソンスキッピングの8症例、およびイントロン保持の4症例、さらにはより複雑なスプライシング異常を観察した(図41F、図46A、図46B、および図46C)。7症例は、転写産物の十分な表現にもかかわらず、LCL内に異常スプライシングを示さなかった。これらのサブセットは偽陽性予測を表し得るけれども、いくつかの潜在的スプライス突然変異は結果として、これらの実験条件の下でLCL内で観察可能でない組織特有の代替スプライシングを生じ得る。

【０２８５】

自閉症スペクトラム障害の患者における予測された潜在的スプライス突然変異の高いバリデーション率(75%)は、RNA-seqアッセイの制限があるにもかかわらず、ほとんどの予測が機能的であることを示している。しかしながら、対照と比較した症例におけるデノボ潜在的スプライスバリアントのエンリッチメント(DDDでは1.5倍、ASDでは1.3倍、図41A)は、デノボタンパク質切り詰めバリアントに対して観察されたエフェクトサイズの38%にすぎない(DDDでは2.5倍、ASDでは1.7倍)(Iossifovら、2014年、McRaeら、2017年、De Rubei

sら、2014年)。これにより、われわれは、多くが正常な転写産物の生成を部分的にしか途絶えさせないので、機能的潜在的スプライス突然変異はタンパク質切り詰め突然変異の古典的形態(ストップゲイン、フレームシフト、および本質的スプライスジヌクレオチド)の臨床的浸透性のおおよそ50%を有すると定量化することを可能にする。実際、FECHにおけるc.315-48T>C(Gouyaら、2002年)およびGAAにおけるc.-32-13T>G(Boerkoelら、1995年)などの、メンデル性疾病における大半のきちんと特徴付けられた潜在的スプライス突然変異のうちのいくつかは、穏やかな表現型または遅い発症年齢に関連付けられている低次形態対立遺伝子である。臨床的浸透度の推定は、比較的許容性のある閾値(Δスコア≧0.1)を満たすすべてのデノボバリアントについて計算され、より強い予測スコアを有するバリアントは、それに対応してより高い浸透度を有することが予想されるであろう。

【０２８６】

ASDおよびDDDコホートにまたがる症例対対照におけるデノボ突然変異の過剰に基づき、250症例が、デノボタンパク質切り詰めバリアントによって説明され得る909症例と比較してデノボ潜在的スプライス突然変異によって説明され得る(図41B)。これは、潜在的スプライス突然変異の低減された浸透度が組み込まれた後に、一般集団内の人毎に稀少タンパク質切り詰めバリアント(～11)と比較して稀少潜在的スプライス突然変異の平均数(～5)のわれわれの以前の推定と一致している。ゲノム上の潜在的スプライス突然変異の広範囲の分布は、神経発生障害における潜在的スプライス突然変異によって説明される症例の割合(9～11%、図41B)が、原発性疾患機構が機能的タンパク質の喪失である他の稀少遺伝性疾患に一般化する可能性が高いことを示唆している。スプライス変更突然変異の解釈を円滑にするために、われわれは、ゲノムワイドでのすべての可能な単一ヌクレオチド置換に対するΔスコア予測を事前計算しており、それらをリソースとして科学団体に提供する。われわれは、このリソースは遺伝的バリエーションのこの以前には過小評価されていたソースの理解を促進すると確信している。

【０２８７】

特定の実装形態

われわれは、ゲノム配列(たとえば、ヌクレオチド配列またはアミノ酸配列)内のスプライス部位を検出するために訓練されたAtrous畳み込みニューラルネットワークを使用することについて製造システム、製造方法、および製造品を説明する。一実装形態の1つまたは複数の特徴がベースの実装形態と組み合わされ得る。相互排他的でない実装形態は、組合せ可能であると教示される。一実装形態の1つまたは複数の特徴が他の実装形態と組み合わされ得る。本開示は、定期的に、これらのオプションについてユーザに通知する。これらのオプションを繰り返す言及のいくつかの実装形態からの省略は、先行する節において教示されている組合せを制限するものとしてみなすべきでなく、これらの言及は、次の実装形態の各々に参照により順に組み込まれる。

【０２８８】

この節では、モジュールおよび段階という用語を交換可能に使用する。

【０２８９】

開示されている技術のシステム実装形態は、メモリに結合されている1つまたは複数のプロセッサを含む。メモリは、ゲノム配列(たとえば、ヌクレオチド配列)内のスプライス部位を識別するスプライス部位検出器を訓練するためのコンピュータ命令をロードされる。

【０２９０】

図30に示されているように、システムは、ドナースプライス部位の少なくとも50000個の訓練例、アクセプタースプライス部位の少なくとも50000個の訓練例、および非スプライシング部位の少なくとも100000個の訓練例上でAtrous畳み込みニューラルネットワーク(略語ACNN)を訓練する。各訓練例は、各側に少なくとも20個のヌクレオチドが隣接する少なくとも1つの標的ヌクレオチドを有する標的ヌクレオチド配列である。

【０２９１】

ACNNは、訓練可能なパラメータが少ししかない大きい受容野を実現可能にするAtrous/D

ilated畳み込みを使用する畳み込みニューラルネットワークである。Atrous/Dilated畳み込みは、Atrous畳み込みレートまたは拡張係数とも呼ばれるあるステップを用いて入力値をスキップすることによって、カーネルがその長さよりも大きい領域にわたって適用される畳み込みである。Atrous/dilated畳み込みは、畳み込みフィルタ/カーネルの要素間の間隔を加え、それによって、畳み込み演算が実行されるときにより大きい間隔における近傍の入力エントリ(たとえば、ヌクレオチド、アミノ酸)が考慮される。これは長距離構成依存性を入力に組み込むことを可能にする。Atrous畳み込みは、部分畳み込み計算を隣接するヌクレオチドが処理されるときに再使用できるように保存する。

【０２９２】

図30に示されているように、ACNNを使用して訓練例を評価するために、システムは、ACNNへの入力として、少なくとも40個の上流構成ヌクレオチドおよび少なくとも40個の下流構成ヌクレオチドがさらに隣接する標的ヌクレオチド配列を提供する。

【０２９３】

次いで、図30に示されているように、この評価に基づき、ACNNは、出力として、標的ヌクレオチド配列内の各ヌクレオチドがドナースプライス部位、アクセプタースプライス部位、または非スプライシング部位である可能性に対するトリプレットスコアを生成する。

【０２９４】

このシステム実装形態および開示されている他のシステムは、任意選択で、次の特徴のうちの1つまたは複数を含む。システムは、開示されている方法に関連して説明されている特徴も含むことができる。簡潔にするため、システム特徴の代替的組合せは、個別には列挙しない。製造システム、製造方法、および製造品に適用可能な特徴は、ベースとなる特徴の法令に定めるクラスのセット毎に繰り返されない。読者は、この節に明記されている特徴が他の法令に定められているクラスにおけるベースとなる特徴とどのように容易に組み合わされ得るかを理解するであろう。

【０２９５】

図25、図26、および図27に示されているように、入力は、各側に2500個のヌクレオチドが隣接する標的ヌクレオチドを有する標的ヌクレオチド配列を含むことができる。そのような実装形態において、標的ヌクレオチド配列は、5000個の上流構成ヌクレオチドおよび5000個の下流構成ヌクレオチドがさらに隣接する。

【０２９６】

入力は、各側に100個のヌクレオチドが隣接する標的ヌクレオチドを有する標的ヌクレオチド配列を含むことができる。そのような実装形態において、標的ヌクレオチド配列は、200個の上流構成ヌクレオチドおよび200個の下流構成ヌクレオチドがさらに隣接する。

【０２９７】

入力は、各側に500個のヌクレオチドが隣接する標的ヌクレオチドを有する標的ヌクレオチド配列を含むことができる。そのような実装形態において、標的ヌクレオチド配列は、1000個の上流構成ヌクレオチドおよび1000個の下流構成ヌクレオチドがさらに隣接する。

【０２９８】

図28に示されているように、システムは、ドナースプライス部位の150000個の訓練例、アクセプタースプライス部位の150000個の訓練例、および非スプライシング部位の800000000個の訓練例上でACNNを訓練することができる。

【０２９９】

図19に示されているように、ACNNは、最低から最高までの順に配置構成されている残差ブロックのグループを含むことができる。残差ブロックの各グループは、残差ブロック内の畳み込みフィルタの数、残差ブロックの畳み込みウィンドウサイズ、および残差ブロックのAtrous畳み込みレートによってパラメータ化される。

【０３００】

図21、図22、図23、および図24に示されているように、ACNNでは、Atrous畳み込みレートは、下位の残差ブロックグループから上位の残差ブロックグループへと非指数関数的に

高まる。

【０３０１】

図21、図22、図23、および図24に示されているように、ACNNにおいて、畳み込みウィンドウサイズは、残差ブロックのグループ間で異なる。

【０３０２】

ACNNは、40個の上流構成ヌクレオチドおよび40個の下流構成ヌクレオチドがさらに隣接する標的ヌクレオチド配列を含む入力を評価するように構成され得る。そのような一実装形態において、ACNNは、4つの残差ブロックおよび少なくとも1つのスキップコネクションからなる1つのグループを含む。各残差ブロックは、32個の畳み込みフィルタ、11畳み込みウィンドウサイズ、および1 Atrous畳み込みレートを有する。ACNNのこの実装形態は、本明細書では「SpliceNet80」と称され、図21に示されている。

【０３０３】

ACNNは、200個の上流構成ヌクレオチドおよび200個の下流構成ヌクレオチドがさらに隣接する標的ヌクレオチド配列を含む入力を評価するように構成され得る。そのような一実装形態において、ACNNは、4つの残差ブロックおよび少なくとも2つのスキップコネクションからなる少なくとも2つのグループを含む。第1のグループ内の各残差ブロックは、32個の畳み込みフィルタ、11畳み込みウィンドウサイズ、および1 Atrous畳み込みレートを有する。第2のグループ内の各残差ブロックは、32個の畳み込みフィルタ、11畳み込みウィンドウサイズ、および4 Atrous畳み込みレートを有する。ACNNのこの実装形態は、本明細書では「SpliceNet400」と称され、図22に示されている。

【０３０４】

ACNNは、1000個の上流構成ヌクレオチドおよび1000個の下流構成ヌクレオチドがさらに隣接する標的ヌクレオチド配列を含む入力を評価するように構成され得る。そのような一実装形態において、ACNNは、4つの残差ブロックおよび少なくとも3つのスキップコネクションからなる少なくとも3つのグループを含む。第1のグループ内の各残差ブロックは、32個の畳み込みフィルタ、11畳み込みウィンドウサイズ、および1 Atrous畳み込みレートを有する。第2のグループ内の各残差ブロックは、32個の畳み込みフィルタ、11畳み込みウィンドウサイズ、および4 Atrous畳み込みレートを有する。第3のグループ内の各残差ブロックは、32個の畳み込みフィルタ、21畳み込みウィンドウサイズ、および19 Atrous畳み込みレートを有する。ACNNのこの実装形態は、本明細書では「SpliceNet2000」と称され、図23に示されている。

【０３０５】

ACNNは、5000個の上流構成ヌクレオチドおよび5000個の下流構成ヌクレオチドがさらに隣接する標的ヌクレオチド配列を含む入力を評価するように構成され得る。そのような一実装形態において、ACNNは、4つの残差ブロックおよび少なくとも4つのスキップコネクションからなる少なくとも4つのグループを含む。第1のグループ内の各残差ブロックは、32個の畳み込みフィルタ、11畳み込みウィンドウサイズ、および1 Atrous畳み込みレートを有する。第2のグループ内の各残差ブロックは、32個の畳み込みフィルタ、11畳み込みウィンドウサイズ、および4 Atrous畳み込みレートを有する。第3のグループ内の各残差ブロックは、32個の畳み込みフィルタ、21畳み込みウィンドウサイズ、および19 Atrous畳み込みレートを有する。第4のグループ内の各残差ブロックは、32個の畳み込みフィルタ、41畳み込みウィンドウサイズ、および25 Atrous畳み込みレートを有する。ACNNのこの実装形態は、本明細書では「SpliceNet10000」と称され、図24に示されている。

【０３０６】

標的ヌクレオチド配列内の各ヌクレオチドに対する各トリプレットスコアは、和が1になるように指数関数的に正規化され得る。そのような一実装形態において、システムは、それぞれのトリプレットスコアにおける最高スコアに基づき標的ヌクレオチド内の各ヌクレオチドをドナースプライス部位、アクセプタースプライス部位、または非スプライシング部位として分類する。

【０３０７】

図35に示されているように、ACNNの入力の次元は(C^{u}+L+C^{d})×4として定義されてよく、C^{u}は上流構成ヌクレオチドの数であり、C^{d}は下流構成ヌクレオチドの数であり、Lは標的ヌクレオチド配列内のヌクレオチドの数である。一実装形態において、入力の次元は(5000+5000+5000)×4である。

【０３０８】

図35に示されているように、ACNNの出力の次元はL×3として定義され得る。一実装形態において、出力の次元は5000×3である。

【０３０９】

図35に示されているように、残差ブロックの各グループは、先行する入力を処理することによって中間出力を生成することができる。中間出力の次元は、(I-[{(W-1)*D}*A])×Nとして定義されるものとしてよく、Iは先行する入力の次元であり、Wは残差ブロックの畳み込みウィンドウサイズであり、Dは残差ブロックのAtrous畳み込みレートであり、Aはグループ内のAtrous畳み込み層の数であり、Nは残差ブロック内の畳み込みフィルタの数である。

【０３１０】

図32に示されているように、ACNNはバッチ式に、エポックにおける訓練例を評価する。訓練例は、ランダムにバッチにサンプリングされる。各バッチは所定のバッチサイズを有する。ACNNは、複数のエポック(たとえば、1～10)にわたって訓練例の評価を反復する。

【０３１１】

入力は、2つの隣接する標的ヌクレオチドを有する標的ヌクレオチド配列を含むことができる。2つの隣接する標的ヌクレオチドは、アデニン(略語A)およびグアニン(略語G)であるものとしてよい。2つの隣接する標的ヌクレオチドは、グアニン(略語G)およびウラシル(略語U)であるものとしてよい。

【０３１２】

システムは、訓練例を疎にエンコードし、ワンホットエンコーディングを入力として与えるワンホットエンコーダ(図29に示されている)を備える。

【０３１３】

ACNNは、残差ブロックの数、スキップコネクションの数、および残差コネクションの数によってパラメータ化され得る。

【０３１４】

ACNNは、先行する入力の空間および特徴次元を再整形する次元変換畳み込み層を備えることができる。

【０３１５】

図20に示されているように、各残差ブロックは、少なくとも1つのバッチ正規化層と、少なくとも1つの正規化線形層(略語ReLU)と、少なくとも1つのAtrous畳み込み層と、少なくとも1つの残差コネクションとを含むことができる。そのような一実装形態において、各残差ブロックは、2つのバッチ正規化層と、2つのReLU非線形層と、2つのAtrous畳み込み層と、1つの残差コネクションとを含む。

【０３１６】

他の実装形態は、上で説明されているシステムの動作を実行するためにプロセッサによって実行可能な命令を記憶する非一時的コンピュータ可読記憶媒体を含み得る。さらに別の実装形態は、上で説明されているシステムの動作を実行する方法を含み得る。

【０３１７】

開示されている技術の別のシステム実装形態は、並列動作し、メモリに結合されている多数のプロセッサ上で稼動する訓練済みスプライス部位予測器を備える。システムは、ドナースプライス部位の少なくとも50000個の訓練例、アクセプタースプライス部位の少なくとも50000個の訓練例、および非スプライシング部位の少なくとも100000個の訓練例について、多数のプロセッサ上で実行する、Atrous畳み込みニューラルネットワーク(略語ACNN)を訓練する。訓練で使用される訓練例の各々は、各側で少なくとも400個のヌクレオチドが隣接する標的ヌクレオチドを含むヌクレオチド配列である。

【０３１８】
システムは、多数のプロセッサのうちの少なくとも1つで実行するACNNの入力段を備え、標的ヌクレオチドの評価のために少なくとも801個のヌクレオチドからなる入力配列を供給する。各標的ヌクレオチドには、各側で少なくとも400個のヌクレオチドが隣接する。他の実装形態では、システムは、多数のプロセッサのうちの少なくとも1つで実行するACNNの入力モジュールを備え、標的ヌクレオチドの評価のために少なくとも801個のヌクレオチドからなる入力配列を供給する。
【０３１９】
システムは、多数のプロセッサのうちの少なくとも1つで実行され、ACNNによる解析結果を標的ヌクレオチドの各々がドナースプライス部位、アクセプタースプライス部位、または非スプライシング部位である可能性に対する分類スコアに翻訳する、ACNNの出力段を備える。他の実装形態では、システムは、多数のプロセッサのうちの少なくとも1つで実行され、ACNNによる解析結果を標的ヌクレオチドの各々がドナースプライス部位、アクセプタースプライス部位、または非スプライシング部位である可能性に対する分類スコアに翻訳する、ACNNの出力モジュールを備える。
【０３２０】
第1のシステム実装形態に対するこの特定の実装形態の節で説明されている特徴の各々は、このシステム実装形態にも等しく適用される。上で示されているように、すべてのシステム特徴は、ここでは繰り返さず、参照により繰り返されると考えられるべきである。
【０３２１】
ACNNは、ドナースプライス部位の150000個の訓練例、アクセプタースプライス部位の150000個の訓練例、および非スプライシング部位の800000000個の訓練例上で訓練され得る。システムの別の実装形態において、ACNNは、最低から最高までの順に配置構成されている残差ブロックのグループを含む。システムのさらに別の実装形態では、残差ブロックの各グループは、残差ブロック内の畳み込みフィルタの数、残差ブロックの畳み込みウィンドウサイズ、および残差ブロックのAtrous畳み込みレートによってパラメータ化される。
【０３２２】
ACNNは、最低から最高までの順に配置構成されている残差ブロックのグループを含むことができる。残差ブロックの各グループは、残差ブロック内の畳み込みフィルタの数、残差ブロックの畳み込みウィンドウサイズ、および残差ブロックのAtrous畳み込みレートによってパラメータ化される。
【０３２３】
ACNNでは、Atrous畳み込みレートは、下位の残差ブロックグループから上位の残差ブロックグループへと非指数関数的に高まる。ACNNでも、畳み込みウィンドウサイズは、残差ブロックのグループ間で異なる。
【０３２４】
ACNNは、図18に示されているように、1つまたは複数の訓練サーバ上で訓練できる。
【０３２５】
訓練済みACNNは、図18に示されているように、要求側クライアントから入力配列を受け取る1つまたは複数のプロダクションサーバ上にデプロイされ得る。そのような一実装形態において、プロダクションサーバは、図18に示されているように、ACNNの入力および出力段を通して入力配列を処理し、クライアントに伝送される出力を生成する。他の実装形態において、プロダクションサーバは、図18に示されているように、ACNNの入力および出力モジュールを通して入力配列を処理し、クライアントに伝送される出力を生成する。
【０３２６】
他の実装形態は、上で説明されているシステムの動作を実行するためにプロセッサによって実行可能な命令を記憶する非一時的コンピュータ可読記憶媒体を含み得る。さらに別の実装形態は、上で説明されているシステムの動作を実行する方法を含み得る。
【０３２７】
開示されている技術の方法実装形態は、ゲノム配列(たとえば、ヌクレオチド配列)内の

スプライス部位を識別するスプライス部位検出器を訓練することを含む。

【０３２８】

この方法は、各側で少なくとも400個のヌクレオチドが各々隣接する標的ヌクレオチドの評価のために、Atrous畳み込みニューラルネットワーク(略語ACNN)に、少なくとも801個のヌクレオチドの入力配列を供給することを含む。

【０３２９】

ACNNは、ドナースプライス部位の少なくとも50000個の訓練例、アクセプタースプライス部位の少なくとも50000個の訓練例、および非スプライシング部位の少なくとも100000個の訓練例上で訓練される。訓練で使用される訓練例の各々は、各側で少なくとも400個のヌクレオチドが隣接する標的ヌクレオチドを含むヌクレオチド配列である。

【０３３０】

この方法は、ACNNによる解析結果を、標的ヌクレオチドの各々がドナースプライス部位、アクセプタースプライス部位、または非スプライシング部位である可能性に対する分類スコアに翻訳することをさらに含む。

【０３３１】

第1のシステム実装形態に対するこの特定の実装形態の節で説明されている特徴の各々は、この方法実装形態にも等しく適用される。上で示されているように、すべてのシステム特徴は、ここでは繰り返さず、参照により繰り返されると考えられるべきである。

【０３３２】

他の実装形態は、上で説明されている方法を実行するためにプロセッサによって実行可能な命令を記憶する非一時的コンピュータ可読記憶媒体を含み得る。さらに別の実装形態は、メモリと、上で説明されている方法を実行する、メモリに記憶されている命令を実行するように動作可能である1つまたは複数のプロセッサとを備えるシステムを含み得る。

【０３３３】

われわれは、ゲノム配列(たとえば、ヌクレオチド配列)内の異常スプライシングを検出するために訓練されたAtrous畳み込みニューラルネットワークを使用することについて製造システム、製造方法、および製造品を説明する。一実装形態の1つまたは複数の特徴がベースの実装形態と組み合わされ得る。相互排他的でない実装形態は、組合せ可能であると教示される。一実装形態の1つまたは複数の特徴が他の実装形態と組み合わされ得る。本開示は、定期的に、これらのオプションについてユーザに通知する。これらのオプションを繰り返す言及のいくつかの実装形態からの省略は、先行する節において教示されている組合せを制限するものとしてみなすべきでなく、これらの言及は、次の実装形態の各々に参照により順に組み込まれる。

【０３３４】

開示されている技術のシステム実装形態は、メモリに結合されている1つまたは複数のプロセッサを含む。メモリは、並列動作し、メモリに結合されている多数のプロセッサ上で稼動する異常スプライシング検出器を実装するコンピュータ命令をロードされる。

【０３３５】

図34に示されているように、システムは、多数のプロセッサ上で稼動する訓練済みAtrous畳み込みニューラルネットワーク(略語ACNN)を含む。ACNNは、訓練可能なパラメータが少ししかない大きい受容野を実現可能にするAtrous/Dilated畳み込みを使用する畳み込みニューラルネットワークである。Atrous/Dilated畳み込みは、Atrous畳み込みレートまたは拡張係数とも呼ばれるあるステップを用いて入力値をスキップすることによって、カーネルがその長さよりも大きい領域にわたって適用される畳み込みである。Atrous/dilated畳み込みは、畳み込みフィルタ/カーネルの要素間の間隔を加え、それによって、畳み込み演算が実行されるときにより大きい間隔における近傍の入力エントリ(たとえば、ヌクレオチド、アミノ酸)が考慮される。これは長距離構成依存性を入力に組み込むことを可能にする。Atrous畳み込みは、部分畳み込み計算を隣接するヌクレオチドが処理されるときに再使用できるように保存する。

【０３３６】

図34に示されているように、ACNNは、入力配列内の標的ヌクレオチドを分類し、標的ヌクレオチドの各々がドナースプライス部位、アクセプタースプライス部位、または非スプライシング部位である可能性に対するスプライス部位スコアを割り当てる。入力配列は、少なくとも801個のヌクレオチドを含み、各標的ヌクレオチドには、各側に少なくとも400個のヌクレオチドが隣接する。

【０３３７】

図34に示されているように、システムは、ACNNを通して参照配列およびバリアント配列を処理し、参照配列およびバリアント配列内の各標的ヌクレオチドがドナースプライス部位、アクセプタースプライス部位、または非スプライシング部位である可能性に対するスプライス部位スコアを生成する、多数のプロセッサのうちの少なくとも1つで稼動する分類器も備える。参照配列およびバリアント配列は、各々、少なくとも101個の標的ヌクレオチドを有し、各標的ヌクレオチドには、各側に少なくとも400個のヌクレオチドが隣接する。図33は、参照配列および代替/バリアント配列を示す。

【０３３８】

次いで、図34に示されているように、参照配列およびバリアント配列内の標的ヌクレオチドのスプライス部位スコアの差から、バリアント配列を生成したバリアントが異常スプライシングを引き起こし、したがって病原性を有するかどうかを決定する。

【０３３９】

この実装形態および開示されている他のシステムは、任意選択で、次の特徴のうちの1つまたは複数を含む。システムは、開示されている方法に関連して説明されている特徴も含むことができる。簡潔にするため、システム特徴の代替的組合せは、個別には列挙しない。製造システム、製造方法、および製造品に適用可能な特徴は、ベースとなる特徴の法令に定めるクラスのセット毎に繰り返されない。読者は、この節に明記されている特徴が他の法令に定められているクラスにおけるベースとなる特徴とどのように容易に組み合わされ得るかを理解するであろう。

【０３４０】

図34に示されているように、スプライス部位スコアの差は、参照配列およびバリアント配列内の標的ヌクレオチドの間で位置毎に決定され得る。

【０３４１】

図34に示されているように、少なくとも1つの標的ヌクレオチド位置について、スプライス部位スコアのグローバルな最大の差が所定の閾値より高いときに、ACNNはバリアントを異常スプライシングを引き起こし、したがって病原性を有すると分類する。

【０３４２】

図17に示されているように、少なくとも1つの標的ヌクレオチド位置について、スプライス部位スコアのグローバルな最大の差が所定の閾値より低いときに、ACNNはバリアントを異常スプライシングを引き起こさず、したがって良性であると分類する。

【０３４３】

閾値は、複数の候補閾値のうちから決定され得る。これは、良性の共通バリアントによって生成される参照配列とバリアント配列の対の第1のセットを処理して異常スプライシング検出の第1のセットを生成することと、病原性稀少バリアントによって生成される参照配列とバリアント配列の対の第2のセットを処理して異常スプライシング検出の第2のセットを生成することと、分類器で使用するために、第2のセット内の異常スプライシング検出のカウントを最大化し、第1のセット内の異常スプライシング検出のカウントを最小化する少なくとも1つの閾値を選択することとを含む。

【０３４４】

一実装形態において、ACNNは、自閉症スペクトラム障害(略語ASD)を引き起こすバリアントを識別する。別の実装形態において、ACNNは、発育遅滞障害(略語DDD)を引き起こすバリアントを識別する。

【０３４５】

図36に示されているように、参照配列およびバリアント配列は、各々、少なくとも101

個の標的ヌクレオチドを有することができ、各標的ヌクレオチドには、各側に少なくとも5000個のヌクレオチドが隣接することができる。

【０３４６】

図36に示されているように、参照配列内の標的ヌクレオチドのスプライス部位スコアは、ACNNの第1の出力においてエンコードされ、バリアント配列内の標的ヌクレオチドのスプライス部位スコアは、ACNNの第2の出力においてエンコードされ得る。一実装形態において、第1の出力は第1の101×3行列としてエンコードされ、第2の出力は第2の101×3行列としてエンコードされる。

【０３４７】

図36に示されているように、そのような一実装形態において、第1の101×3行列内の各行は、参照配列内の標的ヌクレオチドがドナースプライス部位、アクセプタースプライス部位、または非スプライシング部位である可能性に対するスプライス部位スコアを一意的に表す。

【０３４８】

また、図36に示されているように、そのような一実装形態において、第2の101×3行列内の各行は、バリアント配列内の標的ヌクレオチドがドナースプライス部位、アクセプタースプライス部位、または非スプライシング部位である可能性に対するスプライス部位スコアを一意的に表す。

【０３４９】

図36に示されているように、いくつかの実装形態において、第1の101×3行列および第2の101×3行列の各行におけるスプライス部位スコアは、和が1になるように指数関数的に正規化され得る。

【０３５０】

図36に示されているように、分類器は、第1の101×3行列および第2の101×3行列の行同士の比較を実行し、行毎に、スプライス部位スコアの分布の変化を決定することができる。行同士の比較の少なくとも1つのインスタンスについて、分布の変化が所定の閾値より高いときに、ACNNはバリアントを異常スプライシングを引き起こし、したがって病原性を有すると分類する。

【０３５１】

システムは、参照配列およびバリアント配列を疎にエンコードするワンホットエンコーダ(図29に示されている)を備える。

【０３５２】

他のシステムおよび方法実装形態に対するこの特定の実装形態の節で説明されている特徴の各々は、このシステム実装形態にも等しく適用される。上で示されているように、すべてのシステム特徴は、ここでは繰り返さず、参照により繰り返されると考えられるべきである。

【０３５３】

他の実装形態は、上で説明されているシステムの動作を実行するためにプロセッサによって実行可能な命令を記憶する非一時的コンピュータ可読記憶媒体を含み得る。さらに別の実装形態は、上で説明されているシステムの動作を実行する方法を含み得る。

【０３５４】

開示されている技術の方法実装形態は、異常スプライシングを引き起こすゲノムバリアントを検出することを含む。

【０３５５】

この方法は、標的部分配列内の各ヌクレオチドをドナースプライス部位、アクセプタースプライス部位、または非スプライシング部位として分類することによって入力配列の標的部分配列内の差次的スプライシングパターンを検出するように訓練されているAtrous畳み込みニューラルネットワーク(略語ACNN)を通じて参照配列を処理することを含む。

【０３５６】

この方法は、処理に基づき、参照標的部分配列内の各ヌクレオチドをドナースプライス

部位、アクセプタースプライス部位、または非スプライシング部位として分類することによって参照標的部分配列内の第1の差次的スプライシングパターンを検出することを含む。

【０３５７】

この方法は、ACNNを通してバリアント配列を処理することを含む。バリアント配列および参照配列は、バリアント標的部分配列内に配置されている少なくとも1つのバリアントヌクレオチドだけ異なる。

【０３５８】

この方法は、処理に基づき、バリアント標的部分配列内の各ヌクレオチドをドナースプライス部位、アクセプタースプライス部位、または非スプライシング部位として分類することによってバリアント標的部分配列内の第2の差次的スプライシングパターンを検出することを含む。

【０３５９】

この方法は、ヌクレオチド毎に、参照標的部分配列およびバリアント標的部分配列のスプライス部位分類を比較することによって第1の差次的スプライシングパターンと第2の差次的スプライシングパターンとの間の差を決定することを含む。

【０３６０】

この差が所定の閾値より高いときに、この方法は、バリアントを異常スプライシングを引き起こし、したがって病原性を有すると分類することと、分類結果をメモリに記憶することとを含む。

【０３６１】

他のシステムおよび方法実装形態に対するこの特定の実装形態の節で説明されている特徴の各々は、この方法実装形態にも等しく適用される。上で示されているように、すべてのシステム特徴は、ここでは繰り返さず、参照により繰り返されると考えられるべきである。

【０３６２】

差次的スプライシングパターンは、標的部分配列内のスプライシング事象の出現の位置分布を識別することができる。スプライシング事象の例は、潜在的スプライシング、エクソンスキッピング、相互排他的エクソン、代替ドナー部位、代替アクセプター部位、およびイントロン保持のうちの少なくとも1つを含む。

【０３６３】

参照標的部分配列およびバリアント標的部分配列は、ヌクレオチド位置に関して整列され、少なくとも1つのバリアントヌクレオチドの分だけ異なり得る。

【０３６４】

参照標的部分配列およびバリアント標的部分配列は、各々、少なくとも40個のヌクレオチドを有し、各々、各側に少なくとも40個のヌクレオチドが隣接し得る。

【０３６５】

参照標的部分配列およびバリアント標的部分配列は、各々、少なくとも101個のヌクレオチドを有し、各々、各側に少なくとも5000個のヌクレオチドが隣接し得る。

【０３６６】

バリアント標的部分配列は、2つのバリアント含むことができる。

【０３６７】

他の実装形態は、上で説明されている方法を実行するためにプロセッサによって実行可能な命令を記憶する非一時的コンピュータ可読記憶媒体を含み得る。さらに別の実装形態は、メモリと、上で説明されている方法を実行する、メモリに記憶されている命令を実行するように動作可能である1つまたは複数のプロセッサとを備えるシステムを含み得る。

【０３６８】

われわれは、ゲノム配列(たとえば、ヌクレオチド配列)内のスプライス部位および異常スプライシングを検出するために訓練された畳み込みニューラルネットワークを使用することについて製造システム、製造方法、および製造品を説明する。一実装形態の1つまた

は複数の特徴がベースの実装形態と組み合わされ得る。相互排他的でない実装形態は、組合せ可能であると教示される。一実装形態の1つまたは複数の特徴が他の実装形態と組み合わされ得る。本開示は、定期的に、これらのオプションについてユーザに通知する。これらのオプションを繰り返す言及のいくつかの実装形態からの省略は、先行する節において教示されている組合せを制限するものとしてみなすべきでなく、これらの言及は、次の実装形態の各々に参照により順に本明細書に組み込まれる。

【０３６９】

開示されている技術のシステム実装形態は、メモリに結合されている1つまたは複数のプロセッサを含む。メモリは、ゲノム配列(たとえば、ヌクレオチド配列)内のスプライス部位を識別するスプライス部位検出器を訓練するためのコンピュータ命令をロードされる。

【０３７０】

システムは、ドナースプライス部位の少なくとも50000個の訓練例、アクセプタースプライス部位の少なくとも50000個の訓練例、および非スプライシング部位の少なくとも100000個の訓練例上で畳み込みニューラルネットワーク(略語CNN)を訓練する。各訓練例は、各側に少なくとも20個のヌクレオチドが隣接する少なくとも1つの標的ヌクレオチドを有する標的ヌクレオチド配列である。

【０３７１】

CNNを使用して訓練例を評価するために、システムは、CNNへの入力として、少なくとも40個の上流構成ヌクレオチドおよび少なくとも40個の下流構成ヌクレオチドがさらに隣接する標的ヌクレオチド配列を提供する。

【０３７２】

次いで、この評価に基づき、CNNは、出力として、標的ヌクレオチド配列内の各ヌクレオチドがドナースプライス部位、アクセプタースプライス部位、または非スプライシング部位である可能性に対するトリプレットスコアを生成する。

【０３７３】

このシステム実装形態および開示されている他のシステムは、任意選択で、次の特徴のうちの1つまたは複数を含む。システムは、開示されている方法に関連して説明されている特徴も含むことができる。簡潔にするため、システム特徴の代替的組合せは、個別には列挙しない。製造システム、製造方法、および製造品に適用可能な特徴は、ベースとなる特徴の法令に定めるクラスのセット毎に繰り返されない。読者は、この節に明記されている特徴が他の法令に定められているクラスにおけるベースとなる特徴とどのように容易に組み合わされ得るかを理解するであろう。

【０３７４】

入力は、各側に100個のヌクレオチドが隣接する標的ヌクレオチドを有する標的ヌクレオチド配列を含むことができる。そのような実装形態において、標的ヌクレオチド配列は、200個の上流構成ヌクレオチドおよび200個の下流構成ヌクレオチドがさらに隣接する。

【０３７５】

図28に示されているように、システムは、ドナースプライス部位の150000個の訓練例、アクセプタースプライス部位の150000個の訓練例、および非スプライシング部位の1000000個の訓練例上でCNNを訓練することができる。

【０３７６】

図31に示されているように、CNNは、畳み込み層の数、畳み込みフィルタの数、およびサブサンプリング層の数(たとえば、最大プーリングおよび平均プーリング)によってパラメータ化され得る。

【０３７７】

図31に示されているように、CNNは、1つまたは複数の全結合層と、末端分類層とを含むことができる。

【０３７８】

CNNは、先行する入力の空間および特徴次元を再整形する次元変換畳み込み層を備える

ことができる。

【０３７９】

標的ヌクレオチド配列内の各ヌクレオチドに対する各トリプレットスコアは、和が1になるように指数関数的に正規化され得る。そのような一実装形態において、システムは、それぞれのトリプレットスコアにおける最高スコアに基づき標的ヌクレオチド内の各ヌクレオチドをドナースプライス部位、アクセプタースプライス部位、または非スプライシング部位として分類する。

【０３８０】

図32に示されているように、CNNはバッチ式に、エポックにおける訓練例を評価する。訓練例は、ランダムにバッチにサンプリングされる。各バッチは所定のバッチサイズを有する。CNNは、複数のエポック(たとえば、1～10)にわたって訓練例の評価を反復する。

【０３８１】

入力は、2つの隣接する標的ヌクレオチドを有する標的ヌクレオチド配列を含むことができる。2つの隣接する標的ヌクレオチドは、アデニン(略語A)およびグアニン(略語G)であるものとしてよい。2つの隣接する標的ヌクレオチドは、グアニン(略語G)およびウラシル(略語U)であるものとしてよい。

【０３８２】

システムは、訓練例を疎にエンコードし、ワンホットエンコーディングを入力として与えるワンホットエンコーダ(図32に示されている)を備える。

【０３８３】

CNNは、残差ブロックの数、スキップコネクションの数、および残差コネクションの数によってパラメータ化され得る。

【０３８４】

各残差ブロックは、少なくとも1つのバッチ正規化層と、少なくとも1つの正規化線形層(略語ReLU)と、少なくとも1つの次元変換層と、少なくとも1つの残差コネクションとを含むことができる。各残差ブロックは、2つのバッチ正規化層と、2つのReLU非線形層と、2つの次元変換層と、1つの残差コネクションとを含むことができる。

【０３８５】

他のシステムおよび方法実装形態に対するこの特定の実装形態の節で説明されている特徴の各々は、このシステム実装形態にも等しく適用される。上で示されているように、すべてのシステム特徴は、ここでは繰り返さず、参照により繰り返されると考えられるべきである。

【０３８６】

他の実装形態は、上で説明されているシステムの動作を実行するためにプロセッサによって実行可能な命令を記憶する非一時的コンピュータ可読記憶媒体を含み得る。さらに別の実装形態は、上で説明されているシステムの動作を実行する方法を含み得る。

【０３８７】

開示されている技術の別のシステム実装形態は、並列動作し、メモリに結合されている多数のプロセッサ上で稼動する訓練済みスプライス部位予測器を備える。システムは、ドナースプライス部位の少なくとも50000個の訓練例、アクセプタースプライス部位の少なくとも50000個の訓練例、および非スプライシング部位の少なくとも100000個の訓練例について、多数のプロセッサ上で実行する、畳み込みニューラルネットワーク(略語CNN)を訓練する。訓練で使用される訓練例の各々は、各側で少なくとも400個のヌクレオチドが隣接する標的ヌクレオチドを含むヌクレオチド配列である。

【０３８８】

システムは、多数のプロセッサのうちの少なくとも1つで実行するCNNの入力段を備え、標的ヌクレオチドの評価のために少なくとも801個のヌクレオチドからなる入力配列を供給する。各標的ヌクレオチドには、各側で少なくとも400個のヌクレオチドが隣接する。他の実装形態では、システムは、多数のプロセッサのうちの少なくとも1つで実行するCNNの入力モジュールを備え、標的ヌクレオチドの評価のために少なくとも801個のヌクレオ

チドからなる入力配列を供給する。

【０３８９】

システムは、多数のプロセッサのうちの少なくとも1つで実行され、CNNによる解析結果を標的ヌクレオチドの各々がドナースプライス部位、アクセプタースプライス部位、または非スプライシング部位である可能性に対する分類スコアに翻訳する、CNNの出力段を備える。他の実装形態では、システムは、多数のプロセッサのうちの少なくとも1つで実行され、CNNによる解析結果を標的ヌクレオチドの各々がドナースプライス部位、アクセプタースプライス部位、または非スプライシング部位である可能性に対する分類スコアに翻訳する、CNNの出力モジュールを備える。

【０３９０】

他のシステムおよび方法実装形態に対するこの特定の実装形態の節で説明されている特徴の各々は、このシステム実装形態にも等しく適用される。上で示されているように、すべてのシステム特徴は、ここでは繰り返さず、参照により繰り返されると考えられるべきである。

【０３９１】

CNNは、ドナースプライス部位の150000個の訓練例、アクセプタースプライス部位の150000個の訓練例、および非スプライシング部位の800000000個の訓練例上で訓練され得る。

【０３９２】

CNNは、1つまたは複数の訓練サーバ上で訓練できる。

【０３９３】

訓練済みCNNは、要求側クライアントから入力配列を受け取る1つまたは複数のプロダクションサーバ上にデプロイされ得る。そのような一実装形態において、プロダクションサーバは、CNNの入力および出力段を通して入力配列を処理し、クライアントに伝送される出力を生成する。他の実装形態では、プロダクションサーバは、CNNの入力および出力段を通して入力配列を処理し、クライアントに伝送される出力を生成する。

【０３９４】

他の実装形態は、上で説明されているシステムの動作を実行するためにプロセッサによって実行可能な命令を記憶する非一時的コンピュータ可読記憶媒体を含み得る。さらに別の実装形態は、上で説明されているシステムの動作を実行する方法を含み得る。

【０３９５】

開示されている技術の方法実装形態は、ゲノム配列(たとえば、ヌクレオチド配列)内のスプライス部位を識別するスプライス部位検出器を訓練することを含む。この方法は、各側で少なくとも400個のヌクレオチドが各々隣接する標的ヌクレオチドの評価のために、畳み込みニューラルネットワーク(略語CNN)に、少なくとも801個のヌクレオチドの入力配列を供給することを含む。

【０３９６】

CNNは、ドナースプライス部位の少なくとも50000個の訓練例、アクセプタースプライス部位の少なくとも50000個の訓練例、および非スプライシング部位の少なくとも100000個の訓練例上で訓練される。訓練で使用される訓練例の各々は、各側で少なくとも400個のヌクレオチドが隣接する標的ヌクレオチドを含むヌクレオチド配列である。

【０３９７】

この方法は、CNNによる解析結果を、標的ヌクレオチドの各々がドナースプライス部位、アクセプタースプライス部位、または非スプライシング部位である可能性に対する分類スコアに翻訳することをさらに含む。

【０３９８】

他のシステムおよび方法実装形態に対するこの特定の実装形態の節で説明されている特徴の各々は、この方法実装形態にも等しく適用される。上で示されているように、すべてのシステム特徴は、ここでは繰り返さず、参照により繰り返されると考えられるべきである。

【０３９９】

他の実装形態は、上で説明されている方法を実行するためにプロセッサによって実行可能な命令を記憶する非一時的コンピュータ可読記憶媒体を含み得る。さらに別の実装形態は、メモリと、上で説明されている方法を実行する、メモリに記憶されている命令を実行するように動作可能である1つまたは複数のプロセッサとを備えるシステムを含み得る。

【０４００】

開示されている技術のさらに別のシステム実装形態は、メモリに結合されている1つまたは複数のプロセッサを含む。メモリは、並列動作し、メモリに結合されている多数のプロセッサ上で稼動する異常スプライシング検出器を実装するコンピュータ命令をロードされる。

【０４０１】

システムは、多数のプロセッサ上で稼動する訓練済み畳み込みニューラルネットワーク(略語CNN)を含む。

【０４０２】

図34に示されているように、CNNは、入力配列内の標的ヌクレオチドを分類し、標的ヌクレオチドの各々がドナースプライス部位、アクセプタースプライス部位、または非スプライシング部位である可能性に対するスプライス部位スコアを割り当てる。入力配列は、少なくとも801個のヌクレオチドを含み、各標的ヌクレオチドには、各側に少なくとも400個のヌクレオチドが隣接する。

【０４０３】

図34に示されているように、システムは、CNNを通して参照配列およびバリアント配列を処理し、参照配列およびバリアント配列内の各標的ヌクレオチドがドナースプライス部位、アクセプタースプライス部位、または非スプライシング部位である可能性に対するスプライス部位スコアを生成する、多数のプロセッサのうちの少なくとも1つで稼動する分類器も備える。参照配列およびバリアント配列は、各々、少なくとも101個の標的ヌクレオチドを有し、各標的ヌクレオチドには、各側に少なくとも400個のヌクレオチドが隣接する。

【０４０４】

次いで、図34に示されているように、参照配列およびバリアント配列内の標的ヌクレオチドのスプライス部位スコアの差から、バリアント配列を生成したバリアントが異常スプライシングを引き起こし、したがって病原性を有するかどうかを決定する。

【０４０５】

他のシステムおよび方法実装形態に対するこの特定の実装形態の節で説明されている特徴の各々は、このシステム実装形態にも等しく適用される。上で示されているように、すべてのシステム特徴は、ここでは繰り返さず、参照により繰り返されると考えられるべきである。

【０４０６】

スプライス部位スコアの差は、参照配列およびバリアント配列内の標的ヌクレオチドの間で位置毎に決定され得る。

【０４０７】

少なくとも1つの標的ヌクレオチド位置について、スプライス部位スコアのグローバルな最大の差が所定の閾値より高いときに、CNNはバリアントを異常スプライシングを引き起こし、したがって病原性を有すると分類する。

【０４０８】

少なくとも1つの標的ヌクレオチド位置について、スプライス部位スコアのグローバルな最大の差が所定の閾値より低いときに、CNNはバリアントを異常スプライシングを引き起こさず、したがって良性であると分類する。

【０４０９】

閾値は、複数の候補閾値のうちから決定され得る。これは、良性の共通バリアントによって生成される参照配列とバリアント配列の対の第1のセットを処理して異常スプライシング検出の第1のセットを生成することと、病原性稀少バリアントによって生成される参

照配列とバリアント配列の対の第2のセットを処理して異常スプライシング検出の第2のセットを生成することと、分類器で使用するために、第2のセット内の異常スプライシング検出のカウントを最大化し、第1のセット内の異常スプライシング検出のカウントを最小化する少なくとも1つの閾値を選択することとを含む。

【０４１０】

一実装形態において、CNNは、自閉症スペクトラム障害(略語ASD)を引き起こすバリアントを識別する。別の実装形態において、CNNは、発育遅滞障害(略語DDD)を引き起こすバリアントを識別する。

【０４１１】

参照配列およびバリアント配列は、各々、少なくとも101個の標的ヌクレオチドを有し、各標的ヌクレオチドには、各側に少なくとも1000個のヌクレオチドが隣接することができる。

【０４１２】

参照配列内の標的ヌクレオチドのスプライス部位スコアは、CNNの第1の出力においてエンコードされ、バリアント配列内の標的ヌクレオチドのスプライス部位スコアは、CNNの第2の出力においてエンコードされ得る。一実装形態において、第1の出力は第1の101×3行列としてエンコードされ、第2の出力は第2の101×3行列としてエンコードされる。

【０４１３】

そのような一実装形態において、第1の101×3行列内の各行は、参照配列内の標的ヌクレオチドがドナースプライス部位、アクセプタースプライス部位、または非スプライシング部位である可能性に対するスプライス部位スコアを一意的に表す。

【０４１４】

また、そのような一実装形態において、第2の101×3行列内の各行は、バリアント配列内の標的ヌクレオチドがドナースプライス部位、アクセプタースプライス部位、または非スプライシング部位である可能性に対するスプライス部位スコアを一意的に表す。

【０４１５】

いくつかの実装形態において、第1の101×3行列および第2の101×3行列の各行におけるスプライス部位スコアは、和が1になるように指数関数的に正規化され得る。

【０４１６】

分類器は、第1の101×3行列および第2の101×3行列の行同士の比較を実行し、行毎に、スプライス部位スコアの分布の変化を決定することができる。行同士の比較の少なくとも1つのインスタンスについて、分布の変化が所定の閾値より高いときに、CNNはバリアントを異常スプライシングを引き起こし、したがって病原性を有すると分類する。

【０４１７】

システムは、参照配列およびバリアント配列を疎にエンコードするワンホットエンコーダ(図29に示されている)を備える。

【０４１８】

他の実装形態は、上で説明されているシステムの動作を実行するためにプロセッサによって実行可能な命令を記憶する非一時的コンピュータ可読記憶媒体を含み得る。さらに別の実装形態は、上で説明されているシステムの動作を実行する方法を含み得る。

【０４１９】

開示されている技術の方法実装形態は、異常スプライシングを引き起こすゲノムバリアントを検出することを含む。

【０４２０】

この方法は、標的部分配列内の各ヌクレオチドをドナースプライス部位、アクセプタースプライス部位、または非スプライシング部位として分類することによって入力配列の標的部分配列内の差次的スプライシングパターンを検出するように訓練されているAtrous畳み込みニューラルネットワーク(略語CNN)を通じて参照配列を処理することを含む。

【０４２１】

この方法は、処理に基づき、参照標的部分配列内の各ヌクレオチドをドナースプライス

部位、アクセプタースプライス部位、または非スプライシング部位として分類することによって参照標的部分配列内の第1の差次的スプライシングパターンを検出することを含む。

【０４２２】

この方法は、CNNを通してバリアント配列を処理することを含む。バリアント配列および参照配列は、バリアント標的部分配列内に配置されている少なくとも1つのバリアントヌクレオチドだけ異なる。

【０４２３】

この方法は、処理に基づき、バリアント標的部分配列内の各ヌクレオチドをドナースプライス部位、アクセプタースプライス部位、または非スプライシング部位として分類することによってバリアント標的部分配列内の第2の差次的スプライシングパターンを検出することを含む。

【０４２４】

この方法は、ヌクレオチド毎に、参照標的部分配列およびバリアント標的部分配列のスプライス部位分類を比較することによって第1の差次的スプライシングパターンと第2の差次的スプライシングパターンとの間の差を決定することを含む。

【０４２５】

この差が所定の閾値より高いときに、この方法は、バリアントを異常スプライシングを引き起こし、したがって病原性を有すると分類することと、分類結果をメモリに記憶することとを含む。

【０４２６】

他のシステムおよび方法実装形態に対するこの特定の実装形態の節で説明されている特徴の各々は、この方法実装形態にも等しく適用される。上で示されているように、すべてのシステム特徴は、ここでは繰り返さず、参照により繰り返されると考えられるべきである。

【０４２７】

差次的スプライシングパターンは、標的部分配列内のスプライシング事象の出現の位置分布を識別することができる。スプライシング事象の例は、潜在的スプライシング、エクソンスキッピング、相互排他的エクソン、代替ドナー部位、代替アクセプター部位、およびイントロン保持のうちの少なくとも1つを含む。

【０４２８】

参照標的部分配列およびバリアント標的部分配列は、ヌクレオチド位置に関して整列され、少なくとも1つのバリアントヌクレオチドの分だけ異なり得る。

【０４２９】

参照標的部分配列およびバリアント標的部分配列は、各々、少なくとも40個のヌクレオチドを有し、各々、各側に少なくとも40個のヌクレオチドが隣接し得る。

【０４３０】

参照標的部分配列およびバリアント標的部分配列は、各々、少なくとも101個のヌクレオチドを有し、各々、各側に少なくとも1000個のヌクレオチドが隣接し得る。

【０４３１】

バリアント標的部分配列は、2つのバリアント含むことができる。

【０４３２】

他の実装形態は、上で説明されている方法を実行するためにプロセッサによって実行可能な命令を記憶する非一時的コンピュータ可読記憶媒体を含み得る。さらに別の実装形態は、メモリと、上で説明されている方法を実行する、メモリに記憶されている命令を実行するように動作可能である1つまたは複数のプロセッサとを備えるシステムを含み得る。

【０４３３】

先行する説明は、開示されている技術の製造および使用を可能にするために提示されている。開示されている実装形態に対し様々な修正を加えられることは、明白であろうし、また本明細書において定義されている一般原理は、開示された技術の精神または範囲から

逸脱することなく他の実装形態および応用にも適用され得る。したがって、開示された技術は、図示されている実装形態に限定されることを意図されておらず、本明細書で開示された原理および特徴と一致する最も広い範囲を適用されることを意図されている。開示されている技術の範囲は、付属の請求項によって定められる。

【０４３４】

遺伝子当たりエンリッチメント解析

図57は、遺伝子当たりエンリッチメント解析の一実装形態を示している。一実装形態において、異常スプライシング検出器は、異常スプライシングを引き起こすと決定されているバリアントの病原性を決定する遺伝子当たりエンリッチメント解析を実装するようにさらに構成される。遺伝性疾患を患っている個体のコホートからサンプリングされた特定の遺伝子について、遺伝子当たりエンリッチメント解析は、訓練済みACNNを適用して異常スプライシングを引き起こす特定の遺伝子における候補バリアントを識別することと、特定の遺伝子に対する突然変異の、ベースラインとなる数を、候補バリアントの観察されたトリヌクレオチド突然変異率を総和し、その和に伝達カウントおよびコホートのサイズを乗算することに基づき決定することと、訓練済みACNNを適用して異常スプライシングを引き起こす特定の遺伝子におけるデノボバリアントを識別することと、突然変異のベースライン数をデノボバリアントのカウントと比較することとを含む。比較の出力に基づき、遺伝子当たりエンリッチメント解析は、特定の遺伝子が遺伝性疾患に関連付けられていること、およびデノボバリアントが病原性を有することを決定する。いくつかの実装形態において、遺伝性疾患は、自閉症スペクトラム障害(略語ASD)である。他の実装形態において、遺伝性疾患は、発育遅滞障害(略語DDD)である。

【０４３５】

図57に示されている例では、特定の遺伝子内の5個の候補バリアントは、異常スプライシング検出器によって異常スプライシングを引き起こすものとして分類されている。これら5個の候補バリアントは、10^{-8}、10^{-2}、10^{-1}、10^{5}、および10^{1}のそれぞれの観察されたトリヌクレオチド突然変異率を有する。特定の遺伝子に対する突然変異の、ベースラインとなる数は、5個の候補バリアントのそれぞれの観察されたトリヌクレオチド突然変異率を総和し、その和に伝達/染色体カウント(2)およびコホートのサイズ(100)を乗算することに基づき10^{-5}と決定される。これは、次いで、デノボバリアントカウント(3)と比較される。

【０４３６】

いくつかの実装形態において、異常スプライシング検出器は、出力としてp値を生成する統計的検定を使用して比較を実行するようにさらに構成される。

【０４３７】

他の実装形態では、異常スプライシング検出器は、突然変異のベースライン数をデノボバリアントのカウントと比較し、比較の出力に基づき、特定の遺伝子が遺伝性疾患を関連付けられず、デノボバリアントが良性であると決定するようにさらに構成される。

【０４３８】

一実装形態において、候補バリアントの少なくともいくつかは、タンパク質切り詰めバリアントである。

【０４３９】

別の実装形態において、候補バリアントの少なくともいくつかは、ミスセンスバリアントである。

【０４４０】

ゲノムワイドエンリッチメント解析

図58は、ゲノムワイドエンリッチメント解析の一実装形態を示す。別の実装形態において、異常スプライシング検出器は、異常スプライシングを引き起こすと決定されているバリアントの病原性を決定するゲノムワイドエンリッチメント解析を実装するようにさらに構成される。ゲノムワイドエンリッチメント解析は、訓練済みACNNを適用して健康な個体のコホートからサンプリングされた複数の遺伝子内の異常スプライシングを引き起こすデ

ノボバリアントの第1のセットを識別することと、訓練済みACNNを適用して遺伝性疾患を患っている個体のコホートからサンプリングされた複数の遺伝子内の異常スプライシングを引き起こすデノボバリアントの第2のセットを識別することと、第1および第2のセットのそれぞれのカウントを比較し、比較の出力に基づき、デノボバリアントの第2のセットが遺伝性疾患を患っている個体のコホート内でエンリッチされ、したがって病原性を有すると決定することとを含む。いくつかの実装形態において、遺伝性疾患は、自閉症スペクトラム障害(略語ASD)である。他の実装形態において、遺伝性疾患は、発育遅滞障害(略語DDD)である。

【０４４１】

いくつかの実装形態において、異常スプライシング検出器は、出力としてp値を生成する統計的検定を使用して比較を実行するようにさらに構成される。一実装形態において、この比較は、それぞれのコホートサイズによってさらにパラメータ化され得る。

【０４４２】

いくつかの実装形態において、異常スプライシング検出器は、第1および第2のセットのそれぞれのカウントを比較し、比較の出力に基づき、デノボバリアントの第2のセットが遺伝性疾患を患っている個体のコホート内でエンリッチされず、したがって良性であると決定するようにさらに構成される。

【０４４３】

図58に示されている例では、健康なコホートにおける突然変異率(0.001)および影響を受けているコホートにおける突然変異率(0.004)は、個体当たりの突然変異率(4)とともに例示されている。

【０４４４】

論考

重い遺伝性疾患を患っている患者におけるエクソンシーケンシングの診断率が限られているにもかかわらず、臨床的シーケンシングでは稀少コード突然変異に集中しており、解釈の難しさから非コードゲノムにおけるバリエーションを大部分無視している。ここで、われわれは、一次ヌクレオチド配列からスプライシングを正確に予測するディープラーニングネットワークを導入し、それによって、結果として得られるタンパク質上で重大な結果を有するエクソンおよびイントロンの正常なパターン形成を崩壊させる非コード突然変異を識別する。われわれは、予測された潜在的スプライス突然変異がRNA-seqによって高率でバリデーションし、ヒト母集団において強い悪影響を有し、稀少遺伝病の主要原因であることを示している。

【０４４５】

ディープラーニングネットワークをスプライセオソームのコンピュータシミュレーションに基づくモデルとして使用することによって、われわれは、スプライセオソームがインビボで際立った精度を達成することを可能にする特異性決定因子を再構築することができた。われわれは、スプライシング機構への過去40年にわたる研究でなされた発見の多くを再確認し、スプライセオソームが多数の短距離および長距離特異性決定因子をその決定に一体化することを示している。特に、われわれは、大半のスプライスモチーフの認知された縮退は、モチーフレベルの追加の特異性を補償し、不要なものにする以上のことをする、エクソン/イントロン長およびヌクレオソーム位置決めなどの長距離決定因子の存在によって説明されることを見いだしている。われわれの発見は、単にブラックボックス分類器として働くだけでなく生物学的洞察を与えるためのディープラーニングモデルが有望であることを実証している。

【０４４６】

ディープラーニングは、生物学では比較的新しい技術であり、潜在的なトレードオフを有しないわけではない。配列から特徴を自動的に抽出することを学習することによって、ディープラーニングモデルは、人間の専門家ではきちんと説明できない新規の配列決定因子を利用することができるが、スプライセオソームの真の挙動を反映しない特徴をモデルが組み込む得るリスクもある。これらの無関連の特徴は、アノテーションされたエクソン

-イントロン境界を予測する見掛けの精度を高める可能性もあるが、遺伝的バリエーションによって誘発される任意の配列変化のスプライス変更効果を予測する精度を低下させることになる。バリアントの正確な予測は、モデルが真の生物学に一般化できる最も強い証拠をもたらすので、われわれは、3つの完全に直交する方法、すなわち、RNA-seq、ヒト母集団中の自然選択、および事例対対照コホートにおけるデノボバリアントのエンリッチメントを使用して、予測されたスプライス変更バリアントのバリデーションを提供する。これは、無関連の特徴をモデルに組み込むことを完全には除外しないが、その結果得られるモデルは、遺伝病を患っている患者における潜在的スプライス突然変異を識別することなどの実際の適用に対し大きな価値を有するスプライシングの真の生物学にとって十分に信頼できるように見える。

【０４４７】

タンパク質切り詰め突然変異の他のクラスと比較して、潜在的スプライス突然変異の特に興味深い態様は、不完全な浸透性を有するスプライス変更バリアントにより代替スプライシングの広範な現象であり、これは代替スプライス部位に関してカノニカルスプライス部位を弱める傾向があり、その結果、RNA-seqデータにおける異常転写産物および正常転写産物の両方の混合が産生される。これらのバリアントが頻繁に組織特有の代替スプライシングを押し進めるという観察結果は、新規の代替スプライシング多様性を生成する際の潜在的スプライス変異が果たす予期せぬ役割を際立たせている。将来有望な方向は、関連する組織のRNA-seqからスプライス接合アノテーション上でディープラーニングモデルを訓練し、それによって、代替スプライシングの組織特有のモデルを取得することであろう。RNA-seqデータから直接的に導出されるアノテーション上でネットワークを訓練することも、GENCODEアノテーションにおけるギャップを埋めるのに役立ち、バリアント予測におけるモデルの性能を改善する(図52Aおよび図52B)。

【０４４８】

非コードゲノムにおける突然変異が人間に疾病をどのようにもたらすかに関するわれわれの理解はいまだ完全にはほど遠い。小児期神経発生障害の浸透性デノボ潜在的スプライス突然変異の発見は、非コードゲノムの解釈の改善が重大な遺伝性疾患を患っている患者に直接的な恩恵をもたらし得ることを実証している。潜在的スプライス突然変異は、また、癌にも大きな役割を果たし(Jungら、2015年、Sanzら、2010年、Supekら、2014年)、スプライス因子における反復体細胞突然変異は、スプライシング特異性における広範な変更を発生することが示されている(Graubertら、2012年、Shiraiら、2015年、Yoshidaら、2011年)。特にスプライセオソームにおけるタンパク質に直接的影響を及ぼす突然変異の場合に、異なる組織および細胞構成におけるスプライシングの調節を理解するためにまだ多くの研究を必要とする。配列特有の様式でスプライシング欠陥を潜在的に標的とする可能性のあるオリゴヌクレオチド療法にける最近の進歩に照らして(Finkelら、2017年)、この注目すべき過程を支配する調節機構の理解が進むことが、治療的介入のための新規候補の道を開く可能性がある。

【０４４９】

図37A、図37B、図37C、図37D、図37E、図37F、図37G、および図37Hは、深層学習によって一次配列からスプライシングを予測する一実装形態を示す。

【０４５０】

図37Aに関して、pre-mRNA転写産物内の各位置について、SpliceNet-10kは、隣接する配列の10,000個のヌクレオチドを入力として使用し、その位置がスプライスアクセプターであるか、スプライスドナーであるか、またはそのいずれでもないかを予測する。

【０４５１】

図37Bに関して、MaxEntScan(上)およびSpliceNet-10k(下)を使用してスコアを付けられたCFTR遺伝子に対する完全なpre-mRNA転写産物は、予測されたアクセプター(赤色矢印)およびドナー(緑色矢印)部位ならびにエクソンの実際の位置(黒色ボックス)とともに、図示されている。各方法について、われわれは、予測された部位の数を実際の部位の総数に等しくする閾値を適用した。

【０４５２】
図37Cに関して、各エクソンについて、われわれは、RNA-seq上のエクソンの包含率を測定しており、異なる包含率におけるエクソンに対するSpliceNet-10kスコア分布を示している。図示されているのは、エクソンのアクセプターおよびドナースコアの最大値である。
【０４５３】
図37Dに関して、U2SURP遺伝子におけるエクソン9の周りの各ヌクレオチドをコンピュータシミュレーションで突然変異させる場合の影響が示されている。各ヌクレオチドの垂直方向サイズは、そのヌクレオチドが突然変異されたときのアクセプター部位(黒色矢印)の予測された強度の減少を示す(Δスコア)。
【０４５４】
図37Eに関して、ネットワークの精度に対する入力配列構成のサイズの影響が示されている。Top-k精度は、予測された部位の数が存在している部位の実際の数に等しい閾値における正しく予測されたスプライス部位の割合である。PR-AUCは、精度-再現率曲線の下の面積である。われわれは、また、スプライス部位検出のための3つの他のアルゴリズムのtop-k精度およびPR-AUCも示している。
【０４５５】
図37Fに関して、SpliceNet-80nt(ローカルモチーフスコア)およびSpliceNet-10kによって予測されるような、エクソン/イントロン長と隣接するスプライス部位の強度との間の関係が示されている。エクソン長(黄色)およびイントロン長(ピンク色)のゲノムワイド分布は背景に示されている。x軸は対数目盛になっている。
【０４５６】
図37Gに関して、スプライスアクセプターとドナーモチーフの対は150nt離して置かれており、HMGCR遺伝子に沿って進む。図示されているのは、各位置で、K562ヌクレオソーム信号、およびその位置でエクソンを形成する対の可能性であり、SpliceNet-10kによって予測される通りである。
【０４５７】
図37Hに関して、GTExコホートにおいて新規エクソンを形成するためにSpliceNet-10kモデルによって予測されるプライベート突然変異の近くの平均K562およびGM12878ヌクレオソーム信号が示されている。並べかえ検定によるp値が図示されている。
【０４５８】
図38A、図38B、図38C、図38D、図38E、図38F、および図38Gは、RNAシーケンスデータにおける稀少潜在的スプライス突然変異のバリデーションの一実装形態を示す。
【０４５９】
図38Aに関して、突然変異のスプライス変更影響を評価するために、SpliceNet-10kは、ここで、rs397515893、心筋症に関連するMYBPC3イントロンにおける病原性潜在的スプライスバリアントについて示されているように、突然変異のありなしで遺伝子のpre-mRNA配列内の各位置におけるアクセプターおよびドナースコアを予測する。突然変異に対するΔスコア値は、バリアントから50nt以内のスプライス予測スコアの最大の変化である。
【０４６０】
図38Bに関して、われわれは、SpliceNet-10kモデルによりプライベート遺伝的バリアント(GTExコホート内の149人のうちの1人で観察されている)にスコアを付けた。図示されているのは、スプライシングを変更する(Δスコア>0.2、青色)、またはプライベートエクソンスキッピング接合(上)またはプライベートアクセプターおよびドナー部位(下)の付近でスプライシングに影響を及ぼさない(Δスコア<0.01、赤色)ことを予測されたプライベートバリアントのエンリッチメントである。y軸は、置換を通じて取得された予想される数と比較した、プライベートスプライス事象および付近のプライベート遺伝的バリアントが同じ個体に同時に出現する回数を示している。
【０４６１】
図38Cに関して、不完全な浸透性を持つ新規ドナー部位を形成するPYGBにおけるヘテロ

接合同義バリアントの例が示されている。RNA-seqカバレッジ、接合リードカウント、および接合の位置(青色および灰色の矢印)は、バリアントを有する個体および対照個体について図示されている。エフェクトサイズは、バリアントを有する個体とバリアントを有しない個体との間の新規接合(AC)の使用度の差として計算される。以下の積み上げ棒グラフにおいて、われわれは、アノテーションされた接合または新規接合(それぞれ、「スプライシングなし」および「新規接合」)を使用した参照または代替対立遺伝子を有するリードの数を示している。参照リードの総数は、代替リードの総数と著しく異なっていたが(P=0.018、2項検定)、これは、新規接合のところでスプライシングしている転写産物の60%が、おそらくナンセンス変異依存分解(NMD)によりRNA-seqデータにおいて欠損していることを示唆している。

【０４６２】

図38Dに関して、GTEx RNA-seqデータに対してバリデーションしたSpliceNet-10kモデルによって予測された潜在的スプライス突然変異の割合が示されている。本質的なアクセプターまたはドナージヌクレオチド(破線)の切断部のバリデーション率は、カバレッジおよびナンセンス変異依存分解により100%未満である。

【０４６３】

図38Eに関して、バリデーションされた潜在的スプライス予測に対するエフェクトサイズの分布が示されている。破線(50%)は、完全浸透性ヘテロ接合バリアントの予想されるエフェクトサイズに対応する。本質的なアクセプターまたはドナージヌクレオチド切断の測定されたエフェクトサイズは、ナンセンス変異依存分解または説明が付かないイソ型の変化により50%未満である。

【０４６４】

図38Fに関して、異なるΔスコアカットオフにおけるGTExコホート内のスプライス変更プライベートバリアントを検出するときのSpliceNet-10kの感度が示されている。バリアントは、深イントロンバリアント(エクソンから>50nt)およびエクソンの近くのバリアント(重なり合うエクソンまたはエクソン-イントロン境界から≦50nt)に分割される。

【０４６５】

図38Gに関して、SpliceNet-10kのバリデーション率および感度ならびに異なる信頼度カットオフにおけるスプライス部位予測に対する3つの他の方法が示されている。SpliceNet-10k曲線上の3つの点は、Δスコアカットオフ0.2、0.5、および0.8におけるSpliceNet-10kの性能を示している。他の3つのアルゴリズムでは、曲線上の3つの点は、ΔScoreカットオフ0.2、0.5、および0.8においてSpliceNet-10kと同じ数の潜在的スプライスバリアントを予測する閾値における性能を示している。

【０４６６】

図39A、図39B、および図39Cは、潜在的スプライスバリアントが頻繁に組織固有の代替的スプライシングを形成する一実装形態を示す。

【０４６７】

図39Aに関して、新規ドナー部位を形成するCDC25Bにおけるヘテロ接合エクソンバリアントの例が示されている。このバリアントは、GTExコホートにおける単一の個体に対してプライベートであり、線維芽細胞と比較した筋肉中の新規スプライスイソ型のより大きい割合に有利な組織特有の代替スプライシングを示す(フィッシャーの正確確率検定によりP=0.006)。RNA-seqカバレッジ、接合リードカウント、および接合の位置(青色および灰色の矢印)は、筋肉と線維芽細胞の両方においてバリアントを有する個体および対照個体について図示されている。

【０４６８】

図39Bに関して、バリアントを持っているGTExコホート内の3人すべてにわたって一貫性のある組織特有の効果を示すFAM229Bにおけるヘテロ接合エクソンアクセプター形成バリアントの例が図示されている。動脈および肺に対するRNA-seqが、バリアントを有する3人の個体および対照個体について示されている。

【０４６９】

図39Cに関して、均質性についてのカイ二乗検定によって評価される、表現組織にわたる新規接合の著しく不均一な使用に関連付けられているGTExコホート内のスプライス部位形成バリアントの割合が示されている。低から中へのΔスコア値を有するバリデーションされた潜在的スプライスバリアントは結果として組織特有の代替スプライシングをもたらす可能性がより高かった(P=0.015、フィッシャーの正確確率検定)。

【０４７０】

図40A、図40B、図40C、図40D、および図40Eは、予測される潜在的スプライスバリアントがヒト母集団において強い悪影響を及ぼす一実装形態を示す。

【０４７１】

図40Aに関して、自信を持って予測されたスプライス変更効果(Δスコア≧0.8)を有する同義およびイントロンバリアント(既知のエクソン-イントロン境界から≦50nt、および本質的なGTおよびAGジヌクレオチドを除く)は、60,706人の個体に一度だけ観察された稀少バリアントに関するヒト母集団における共通対立遺伝子頻度(≧0.1%)で強く枯渇している。4.58のオッズ比(カイ二乗検定により$P<10^{-127}$)は、最近現れた予測された潜在的スプライスバリアントの78%が自然選択により除去される十分な悪影響を有していることを示している。

【０４７２】

図40Bに関して、(A)のように計算された、タンパク質切り詰めバリアントと悪影響を有するExACデータセット内の予測された同義およびイントロン潜在的スプライスバリアントの割合が示されている。

【０４７３】

図40Cに関して、バリアントがフレームシフトを引き起こすと予想されるかどうかに基づき分割された、悪影響を有するExACデータセット内の同義およびイントロン潜在的スプライス利得バリアントの割合が示されている(Δスコア≧0.8)。

【０４７４】

図40Dに関して、タンパク質切り詰めバリアントと悪影響を有するgnomADデータセット内の予測された深イントロン(既知のエクソン-イントロン境界から>50nt)潜在的スプライスバリアントの割合が示されている。

【０４７５】

図40Eに関して、稀少(遺伝子頻度<0.1%)タンパク質切り詰めバリアントおよび個別ヒトゲノム毎の稀少機能的潜在的スプライスバリアントの平均数が示されている。機能的であると予想される潜在的スプライス突然変異の数は、悪影響を有する予測の割合に基づき推定される。予測の総数は、より高い。

【０４７６】

図41A、図41B、図41C、図41D、図41E、および図41Fは、稀少遺伝病の患者におけるデノボ潜在的スプライス突然変異の一実装形態を示す。

【０４７７】

図41Aに関して、Deciphering Developmental Disordersコホート(DDD)からの患者、Simons Simplex CollectionおよびAutism Sequencing Consortiumからの自閉症スペクトラム障害(ASD)を患っている個体、さらには健康な対照に対する1人当たりの予測された潜在的スプライスデノボ突然変異が示されている。健康な対照を超えるDDDおよびASDコホートにおけるエンリッチメントが図示されており、コホート間でバリアント確認を調整している。エラーバーは、95%の信頼区間を示している。

【０４７８】

図41Bに関して、健康な対照と比較した各カテゴリのエンリッチメントに基づく、DDDおよびASDコホートに対する機能的カテゴリによる病原性デノボ突然変異の推定された割合が示されている。

【０４７９】

図41Cに関して、異なるΔスコア閾値における健康な対照と比較したDDDおよびASDコホート内の潜在的スプライスデノボ突然変異のエンリッチメントおよび過剰が示されている

。

【０４８０】

図41Dに関して、予測された潜在的スプライス突然変異がエンリッチメント解析においてタンパク質コード突然変異と一緒に含まれたときの、DDDおよびASDコホート(FDR<0.01)におけるデノボ突然変異に対してエンリッチされた新規候補疾病遺伝子のリストが示されている。複数の個体に存在していた表現型が図示されている。

【０４８１】

図41Eに関して、結果としてそれぞれイントロン保持、エクソンスキッピング、およびエクソン伸長を生じる、RNA-seq上でバリデーションする自閉症患者の予測されたデノボ潜在的スプライス突然変異の3つの例が示されている。各例について、影響のある個体に対するRNA-seqカバレッジおよび接合カウントは、上に示されており、突然変異のない対照個体は、下に示されている。配列は、遺伝子の転写に関してセンス鎖上に示されている。青色および灰色の矢印は、それぞれバリアントを有する個体および対照個体における接合の位置を示している。

【０４８２】

図41Fに関して、RNA-seqによる実験的バリデーションについて選択された36個の予測された潜在的スプライス部位に対するバリデーション状況が示されている。

【０４８３】

実験モデルおよび被検体の詳細

36人の自閉症患者に対する被検体詳細が、Iossifovら、Nature 2014年(Table S1)によってすでに公表されており、われわれの論文のTable S4の第1欄の匿名識別子を使用して相互参照できる。

【０４８４】

方法詳細

I. スプライス予測のためのディープラーニング

SpliceNetのアーキテクチャ

われわれは、pre-mRNAヌクレオチド配列からスプライシングを計算により予測するためにいくつかの超深層畳み込みニューラルネットワークベースのモデルを訓練した。われわれは、40、200、1,000、5,000個のヌクレオチドを注目する位置の各側でそれぞれ入力として使用し、位置がスプライスアクセプターおよびドナーである確率を出力する4つのアーキテクチャ、すなわち、SpliceNet-80nt、SpliceNet-400nt、SpliceNet-2k、およびSpliceNet-10kを設計した。より正確には、モデルへ入力は、ワンホットエンコードされたヌクレオチドの配列であり、A、C、G、およびT(または同等のU)は、それぞれ、[1, 0, 0, 0]、[0, 1, 0, 0]、[0, 0, 1, 0]、および[0, 0, 0, 1]としてエンコードされ、モデルの出力は、注目する位置がスプライスアクセプターである、スプライスドナーである、およびいずれでもない確率に対応する、足して1になる3つのスコアからなる。

【０４８５】

SpliceNetアーキテクチャの基本ユニットは、残差ブロック(Heら、2016b)であり、これはバッチ正規化層(IoffeおよびSzegedy、2015年)、正規化線形ユニット(ReLU)、および特定の様式で編成された畳み込みユニットからなる(図21、図22、図23、および図24)。残差ブロックは、深層ニューラルネットワークを設計するときに一般に使用される。残差ブロックを開発する前に、次々に積み重ねられた多数の畳み込みユニットからなる深層ニューラルネットワークは、爆発的に増大する/消失する勾配(GlorotおよびBengio、2010年)、およびそのようなニューラルネットワークの深さを大きくし多くの場合に結果としてより高い訓練誤差を生じさせる(Heら、2016a)という問題により訓練が非常に困難であった。計算実験の包括的なセットを通じて、次々に積み重ねられた多くの残差ブロックからなるアーキテクチャは、これらの問題を克服することが示された(Heら、2016a)。

【０４８６】

完全なSpliceNetアーキテクチャは、図21、図22、図23、および図24において提示されている。アーキテクチャは、入力層を最後から2番目の層に接続するK個の積み重ねられた

残差ブロックと、最後から2番目の層を出力層に接続するソフトマックス活性化を有する畳み込みユニットとからなる。残差ブロックは、i番目の残差ブロックの出力がi+1番目の残差ブロックの入力に接続されるように積み重ねられる。さらに、4番目毎の残差ブロックの出力は、最後から2番目の層の入力に付加される。そのような「スキップコネクション」は、訓練中に収束速度を高めるために深層ニューラルネットワークにおいて一般に使用されている(Oordら、2016年)。

【０４８７】

各残差ブロックは、3つのハイパーパラメータN、W、およびDを有し、Nは畳み込みカーネルの数を表し、Wはウィンドウサイズを表し、Dは各畳み込みカーネルの拡張率(YuおよびKoltun、2016年)を表す。ウィンドウサイズWおよび拡張率Dの畳み込みカーネルは(W-1)D個の近傍位置に及ぶ特徴を抽出するので、ハイパーパラメータN、W、およびDを有する残差ブロックは、2(W-1)D個の近傍位置に及ぶ特徴を抽出する。したがって、SpliceNetアーキテクチャの全近傍範囲は、

【数４０】

$$S = \sum_{i=1}^{K} 2(W_i - 1)D_i$$

で与えられ、N_i、W_i、およびD_iはi番目の残差ブロックのハイパーパラメータである。SpliceNet-80nt、SpliceNet-400nt、SpliceNet-2k、およびSpliceNet-10kアーキテクチャについて、残差ブロックの数および各残差ブロックに対するハイパーパラメータは、Sがそれぞれ80、400、2,000、および10,000に等しくなるように選択された。

【０４８８】

SpliceNetアーキテクチャは、畳み込みユニットに加えて正規化および非線形活性化ユニットしか有していない。その結果、モデルは、可変配列長の配列間モードで使用され得る(Oordら、2016年)。たとえば、SpliceNet-10kモデル(S=10,000)は、長さS/2+l+S/2のワンホットエンコードされたヌクレオチド配列であり、出力はl×3行列であり、これは入力内のl個の中心位置、すなわち、最初と最後のS/2個のヌクレオチドを除外した後に残っている位置の3つのスコアに対応する。この特徴は、訓練さらにはテストで膨大な計算量削減を達成するために活用できる。これは、互いに近い位置に対する計算の大部分が共通であるという事実によるものであり、共有される計算は、配列間モードで使用されるときにモデルによって一度だけ行うだけでよい。

【０４８９】

われわれのモデルは残差ブロックのアーキテクチャを採用しており、これは画像分類において成功したため広く使用されるようになっている。残差ブロックは、前の方の層からの情報が残差ブロックを飛ばすことを可能にするスキップコネクションがちりばめられている、畳み込みの繰り返しユニットを含む。各残差ブロックにおいて、入力層は、最初にバッチ正規化され、その後に正規化線形ユニット(ReLU)を使用する活性化層が続く。次いで、活性化は1D畳み込み層に通される。1D畳み込み層からのこの中間出力は、再びバッチ正規化され、ReLU活性化され、その後に別の1D畳み込み層が続く。第2の1D畳み込みの終わりに、われわれは、残差ブロック内に元の入力を含む出力を総和し、これは元の入力情報が残差ブロックをバイパスすることを可能にすることによってスキップコネクションとして働く。著者らによって深層残差学習ネットワークと呼ばれるそのようなアーキテクチャにおいて、入力は元の状態に保持され、残差コネクションはモデルからの非線形活性化を有さず、より深いネットワークの効果的訓練を可能にする。

【０４９０】

残差ブロックに続き、ソフトマックス層が各アミノ酸に対する3つの状態の確率を計算し、そのうち最大のソフトマックス確率がアミノ酸の状態を決定する。モデルはADAMオプ

ティマイザを使用して全タンパク質配列に対する累積多クラス交差エントロピー損失関数により訓練される。

【０４９１】

Atrous/Dilated畳み込みは、訓練可能なパラメータが少ししかない大きい受容野を実現可能にする。Atrous/Dilated畳み込みは、Atrous畳み込みレートまたは拡張係数とも呼ばれるあるステップを用いて入力値をスキップすることによって、カーネルがその長さよりも大きい領域にわたって適用される畳み込みである。Atrous/dilated畳み込みは、畳み込みフィルタ/カーネルの要素間の間隔を加え、それによって、畳み込み演算が実行されるときにより大きい間隔における近傍の入力エントリ(たとえば、ヌクレオチド、アミノ酸)が考慮される。これは長距離構成依存性を入力に組み込むことを可能にする。Atrous畳み込みは、部分畳み込み計算を隣接するヌクレオチドが処理されるときに再使用できるように保存する。

【０４９２】

図示されている例では1D畳み込みを使用している。他の実装形態において、モデルは、2D畳み込み、3D畳み込み、DilatedまたはAtrous畳み込み、転置畳み込み、分離可能畳み込み、および深さ方向分離可能畳み込みなどの異なる種類の畳み込みを使用することができる。いくつかの層では、シグモイドまたは双曲正接などの飽和非線形性と比較して確率勾配降下の収束を大きく加速するReLU活性化関数も使用する。開示されている技術によって使用できる活性化関数の他の例は、パラメトリックReLU、Leaky ReLU、および指数関数的線形ユニット(ELU)を含む。

【０４９３】

いくつかの層は、バッチ正規化も使用する(IoffeおよびSzegedy 2015年)。バッチ正規化に関して、畳み込みニューラルネットワーク(CNN)における各層の分布は訓練中に変化し、層同士で異なる。これは、最適化アルゴリズムの収束速度を低下させる。バッチ正規化は、この問題を克服するための技術である。xを有するバッチ正規化層の入力およびzを使用する出力を示すことで、バッチ正規化はx上で以下の変換を適用する。

【０４９４】

【数４１】

$$z = \frac{x - \mu}{\sqrt{\sigma^2 + \varepsilon}} \gamma + \beta$$

【０４９５】

バッチ正規化は、μおよびσを使用して入力x上で平均分散正規化を適用し、γおよびβを使用してそれを線形スケーリングし、シフトする。正規化パラメータμおよびσは、指数関数的移動平均と呼ばれる方法を使用して訓練セット上で現在の層について計算される。言い換えると、これらは訓練可能パラメータではない。対照的に、γおよびβは訓練可能パラメータである。訓練中に計算されたμおよびσに対する値は、推論時にフォワードパスで使用される。

【０４９６】

モデルの訓練およびテスト

われわれは、UCSCテーブルブラウザからGENCODE(Harrowら、2012年)V24lift37遺伝子アノテーションテーブルをダウンロードし、20,287個のタンパク質コード遺伝子アノテーションを抽出し、複数のイソ型が利用可能なときに主転写産物を選択した。われわれは、スプライス接合を有していなかった遺伝子を取り除き、次のようにして残りを訓練およびテストセット遺伝子に分割した。染色体2、4、6、8、10～22、XおよびYに属している遺伝子は、モデルを訓練するために使用された(13,384個の遺伝子、130,796個のドナー-アクセプター対)。われわれは、訓練遺伝子の10%をランダムに選択し、それらを訓練中に早期中

止のポイントを決定するために使用し、残りはモデルを訓練するために使用された。モデルをテストするために、われわれは、パラログ(1,652個の遺伝子、14,289個のドナー-アクセプター対)を有していなかった染色体1、3、5、7、および9からの遺伝子を使用した。この目的のために、われわれは、http://grch37.ensembl.org/biomart/martviewからのヒト遺伝子パラログリストを参照した。

【０４９７】

われわれは、次の手順を使用して、サイズl=5,000のチャンクを有する配列間モードのモデルを訓練し、テストした。各遺伝子について、カノニカル転写開始部位と終了部位との間のmRNA転写産物配列は、hg19/GRCh37アセンブリから抽出された。入力mRNA転写産物配列は、次のようにワンホットエンコードされた。すなわち、A、C、G、T/Uそれぞれは[1, 0, 0, 0]、[0, 1, 0, 0]、[0, 0, 1, 0]、[0, 0, 0, 1]にマッピングされた。ワンホットエンコードされたヌクレオチド配列は、長さが5,000の倍数になるまでゼロパディングされ、次いで、長さS/2の隣接配列で開始および終了においてさらにゼロパディングされ、Sは、SpliceNet-80nt、SpliceNet-400nt、SpliceNet-2k、およびSpliceNet-10kモデルに対してそれぞれ80、400、2,000、および10,000に等しい。次いで、パディングされたヌクレオチド配列は、i番目のブロックが5,000(i-1)-S/2+1から5,000i+S/2のヌクレオチド位置からなるように長さS/2+5,000+S/2のブロックに分割された。同様に、スプライス出力標識配列は、次のようにワンホットエンコードされた。スプライス部位、スプライスアクセプター(対応するエクソンの第1ヌクレオチド)、およびスプライスドナー(対応するエクソンの最後のヌクレオチド)はそれぞれ[1, 0, 0]、[0, 1, 0]、および[0, 0, 1]にマッピングされなかった。ワンホットエンコードされたスプライス出力標識配列は、長さが5,000の倍数になるまでゼロパディングされ、次いで、i番目のブロックが5,000(i-1)+1から5,000iの位置からなるように長さ5,000のブロックに分割された。ワンホットエンコードされたヌクレオチド配列および対応するワンホットエンコードされた標識配列は、それぞれ、モデルへの入力およびモデルの標的出力として使用された。

【０４９８】

モデルは、10エポックの間、バッチサイズ12により2個のNVIDIA GeForce GTX 1080 Ti GPU上で訓練された。標的出力と予測された出力との間の多カテゴリ交差エントロピー損失は、訓練中にAdamオプティマイザ(KingmaおよびBa、2015年)を使用して最小化された。オプティマイザの学習速度は最初の6エポックで0.001に設定され、次いで、その後のエポック毎に1/2に減らされた。各アーキテクチャについて、われわれは、訓練手順を5回繰り返し、5個の訓練済みモデルを取得した(図53Aおよび図53B)。テスト中、各入力は5個の訓練済みモデルすべてを使用して評価され、その出力の平均値が予測された出力として使用された。われわれは、これらのモデルを図37Aの解析および他の関係する図に使用した。

【０４９９】

スプライス変更バリアントの識別を伴う図38A～図38G、図39A～図39C、図40A～図40E、および図41A～図41Fの解析結果に関して、われわれは、GENCODEアノテーションの訓練セットを、染色体2、4、6、8、10～22、X、Y(67,012個のスプライスドナーおよび62,911個のスプライスアクセプター)上のGTExコホートにおいて一般的に観察される新規スプライス接合も含むように増補した。これは、訓練セット内のスプライス接合アノテーションの数を～50%増やした。組み合わされたデータセット上でネットワークを訓練することで、特に、深イントロンスプライス変更バリアントを予測するために、GENCODEアノテーション単独(図52Aおよび図52B)で訓練されたネットワークと比較してRNA-seqデータ内のスプライス変更バリアントを検出する感度を改善しており、われわれは、このネットワークをバリアントの評価を伴う解析に使用した(図38A～図38G、図39A～図39C、図40A～図40E、および図41A～図41Fならびに関連する図)。GTEx RNA-seqデータセットが訓練と評価との間の重なりを含まなかったことを確認するために、われわれは、訓練データセット内の5人またはそれ以上の個体に存在している接合のみを含め、4人以下に存在しているバリアント上でネットワークの性能を評価するのみであった。新規スプライス接合識別の詳細は、方法のGTEx解析の節の「スプライス接合の検出」において説明されている。

【０５００】
Top-k精度
正しく分類された位置のパーセンテージのような精度測定基準は、位置の大半がスプライス部位でないという事実により大部分効果がない。われわれは、その代わりに、そのような設定において有効である2つの測定基準、すなわち、top-k精度および精度-再現率曲線の下の面積を使用してモデルを評価した。特定のクラスに対するTop-k精度は次のように定義される。テストセットがクラスに属するk個の位置を有すると仮定する。われわれは、ちょうどk個のテストセット位置がそのクラスに属すものとして予測されるように閾値を選択する。真にこのクラスに属すこれらのk個の予測された位置の割合はTop-k精度として報告される。実際、これは、精度および再現率が同じ値を有するように閾値が選択されたときの精度に等しい。
【０５０１】
lincRNA上のモデル評価
われわれは、GENCODE V24lift37アノテーションに基づくすべてのlincRNA転写産物のリストを取得した。タンパク質コード遺伝子と異なり、lincRNAは、GENCODEアノテーション内の主転写産物を割り当てられない。バリデーションセット内の冗長性を最小にするために、われわれは、lincRNA遺伝子毎に最長の総エクソン配列で転写産物を識別し、これを遺伝子に対するカノニカル転写産物と呼んだ。lincRNAアノテーションは、タンパク質コード遺伝子に対するアノテーションに比べて信頼性が低いと予想されるので、またそのようミスアノテーションはTop-k精度のわれわれの推定に影響を及ぼすので、われわれは、GTExデータを使用して潜在的アノテーション問題のあるlincRNAを排除した(これらのデータの詳細については以下の「GTExデータセット上の解析」の節を参照)。各lincRNAについて、われわれは、すべてのGTExサンプルにわたるlincRNAの長さにわたってマッピングされたすべての分割リードをカウントした(詳細について以下の「スプライス接合の検出」を参照)。これは、アノテーションされるか、または新規の接合のいずれかを使用するlincRNAの全接合スパニングリードの推定であった。われわれは、また、カノニカル転写産物の接合に及ぶリードの数をカウントした。われわれは、すべてのGTExサンプルにわたる接合スパニングリードの少なくとも95%がカノニカル転写産物に対応しているlincRNAのみを考察した。われわれは、また、カノニカル転写産物のすべての接合がGTExコホート内で少なくとも一回観察されることも要求した(長さ<10ntのイントロンに及ぶ接合を除く)。Top-k精度を計算するために、われわれは、上記のフィルタを通ったlincRNAのカノニカル転写産物の接合のみを考慮した(781個の転写産物、1047個の接合)。
【０５０２】
pre-mRNA配列からのスプライス接合の識別
図37Bにおいて、MaxEntScanおよびSpliceNet-10kの性能を配列からの遺伝子のカノニカルエクソン境界を識別することに関して比較している。われわれは、われわれのテストセット内にあり、26個のカノニカルスプライスアクセプターおよびドナーを有する、CFTR遺伝子を、ケーススタディとして使用し、MaxEntScanおよびSpliceNet-10kを使用してカノニカル転写開始部位(chr7:117,120,017)からカノニカル転写終了部位(chr7:117,308,719)までの188,703個の位置の各々についてアクセプターおよびドナーのスコアを取得した。位置は、対応するスコアがTop-k精度を評価しながら選択された閾値より大きい場合にスプライスアクセプターまたはドナーとして分類された。MaxEntScanは、49個のスプライスアクセプターおよび22個のスプライスドナーを予測し、そのうち9および5個はそれぞれ真のスプライスアクセプターおよびドナーである。視覚化がより適切になされるように、われわれは、MaxEntScanのログ前スコアを示している(最大2,500にクリップされている)。SpliceNet-10kは、26個のスプライスアクセプターおよび26個のスプライスドナー部位を予測したが、これらはすべて正しい。図42Bでは、われわれは、LINC00467遺伝子を使用して解析を繰り返した。
【０５０３】
GENCODEアノテーションされたスプライス接合におけるエクソン包含の推定

われわれは、GTEx RNA-seqデータからすべてのGENCODEアノテーションされたエクソンの包含率を計算した(図37C)。各エクソンについて、各遺伝子の最初と最後のエクソンを除外して、われわれは包含率を以下のように計算した。

【０５０４】

【数４２】

$$\frac{(L+R)/2}{S+(L+R)/2}$$

【０５０５】

ここで、Lはすべてのサンプルにわたって前のカノニカルエクソンから考察対象のエクソンまで接合の全リードカウントであり、Rは考察対象のエクソンから次のカノニカルエクソンまで接合の全リードカウントであり、Sは前のカノニカルエクソンから次のカノニカルエクソンまでスキッピング接合の全リードカウントである。

【０５０６】

スプライス部位認識への様々なヌクレオチドの有意性

図37Dにおいて、われわれは、位置をスプライスアクセプターとして分類することに向けてSpliceNet-10kによって重要とみなされるヌクレオチドを識別している。このために、われわれは、われわれのテストセットの中にある、U2SURP遺伝子中のchr3:142,740,192のスプライスアクセプターを考察した。スプライスアクセプターに関するヌクレオチドの「重要度スコア」は次のように定義される。s_{ref}は考察対象のスプライスアクセプターのアクセプタースコアを表すものとする。アクセプタースコアは、考察対象のヌクレオチドをA、C、G、およびTで置き換えることによって再計算される。これらのスコアをそれぞれs_A、s_C、s_G、およびs_Tで表すものとする。ヌクレオチドの重要度スコアは以下と推定される。

【０５０７】

【数４３】

$$s_{ref} - \frac{s_A + s_C + s_G + s_T}{4}$$

【０５０８】

この手順は、コンピュータ内突然変異生成と呼ばれることが多い(ZhouおよびTroyanskaya、2015年)。われわれは、各ヌクレオチドの高さがchr3:142,740,192におけるスプライスアクセプターに関する重要度スコアとなるようにchr3:142,740,137からchr3:142,740,263までの127個のヌクレオチドをプロットした。プロット機能は、DeepLIFT(Shrikumarら、2017年)ソフトウェアから採用した。

【０５０９】

スプライシングに対するTACTAACおよびGAAGAAモチーフの効果

アクセプター強度に対する分岐点配列の位置の影響を研究するために、われわれは、最初に、SpliceNet-10kを使用して14,289個のテストセットスプライスアクセプターのアクセプタースコアを取得した。y_{ref}はこれらのスコアを含むベクトルを表すものとする。0から100までの範囲内のiの各値について、われわれは次のことを行った。各テストセットスプライスアクセプターについて、われわれは、スプライスアクセプターの前のiからi-6までの位置のヌクレオチドをTACTAACで置き換え、SpliceNet-10kを使用してアクセプタースコアを再計算した。これらのスコアを含むベクトルは、$y_{alt,i}$で表される。われわれは

、図43Aにおいて次の数量をiの関数としてプロットしている。
$mean(y_{alt,i}-y_{ref})$
【０５１０】
図43Bでは、われわれは、SR-タンパク質モチーフGAAGAAを使用して同じ手順を繰り返した。この場合、われわれは、また、スプライスアクセプターの後に存在しているときのモチーフの影響さらにはドナー強度に対する影響も研究した。GAAGAAおよびTACTAACは、k-mer空間内の包括的探索に基づき、アクセプターおよびドナー強度に対する最大の影響を持つモチーフであった。
【０５１１】
スプライシングにおけるエクソンおよびイントロン長の役割
スプライシングに対するエクソン長の効果を調べるために、われわれは、最初または最後のいずれかのエクソンであったテストセットエクソンをフィルタで除去した。このフィルタ処理ステップで、14,289個のエクソンから1,652個を取り除いた。われわれは、長さが大きくなる順に残りの12,637個のエクソンをソートした。それらの各々について、われわれは、SpliceNet-80ntを使用してスプライスアクセプター部位においてアクセプタースコアおよびスプライスドナー部位におけるドナースコアを平均することによってスプライシングスコアを計算した。われわれは、図37Fにおいてスプライシングスコアをエクソン長の関数としてプロットしている。プロットする前に、われわれは次の平滑化手順を適用した。xはエクソンの長さを含むベクトルを表し、yはその対応するスプライシングスコアを含むベクトルを表すものとする。われわれは、サイズ2,500の平均化ウィンドウを使用してxおよびyの両方を平滑化した。
【０５１２】
われわれは、SpliceNet-10kを使用してスプライシングスコアを計算することによってこの解析を繰り返した。背景で、われわれは、この解析について考察した12,637個のエクソンの長さのヒストグラムを図示している。われわれは、類似の解析を適用してスプライシングに対するイントロン長の効果を調べたが、主要な違いは最初と最後のエクソンを除外する必要がなかったことである。
【０５１３】
スプライシングにおけるヌクレオソームの役割
われわれは、UCSCゲノムブラウザからK562細胞株に対するヌクレオソームデータをダウンロードした。われわれは、われわれのテストセットの中にある、HMGR遺伝子を事例として使用し、SpliceNet-10kスコアに対するヌクレオソーム位置決めの影響を実証した。遺伝子内の各位置pについて、われわれは、次のようにして「プラントスプライシングスコア」を計算した。
・　位置p+74からp+81の8個のヌクレオチドは、ドナーモチーフAGGTAAGGによって置き換えられた。
・　位置p-78からp-75の4個のヌクレオチドは、アクセプターモチーフTAGGによって置き換えられた。
・　位置p-98からp-79の20個のヌクレオチドは、ポリピリミジントラクトCCTCCTTTTTCCTCGCCCTCによって置き換えられた。
・　位置p-105からp-99の7個のヌクレオチドは、分岐点配列CACTAACによって置き換えられた。
・　SpliceNet-10kによって予測されたp-75におけるアクセプタースコアとp+75におけるドナースコアとの平均はプラントスプライシングスコアとして使用されている。
【０５１４】
K562ヌクレオソーム信号さらにはchr5:74,652,154からchr5:74,657,153までの5,000個の位置に対するプラントスプライシングスコアが図37Gに示されている。
【０５１５】
これらの2つのトラックの間のゲノムワイドのスピアマン相関を計算するために、われわれは、すべてのカノニカル遺伝子から少なくとも100,000ntだけ離れていた1000000個の

遺伝子間位置をランダムに選択する。これらの位置の各々について、われわれは、プラントスプライシングスコアさらには平均K562ヌクレオソーム信号を計算した(ウィンドウサイズ50が平均化に使用された)。1000000個の位置にわたるこれら2つの値の間の相関は図37Gに示されている。われわれは、さらに、ビンサイズ0.02のGC内容(プラントアクセプターとドナーモチーフとの間にあるヌクレオチドを使用して推定される)に基づきこれらの位置を部分分類した。われわれは、図44Aに各ビンに対するゲノムワイドのスピアマン相関を示している。

【０５１６】

14,289個のテストセットスプライスアクセプターの各々について、われわれは、各側で50個のヌクレオチド内のヌクレオソームデータを抽出し、そのヌクレオソームエンリッチメントをイントロン側の平均信号によって除算されるエクソン側の平均信号として計算した。われわれは、ヌクレオソームエンリッチメントの昇順でスプライスアクセプターをソートし、SpliceNet-80ntを使用してそのアクセプタースコアを計算した。アクセプタースコアは、図44Bにおいてヌクレオソームエンリッチメントの関数としてプロットされている。プロットする前に、図37Fで使用されている平滑化手順が適用された。われわれは、SpliceNet-10kを使用して、また14,289個のテストセットスプライスドナーに対してこの解析を繰り返した。

【０５１７】

新規エクソンにおけるヌクレオソーム信号のエンリッチメント

図37Hでは、われわれは、予測された新規エクソンの周りのヌクレオソーム信号を見たかった。われわれが信頼性の高い新規エクソンを見ていたと確認するために、われわれは、予測された利得接合がバリアントを有する個体に対して完全にプライベートであったシングルトンバリアント(単一のGTEx個体に存在するバリアント)のみを選択した。それに加えて、付近のエクソンから交絡効果を取り除くために、われわれは、アノテーションされたエクソンから少なくとも750nt離れているイントロンバリアントのみを見た。われわれは、UCSCブラウザからGM12878およびK562細胞株に対するヌクレオソーム信号をダウンロードし、予測された新規アクセプターまたはドナー部位の各々から750ntの範囲内のヌクレオソーム信号を抽出した。われわれは、2つの細胞株の間のヌクレオソーム信号を平均化し、マイナス鎖上の遺伝子と重なり合うバリアントに対する信号ベクトルをフリップさせた。われわれは、アクセプター部位からの信号を右に70ntだけシフトし、ドナー部位からの信号を左に70ntだけシフトした。シフトした後、アクセプター部位およびドナー部位の両方に対するヌクレオソーム信号は長さ140ntの理想化されたエクソンの真ん中にセンタリングされたが、これはGENCODE v19アノテーション内のエクソンの長さ中央値である。われわれは、最後に、すべてのシフトされた信号を平均化し、各位置を中心とする11ntのウィンドウ内で平均を計算することによって結果として得られる信号を平滑化した。

【０５１８】

関連付けをテストするために、われわれは、アノテーションされたエクソンから少なくとも750nt離れ、スプライシングに対して効果を有しないとモデルによって予測されたランダムシングルトンSNVを選択した(Δスコア<0.01)。われわれは、そのようなSNVの1000個のランダムサンプルを作成し、各サンプルは図37Hに使用されたスプライス部位利得部位のセットと同じ数のSNVを有していた(128個の部位)。各ランダムサンプルについて、われわれは、上で説明されているように平滑化された平均信号を計算した。ランダムSNVは新規エクソンを形成すると予測されなかったので、われわれは、SNVそれ自体に各SNVからのヌクレオソーム信号をセンタリングし、左に70ntまたは右に70ntのいずれかにランダムにシフトした。次いで、われわれは、図37Hの真ん中の塩基のヌクレオソーム信号とその塩基での1000回のシミュレーションから取得された信号とを比較した。経験的p値は、スプライス部位利得バリアントについて観察された値以上の中間値を有していたシミュレートされたセットの割合として計算された。

【０５１９】

エクソン密度の差に対するネットワークのロバスト性

ネットワークの予測の一般化可能性を調べるために、われわれは、エクソン密度が変化する領域においてSpliceNet-10kを評価した。われわれは、最初に、10,000ヌクレオチドウィンドウ(各側に5,000個のヌクレオチド)内に存在しているカノニカルエクソンの数に応じてテストセット位置を5つのカテゴリに分けた(図54)。エクソンカウントが各位置に対して整数値となるように、われわれは、ウィンドウ内に存在するエクソン開始の数を代理として使用した。各カテゴリについて、われわれは、Top-k精度および精度-再現率曲線の下の面積を計算した。位置の数およびkの値は、異なるカテゴリでは異なっている(以下の表に詳述)。

【０５２０】

【表２】

エクソンカウント	位置の数	スプライス アクセプターの数	スプライス ドナーの数
１エクソン	15,870,045	1,712	1,878
２エクソン	10,030,710	2,294	2,209
３エクソン	6,927,885	2,351	2,273
４エクソン	4,621,341	2,095	2,042
≧５エクソン	7,247,582	5,679	5,582

【０５２１】

アンサンブル内の5つのモデルの各々に対するネットワークのロバスト性

複数のモデルを訓練し、その予測値の平均を出力として使用することは、よりよい予測性能を得るための機械学習における一般的な戦略であり、アンサンブル学習と称される。図53Aにおいて、われわれは、アンサンブルの構築するためにわれわれが訓練した5つのSpliceNet-10kモデルのTop-k精度および精度-再現率曲線の下の面積を示している。結果は、明らかに、訓練プロセスの安定性を示している。

【０５２２】

われわれは、また、予測の間のピアソン相関も計算した。ゲノム内のほとんどの位置はスプライス部位でないので、ほとんどのモデルの予測の間の相関は1に近くなり、解析結果を無意味なものにする。この問題を克服するために、われわれは、少なくとも1つのモデルによって0.01以上のアクセプターまたはドナースコアを割り当てられたテストセット内の位置のみを考慮した。この基準は、53,272個の位置(スプライス部位および非スプライス部位のおおよそ等しい数)で満たされた。これらの結果は、図53Bに要約されている。モデルの予測の間の非常に高いピアソン相関は、そのロバスト性をさらに例示している。

【０５２３】

われわれは、図53Cにおいて性能に対するアンサンブルを構築するために使用されるモデルの数の効果を示している。これらの結果は、モデルの数が増加するにつれ性能が改善することを示しているが、収穫は減る。

【０５２４】

II. GTEx RNA-seqデータセット上の解析

単一ヌクレオチドバリアントのΔスコア

われわれは、次のようにして単一ヌクレオチドバリアントによるスプライシング変化を定量化した。われわれは、最初に参照ヌクレオチドを使用し、バリアントの周りの101個の位置に対するアクセプターおよびドナースコアを計算した(各側に50個の位置)。これらのスコアは、それぞれ、ベクトルa_{ref}およびd_{ref}によって表される。われわれは、次いで、代替ヌクレオチドを使用し、アクセプターおよびドナースコアを再計算した。これらのスコアはそれぞれベクトルa_{alt}およびd_{alt}によって表されるものとする。

われわれは、次の4つの量を評価した。
Δスコア(アクセプター利得)=max(a_{alt}-a_{ref})
Δスコア(アクセプター損失)=max(a_{ref}-a_{alt})
Δスコア(ドナー利得)=max(d_{alt}-d_{ref})
Δスコア(ドナー損失)=max(d_{ref}-d_{alt})
【０５２５】
これら4つのスコアの最大値は、バリアントのΔスコアと呼ばれる。
【０５２６】
バリアントの品質管理およびフィルタ処理の基準
われわれは、GTEx VCFおよびRNA-seqデータをdbGaPからダウンロードした(研究アクセッションphs000424.v6.p1; https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000424.v6.p1)。
【０５２７】
われわれは、GTExコホートにおけるせいぜい4人の個体に出現した常染色体SNV上のSpliceNetの性能を評価した。特に、バリアントは、少なくとも1人の個体Aにおいて次の基準を満たした場合に考察された。
1. バリアントはフィルタ処理されなかった(VCFのFILTERフィールドはPASSであった)。
2. バリアントは個体AのVCFのINFOフィールド内でMULTI_ALLELICとマークされず、VCFはALTフィールド内で単一の対立遺伝子を含んでいた。
3. 個体Aはバリアントに対してヘテロ接合体であった。
4. 比alt_depth / (alt_depth + ref_depth)は、0.25から0.75の間であり、alt_depthおよびref_depthはそれぞれ個体Aにおける代替および参照対立遺伝子を支持するリードの数である。
5. 全深さ、alt_depth + ref_depthは、個体AのVCFにおいて20から300の間であった。
6. バリアントは遺伝子本体領域と重なっていた。遺伝子本体は、GENCODE(V24lift37)からのカノニカル転写産物の転写の開始と終了との間の領域として定義された。
【０５２８】
少なくとも1人の個体においてこれらの基準を満たすバリアントに対して、われわれは、バリアントが出現した(上記の基準を満たしていなかったとしても)すべての個体をバリアントを有しているとみなした。われわれは、単一の個体に出現するバリアントをシングルトン、2～4人の個体に出現するバリアントを共通と称する。われわれは、訓練データセットと重なり合うのを防ぐために、5人またはそれ以上の個体に出現するバリアントを評価しなかった。
【０５２９】
RNA-seqリードアライメント
われわれは、OLego(Wuら、2013年)を使用して、hg19参照に対してGTExサンプルのリードをマッピングし、クエリリードと参照との間のせいぜい4の編集距離(パラメータ -M 4)を可能にした。OLegoは完全にデノボで動作することができ、遺伝子アノテーションを必要としないことに留意されたい。Olegoは、分割されたリードの末端のところにスプライシングモチーフが存在するかどうかを調べるので、そのアライメントは、それぞれスプライス部位を切断または作成するSNVの周りで参照の方へ、または参照に対抗してバイアスされ得る。そのようなバイアスを排除するために、われわれは、PASSフィルタによりhg19参照内に個体のすべてのSNVを挿入することによって、各GTEx個体に対する代替参照配列をさらに作成した。われわれは、OLegoを同じパラメータで使用し、各個体からのすべてのサンプルをその個体の代替参照配列に対してマッピングした。各サンプルについて、次いで、われわれは、各リード対に対する最良のアライメントをピックすることによって、アライメントの2つのセットを組み合わせた(hg19参照に対して、および個体の代替参照に対して)。リード対Pに対する最良のアライメントを選択するために、われわれは、次の手順を使用した。

1．Pの両方のリードがアライメントの両方のセット内でアンマッピングされた場合、われわれは、hg19またはPの代替アライメントをランダムに選択する。
2．Pがアライメントの一方のセットにおけるアンマッピングされた末端を他に比べて多く有していた(たとえば、Pの両端は代替参照に対してマッピングされたが、hg19に対してはただ1つの末端がマッピングされた)場合、われわれは、Pの両端がマッピングされたアライメントを選択する。
3．Pの両端がアライメントの両方のセット内でマッピングされた場合、われわれは、最小の総ミスマッチとのアライメント、またはミスマッチの数が同じであった場合にランダム1を選択する。

【０５３０】

整列されたRNA-seqデータ内のスプライス接合の検出

われわれは、leafcutterパッケージのユーティリティであるleafcutter_cluster(Liら、2018年)を使用して各サンプル中のスプライス接合を検出し、カウントした。われわれは、単一の分割されたリードが接合を支持することを要求し、500Kbの最大イントロン長を仮定した(パラメータ -m 1 -l 500000)。ディープラーニングモデルを訓練するための接合の高信頼度セットを得るために、われわれは、すべてのサンプル上ですべてのleafcutter接合の合併をコンパイルし、次の基準のうちのどれかを満たす接合を考察対象から外した。

1．接合のいずれかの末端がENCODEブラックリスト領域(UCSCゲノムブラウザからのhg19内のテーブルwgEncodeDacMapabilityConsensusExcludable)または単純な反復(UCSCゲノムブラウザからのhg19におけるSimple Repeatsトラック)と重なった。
2．接合の両端が、非カノニカルエクソン上にあった(GENCODE version V24lift37からのカノニカル転写産物に基づく)。
3．接合の2つの末端が異なる遺伝子上にあった、またはいずれかの末端が非遺伝的領域内にあった。
4．いずれかの末端が本質的なGT/AGジヌクレオチドを欠いた。

【０５３１】

5人またはそれ以上の個体に存在していた接合は、バリアント予測に関する解析のためGENCODEアノテーションスプライス接合のリストを増補するために使用された(図38A～図38G、図39A～図39C、図40A～図40E、および図41A～図41F)。モデルを訓練するために使用されるスプライス接合のリストを含むファイルへのリンクがKey Resourcesテーブル内に提供される。

【０５３２】

われわれは、leafcutterによって検出された接合を使用して訓練データセットを増補したけれども、われわれは、緩和されたパラメータを使用したにもかかわらず、leafcutterがRNA-seqデータにおける良好な支持で多くの接合フィルタリングしていたことに気付いている。これは、人工的に、われわれのバリデーション率を下げた。したがって、GTEx RNA-seqバリデーション解析(図38A～図38Gおよび図39A～図39C)についいては、われわれは、RNA-seqリードデータから直接的に接合および接合カウントのセットを再計算した。われわれは、MAPQにより少なくとも10回、接合の各側で整列された少なくとも5ntですべての非重複分割マッピング済みリードをカウントした。リードは2つよりも多いエクソンに及ぶことを許されており、その場合、リードは、両側でマッピングされた配列の少なくとも5ntで各接合の方へカウントされた。

【０５３３】

プライベート接合の定義

接合は、次の基準のうちの少なくとも1つを満たした場合に個体Aにおいてプライベートと考えられた。

1．接合はAからの少なくとも1つのサンプルにおいて少なくとも3つのリードを有しており、他の任意の個体では決して観察されなかった。
2．次の2つの基準のうちの両方を満たした少なくとも2つの組織があった。

a. 組織内の個体Aからのサンプル内の接合の平均リードカウントは少なくとも10であった。

b. 個体Aはその組織内の他の任意の個体に比べて平均して少なくとも2倍多い正規化されたリードを有していた。ここで、サンプル内の接合の正規化されたリードカウントは、対応する遺伝子に対するすべての接合にわたってリードの総数によって正規化された接合のリードの数として定義された。

【０５３４】

他の個体(Aでない)からの5個より少ないサンプルを有する組織はこのテストでは無視された。

【０５３５】

プライベート接合の周りのシングルトンSNVのエンリッチメント

プライベート接合がアノテーションされたちょうど1つの末端を有していた場合、GENCODEアノテーションに基づき、われわれは、それをアクセプターまたはドナー利得に対する候補としてみなし、アノテーションされていない末端から150nt以内の同じ個体においてプライベートであったシングルトンSNV(単一のGTEx個体に出現するSNV)を探索した。プライベート接合がアノテーションされた両方の末端を有していた場合、われわれは、それを、GENCODEアノテーションに基づき同じ遺伝子の少なくとも1つただし3個以下のエクソンをスキップした場合にプライベートエクソンスキッピング事象に対する候補とみなした。次いで、われわれは、スキップされたエクソンの各々の末端から150nt以内でシングルトンSNVを探索した。GENCODEエクソンアノテーションに両方の末端がないプライベート接合は、これらのうちの実質的割合がアライメント誤差であるので、無視された。

【０５３６】

新規プライベートアクセプターまたはドナー(図38B、下)の周りでシングルトンSNVのエンリッチメントを計算するために、われわれは、プライベート接合に関する各位置でシングルトンSNVのカウントを集計した。重なり合う遺伝子がマイナス鎖上にあった場合、相対的位置はフリップされた。われわれは、SNVを2つのグループに分割した。すなわち、1つはプライベート接合を持つ個体においてプライベートであったSNV、および異なる個体においてプライベートであったSNVである。結果として得られる信号を平滑化するために、われわれは、各位置を中心とする7ntのウィンドウにおいてカウントを平均した。次いで、われわれは、第1のグループ(同じ個体においてプライベート)からの平滑化されたカウントと第2のグループ(異なる個体においてプライベート)の平滑化されたカウントとの比を計算した。新規プライベートエクソンスキップ(図38B、上)について、われわれは類似の手順に従い、スキップされたエクソンの末端の周りでシングルトンSNVのカウントを集計した。

【０５３７】

GTEx RNA-seqデータ内のモデル予測のバリデーション

プライベートバリアント(GTExコホート内の1人の個体に出現する)または共通バリアント(GTExコホート内の2人から4人の個体に出現する)のいずれかについて、われわれは、参照および代替対立遺伝子に対するディープラーニングモデルの予測を取得し、Δスコアを計算した。われわれは、モデルがそうであるべき異常(新規または切断)接合を予測した配置も取得した。次いで、われわれは、予測された配置においてバリアントを有する個体におけるスプライシング異常を支持するRNA-seqデータ内に証拠があるかどうかを決定することに努めた。多くの場合において、モデルは、同じバリアントに対する複数の効果を予測することができ、たとえば、アノテーションされたスプライスドナーを切断するバリアントは、また、図45のような次善のドナーの使用度を高めることが可能であり、その場合、モデルはアノテーションされたスプライス部位でのドナー損失および次善の部位でのドナー利得の両方を予測することもあるであろう。しかしながら、バリデーションの目的では、われわれは、各バリアントに対して最高の予測されたΔスコアを持つ効果のみを考察した。したがって、各バリアントについて、われわれは、予測されたスプライス部位形成およびスプライス部位切断効果を別々に考察した。5人より少ない個体に出現する接合は

、モデルが訓練された新規接合上でそのモデルを評価することを回避するために、モデル訓練時に除外された。

【０５３８】

プライベートスプライス接合に基づく予測された潜在的スプライス突然変異のバリデーション

新規接合形成を引き起こすと予測された各プライベートバリアントについて、われわれは、ネットワークを使用して新しく作成された異常スプライス接合の位置を予測し、そのような新規接合がSNVを持つ個体のみに出現し、他のどのGTEx個体にも出現しなかった場合バリデーションするRNA-seqデータを見た。同様に、エクソンXのスプライス部位に影響を及ぼすスプライス部位損失を引き起こすと予測されたバリアントについて、われわれは、前のカノニカルエクソン(GENCODEアノテーションに基づくXの上流にあるもの)から、バリアントを有する個体のみに出現し、GTExにおける他のどの個体にも出現しなかった次のカノニカルエクソン(Xの下流にあるもの)へ、新規エクソンスキッピング事象を探した。われわれは、モデルによって失われると予測されたスプライス部位がGENCODEにおいてアノテーションされないか、またはバリアントを有しないGTEx個体において決して観察されなかった場合に予測された損失を除外した。われわれは、また、利得を得ると予測されたスプライス部位がGENCODEにおいてすでにアノテーションされていた場合に予測された利得を除外した。この解析を共通バリアント(2人から4人の個体に存在している)に拡大適用するために、われわれは、また、バリアントを有する個体の少なくとも半分に存在し、バリアントを有しないすべての個体には存在していない新規接合をバリデーションした。

【０５３９】

予測された異常スプライス事象がバリアントを有する個体にプライベートであるという要求条件を使用することで、われわれは、予測された高スコア(Δスコア≧0.5)アクセプターおよびドナー利得の40%、ただし予測された高スコア損失の3.4%のみおよび本質的なGTまたはAG切断の5.6%(置換に基づき<0.2%の誤バリデーション率で--「誤バリデーション率の推定」の節を参照)をバリデーションすることが可能である。利得および損失のバリデーション率における食い違いの理由は2つある。第1に、利得とは異なり、エクソンスキッピング事象は、バリアントを有する個体に対して全体的にプライベートであることはめったにないが、それは、エクソンが低いベースラインレベルでスキップされることが多いからであり、これは十分に深いRNA-seqで観察され得る。第2に、スプライス部位損失は、イントロン補助を高めることまたは代替次善スプライス部位の使用度を高めることなど、エクソンスキッピングを高めることに加えて他の効果を有することができる。これらの理由から、われわれは、モデルの予測をバリデーションすることに対してプライベート新規接合に完全には頼らず、われわれは、また、バリアントを有する個体において影響を受けると予測された接合の使用度の増減に対する定量的証拠に基づきバリアントをバリデーションした。

【０５４０】

定量的基準を通しての予測された潜在的スプライス突然変異のバリデーション

サンプルsからの接合jについて、われわれは正規化された接合カウントc_{js}を取得した。

【０５４１】

【数４４】

$$c_{js} = \operatorname{asinh}\left(\frac{r_{js}}{\sum_g r_{gs}}\right) (1)$$

【０５４２】

ここで、r_{js}はサンプルsにおける接合jに対する未処理接合カウントであり、分母の中

の和は、jと同じ遺伝子のアノテーションされたアクセプターとドナーとの間の他のすべての接合にわたって取られる(GENCODE v19からのアノテーションを使用する)。asinh変換は、

【数４５】

$$\mathrm{asinh}(x) = \ln(x + \sqrt{x^2 + 1})$$

として定義される。これは、RNA-seqデータを変換するために使用されることが多い対数変換に類似しているが(Lonsdaleら、2013年)、これは0で定義され、したがって、多くの接合、特に新規接合が低またはゼロのカウントを有するので、実質的に値を歪ませたであろう、疑似カウントの必要がなくなる。asinh変換は、大きな値に対して対数変換のように振る舞うが、小さな値に対しては線形に近い。このような理由から、これは少数の大きな値が信号を支配するのを防ぐためにゼロに近い多数の値を含むデータセット(RNA-seqまたはChIP-seqデータセット)において使用されることが多い(Azadら、2016年、Herringら、2018年、Hoffmanら、2012年、Kasowskiら、2013年、SEQC/MAQC-III Consortium, 2014年)。以下で説明されているように、「バリデーションに対する考慮基準」の節において、式(1)の中の分母が200未満であるサンプルは、すべてのバリデーション解析のために除外され、それにより数値的な問題を回避した。

【０５４３】

個体のセットI内に出現するSNVによって引き起こされると予測された各利得または損失接合jについて、われわれは、各組織tにおいてzスコアを計算した。

【０５４４】

【数４６】

$$z_{jt} = \frac{mean_{s \in A_t}(c_{js}) - mean_{s' \in U_t}(c_{js'})}{std_{s' \in U_t}(c_{js'})} \quad (2)$$

【０５４５】

ここで、A_tは組織t内のIにおける個体からのサンプルセットであり、U_tは組織t内の他のすべての個体からのサンプルセットである。同じ個体および組織に対してGTExデータセット内に複数のサンプルがある可能性があることに留意されたい。以前のようにc_{js}はサンプルs内の接合jに対するカウントである。予測された損失について、われわれは、以下のように影響を受けると推測されるエクソンをスキップする接合kに対する類似のzスコアも計算した。

【０５４６】

【数４７】

$$z_{kt} = \frac{mean_{s' \in U_t}(c_{ks'}) - mean_{s \in A_t}(c_{ks})}{std_{s' \in U_t}(c_{ks'})} \quad (3)$$

【０５４７】

結果としてスキッピングを引き起こした損失は、損失接合の相対的減少およびスキッピングの相対的増加を引き起こすであろうことに留意されたい。このことは、分子z_{jt}およ

びz_{kt}における差の反転を正当化し、したがってこれらのスコアは両方とも、実際のスプライス部位損失に対してマイナスになる傾向がある。

【０５４８】

最後に、われわれは、すべての考察されている組織にわたって中央値zスコアを計算した。損失について、われわれは、式(2)および(3)から別々にzスコアの各々の中央値を計算した。アクセプターまたはドナー損失予測は、次のうちのどれかが真である場合にバリデーションされたと考えられた。

1. 接合の相対的損失を定量化する、式(2)からのzスコアの中央値は置換データにおける対応する値の第5百分位(-1.46)未満であり、スキッピングにおける相対的変化を定量化する、式(3)からのzスコアの中央値は非正(ゼロ、負、または欠損であり、これはスキッピング接合が任意の個体において観察されなかった場合である)であった。言い換えると、影響を受ける接合の使用度の低減に対する強い証拠があり、影響を受ける個体におけるスキッピングの減少を示唆する証拠はなかった。
2. 式(3)からのzスコアの中央値は置換データ内の対応する値の第5百分位(-0.74)未満であり、式(3)からのzスコアの中央値は非正であった。
3. 式(2)からのzスコアの中央値は置換データ内の対応する値の第1百分位(-2.54)未満であった。
4. 式(3)からのzスコアの中央値は置換データ内の対応する値の第1百分位(-4.08)未満であった。
5. 影響を受けるエクソンをスキップする接合はバリアントを有する個体の少なくとも半分で観察され、他の個体では観察されなかった(上記の「プライベートスプライス接合に基づく予測された潜在的スプライス突然変異のバリデーション」において説明されているように)。

【０５４９】

上記のカットオフを得るために使用される置換の説明は、「誤バリデーション率の推定」の節で述べられている。

【０５５０】

経験的に、われわれは、「プライベートスプライス接合に基づく予測された潜在的スプライス突然変異のバリデーション」の節で説明されているように、損失は結果として利得に比べて多くの混合効果をもたらす傾向があるので、利得に比べて損失に対するより厳格なバリデーション基準を適用する必要があることを観察した。プライベートSNVの近くの新規接合を観察することは偶然に行われる可能性がほとんどなく、したがって接合のわずかな証拠であってもバリデーションに十分であろう。対照的に、大半の予測された損失の結果、既存の接合が弱まり、そのような弱まりは、利得によって引き起こされるオンオフ変化以上に検出が難しく、RNA-seqデータ内のノイズに帰因する可能性が高い。

【０５５１】

バリデーション解析のための包含基準

カウントが低いか、またはカバレッジが劣っている場合のzスコアの計算を回避するために、われわれは、バリデーション解析にバリアントをフィルタ処理するために次の基準を使用した。

1. サンプルは遺伝子を表現した場合のみ上記のzスコアの計算について考慮された(式(1)において$\Sigma_g r_{gs}>200$)。
2. 組織は、バリアントのない個体におけるそれぞれ損失または「参照」接合の平均カウントが10未満であった場合に損失または利得zスコアの計算について考慮されなかった。「参照」接合は、GENCODEアノテーションに基づき、新規接合の利得の前に使用されるカノニカル接合である(詳細についてはエフェクトサイズ計算の節を参照)。直感は、対照個体において表現されない接合に影響を及ぼすスプライス損失バリアントをバリデーションすることを試みるべきでないということである。同様に、われわれは、対照個体が影響を受ける部位に及ぶ転写産物を十分に表現しなかった場合にスプライス利得バリアントをバリデーションすることを試みるべきでない。

3. 予測されたスプライス部位損失の場合に、バリアントを有しない個体からのサンプルは、損失接合の少なくとも10カウントを有していた場合にのみ考慮された。予測されたアクセプターまたはドナー位得の場合に、対照個体からのサンプルは、「参照」接合の少なくとも10カウントを有していた場合にのみ考慮された。直感は、影響を受ける接合の大きな平均表現を有する組織(すなわち、基準2を満たす)であっても、異なるサンプルは、大きく異なるシーケンシング深度を有することが可能であり、したがって十分な表現を有する対照サンプルのみが含まれるべきである。

4. 組織は、バリアントを有する個体からの上記の基準を満たす少なくとも1つのサンプル、さらには少なくとも2つの明確に異なる対照個体からの上記の基準を満たす少なくとも5つのサンプルがあった場合のみ考慮された。

【０５５２】

考慮するため基準を満たす組織がなかったバリアントは、非確認可能とみなされ、バリデーション率を計算するときに除外された。スプライス利得バリアントについて、われわれは、すでに存在しているGENCODEアノテーションされたスプライス部位のところに出現するものをフィルタ処理した。同様に、スプライス損失バリアントについて、われわれは、既存のGENCODEアノテーションされたスプライス部位のスコアを減らすもののみを考察した。全体として、高スコア(Δスコア≧0.5)の予測された利得および損失の55%および44%はそれぞれ確認可能と考えられ、バリデーション解析に使用された。

【０５５３】

誤バリデーション率の推定

上記の手順が妥当な真のバリデーション率を有していることを確認するために、われわれは、最初に、1～4GTEx個体に出現するSNVを見て、本質的なGT/AGジヌクレオチドを切断した。われわれは、バリデーション率が100%に近くなるようにそのような突然変異がほとんど確実にスプライシングに影響を及ぼすと論じた。そのような切断のうち、39%は上で説明されている基準に基づき確認可能であり、確認可能なもののうちで、バリデーション率は81%であった。誤バリデーション率を推定するために、われわれは、SNVデータの個体の標識を置換した。k人のGTEx個体に出現した各SNVについて、われわれは、k人のGTEx個体のランダムサブセットを選び、SNVをそれらに割り当てた。われわれは、10個のそのようなランダム化されたデータセットを作成し、それらの上でバリデーションプロセスを繰り返した。置換データセットにおけるバリデーション率は、利得については1.7～2.1%であり、損失については4.3～6.9%であり、中央値はそれぞれ1.8%および5.7%であった。損失に対する誤バリデーション率が高いこと、また本質的な切断のバリデーション率が比較的低いことは、「プライベートスプライス接合に基づく予測された潜在的スプライス突然変異のバリデーション」の節で強調されているようにスプライス部位損失をバリデーションすることが困難であることによる。

【０５５４】

RNA-seqデータにおける潜在的スプライスバリアントのエフェクトサイズの計算

われわれは、バリアントの「エフェクトサイズ」をバリアントによりスプライシングパターンを変化させた影響を受ける遺伝子の転写産物の割合として定義した(たとえば、新規アクセプターまたはドナーに切り替えた割合)。予測されたスプライス利得バリアントに対する参照例として、図38Cのバリアントを考察する。予測された利得ドナーAについて、われわれは、最初に、最も近いアノテーションされたアクセプターCへの接合(AC)を識別した。われわれは、次いで、「参照」接合(BC)を識別し、B≠AはAに最も近いアノテーションされたドナーある。次いで、各サンプルsにおいて、われわれは、参照接合(BC)と比較して新規接合(AC)の相対的使用度を以下のように計算した。

【０５５５】

【数４８】

$$u_{(AB)s} = \frac{r_{(AC)s}}{r_{(AC)s} + r_{(BC)s}} \quad (4)$$

【０５５６】

ここで、$r_{(AC)s}$はサンプルs内の接合(AC)の未処理リードカウントである。各組織について、われわれは、バリアントを有する個体と他のすべての個体との間の接合(AC)の使用度の変化を以下のように計算した。

【０５５７】

【数４９】

$$mean_{s \in A_t} u_{(AC)s} - mean_{s' \in U_t} u_{(AC)s'} \quad (5)$$

【０５５８】

ここで、A_tは組織t内のバリアントを有する個体からのサンプルセットであり、U_tは組織t内の他の個体からのサンプルセットである。最終エフェクトサイズは、考察されたすべての組織にわたって上記の差の中央値として計算された。計算は、利得アクセプターの場合またはスプライス部位形成バリアントがイントロンであった場合に類似していた。エフェクトサイズ計算の簡略化されたバージョン(バリアントを有する個体およびバリアントを有しない個体からの単一のサンプルを仮定して)は図38Cに示されている。

【０５５９】

予測された損失について、われわれは、最初に、影響を受けるエクソンをスキップした転写産物の割合を計算した。計算は、図45に示されている。ドナーCの予測された損失について、われわれは、次の下流のアノテーションされたエクソンへの接合(CE)、さらには影響を受けると推測されたものへの上流もエクソンからの接合(AB)を識別した。われわれは、影響を受けるエクソンをスキップした転写産物の割合を以下のように定量化した。

【０５６０】

【数５０】

$$k_{(AE)s} = \frac{r_{(AE)s}}{r_{(AE)s} + mean(r_{(AB)s} + r_{(CE)s})} \quad (6)$$

【０５６１】

次いで、利得に関して、われわれは、バリアントを有する個体からのサンプルとバリアントを有しない個体からのサンプルとの間のスキップされた割合の変化を以下のように計算した。

【０５６２】

【数５１】

$$mean_{s\in A_t}k_{(AE)s} - mean_{s'\in U_t}k_{(AE)s'} \quad (7)$$

【０５６３】

上で計算されたようなスキップされた転写産物の割合はアクセプターまたはドナー損失の効果を完全には捕捉せず、たとえば、切断は次善スプライス部位のイントロン保持または使用度のレベルを高める可能性がある。これらの効果のうちのいくつかを説明するために、われわれは、また、同じアクセプターEを持つ他の接合の使用度に関して損失接合(CE)の使用度を以下のように計算した。

【０５６４】

【数５２】

$$l_{(CE)s} = \frac{r_{(CE)s}}{\sum r_{(\cdot E)s}} \quad (8)$$

【０５６５】

ここで、$\Sigma r_{(\cdot E)s}$は、任意の(アノテーションされたまたは新規の)アクセプターからドナーEへのすべての接合の和である。これは、図45の例で示されているように、影響を受ける接合(CE)、スキッピング接合(AE)、さらにはCの損失を補償した他の次善ドナーからの潜在的接合を含む。次いで、われわれは、影響を受ける接合の相対的使用度の変化を以下のように計算した。

【０５６６】

【数５３】

$$mean_{s'\in U_t}l_{(CE)s'} - \ mean_{s\in A_t}l_{(CE)s} \quad (9)$$

【０５６７】

バリアントを有する個体の利得またはスキッピング接合の使用度の増加を測定する、(5)および(7)とは異なり、(9)において、われわれは、損失接合の使用度の減少、したがって、差の2つの部分の反転を測定することを望んでいることに留意されたい。各組織について、エフェクトサイズは(7)および(9)の最大値として計算された。利得に関しては、バリアントに対する最終的なエフェクトサイズは組織にわたってエフェクトサイズの中央値であった。

【０５６８】

エフェクトサイズ解析のための包含基準

バリアントは、前の節で説明されている基準に基づきバリデーションされたとみなされた場合のみエフェクトサイズ計算について考察された。非常に小さな数での異常転写産物の割合を計算することを回避するために、われわれは、異常および参照接合のカウントが両方とも少なくとも10であるサンプルのみを考察した。大半の潜在的スプライスバリアントはイントロン内にあるので、エフェクトサイズは、バリアントと重なり合う参照および代替リードの数をカウントすることでは直接計算することはできない。したがって、損失のエフェクトサイズは、正常なスプライス接合の相対的使用度の減少から間接的に計算される。新規接合利得のエフェクトサイズについては、異常転写産物はナンセンス変異依存分解の影響を受け、観察されたエフェクトサイズを減少させ得る。これらの測定の制限が

あるにもかかわらず、われわれは、両方の利得および損失事象にわたって低いスコアリングの潜在的スプライスバリアントに対してより小さいエフェクトサイズに向かう一貫した傾向を観察している。

【０５６９】

完全浸透性を有するヘテロ接合プライベートSNVの予想されるエフェクトサイズ

バリアントを有する個体のバリアントハプロタイプからのすべての転写産物が新規接合に切り替わることを引き起こす完全浸透性を有するスプライス部位形成バリアントについて、新規接合が対照個体に出現しないと仮定すると、予想されるエフェクトサイズは式(5)により0.5となる。

【０５７０】

同様に、ヘテロ接合SNVが新規エクソンスキッピング事象を引き起こし、影響を受けるハプロタイプのすべての転写産物がスキッピング接合に切り替わった場合、式(7)における予想されるエフェクトサイズは0.5である。バリアントを有する個体からのすべての転写産物が異なる接合(スキッピング接合、または別の補償する接合のいずれか)に切り替わった場合、式(8)の中の比は、バリアントを有する個体からのサンプルでは0.5、他の個体からのサンプルでは1となり、したがって式(9)の中の差は0.5となる。これは、バリアントを有しない個体においてアクセプターEへのスキッピングまたは他の接合がなかったと仮定している。これは、また、スプライス部位切断がイントロン保持をトリガしないことも仮定する。実際、イントロン保持の少なくとも低いレベルは、多くの場合に、スプライス部位切断に関連付けられている。さらに、エクソンスキッピングは広範であり、スプライス変更バリアントが存在しない場合であってもそうである。これは、測定されたエフェクトサイズが、バリアントが本質的なGT/AGジヌクレオチドを切断する場合であっても0.5未満である理由を説明する。

【０５７１】

完全浸透性を有するヘテロ接合バリアントに対する0.5のエフェクトサイズの予想でも、バリアントはナンセンス変異依存分解(NMD)をトリガしなかったと仮定する。NMDの存在下で、式(4)、(6)、および(8)の分子および分母は両方とも小さくなり、したがって観察されたエフェクトサイズを減少させる。

【０５７２】

ナンセンス変異依存分解(NMD)を通じて分解される転写産物の割合

図38Cについて、バリアントはエクソンなので、われわれは、バリアントに及ぶリードの数をカウントすることができ、参照または代替対立遺伝子を有していた(それぞれ「Ref(スプライシングなし)」および「Alt(スプライシングなし)」)。われわれは、また、新規スプライス部位のところでスプライスし、代替対立遺伝子(「Alt(新規接合)」)を持っていると推測された、リードの数をカウントした。図38Cの例では、またわれわれが見ていた他のケースの多くにおいて、われわれは、代替対立遺伝子(「Alt(スプライシングなし)」および「Alt(新規接合)」の和)を有するハプロタイプに由来するリードの総数が参照対立遺伝子(「Ref(スプライシングなし)」)を有するリードの数より少ないことを観察した。われわれは、参照および代替ハプロタイプの両方にマッピングし、リードの数が各対立遺伝子を有する転写産物の数に比例することを仮定することによって、われわれがリードマッピング中に参照バイアスを排除したと確信しているので、われわれは、参照対立遺伝子がバリアント軌跡におけるリードの半分を占めることを予想していた。われわれは、「欠損」代替対立遺伝子リードが新規接合のところでスプライスし、ナンセンス変異依存分解(NMD)を通じて分解した代替対立遺伝子ハプロタイプからの転写産物に対応すると仮定している。われわれは、このグループを「Alt(NMD)」と呼んだ。

【０５７３】

参照および代替リードの観察された数の差が有意かどうかを決定するために、われわれは、成功確率0.5で2項分布に従うAlt(スプライシングなし)+Alt(新規接合)(またはそれより少ない)リードおよびAlt(スプライシングなし)+Alt(新規接合)+Ref(スプライシングなし)の試行の総数を観察する確率を計算した。これは、われわれは潜在的に分解する転写

産物をカウントしないことにより「試行」の総数を過小評価しているので控えめなp値である。図38CにおけるNMD転写産物の割合は、新規接合(Alt(NMD)+Alt(新規接合))でスプライスするリードの総数にわたって「Alt(NMD)」リードの数として計算された。

【０５７４】

潜在的スプライス接合を検出する際のネットワークの感度

SpliceNetモデル(図38F)の感度を評価するために、われわれは、影響を受けるスプライス部位(すなわち、新規または切断アクセプターもしくはドナー)から20nt以内にあるがアノテーションされたエクソンの本質的なGT/AGジヌクレオチドと重なり合わず、少なくとも0.3の推定されたエフェクトサイズ(「エフェクトサイズの計算」の節を参照)を有しているSNVを使用した。すべての感度プロットにおいて、SNVは、アノテーションされたエクソンと重なり合うか、またはアノテーションされたエクソンの境界から50nt以内にある場合に「近エクソン」であると定義された。他のすべてのSNVは、「深イントロン」と考えられた。強く支持されている潜在的スプライス部位のこの真のデータセットを使用することで、われわれは、変化するΔスコア閾値でわれわれのモデルを評価し、そのカットオフにおいてモデルによって予測される真のデータセット内の潜在的スプライス部位の割合を報告した。

【０５７５】

既存のスプライシング予測モデルとの比較

われわれは、様々な測定基準に関してSpliceNet-10k、MaxEntScan(YeoおよびBurge、2004年)、GeneSplicer(Perteaら、2001年)、およびNNSplice(Reeseら、1997年)の一対一の比較を実行した。われわれは、MaxEntScanおよびGeneSplicerソフトウェアをhttp://genes.mit.edu/burgelab/maxent/download/およびhttp://www.cs.jhu.edu/~genomics/GeneSplicer/からそれぞれダウンロードした。NNSpliceは、ダウンロード可能ソフトウェアとして利用可能でなく、したがって、われわれは、http://www.fruitfly.org/data/seq_tools/datasets/Human/GENIE_96/splicesets/から訓練およびテストセットをダウンロードし、(Reeseら、1997年)で述べられている最良の実行アーキテクチャでモデルを訓練した。サニティチェックとして、われわれは、(Reeseら、1997年)で報告されているテストセット測定基準を再現した。

【０５７６】

これらのアルゴリズムのTop-k精度および精度-再現率曲線の下面積を評価するために、われわれは、各アルゴリズムでテストセット遺伝子およびlincRNA内のすべての位置にスコアを付けた(図37D)。

【０５７７】

MaxEntScanおよびGeneSplicer出力は対数オッズ比に対応するが、NNSpliceおよびSpliceNet-10k出力は確率に対応している。われわれがMaxEntScanおよびGeneSplicerに最も高い成功確率を与えたことを確認するために、われわれは、それらを既定の出力とともに、さらにはわれわれが最初に確率に対応するように出力を変換する場合の変換済み出力とともに使用してΔスコアを計算した。より正確には、MaxEntScanの既定の出力は、以下に対応している。

【０５７８】

【数５４】

$$x = \log_2 \frac{p(\text{スプライス部位})}{p(\text{スプライス部位なし})}$$

【０５７９】

これは、変換

【数５５】

$$\frac{2^x}{2^x+1}$$

の後に、所望の量に対応する。われわれは、GeneSplicerソフトウェアを2回、1回目はRETURN_TRUE_PROBフラグを0に設定して、もう1回はそれを1に設定することによってコンパイルした。われわれは、RNA-seqデータ(MaxEntScan:変換済み出力、GeneSplicer:既定の出力)に対して最良のバリデーション率を出す出力戦略を選んだ。

【０５８０】

様々なアルゴリズムのバリデーション率および感度を比較するために(図38G)、われわれは、すべてのアルゴリズムがゲノムワイドの同じ数の利得および損失を予測したカットオフを見つけ出した。すなわち、SpliceNet-10kのΔスコア値上の各カットオフについて、われわれは、各競合するアルゴリズムがSpiceNet-10kと同じ数の利得予測および同じ数の損失予測を行うカットオフを見つけ出した。選択されたカットオフは、Table S2(表S2)に示されている。

【０５８１】

シングルトン対共通バリアントに対するバリアント予測の比較

われわれは、シングルトンSNVおよび2～4人のGTEx個体に出現するSNVについて別々にバリデーションおよび感度解析(「感度解析」および「モデル予測のバリデーション」の説に説明したように)を実行した(図46A、図46B、および図46C)。バリデーション率がシングルトンと共通バリアントとの間で著しく異なっているかどうかをテストするために、われわれは、各Δスコアグループ(0.2～0.35, 0.35～0.5, 0.5～0.8, 0.8～1)内のバリデーション率を比較する、また各予測された効果(アクセプターまたはドナー利得または損失)について、フィッシャーの正確確率検定を実行した。ボンフェローニ補正を行って16回の検定を検討した後、すべてのP値は0.05より大きかった。われわれは、同様に、シングルトンまたは共通バリアントを検出するための感度を比較した。われわれは、フィッシャーの正確確率検定を使用して、バリデーション率がバリアントの2つのグループの間で著しく異なっているかどうかを検定した。われわれは、深イントロンバリアントおよびエクソンに近いバリアントを別々に考察し、ボンフェローニ補正を2つの検定に対して実行した。0.05カットオフを使用したときにP値のどれも有意でなかった。したがって、われわれは、シングルトンおよび共通GTExバリアントを組み合わせて、図48A、図48B、図48C、図48D、図48E、図48F、および図48Gならびに図39A、図39B、および図39Cに提示されている解析に関して一緒に考察した。

【０５８２】

訓練対テスト染色体上のバリアント予測の比較

われわれは、訓練時に使用される染色体上のバリアントおよび染色体の残りの上のバリアントの間でSpliceNet-10kのRNA-seqおよび感度についてバリデーション率を比較した(図48Aおよび図48B)。すべてのP値は、ボンフェローニ補正の後に0.05より大きかった。われわれは、また、以下の「悪影響を有するバリアントの割合」で説明されているように、訓練およびテスト染色体上でバリアントに対して別々に悪影響もあるバリアントの割合を計算した(図48C)。各Δスコアグループおよびバリアントの各種類について、われわれは、フィッシャーの正確確率検定を使用して、訓練染色体とテスト染色体との間で共通バリアントと稀少バリアントの数を比較した。12回の検定についてボンフェローニ補正を行った後、すべてのP値は0.05より大きかった。最後に、われわれは、「コホート毎のデノボ突然変異のエンリッチメント」の節で説明されているような訓練およびテスト染色体(図48D)上で潜在的スプライスデノボバリアントの数を計算した。

【０５８３】

異なる種類の潜在的スプライスバリアントにわたるバリアント予測の比較

われわれは、予測されたスプライス部位形成バリアントを3つのグループ、すなわち、新規GTまたはAGスプライスジヌクレオチドを形成するバリアント、スプライシングモチーフの残りと重なり合うバリアント(エクソン内への最大3ntおよびイントロン内への8ntのエクソン-イントロン境界の周りの位置)、およびスプライスモチーフの外側のバリアントに分割した(図47Aおよび図47B)。各ΔスコアグループについてバリアントタイプのP値を比較した。各Δスコアグループ(0.2～0.35, 0.35～0.5, 0.5～0.8, 0.8～1)について、われわれは、χ^2検定を実行して、バリデーション率が3種類のスプライス部位形成バリアントにわたって均一であるという仮説を検定した。すべての検定は、複数の仮説補正の前であってもP値>0.3が得られた。3種類のバリアントの間のエフェクトサイズ分布を比較するために、われわれは、マン・ホイットニーのU検定を使用し、各Δスコアグループについてバリアントタイプの3つの対すべて(合計4×3=12回の検定)を比較した。12回の検定についてボンフェローニ補正を行った後、すべてのP値は>0.3であった。

【０５８４】

組織特有のスプライス利得バリアントの検出

図39Cについて、われわれは、新規接合の使用率が影響を受ける遺伝子を表現する組織にわたって均一であるかどうかを検定することを望んでいた。われわれは、新規プライベートスプライス部位を形成したSNV、すなわち、バリアントを有する個体の少なくとも半分にのみ出現し、他の個体には出現しない利得スプライス接合を結果としてもたらすSNVに集中した。各そのような新規接合jについて、われわれは、各組織tにおいて、組織内にバリアントを有する個体からのすべてのサンプルにわたって接合の総カウント

【数５６】

$$\sum_{s\in A_t} r_{js}$$

を計算した。ここで、A_tは組織t内のバリアントを有する個体からのサンプルセットである。同様に、われわれは、同じサンプルに対する遺伝子のすべてのアノテーションされた接合の総カウント

【数５７】

$$\sum_{s\in A_t}\sum_{g} r_{gs}$$

を計算したが、ここで、gは遺伝子のアノテーションされた接合のインデックスである。次いで、遺伝子の背景カウントに対して正規化された、組織t内の新規接合の相対的使用度は、以下として測定され得る。

【０５８５】

【数５８】

$$m_t = \frac{\sum_{s\in A_t} r_{js}}{\sum_{s\in A_t}(r_{js} + \sum_g r_{gs})}$$

【０５８６】

われわれは、また、組織にわたって接合の平均使用度を以下のように計算した。

【０５８７】

【数５９】

$$m = \frac{\sum_t \sum_{s \in A_t} r_{js}}{\sum_t \sum_{s \in A_t} (r_{js} + \sum_g r_{gs})}$$

【０５８８】

われわれは、接合の相対的使用度が組織にわたって均一であり、mに等しいという仮説を検定することを望んでいた。われわれは、したがって、観察された組織カウント

【数６０】

$$\sum_{s \in A_t} r_{js}$$

を均一な率

【数６１】

$$m \sum_{s \in A_t} (r_{js} + \sum_g r_{gs})$$

の仮定の下で予想されたカウントと比較するχ^{2}検定を実行した。スプライス部位形成バリアントは、ボンフェローニ補正されたχ^{2}p値が10^{-2}より小さかった場合に組織特有であると考察された。検定に対する自由度はT-1であり、Tは考察されている組織の数である。バリデーションの節で説明されている考察基準を満たした組織のみが検定において使用された。さらに、カウントが小さい場合のケースを回避するために、均一性検定が力不足である場合、われわれは、少なくとも3つの考察されている組織、平均して組織毎に少なくとも1つの異常リード(すなわち、m>1)、およびすべての考察されている組織にわたって合計した少なくとも15個の異常リード(すなわち、

【数６２】

$$\sum_t \sum_{s \in A_t} r_{js} > 15$$

で均一性バリアントに対してのみ検定した。われわれは、バリアントのこのクラスが一般的に低いエフェクトサイズおよび低い接合カウントを有するので、ΔスコアがΔ0.35未満であるすべてのバリアントを無視した。われわれは、組織特有のバリアントの割合がこのクラスに対して非常に低いことを観察したが、われわれはこれがパワー問題によるものであったと確信している。

【０５８９】

III. ExACおよびgnomADデータセット上の解析

バリアントフィルタ処理

われわれは、Sites VCF release 0.3ファイル(60,706個のエクソーム)をExACブラウザ(Lekら、2016年)から、Sites VCF release 2.0.1ファイル(15,496個の全ゲノム)をgnomADブラウザからダウンロードした。われわれは、SpliceNet-10kを評価するためにそれらからバリアントのフィルタ処理済みリストを作成した。特に、次の基準を満たしたバリアントが考察された。

・ FILTERフィールドはPASSであった。

・ バリアントは単一ヌクレオチドバリアントであり、ただ1つの代替ヌクレオチドがあった。

・ ANフィールド(コールされた遺伝子型における対立遺伝子の総数)は少なくとも10,000の値を有していた。

・ バリアントは、カノニカルGENCODE転写産物の転写開始部位と終了部位との間にあった。

【０５９０】

合計7,615,051および73,099,995個のバリアントが、ExACおよびgnomADデータセットにおいてそれぞれこれらのフィルタを通った。

【０５９１】

悪影響を有するバリアントの割合

この解析について、われわれは、コホートにおいてシングルトンまたは共通であったExACおよびgnomADフィルタ処理済みリスト内のバリアントのみを考察した(対立遺伝子頻度(AF)≧0.1%)。われわれは、GENCODEカノニカルアノテーションに従ってそのゲノム位置に基づきこれらのバリアントを下位のクラスに分類した。

・ エクソン:このグループは、同義ExACバリアント(676,594個のシングルトンおよび66,524個の共通)からなる。ミスセンスバリアントは、このグループ内のバリアントの悪影響の大部分はスプライシング変化によるものであったことを確認するためにここでは考察されていない。

・ 近イントロン:このグループは、カノニカルエクソン境界から3～50ntの範囲内にあるイントロンExACバリアントからなる。より正確には、アクセプター利得/損失およびドナー利得/損失バリアントの解析のために、スプライスアクセプターおよびドナーからそれぞれ3～50ntのところにあるバリアントのみが考察された(アクセプター利得/損失に対しては575,636個のシングルトンおよび48,362個の共通、ドナー利得/損失に対しては567,774個のシングルトンおよび50,614個の共通)。

・ 深イントロン:このグループは、カノニカルエクソン境界から50ntを超えて離れているイントロンgnomADバリアントからなる(34,150,431個のシングルトンおよび8,215,361個の共通)。

【０５９２】

各バリアントについて、われわれは、SpliceNet-10kを使用して4つのスプライスタイプのΔスコアを計算した。次いで、各スプライスタイプについて、われわれは、2つの行が予測されたスプライス変更バリアント(スプライスタイプに対して適切な範囲にあるΔスコア)対予測されたスプライス変更でないバリアント(すべてのスプライスタイプに対してΔスコア<0.1)に対応しており、2つの列がシングルトン対共通バリアントに対応していた場合の$2\times 2\ \chi^2$分割表を構築した。スプライス利得バリアントについて、われわれは、すでに存在しているGENCODEアノテーションされたスプライス部位のところに出現するものをフィルタ処理した。同様に、スプライス損失バリアントについて、われわれは、既存のGENCODEアノテーションされたスプライス部位のスコアを減らすもののみを考察した。オッズ比が計算され、悪影響を有するバリアントの割合は、以下と推定された。

【０５９３】

【数６３】

$$\left(1-\frac{1}{\text{オッズ比}}\right)\times 100\%$$

【０５９４】

ExACおよびgnomADフィルタ処理済みリスト内のタンパク質切り詰めバリアントは、次のように識別された。

・ ナンセンス:VEP(McLarenら、2016年)の結果は「stop_gained」(ExACにおいて44,04

6個のシングルトンおよび722個の共通、gnomADにおいて20,660個のシングルトンおよび970個の共通)であった。

・ フレームシフト:VEPの結果は「frameshift_variant」であった。バリアントフィルタ処理時の単一ヌクレオチドバリアント基準は、このグループを形成するために緩和された(ExACにおいて48,265個のシングルトンおよび896個の共通、gnomADにおいて30,342個のシングルトンおよび1,472個の共通)。

・ 本質的なアクセプター/ドナー損失:バリアントは、カノニカルイントロンの最初または最後の2つの位置にあった(ExACにおいて29,240個のシングルトンおよび481個の共通、gnomADにおいて12,387個のシングルトンおよび746個の共通)。

【０５９５】

タンパク質切り詰めバリアントに対する$2\times2\ \chi^2$分割表は、ExACおよびgnomADフィルタ処理済みリストに対して構築され、悪影響を有するバリアントの割合を推定するために使用された。ここで、2つの行はタンパク質切り詰めバリアント対同義バリアントに対応し、2つの列は前のようにシングルトン対共通バリアントに対応した。

【０５９６】

ExAC(エクソンおよび近イントロン)およびgnomAD(深イントロン)バリアントに対する結果は、それぞれ図40Bおよび図40Dに示されている。

【０５９７】

フレームシフト対インフレームスプライス利得

この解析について、われわれは、エクソン(同義のみ)または近イントロンであった、またコホートにおいてシングルトンまたは共通(AF≧0.1%)であったExACバリアントに特に注目した。アクセプター利得バリアントをインフレームまたはフレームシフトとして分類するために、われわれは、カノニカルスプライスアクセプターと新しく作成されたスプライスアクセプターとの間の距離を測定し、それが3の倍数かどうかをチェックした。われわれは、カノニカルスプライスドナーと新しく作成されたスプライスドナーとの間の距離を測定することによって同様にドナー利得バリアントを分類した。

【０５９８】

悪影響のあるインフレームスプライス利得バリアントの割合は、2つの行が予測されたインフレームスプライス利得バリアント(アクセプターまたはドナー利得に対してΔスコア≧0.8)対予測されたスプライス変更でないバリアント(すべてのスプライスタイプに対してΔスコア<0.1)に対応しており、2つの列がシングルトン対共通バリアントに対応していた場合の$2\times2\ \chi^2$分割表から推定された。この手順は、分割表内の第1の行を予測されたフレームシフトスプライス利得バリアントで置き換えることによってフレームシフトスプライス利得バリアントについて繰り返された。

【０５９９】

図40Cに示されているp値を計算するために、われわれは、予測されたスプライス利得バリアントのみを使用して$2\times2\ \chi^2$分割表を構築した。ここで、2つの行はインフレーム対フレームシフトスプライス利得バリアントに対応し、2つの列は前のようにシングルトン対共通バリアントに対応した。

【０６００】

個体毎の潜在的スプライスバリアントの数

個体毎の稀少機能的潜在的スプライスバリアントの数を推定するために(図40E)、われわれは、最初に、対立遺伝子頻度に等しい確率で各対立遺伝子内に各gnomADバリアントを含めることによって100人のgnomAD個体をシミュレートした。言い換えると、各バリアントは、2倍性を模倣するために各個体に対して独立して2回サンプリングされた。われわれは、Δスコアがそれぞれ0.2、0.2、および0.5以上であった1人当たり稀少(AF<0.1%)エクソン(同義のみ)、近イントロン、および深イントロンバリアントをカウントした。これらは、予測されたバリアントの少なくとも40%が悪影響を有することを確認しながら感度を最適化する比較的許容性の高いΔスコア閾値である。これらのカットオフにおいて、われわれは、1人当たり7.92の同義/近イントロンおよび3.03の深イントロン稀少潜在的スプラ

イスバリアントの平均を取得した。これらのバリアントはすべてが機能的であるわけではないので、われわれは、これらのカウントにこれらのカットオフにおいて悪影響のあるバリアントの割合を乗算した。

【０６０１】

IV. DDDおよびASDデータセット上の解析

潜在的スプライシングデノボ突然変異

われわれは、公開されているデノボ突然変異(DNM)を取得した。これらは、自閉症スペクトラム障害を患っている3953人の遺伝発端者(Dongら、2014年、Iossifovら、2014年、De Rubeisら、2014年)、Deciphering Developmental Disordersコホートからの4293人の遺伝発端者(McRaeら、2017年)、および2073人の健康な対照(Iossifovら、2014年)を含んでいた。低品質DNMは、解析から除外された(ASDおよび健康な対照:信頼度==lowConf、DDD:PP(DNM)<0.00781、(McRaeら、2017年))。DNMはネットワークにより評価され、われわれはΔスコア(上記の方法を参照)を使用して構成に応じて潜在的スプライス突然変異を分類した。われわれは、synonymous_variant、splice_region_variant、intron_variant、5_prime_UTR_variant、3_prime_UTR_variant、またはmissense_variantのVEP結果によりアノテーションされた突然変異のみを考察した。われわれは、図41A、図41B、図41C、図41D、図41E、および図41Fならびに図50Aおよび図50Bに対するΔスコア>0.1を有する部位と、図49A、図49B、および図49Cに対するΔスコア>0.2を有する部位とを使用した。

【０６０２】

図20、図21、図22、図23、および図24は、SpliceNet-80nt、SpliceNet-400nt、SpliceNet-2k、およびSpliceNet-10kアーキテクチャの詳細な説明を示している。これら4つのアーキテクチャは、入力として注目する位置の各側でそれぞれ長さ40、200、1,000、および5,000の隣接ヌクレオチド配列を使用し、位置がスプライスアクセプターである、スプライスドナーである、およびいずれでもない確率を出力する。これらのアーキテクチャは、もっぱら、畳み込み層Conv(N、W、D)からなり、N、W、およびDは、それぞれ、層内の畳み込みカーネルの数、ウィンドウサイズ、および各畳み込みカーネルの拡張率である。

【０６０３】

図42Aおよび図42Bは、lincRNAに対する様々なスプライシング予測アルゴリズムの評価を示す。図42Aは、lincRNA上で評価したときに様々なスプライシング予測アルゴリズムのTop-k精度および精度-再現率曲線の下の面積を示している。図42Bは、MaxEntScanおよびSpliceNet-10kを使用してスコアを付けられたLINC00467 遺伝子に対する完全なpre-mRNA転写産物を、予測されたアクセプター(赤色矢印)およびドナー(緑色矢印)部位ならびにエクソンの実際の位置とともに、図示している。

【０６０４】

図43Aおよび図43Bは、TACTAAC分岐点およびGAAGAAエクソン内スプライスエンハンサーモチーフの位置依存効果を示す。図43Aに関して、最適な分岐点配列TACTAACは、14,289個のテストセットスプライスアクセプターの各々からの様々な距離で導入され、アクセプタースコアはSpliceNet-10kを使用して計算された。予測されたアクセプタースコアの平均変化は、スプライスアクセプターからの距離の関数としてプロットされる。予測されたスコアは、スプライスアクセプターからの距離が20から45ntの範囲内にあるときに増加し、20nt未満の距離において、TACTAACはポリピリミジントラクトを切断し、これにより予測されたアクセプタースコアは非常に低い。

【０６０５】

図43Bに関して、SRタンパク質六量体モチーフGAAGAAは、14,289個のテストセットスプライスアクセプターおよびドナーの各々からの様々な距離で同様に導入された。予測されたSpliceNet-10kアクセプターおよびドナースコアの平均変化は、それぞれスプライスアクセプターおよびドナーからの距離の関数としてプロットされる。予測されたスコアは、モチーフがエクソン側にありスプライス部位から～50nt未満であるときに増加する。エクソン内へより大きい距離において、GAAGAAモチーフは考察対象のスプライスアクセプターまたはドナー部位の使用を有利に扱わない傾向があるが、それは、より近位にあるアクセ

プターまたはドナーモチーフを優先的に支持することになるからである。非常に低いアクセプターおよびドナースコアは、GAAGAAがイントロンに非常に近い位置に置かれたときに、伸長されたアクセプターまたはドナースプライスモチーフの切断によるものである。

【０６０６】

図44Aおよび図44Bは、スプライシングに対するヌクレオソームの影響を示す。図44Aに関して、百万個の無作為に選択された遺伝子間位置において、150nt相隔てて並ぶ強いアクセプターおよびドナーチーフは導入され、エクソン包含の確率は、SpliceNet-10kを使用して計算された。SpliceNet-10k予測とヌクレオソーム位置決めとの間の相関がGC組成とは無関係に生じることを示すために、位置は、GC内容(導入されたスプライス部位の間の150個のヌクレオチドを使用して計算される)に基づきビンに入れられ、SpliceNet-10k予測とヌクレオソーム信号との間のスピアマン相関は各ビンについてプロットされる。

【０６０７】

図44Bに関して、テストセットからのスプライスアクセプターおよびドナー部位は、SpliceNet-80nt(ローカルモチーフスコアと呼ばれる)とSpliceNet-10kとを使用してスコアを付けられており、スコアはヌクレオソームエンリッチメントの関数としてプロットされる。ヌクレオソームエンリッチメントは、スプライス部位のイントロン側で50ntにわたって平均されたヌクレオソーム信号によって除算されたスプライス部位のエクソン側で50ntにわたって平均されたヌクレオソーム信号として計算される。SpliceNet-80ntスコアはモチーフ強度の代用であり、ヌクレオソームエンリッチメントと負の相関を有するが、SpliceNet-10kスコアはヌクレオソームエンリッチメントと正の相関を有する。これは、ヌクレオソーム位置決めが弱いローカルスプライスモチーフを補償することができる長距離特異性決定因子であることを示唆している。

【０６０８】

図45は、複合効果を有するスプライス破断バリアントについてのエフェクトサイズを計算する例を示す。イントロンバリアントchr9:386429 A>Gは正常なドナー部位(C)を切断し、すでに抑制されているイントロン下流ドナー(D)を活性化する。図示されているのは、RNA-seqカバレッジならびにバリアントを有する個体および対照個体からの全血における接合リードカウントである。バリアントを有する個体および対照個体におけるドナー部位は、それぞれ、青色矢印および灰色矢印でマークされる。赤色太文字は接合エンドポイントに対応する。見やすくするために、エクソン長はイントロン長に比べて4倍誇張されている。エフェクトサイズを推定するために、われわれは、エクソンスキッピング接合(AE)の使用の増大および同じドナーEを含む他のすべての接合に関して切断された接合(CE)の使用度の減少の両方を計算する。最終的なエフェクトサイズは、2つの値の最大値(0.39)である。増加させた量のイントロン保持も、突然変異サンプル内に存在している。これらの可変効果は、エクソンスキッピング事象において共通であり、アクセプターまたはドナー部位損失を引き起こすと予測されている稀少バリアントをバリデーションする複雑さを増大させる。

【０６０９】

図46A、図46B、および図46Cは、シングルトンおよび共通バリアントに対するSpliceNet-10kモデルの評価を示す。図46Aに関して、GTEx RNA-seqデータに対してバリデーションしたSpliceNet-10kによって予測された潜在的スプライス突然変異の割合が示されている。モデルは、GTExコホート内のせいぜい4人の個体に出現するすべてのバリアント上で評価された。予測されたスプライス変更効果を有するバリアントは、RNA-seqデータと突き合わせてバリデーションされた。このバリデーション率は、単一のGTEx個体(左)に出現するバリアントおよび2から4人GTEx個体(右)に出現するバリアントについて別々に図示されている。予測は、Δスコアでグループ化される。われわれは、各Δスコアグループ内のバリアントの4つのクラス(アクセプターまたはドナーの利得または損失)の各々に対するシングルトンと共通バリアントとの間でバリデーション率を比較した。差は、有意でない($P>0.05$、16回の検定に対するボンフェローニ補正を含むフィッシャーの正確確率検定)。

【０６１０】

図46Bに関して、異なるΔスコアカットオフにおけるGTExコホート内のスプライス変更バリアントを検出するときのSpliceNet-10kの感度が示されている。モデルの感度は、シングルトン(左)および共通(右)バリアントについて別々に図示されている。0.2のΔスコアカットオフにおけるシングルトンと共通バリアントとの間の感度の差は、いずれかのバリアント、近エクソンまたは深イントロンバリアントについて有意でない(P>0.05、2回の検定に対するボンフェローニ補正を含むフィッシャーの正確確率検定)。

【０６１１】

図46Cに関して、バリデーションされたシングルトンおよび共通バリアントに対するΔスコア値の分布が示されている。P値は、マン・ホイットニーのU検定に対する値であり、シングルトンおよび共通バリアントのスコアを比較するものである。共通バリアントは、大きな効果を持つスプライス切断突然変異をフィルタ除去する自然選択により、著しく弱いΔスコア値を有する。

【０６１２】

図47Aおよび図47Bは、バリアントの位置によって分割されたスプライス部位形成バリアントのバリデーション率およびエフェクトサイズを示す。予測されたスプライス部位形成バリアントは、バリアントが新しい本質的なGTまたはAGスプライスジヌクレオチドを生成したかどうか、これがスプライスモチーフの残り部分と重なり合っているかどうか(本質的なジヌクレオチドを除く、エクソン内へ最大3nt、イントロン内へ8ntのエクソン-イントロン境界の周りのすべての位置)、またはそれがスプライスモチーフの外側にあるかどうかに基づきグループ化された。

【０６１３】

図47Aに関して、スプライス部位形成バリアントの3つのカテゴリの各々に対するバリデーション率が示されている。各カテゴリ内のバリアントの総数は、バーの上に示されている。各Δスコアグループ内で、バリアントの3つのグループの間のバリデーション率の差は、有意でない(P>0.3、均一性のχ^2検定)。

【０６１４】

図47Bに関して、スプライス部位形成バリアントの3つのカテゴリの各々に対するエフェクトサイズの分布が示されている。各Δスコアグループ内で、バリアントの3つのグループの間のエフェクトサイズの差は、有意でない(P>0.3、ボンフェローニ補正を含むマン・ホイットニーのU検定)。

【０６１５】

図48A、図48B、図48C、および図48Dは、訓練およびテスト染色体に対するSpliceNet-10kモデルの評価を示す。図48Aに関して、GTEx RNA-seqデータに対してバリデーションしたSpliceNet-10kモデルによって予測された潜在的スプライス突然変異の割合が示されている。バリデーション率は、訓練中に使用される染色体(chr1、chr3、chr5、chr7、およびchr9を除くすべての染色体、左)および染色体の残り部分(右)上のバリアントに対して別々に示される。予測は、Δスコアでグループ化される。われわれは、各Δスコアグループ内のバリアントの4つのクラス(アクセプターまたはドナーの利得または損失)の各々に対する訓練染色体とテスト染色体との間でバリデーション率を比較した。これは、訓練染色体とテスト染色体との間で予測されたΔスコア値の分布の潜在的な差を説明する。バリデーション率の差は、有意でない(P＞0.05、16回の検定に対するボンフェローニ補正を含むフィッシャーの正確確率検定)。

【０６１６】

図48Bに関して、異なるΔスコアカットオフにおけるGTExコホート内のスプライス変更バリアントを検出するときのSpliceNet-10kの感度が示されている。モデルの感度は、訓練に使用される染色体(左)および染色体の残り部分(右)上のバリアントについて別々に図示されている。われわれは、フィッシャーの正確確率検定を使用して、訓練染色体とテスト染色との間の0.2のΔスコアカットオフでモデルの感度を比較した。これらの差は、エクソンの近くのバリアントまたは深イントロンバリアントのいずれかに対して有でない(2回の検定に対するボンフェローニ補正の後にP>0.05)。

【０６１７】
図48Cに関して、訓練に使用される染色体(左)および染色体の残り部分(右)上のバリアントについて別々に計算された、悪影響を有するExACデータセット内の予測された同義バリアントおよびイントロン潜在的スプライスバリアントの割合が示されている。割合およびP値は、図4Aに示されているように計算される。われわれは、各Δスコアグループ内のバリアントの4つのクラス(アクセプターまたはドナーの利得または損失)の各々に対する訓練染色体とテスト染色体との間で共通バリアントおよび稀少バリアントの数を比較した。差は、有意でない(P＞0.05、12回の検定に対するボンフェローニ補正を含むフィッシャーの正確確率検定)。
【０６１８】
図48Dに関して、DDD、ASD、および対照コホートに対する1人当たりの予測された潜在的スプライスデノボ突然変異(DNM)が、訓練に使用される染色体(左)および染色体の残り部分(右)上のバリアントについて別々に図示されている。エラーバーは、95%の信頼区間(CI)を示している。1人当たりの潜在的スプライスデノボバリアントの数は、訓練セットのサイズのおおよそ半分なのでテストセットに対しては小さい。数はサンプルサイズが小さいのでノイズが多い。
【０６１９】
図49A、図49B、図49Cは、同義領域部位、イントロン領域部位、または非翻訳領域部位のみからの、稀少遺伝病の患者におけるデノボ潜在的スプライス変異体を示す。図49Aに関して、Deciphering Developmental Disordersコホート(DDD)からの患者、Simons Simplex CollectionおよびAutism Sequencing Consortiumからの自閉症スペクトラム障害(ASD)を患っている個体、さらには健康な対照に対する1人当たりの、潜在的スプライスΔスコア>0.2である予測された潜在的スプライスデノボ突然変異(DNM)が示されている。健康な対照を超えるDDDおよびASDコホートにおけるエンリッチメントが図示されており、コホート間でバリアント確認を調整している。エラーバーは、95%の信頼区間を示している。
【０６２０】
図49Bに関して、健康な対照と比較した各カテゴリのエンリッチメントに基づく、DDDおよびASDコホートに対する機能的カテゴリによる病原性DNMの推定された割合が示されている。潜在的スプライス割合は、ミスセンスおよびより深いイントロン部位の欠如について調整される。
【０６２１】
図49Cに関して、異なるΔスコア閾値における健康な対照と比較したDDDおよびASDコホート内の潜在的スプライスDNMのエンリッチメントおよび過剰が示されている。潜在的スプライス過剰は、ミスセンスおよびより深いイントロン部位の欠如について調整される。
【０６２２】
図50Aおよび図50Bは、ASDにおける潜在的スプライスデノボ突然変異を病原性DNMの割合として示す。図50Aに関して、潜在的スプライス部位を予測するために異なるΔスコア閾値におけるASD遺伝発端者内の潜在的スプライスDNMのエンリッチメントおよび過剰が示されている。
【０６２３】
図50Bに関して、潜在的スプライス部位を予測するために異なるΔスコア閾値を使用して、病原性DNM(タンパク質コード突然変異を含む)のすべてのクラスの割合としての潜在的スプライス部位に帰因する病原性DNMの割合が示されている。より許容性の高いΔスコア閾値は、低いオッズ比を有することのトレードオフで、背景予想の上で識別された潜在的スプライス部位の数を増加させる。
【０６２４】
図51A、図51B、図51C、図51D、図51E、図51F、図51G、図51H、図51I、および図51Jは、ASD患者における予測された潜在的スプライスデノボ突然変異のRNA-reqバリデーションを示す。RNA-seqによる実験的バリデーションについて選択された36個の予測された潜在的スプライス部位からのRNA表現のカバレッジおよびスプライス接合カウントが示されてい

る。各サンプルについて、影響のある個体に対するRNA-seqカバレッジおよび接合カウントは、上に示されており、突然変異のない対照個体は、下に示されている。プロットは、バリデーション状況およびスプライス異常タイプ別にグループ化される。

【０６２５】

図52Aおよび図52Bは、カノニカル転写産物のみで訓練されたモデルのRNA-seqに対するバリデーション率および感度を示す。図52Aに関して、われわれは、カノニカルGENCODE転写産物からの接合のみを使用してSpliceNet-10kモデルを訓練し、このモデルの性能をGTExコホート内の少なくとも5人の個体に出現するカノニカル接合およびスプライス接合の両方で訓練されたモデルと比較した。われわれは、各Δスコアグループ内のバリアントの4つのクラス(アクセプターまたはドナーの利得または損失)の各々に対する2つのモデルのバリデーション率を比較した。2つのモデルの間のバリデーション率の差は、有意でない(P＞0.05、16回の検定に対するボンフェローニ補正を含むフィッシャーの正確確率検定)。

【０６２６】

図52Bに関して、異なるΔスコアカットオフにおけるGTExコホート内のスプライス変更バリアントを検出するときにカノニカル接合上で訓練されたモデルの感度が示されている。深イントロン領域内のこのモデルの感度は、図2のモデルより低い(P<0.001、ボンフェローニ補正を含むフィッシャーの正確確率検定)。エクソンの近くの感度は著しく異ならない。

【０６２７】

図53A、図53B、および図53Cは、アンサンブルモデリングがSpliceNet-10k性能を向上させることを示す。図53Aにおいて、5つの個別のSpliceNet-10kモデルのTop-k精度および精度-再現率曲線の下の面積が図示されている。モデルは同じアーキテクチャを有しており、同じデータセットを使用して訓練された。しながら、これらは、パラメータ初期化、データシャッフリングなどの、訓練プロセスに関わる様々なランダム態様により互いに異なる。

【０６２８】

図53Bに関して、5つの個別のSpliceNet-10kモデルの予測は、相関性が高い。この研究のために、われわれは、少なくとも1つのモデルによって0.01以上のアクセプターまたはドナースコアを割り当てられたテストセット内の位置のみを考慮した。サブプロット(i, j)は、モデル#jの予測に対してモデル#iの予測をプロットすることによって構築される(対応するピアソン相関はサブプロットの上に表示される)。

【０６２９】

図53Cに関して、性能は、SpliceNet-10kアンサンブルを構築するために使用されるモデルの数が1から5に増やされると改善する。

【０６３０】

図54Aおよび図54Bは、エクソン密度が変化する領域におけるSpliceNet-10kの評価を示す。図54Aに関して、テストセット位置は、10,000ヌクレオチドウィンドウ内に存在しているカノニカルエクソンの数に応じて5つのビンに分類された。各ビンについて、われわれは、SpliceNet-10kに対してTop-k精度および精度-再現率曲線の下の面積を計算した。

【０６３１】

図54Bに関して、われわれは、MaxEntScanを比較として用いて解析を繰り返した。両方のモデルの性能は、陽性テストケースの数が陰性テストケースの数に比べて増加するので、Top-k精度および精度-再現率AUCによって測定されるように、より高いエクソン密度で改善する。

【０６３２】

コホート毎のデノボ突然変異のエンリッチメント

候補潜在的スプライスDNMは3つのコホートの各々でカウントされた。DDDコホートはエクソンから8nt>離れているイントロンDNMを報告せず、したがって、エクソンから>8ntの領域は、エンリッチメント解析がDDDコホートとASDコホートとの間の同等の比較を可能にするためにすべてのコホートから除外された(図41A)。われわれは、また、二重潜在的ス

プライシングおよびタンパク質コード機能結果を含む突然変異を除外する別々の解析を実行して、エンリッチメントが影響を受けるコホート内のタンパク質コード効果を有する突然変異のエンリッチメントによるものであることを実証した(図49A、図49B、および図49C)。健康な対照のコホートをベースラインとして使用して、コホート間の個体毎の同義DNMの率を正規化することによってコホート間のDNMの異なる確認のためカウントがスケーリングされた。われわれは、E検定を使用してポアソン率を比較してコホート毎の潜在的スプライスDNMの率を比較した(KrishnamoorthyおよびThomson、2004年)。

【０６３３】

予想を上回るエンリッチメントに対するプロットされた率(図41C)は、トリヌクレオチド配列構成モデルを使用してエクソンから9から50nt離れて生じると予想されるすべての潜在的スプライスDNMの割合だけ上方にスケーリングすることによってエクソンから>8ntのDNMの欠如に対して調整された(下の、遺伝子毎のデノボ突然変異のエンリッチメントを参照)。潜在的部位の無症状のみの診断の割合および過剰(図49Bおよび図49C)も、ミスセンス部位対同義部位で生じると予想される潜在的スプライス部位の割合だけ潜在的カウントをスケーリングすることによってミスセンス部位の欠如に対して調整された。エンリッチメントに対するΔスコア閾値の影響は、一定範囲のカットオフにわたってDDDコホート内の潜在的スプライスDNMのエンリッチメントを計算することによって評価された。これらの観察結果の各々について、予想されるオッズ比が、潜在的スプライスDNMの過剰とともに、計算された。

【０６３４】

病原性DNMの割合

ベースライン突然変異率と比較したDNMの過剰は、コホート内の発病率と考えてよい。われわれは、健康な対照のコホートの背景に対して、ASDおよびDDDコホート内の機能型によるDNMの過剰を推定した(図41B)。DNMカウントは、上で説明されているように個体毎に同義DNMの率に正規化された。DDD潜在的スプライスカウントは、上で説明されているようにイントロンから9～50nt離れているDNMの欠如について調整された。ASDおよびDDDの両方のコホートについて、われわれは、また、陰性選択解析から近イントロン(<50n)潜在的スプライスバリアント対深イントロン(>50nt)潜在的スプライスバリアントの比を使用して、エクソンから>50nt離れている深イントロンバリアントの欠損確認について調整した(図38G)。

【０６３５】

遺伝子毎のデノボ突然変異のエンリッチメント

われわれは、トリヌクレオチド配列構成モデル(Samochaら、2014年)を使用してゲノム内のすべてのバリアントに対するヌル突然変異率を決定した。われわれはネットワークを使用してエクソン内の、およびイントロン内最大8ntの可能なすべての単一ヌクレオチド置換に対するΔスコアを予測した。ヌル突然変異率モデルに基づき、われわれは、遺伝子毎のデノボ潜在的スプライス突然変異の予想された数を取得した(Δスコア>0.2をカットオフとして使用して)。

【０６３６】

DDD研究(McRaeら、2017年)に関して、遺伝子は2つのモデル、すなわち、タンパク質切り詰め(PTV)DNMのみを考慮したモデルおよびすべてのタンパク質変更DNM(PTV、ミスセンス、およびインフレームインデル)を考慮したモデルの下で、可能性と比較してDNMのエンリッチメントについて評価された。各遺伝子について、われわれは、最も重要なモデルを選択し、多重仮説検証に対してP値を調整した。これらのテストは、われわれが潜在的スプライスDNMまたは潜在的スプライス率(既定のテスト、元のDDD研究において使用された)を考察しなかった場合に1回、われわれがまた潜在的スプライスDNMおよびその突然変異率をカウントしなかった場合に1回行われた。われわれは、潜在的スプライスDNMを含めたときにFDR調整済みP値<0.01を有するが、潜在的スプライスDNMを含めないとき(既定のテスト)にFDR調整済みP値>0.01を有する遺伝子として識別された追加の候補遺伝子を報告している。エンリッチメントテストは、ASDコホートについて同様に実行された。

【０６３７】
予測された潜在的スプライス部位のバリデーション
われわれは、リンパ芽球様細胞株内に少なくともRPKM>1 RNA-seq表現を有する、Simons Simplex Collection内の影響を受ける遺伝発端者から高信頼度デノボを選択した。われわれは、スプライス損失バリアントに対するΔスコア閾値>0.1およびスプライス利得バリアントに対するΔスコア閾値>0.5に基づきバリデーションに対してデノボ潜在的スプライスバリアントを選択した。細胞株は前々から調達されている必要があるので、これらの閾値は、われわれが論文の別のところで採用した閾値と比較して、われわれの方法の以前の反復を反映しており(図38Gならびに図41A、図41B、図41C、および図41D)、ネットワークはモデル訓練に対するGTEx新規スプライス接合を含まなかった。

【０６３８】
リンパ芽球様細胞株は、これらの遺伝発端者に対するSSCから取得された。細胞は、培養基(RPMI 1640、2mMのL-グルタミン酸、15%のウシ胎仔血清)で最大細胞密度1×10^{6}細胞/mlになるまで培養された。細胞が最高密度に達したときに、これらは4、5回ピペット操作し、200,000～500,000生存細胞/mlの密度まで播種することによって細胞を解離して継代された。細胞は、37℃、5%のCO_2の条件の下で10日間かけて成長させた。次いで、約5×10^{5}個の細胞が分離され、4℃で5分間、300×gで遠沈された。RNAは、メーカーのプロトコルに従ってRNeasy(登録商標)Plus Micro Kit (QIAGEN)を使用して抽出された。RNAクオリティは、Agilent RNA 6000 Nano Kit(Agilent Technologies)を使用して評価され、Bioanalyzer 2100 (Agilent Technologies)にかけられた。RNA-seqライブラリはRibo-Zero Gold Set A (Illumina)を使用するTruSeq(登録商標)Stranded Total RNA Library Prep Kitによって生成された。ライブラリは、2億7000万～3億8800万リード(中央値3億5800万リード)のカバレッジで150-nt単一リードシーケンシングを使用してCenter for Advanced Technology (UCSF)のHiSeq 4000計測器でシーケンシングされた。

【０６３９】
各患者に対するシーケンシングリードは、患者のデノボバリアント(Iossifovら、2014年)を対応する代替対立遺伝子で置換することによってhg19から作成された参照に対してOLego(Wuら、2013年)で整列された。シーケンシングカバレッジ、スプライス接合使用度、および転写産物配置は、MISO(Katzら、2010年)からのsashimiプロットでプロットされた。われわれは、モデル予測の節のバリデーションにおいて上で説明されたように予測された潜在的スプライス部位を評価した。13個の新規スプライス部位(9個の新規接合、4個のエクソンスキッピング)が確認され、これらは潜在的スプライス部位を含むサンプル中でのみ観察され、149個のGTExサンプルのどれかにおいてまたは他の35個のシーケンシングされたサンプルにおいて観察されなかった。4つの追加のエクソンスキッピング事象について、低レベルのエクソンスキッピングはGTExにおいて観察されることが多かった。これらの場合に、われわれは、スキッピング接合を使用したリードの割合を計算し、この割合が他のサンプルと比較したサンプルを含む潜在的スプライス部位内で最高であることを検証した。4つの追加のケースは、他のサンプルでは存在しないか、かなり低かった顕著なイントロン保持に基づきバリデーションされた。対照サンプル内の控えめなイントロン保持は、DDX11およびWDR4における事象を解決することを妨げた。2つの事象(CASDおよびGASAPにおける)は、バリアントがシーケンシングリード内に存在していなかったので、失敗バリデーションとして分類された。

【０６４０】
データおよびソフトウェアの利用可能性
訓練およびテストデータ、参照ゲノム内のすべての単一ヌクレオチド置換に対する予測スコア、RNA-seqバリデーション結果、RNA-seq接合、およびソースコードは以下のところで公開されてホストされている。
https://basespace.illumina.com/s/5u6ThOblecrh

【０６４１】
36個のリンパ芽球様細胞株に対するRNA-seqデータは、アクセッション番号E-MTAB-xxxx

の下でEMBL-EBI(www.ebi.ac.uk/arrayexpress)のArrayExpressデータベース内に預けられている。

【０６４２】

予測スコアおよびソースコードは、オープンソース修正済みApache License v2.0の下で公開され、学術的および非商用ソフトウェアアプリケーションでの使用には無料である。現場の懸念事項となっている循環の問題を低減するために、著者らは、この方法からの予測スコアは他の分類器のコンポーネントとして組み込まれないことを明示的に要求し、その代わりに、利害関係者は提供されるソースコードおよびデータを使用して自分たちのディープラーニングモデルを直接訓練し改善することを求める。

【０６４３】

キーリソースの表

【０６４４】

【表３】

試薬またはリソース	ソース	識別子
預けられたデータ		
GTEx コホートに対する RNA-seq データおよびバリアントコール	https://www.ncbi.nlm.nih.gov/projects/gap	dbGAP accession: phs000424.v6.p1
自閉症患者および健康な対照に対するデノボ突然変異	(Iossifov ら、2014 年)	N/A
Deciphering Developmental Disorders コホートからのデノボ突然変異	(McRae ら、2017 年)	N/A
カノニカル SpliceNet モデルを訓練するために使用される GENCODE 主転写産物からのスプライス接合	当研究	https://basespace.illumina.com/s/5u6ThOblecrh
訓練データセットを増補するために使用される GTEx からのスプライス接合	当研究	https://basespace.illumina.com/s/5u6ThOblecrh
パラログを除外して、モデルをテストするために使用される GENCODE 主転写産物からのスプライス接合	当研究	https://basespace.illumina.com/s/5u6ThOblecrh
モデルをテストするために使用される lincRNA のスプライス接合	当研究	https://basespace.illumina.com/s/5u6ThOblecrh
カノニカルモデルの予測	当研究	https://basespace.illumina.com/s/5u6ThOblecrh
GTEx 補足モデルの予測	当研究	https://basespace.illumina.com/s/5u6ThOblecrh
すべての GTEx v6.p1 サンプルにおけるすべての GTEx 接合	当研究	https://basespace.illumina.com/s/5u6ThOblecrh
Δスコア>0.1 であるバリデーションされた GTEx プライベートバリアントのリスト	当研究	https://basespace.illumina.com/s/5u6ThOblecrh
36 人の自閉症患者における RNA-seq に対する整列済み BAM ファイル	当研究	ArrayExpress accession: E-MTAB-xxxx
ソフトウェアおよびアルゴリズム		
SpliceNet ソースコード	当研究	https://basespace.illumina.com/s/5u6ThOblecrh

【０６４５】

補足表題

Table S1は、エフェクトサイズ計算および組織特有のスプライシング効果を実証するために使用されるGTExサンプルを示している。図38A、図38B、図38C、図38D、図38E、図38F

、および図38G、図39A、図39B、および図45に関係している。

【０６４６】

Table S2は、すべてのアルゴリズムが同じ数の利得および損失をゲノムワイドで予測するSpliceNet-10k、GeneSplicer、MaxEntScan、およびNNSpliceに対するマッチした信頼度カットオフを示している。図38Gに関係する。

【０６４７】

Table S3は、各コホートにおける予測された潜在的スプライスDNMのカウントを示している。図41A、図41B、図41C、図41D、図41E、および図41Fに関係しており、以下に提示される。

【０６４８】

【表４】

			最大 8nt までのエクソン+イントロン		エクソンから>8nt のイントロン	
コホート	遺伝発端者(n)	遺伝発端者毎の同義デノボ	未調整	同義に合わせて正規化済み	未調整	同義に合わせて正規化済み
DDD	4293	0.28744	347	298.7	14	12.1
ASD	3953	0.24462	236	238.7	64	64.7
対照	2073	0.24747	98	98	20	20

【０６４９】

Table S4は、各突然変異カテゴリに対する遺伝子毎の予想されるデノボ突然変異率を示している。図41A、図41B、図41C、図41D、図41E、および図41Fに関係する。

【０６５０】

Table S5は、DDDおよびASDにおける遺伝子エンリッチメントに対するp値を示している。図41A、図41B、図41C、図41D、図41E、および図41Fに関係する。

【０６５１】

Table S6は、自閉症患者における36個の予測された潜在的スプライスDNMに対するバリデーション結果を示している。図41A、図41B、図41C、図41D、図41E、および図41Fに関係する。

【０６５２】

コンピュータシステム

図59は、開示された技術を実施するために使用することができるコンピュータシステムの簡略ブロック図である。コンピュータシステムは、典型的には、バスサブシステムを介して多数の周辺デバイスと通信する少なくとも1つのプロセッサを備える。これらの周辺デバイスは、たとえば、メモリデバイスおよびファイル記憶装置サブシステム、ユーザインターフェース入力デバイス、ユーザインターフェース出力デバイス、およびネットワークインターフェースサブシステムを含む、記憶装置サブシステムを含むことができる。入力および出力デバイスは、ユーザがコンピュータシステムをインタラクティブに操作することを可能にする。ネットワークインターフェースサブシステムは、他のコンピュータシステム内の対応するインターフェースデバイスへのインターフェースを含む、外部ネットワークへのインターフェースを提供する。

【０６５３】

一実装形態において、ACNNおよびCNNなどのニューラルネットワークは、記憶装置サブシステムおよびユーザインターフェース入力デバイスに通信可能に結合される。

【０６５４】

ユーザインターフェース入力デバイスは、キーボード、マウス、トラックボール、タッ

チパッド、もしくはグラフィックスタブレットなどのポインティングデバイス、スキャナ、ディスプレイに組み込まれたタッチスクリーン、音声認識システムおよびマイクロフォンなどの音声入力デバイス、および他の種類の入力デバイスを含むことができる。一般に、「入力デバイス」という用語の使用は、情報をコンピュータシステム内に入力するためのすべての可能な種類のデバイスおよび方法を含むことが意図されている。

【０６５５】

ユーザインターフェース出力デバイスは、表示サブシステム、プリンタ、ファックス機、または音声出力デバイスなどの非視覚的ディスプレイを含むことができる。表示サブシステムは、陰極線管(CRT)、液晶ディスプレイ(LCD)などのフラットパネルデバイス、プロジェクションデバイス、または可視画像を生成するための他の何らかのメカニズムを含むことができる。表示サブシステムも、オーディオ出力デバイスなどの非視覚的ディスプレイを備えることができる。一般に、「出力デバイス」という用語の使用は、情報をコンピュータシステムからユーザまたは別のマシンもしくはコンピュータシステムに出力するためのすべての可能な種類のデバイスおよび方法を含むことが意図されている。

【０６５６】

記憶装置サブシステムは、本明細書で説明されているモジュールおよび方法のうちのいくつかまたはすべての機能を実現するプログラミングおよびデータ構造を記憶する。これらのソフトウェアモジュールは、一般的に、プロセッサによって、単独で、または他のプロセッサと組み合わせて、実行される。

【０６５７】

記憶装置サブシステムにおいて使用されるメモリは、プログラム実行時に命令およびデータを記憶するための主ランダムアクセスメモリ(RAM)、ならびに固定された命令が記憶されるリードオンリーメモリ(ROM)を含む多数のメモリを備えることができる。ファイル記憶装置サブシステムは、プログラムおよびデータファイル用の永続的記憶域を備えることができ、ハードディスクドライブ、関連する取り外し可能媒体を伴ったフロッピーディスクドライブ、CD-ROMドライブ、光ドライブ、または取り外し可能メディアカートリッジを含むものとしてよい。いくつかの実装形態の機能を実装するモジュールは、ファイル記憶装置サブシステムによって、記憶装置サブシステム、またはプロセッサによってアクセス可能な他のマシン内に記憶され得る。

【０６５８】

バスサブシステムは、コンピュータシステムの様々なコンポーネントおよびサブシステムに意図された通りに互いに通信させるメカニズムを備える。バスサブシステムは、単一のバスとして概略が図示されているけれども、バスサブシステムの代替的実装形態では、複数のバスを使用してよい。

【０６５９】

コンピュータシステムそれ自体は、パーソナルコンピュータ、ポータブルコンピュータ、ワークステーション、コンピュータ端末、ネットワークコンピュータ、テレビ、メインフレーム、サーバファーム、緩くネットワークで接続されたコンピュータで設定されている広域分散コンピュータ群、または任意の他のデータ処理システムもしくはユーザデバイスを含む様々な種類のものであってよい。コンピュータおよびネットワークはその性質上絶えず変化し続けるので、図59に示されているコンピュータシステムの説明は、開示されている技術を例示することを目的とする特定の例としてのみ意図されている。コンピュータシステムの他の多くの構成は、図59に示されているコンピュータシステムよりも多い、または少ない構成要素を有するものが可能である。

【０６６０】

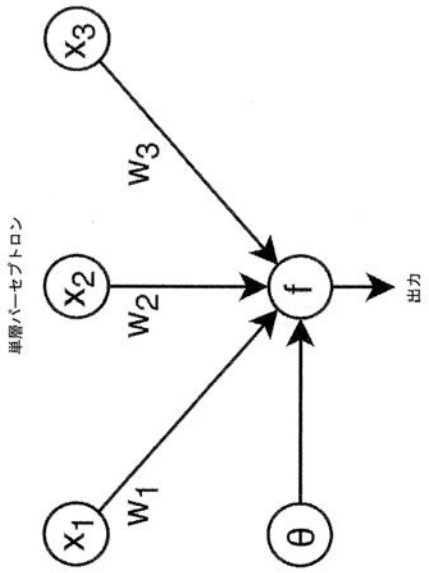
ディープラーニングプロセッサは、GPUまたはFPGAであってよく、Google Cloud Platform、Xilinx、およびCirrascaleなどのディープラーニングクラウドプラットフォームによってホストされてよい。ディープラーニングプロセッサの例は、GoogleのTensor Processing Unit(TPU)、GX4 Rackmount Series、GX8 Rackmount Seriesのようなラックマウントソリューション、NVIDIA DGX-1、MicrosoftのStratix V FPGA、GraphcoreのIntelligent

Processor Unit (IPU)、SnapdragonプロセッサのQualcommのZeroth platform、NVIDIAのVolta、NVIDIAのDRIVE PX、NVIDIAのJETSON TX1/TX2 MODULE、IntelのNirvana、Movidius VPU、Fujitsu DPI、ARMのDynamicIQ、IBM TrueNorth、およびその他を含む。

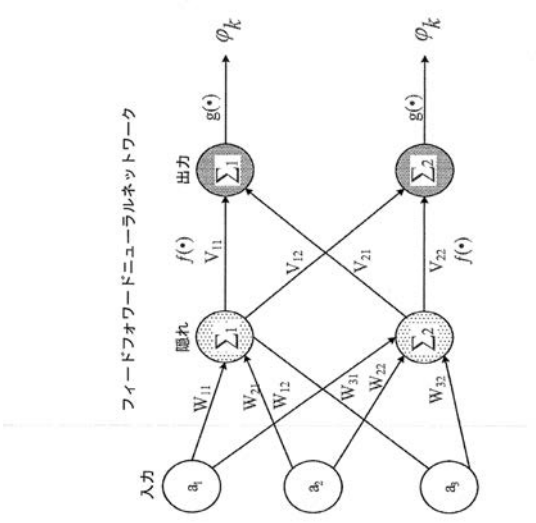
【０６６１】

先行する説明は、開示されている技術の製造および使用を可能にするために提示されている。開示されている実装形態に対し様々な修正を加えられることは、明白であろうし、また本明細書において定義されている一般原理は、開示された技術の精神または範囲から逸脱することなく他の実装形態および応用にも適用され得る。したがって、開示された技術は、図示されている実装形態に限定されることを意図されておらず、本明細書で開示された原理および特徴と一致する最も広い範囲を適用されることを意図されている。開示視されている技術の範囲は、付属の請求項によって定められる。

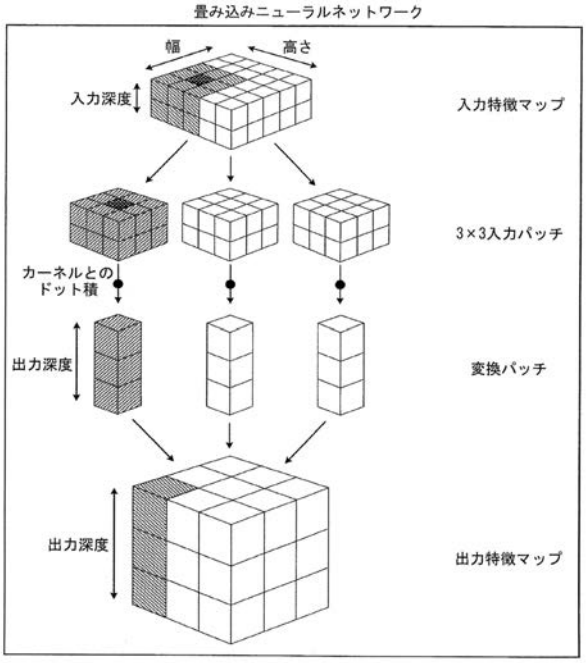
【図１】



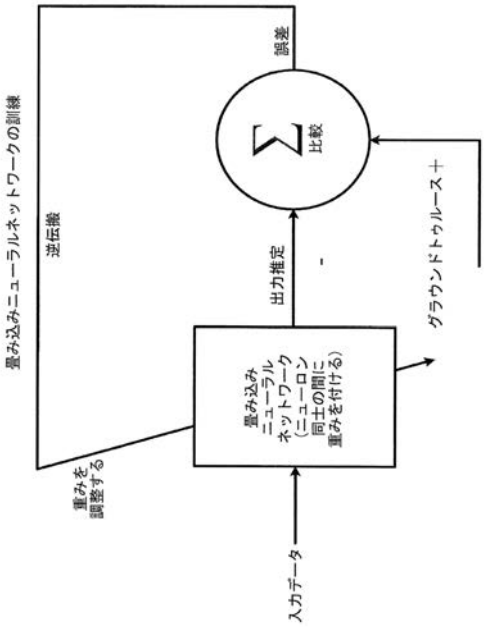
【図２】



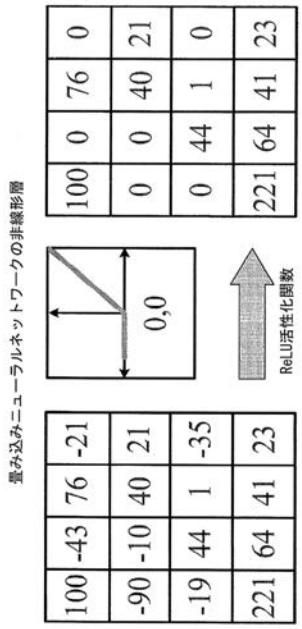
【図３】


畳み込みニューラルネットワーク
幅
高さ
入力深度
入力特徴マップ
3×3入力パッチ
カーネルとの
ドット積
出力深度
変換パッチ
出力深度
出力特徴マップ

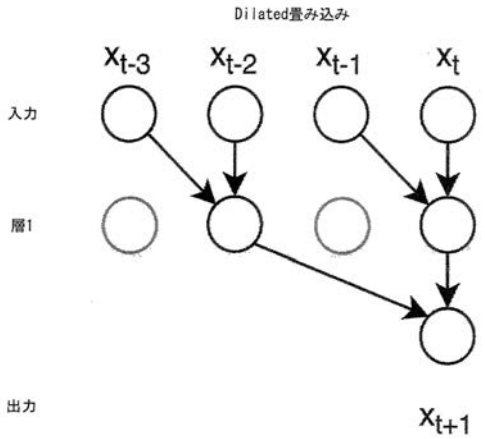
【図４】


誤差
Σ
比較
畳み込みニューラルネットワークの訓練
逆伝搬
出力推定
-
グラウンドトゥルース＋
畳み込み
ニューラル
ネットワーク
(ニューロン
同士の間に
重みを付ける)
重みを
調整する
入力データ

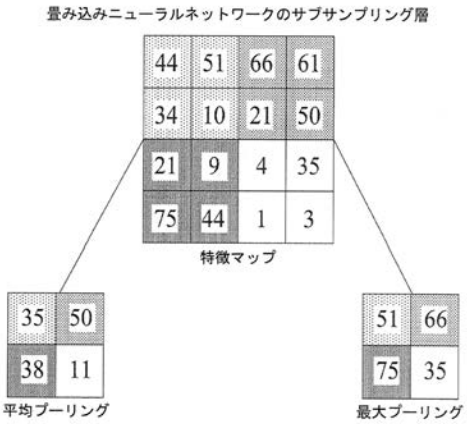
【図５】


畳み込みニューラルネットワークの非線形層
100 0 0 76
0 0 0 0
0 44 40 1
221 64 41 23
0,0
ReLU活性化関数
100 -43 -21 76
-90 -10 40 1
-19 44 21 -35
221 64 41 23

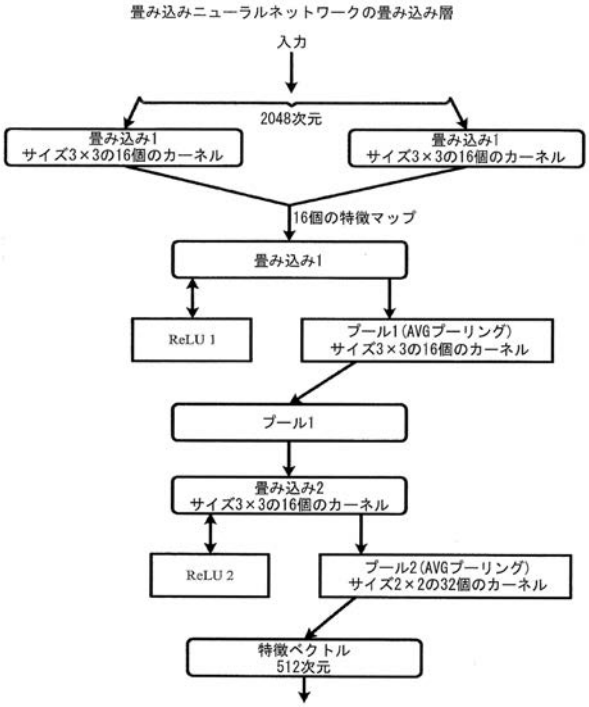
【図６】


Dilated畳み込み
x_{t-3}
x_{t-2}
x_{t-1}
x_t
入力
層1
出力
x_{t+1}

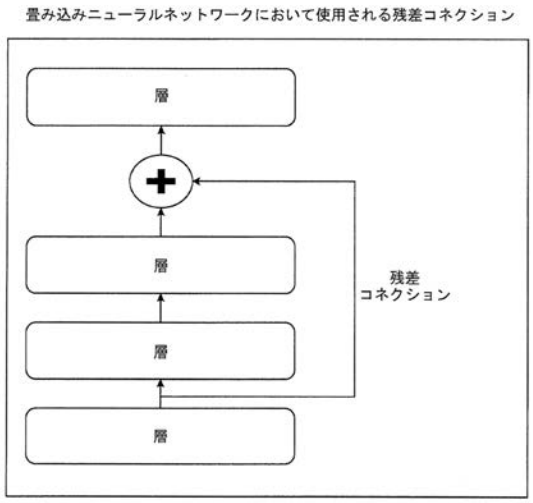
【図７】


畳み込みニューラルネットワークのサブサンプリング層
44 51 66 61
34 10 21 50
21 9 4 35
75 44 1 3
特徴マップ
35 50
38 11
平均プーリング
51 66
75 35
最大プーリング

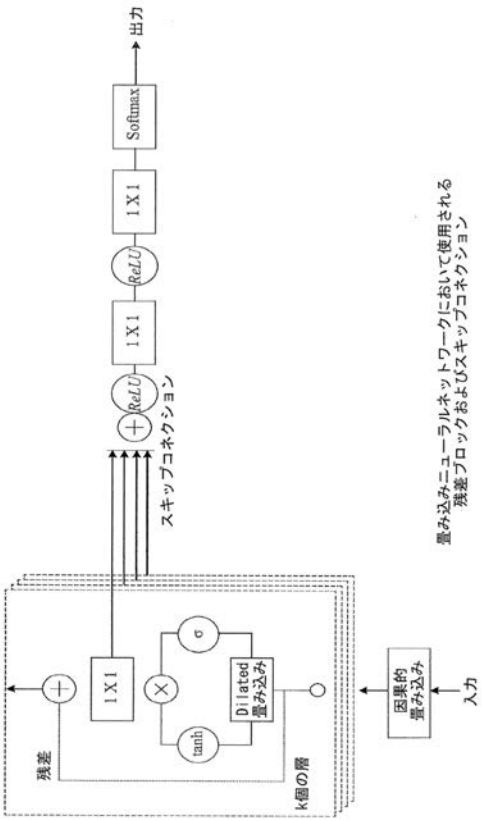
【図８】


畳み込みニューラルネットワークの畳み込み層
入力
2048次元
畳み込み1
サイズ3×3の16個のカーネル
畳み込み1
サイズ3×3の16個のカーネル
16個の特徴マップ
畳み込み1
ReLU 1
プール1(AVGプーリング)
サイズ3×3の16個のカーネル
プール1
畳み込み2
サイズ3×3の16個のカーネル
ReLU 2
プール2(AVGプーリング)
サイズ2×2の32個のカーネル
特徴ベクトル
512次元

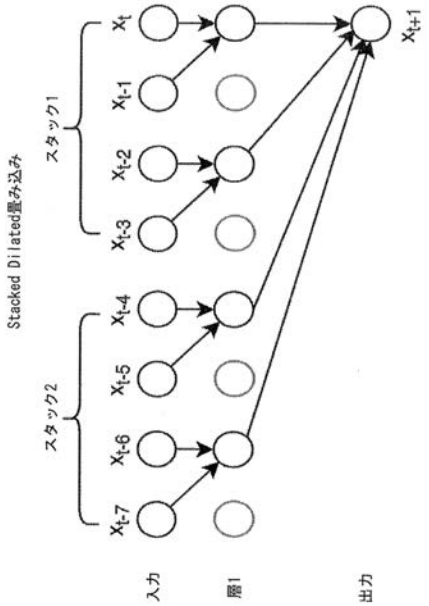
【図９】


畳み込みニューラルネットワークにおいて使用される残差コネクション
層
層
層
層
残差
コネクション

【図１０】


出力
Softmax
1X1
ReLU
1X1
ReLU
スキップコネクション
畳み込みニューラルネットワークにおいて使用される
残差ブロックおよびスキップコネクション
σ
1X1
Dilated
畳み込み
因果的
畳み込み
入力
tanh
残差
k個の層

【図１１】



【図１２】

畳み込みニューラルネットワークによるバッチ正規化フォワードパス

$$\mu_{\mathcal{B}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i^{(\ell-1)}$$

$$\sigma_{\mathcal{B}}^2 = \frac{1}{n}\sum_{i=1}^{n} (\boldsymbol{x}_i^{(\ell-1)} - \mu_{\mathcal{B}})^2$$

$$\hat{\boldsymbol{x}}^{(\ell-1)} = \frac{\boldsymbol{x}^{(\ell-1)} - \mu_{\mathcal{B}}}{\sqrt{\sigma_{\mathcal{B}}^2 + \epsilon}}$$

$$\boldsymbol{x}^{(\ell)} = \boldsymbol{\gamma}^{(\ell)} \hat{\boldsymbol{x}}^{(\ell-1)} + \boldsymbol{\beta}^{(\ell)}$$

【図１３】

バッチ正規化―畳み込みニューラルネットワークによる推論

$$\hat{\boldsymbol{x}}^{(\ell-1)} = \frac{\boldsymbol{x}^{(\ell-1)} - \mu_{\mathcal{D}}}{\sqrt{\sigma_{\mathcal{D}}^2 + \epsilon}}$$

$$\boldsymbol{x}_i^{(\ell)} = \boldsymbol{\gamma}^{(\ell)} \hat{\boldsymbol{x}}_i^{(\ell-1)} + \boldsymbol{\beta}^{(\ell)}$$

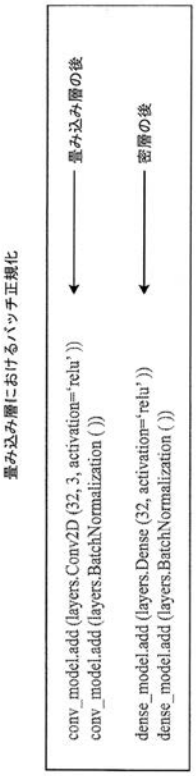
【図１４】

畳み込みニューラルネットワークによるバッチ正規化バックワードパス

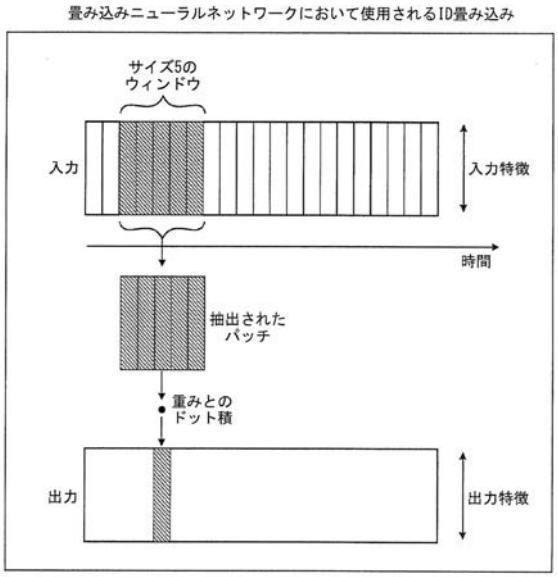
$$\nabla_{\boldsymbol{\gamma}^{(\ell)}} \mathcal{L} = \sum_{i=1}^{n} \left(\nabla_{\boldsymbol{x}^{(\ell+1)}} \mathcal{L}\right)_i \cdot \hat{\boldsymbol{x}}_i^{(\ell)}$$

$$\nabla_{\boldsymbol{\beta}^{(\ell)}} \mathcal{L} = \sum_{i=1}^{n} \left(\nabla_{\boldsymbol{x}^{(\ell+1)}} \mathcal{L}\right)_i$$

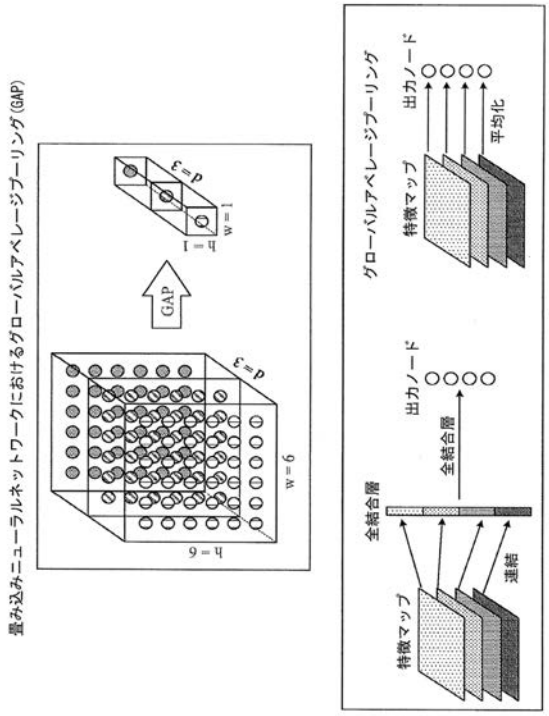
【図１５】



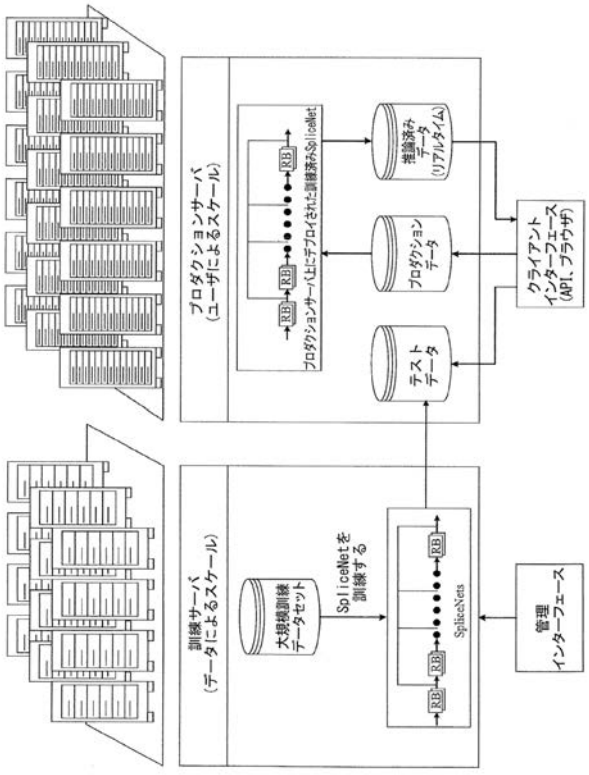
【図１６】



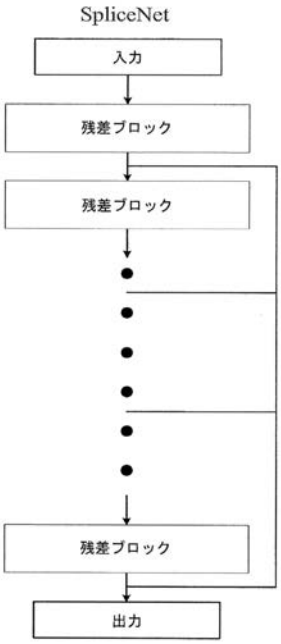
【図１７】



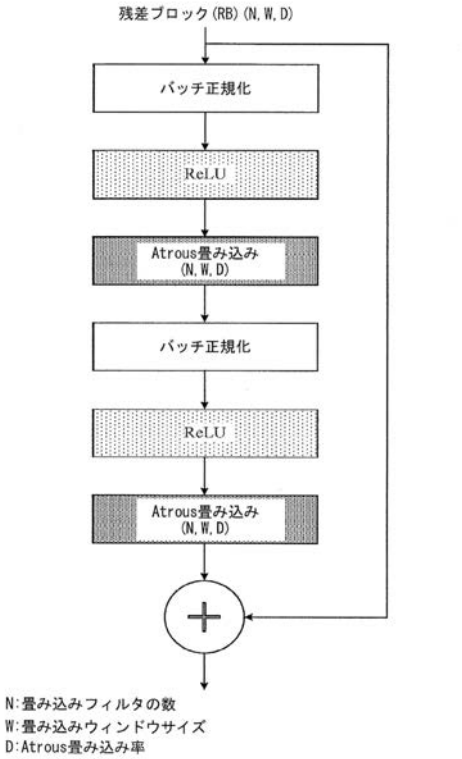
【図１８】



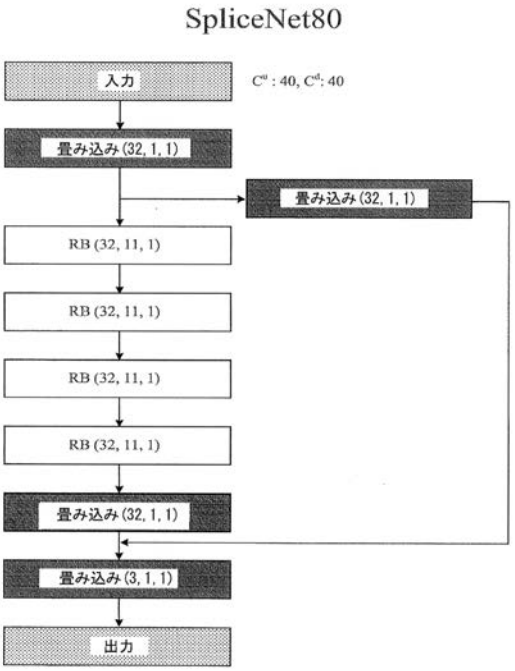
【図１９】



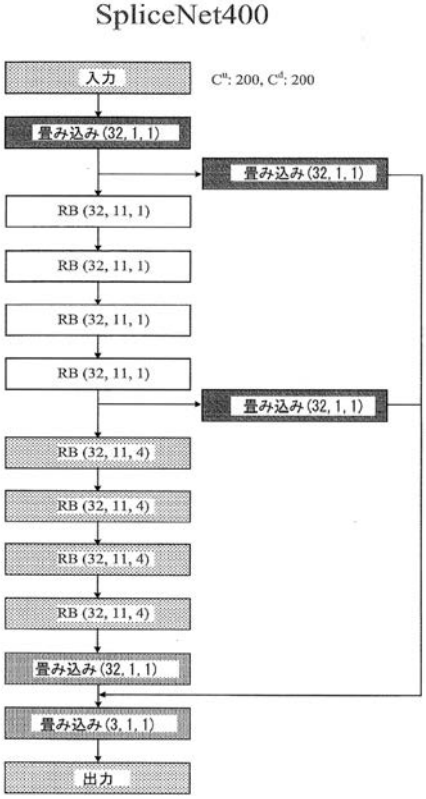
【図２０】



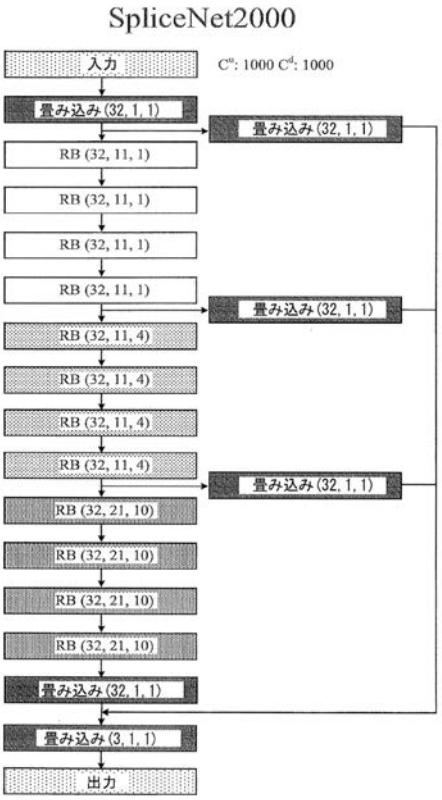
【図２１】



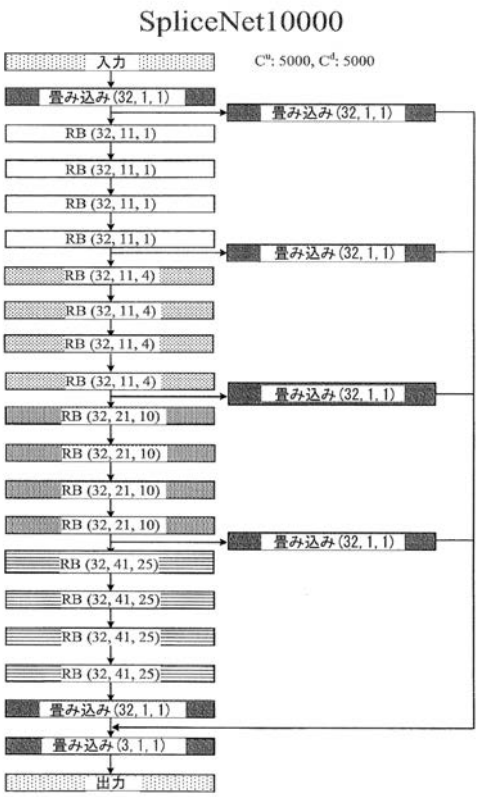
【図２２】



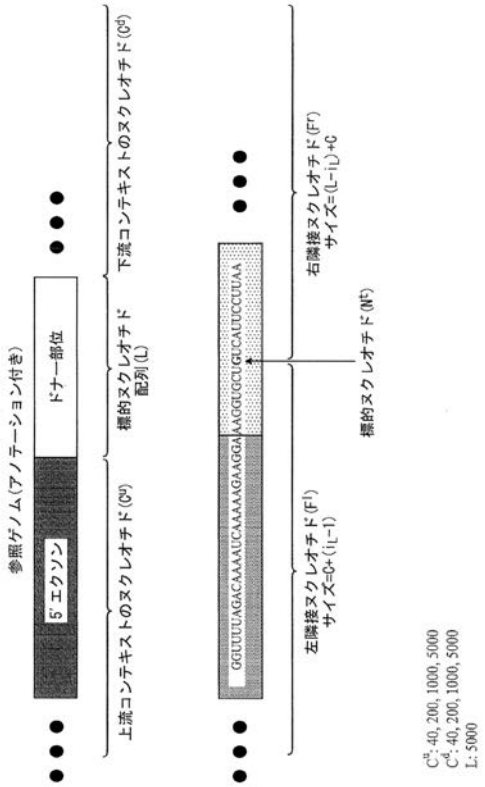
【図２３】



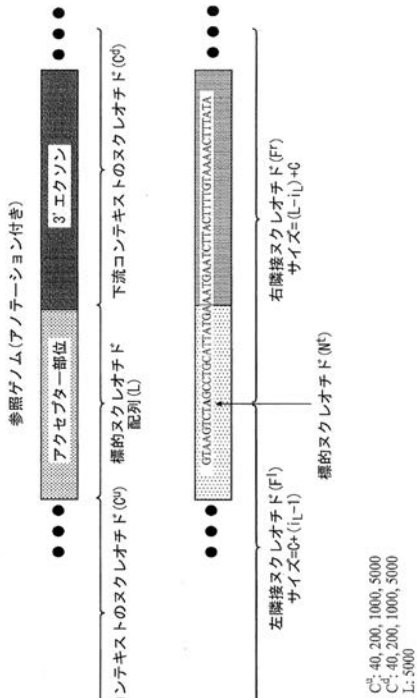
【図２４】



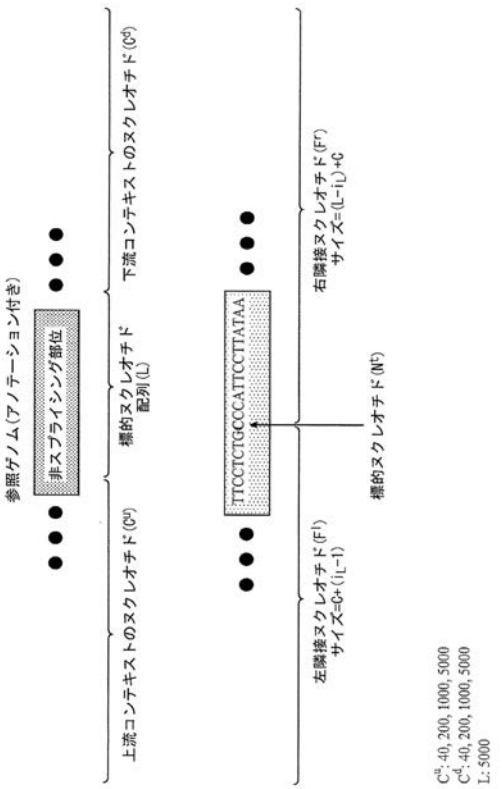
【図２５】



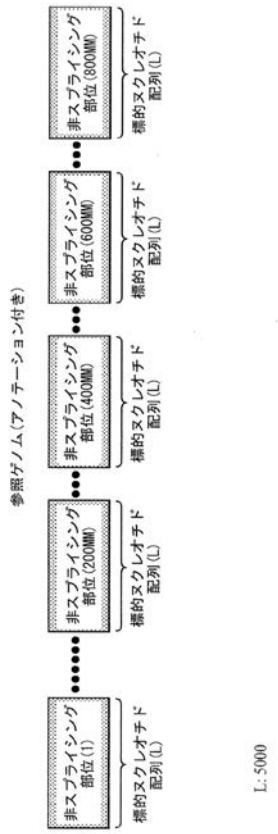
【図２６】



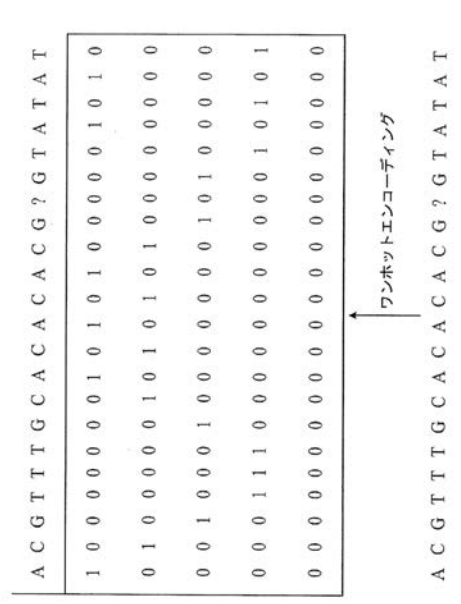
【図２７】



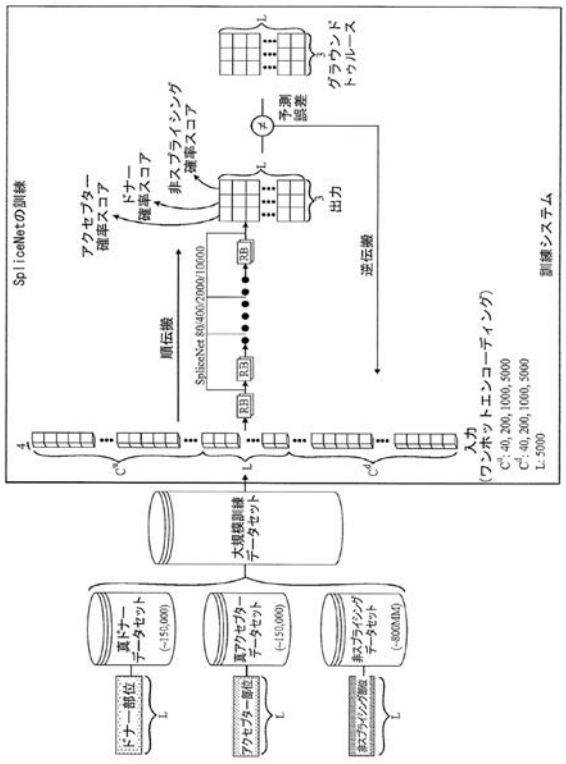
【図２８】



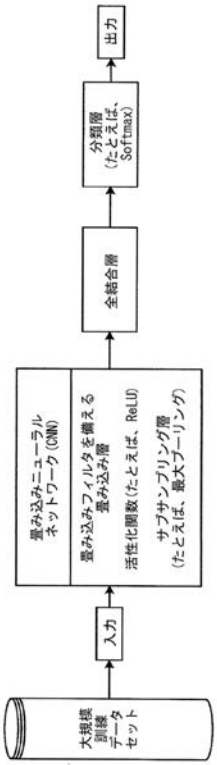
【図２９】



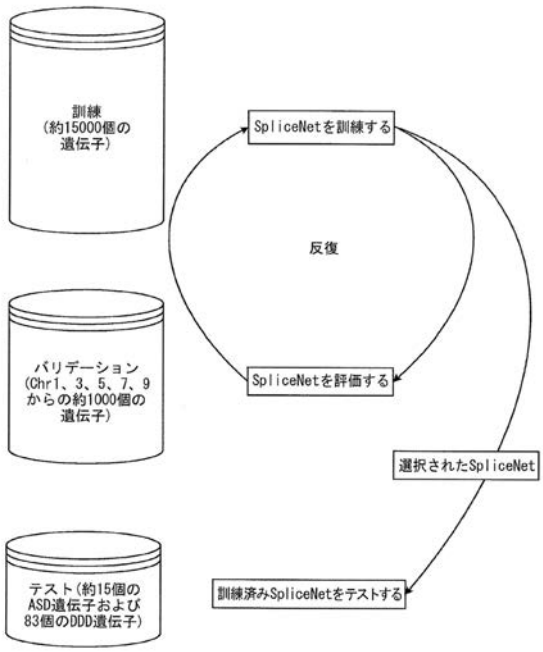
【図３０】



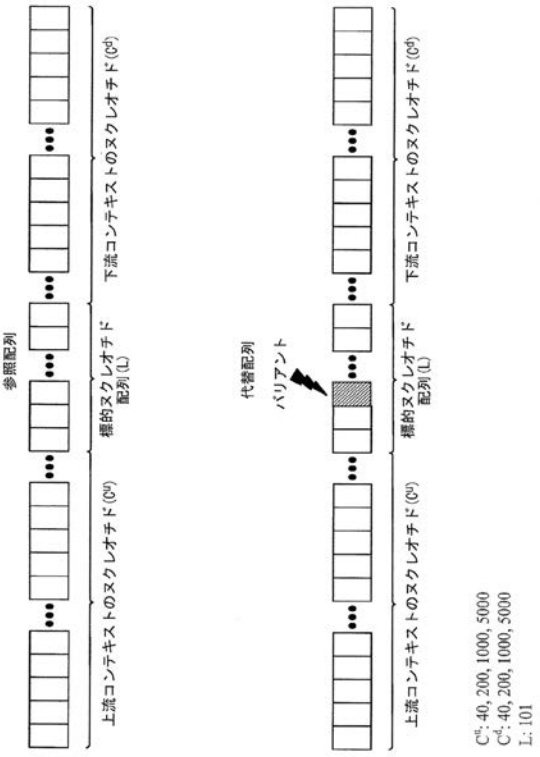
【図３１】



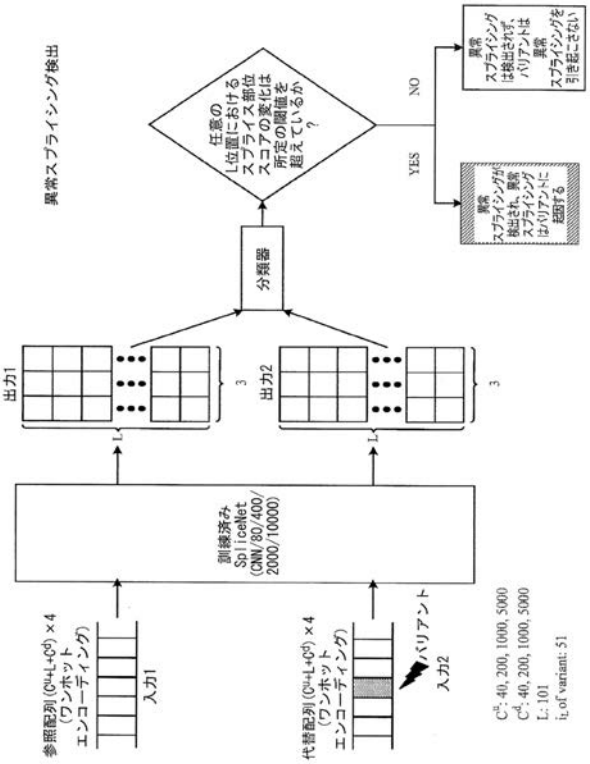
【図３２】



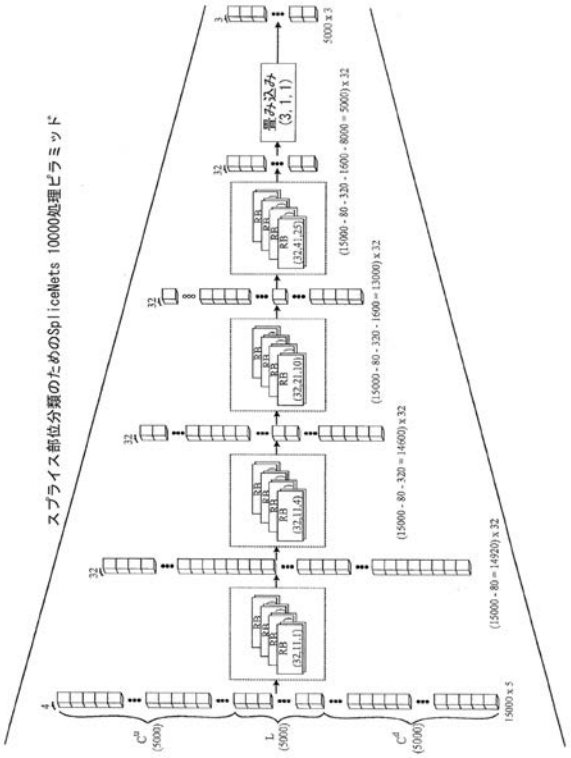
【図３３】



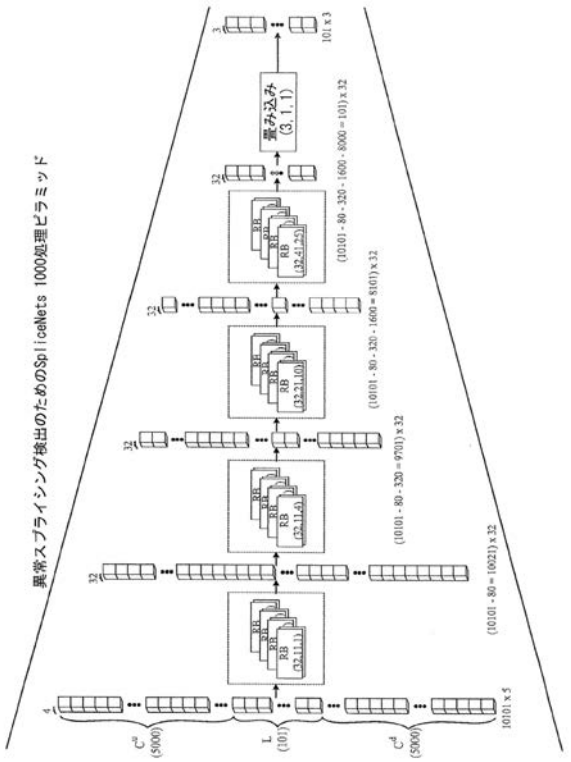
【図３４】



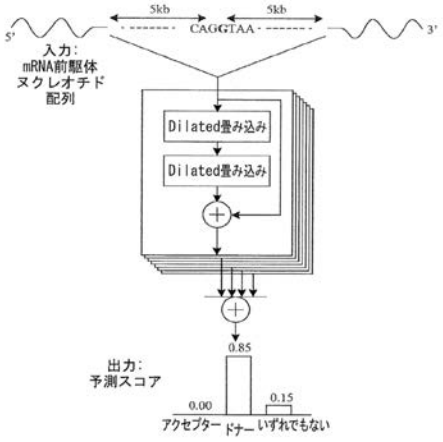
【図３５】



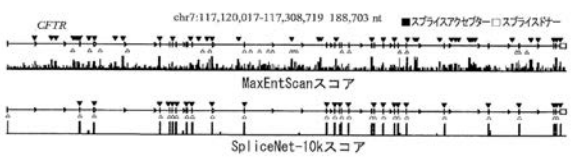
【図３６】



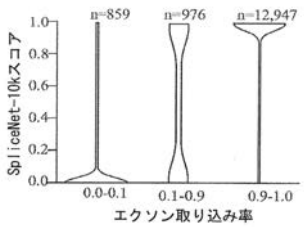
【図３７Ａ】



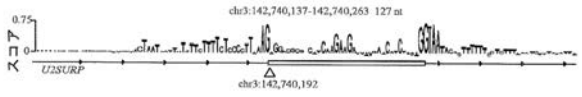
【図３７Ｂ】



【図３７Ｃ】



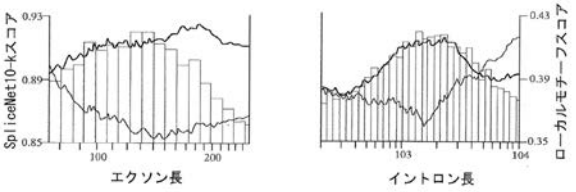
【図３７Ｄ】



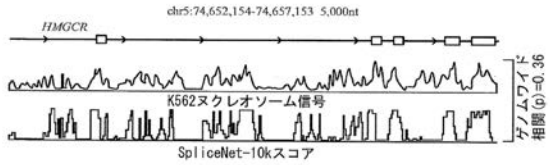
【図３７Ｅ】

	Top-k精度	PR-AUC
SpliceNet-80nt	0.57	0.60
SpliceNet-400nt	0.90	0.95
SpliceNet-2k	0.93	0.97
SpliceNet-10k	0.95	0.98
GeneSplicer	0.30	0.23
MaxEntScan	0.22	0.15
NNSplice	0.22	0.15

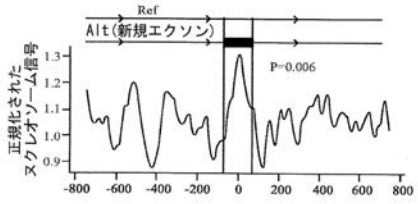
【図３７Ｆ】



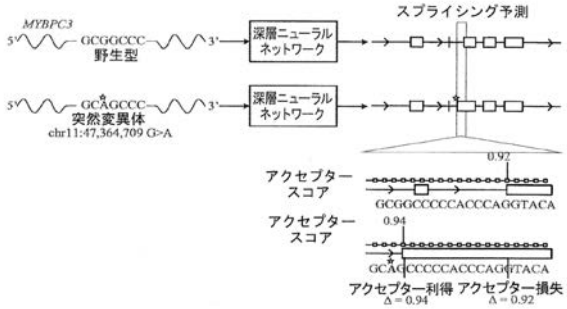
【図３７Ｇ】



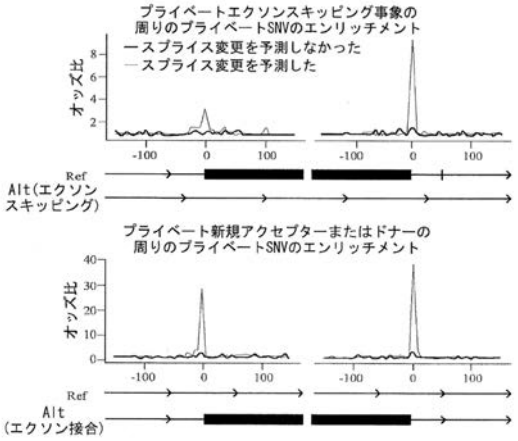
【図３７Ｈ】



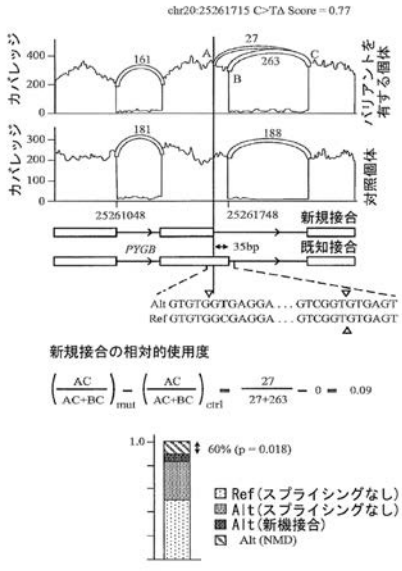
【図３８Ａ】



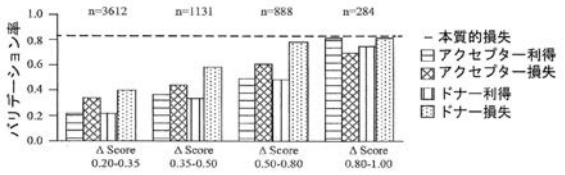
【図３８Ｂ】



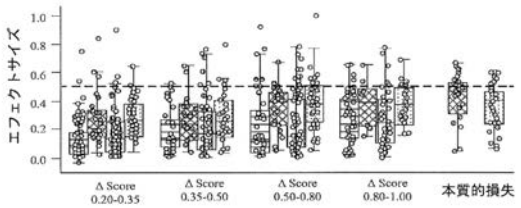
【図３８Ｃ】



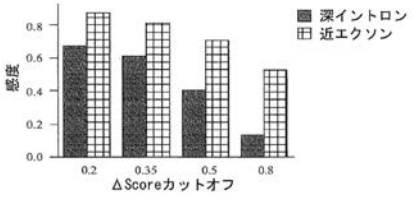
【図３８Ｄ】



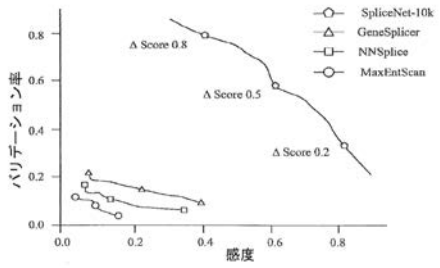
【図３８Ｅ】



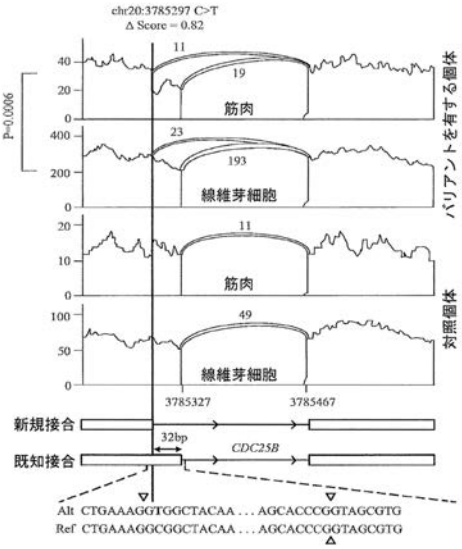
【図３８Ｆ】



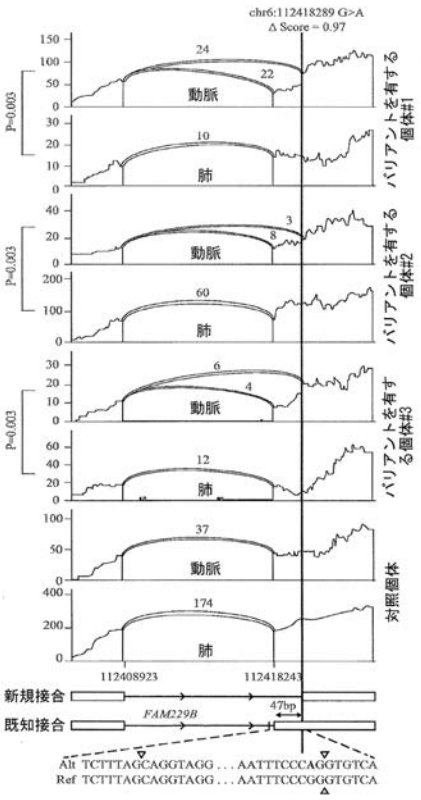
【図３８Ｇ】



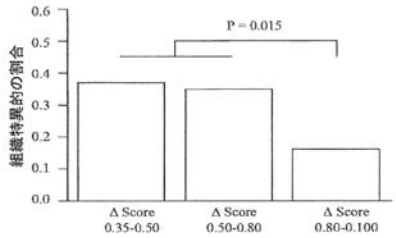
【図３９Ａ】



【図３９Ｂ】



【図３９Ｃ】



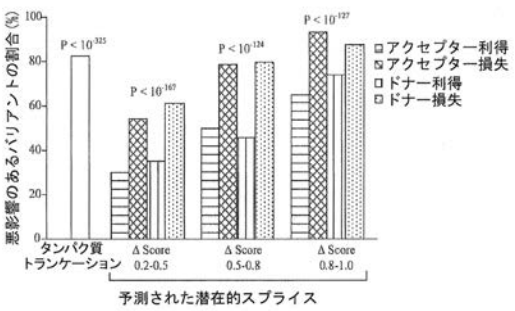
【図４０Ａ】

ExAC近エクソンバリアント

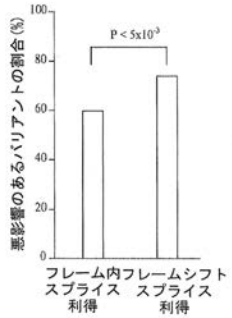
	シングルトン	共通 (AF ≥0.1%)
ΔScore≧0.8 であるSNV	10,369	212
ΔScore<0.1 であるSNV	1,687,004	158,177

オッズ比 (OR) = 4.58 ($P < 10^{-127}$)
悪影響のあるバリアントの割合 = (1-1/OR)

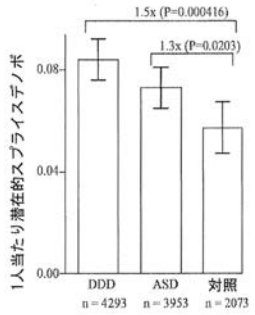
【図４０Ｂ】



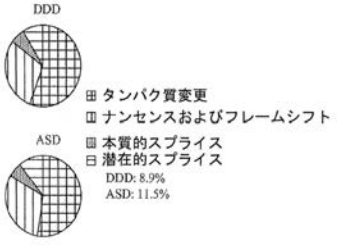
【図４０Ｃ】



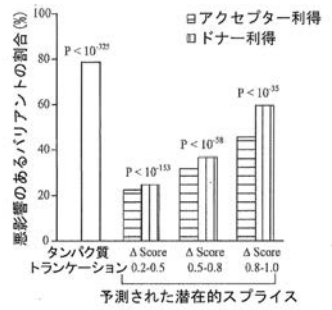
【図４１Ａ】



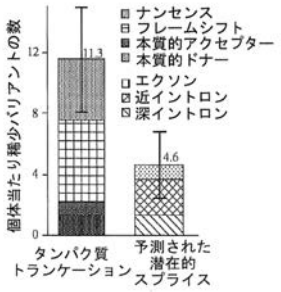
【図４１Ｂ】



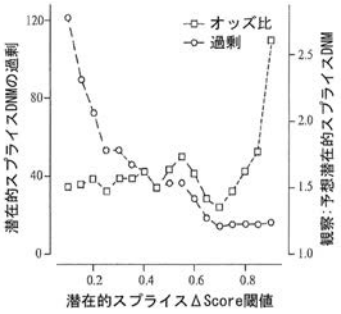
【図４０Ｄ】



【図４０Ｅ】



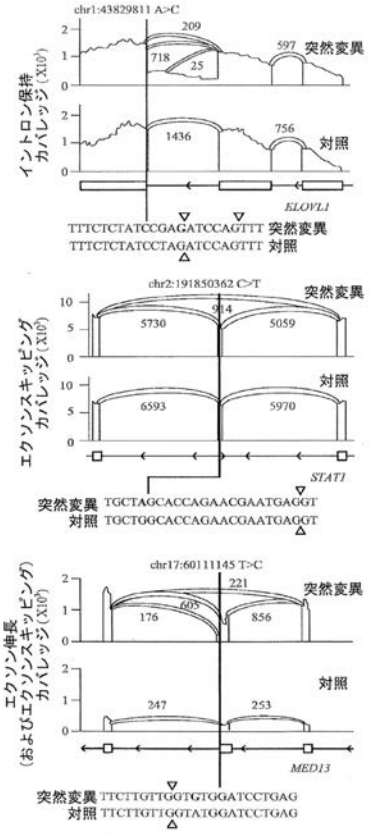
【図４１Ｃ】



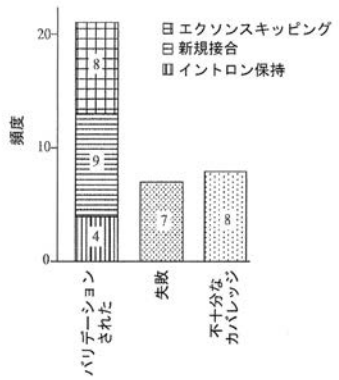
【図４１Ｄ】

	記号	PTV	ミスセンス	潜在的	P	表現型
DDD	*CTNND1*	2	0	1	4.7×10^{-5}	発育遅延
	HNRNPK	2	1	1	9.7×10^{-6}	発育遅延、脊柱側弯症、口腔顔面裂、膀胱尿管逆流現象
	PHF5A	1	0	1	3.1×10^{-3}	発育遅延
	SNAP25	1	0	2	6.8×10^{-5}	アプサンス発作、エピカンサス
	TRIP12	3	2	1	8.5×10^{-6}	言語遅滞
ASD	*KDM6B*	3	1	1	4.6×10^{-6}	自閉症
	SLC6A8	2	0	1	5.8×10^{-6}	自閉症

【図４１Ｅ】



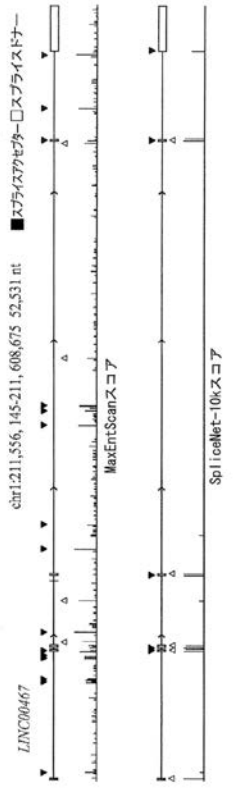
【図４１Ｆ】



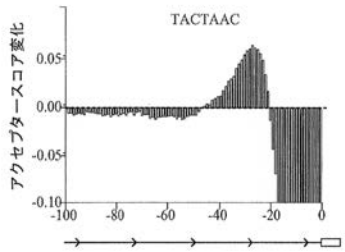
【図４２Ａ】

	Top-k精度		精度-再現率AUC	
	アクセプター	ドナー	アクセプター	ドナー
SpliceNet-80nt	0.5129	0.5110	0.5195	0.5273
SpliceNet-400nt	0.6848	0.7221	0.7320	0.7982
SpliceNet-2k	0.8090	0.8405	0.8742	0.9047
SpliceNet-10k	0.8223	0.8586	0.8896	0.9107
GeneSplicer	0.3238	0.3381	0.2234	0.2982
MaxEntScan	0.2932	0.3687	0.2233	0.2937
NNSplice	0.2655	0.3687	0.1702	0.3003

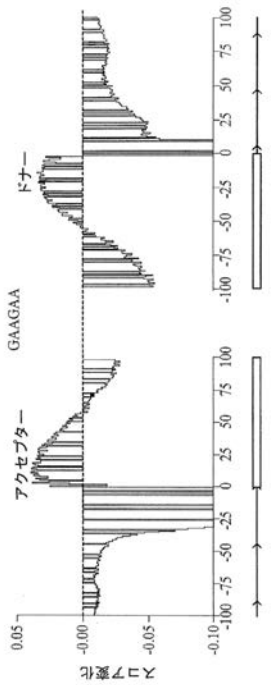
【図４２Ｂ】



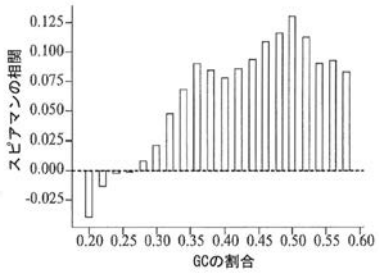
【図４３Ａ】



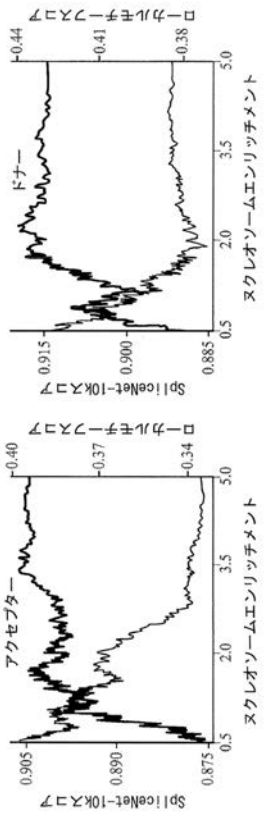
【図４３Ｂ】



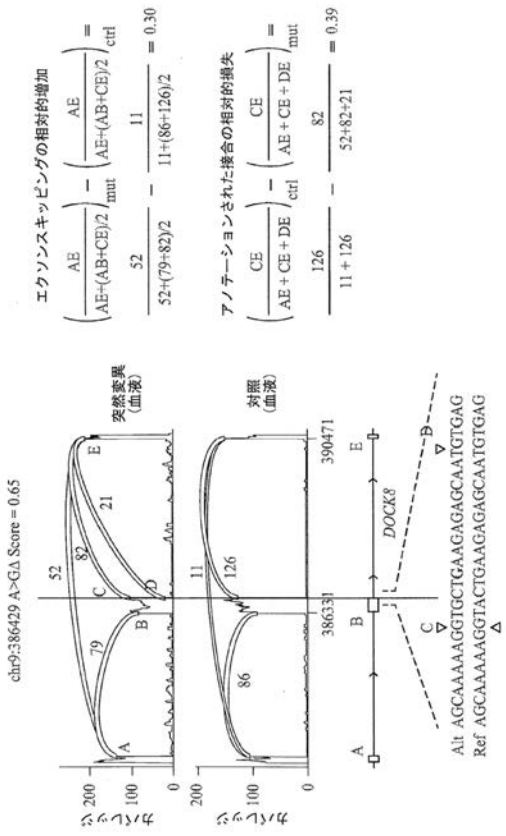
【図４４Ａ】



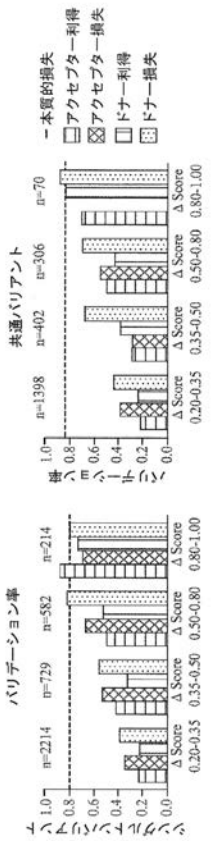
【図４４Ｂ】



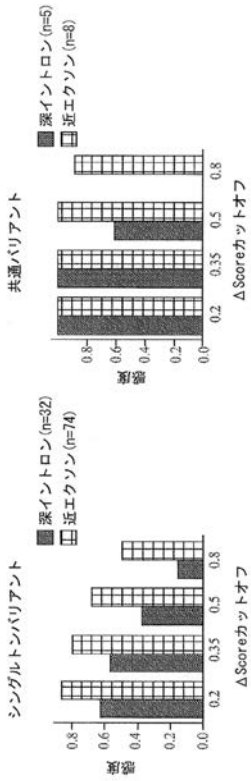
【図４５】



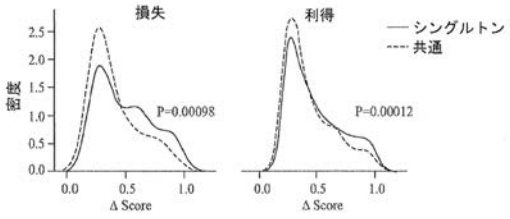
【図４６Ａ】



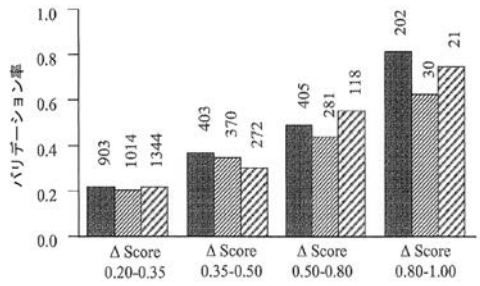
【図４６Ｂ】



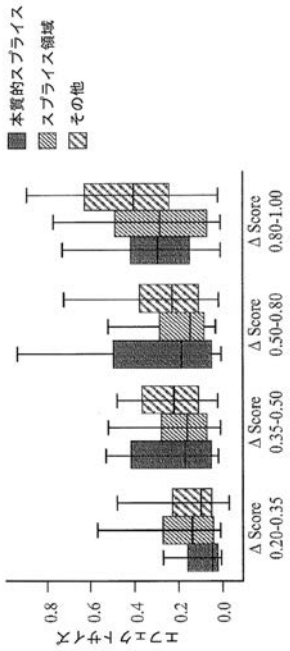
【図４６Ｃ】



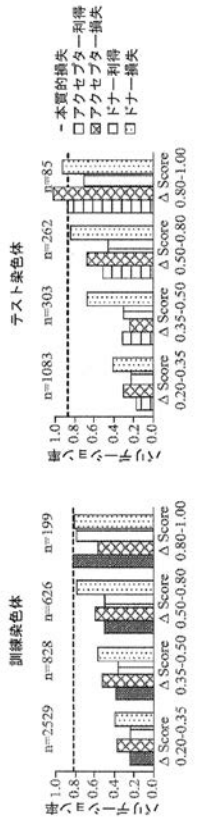
【図４７Ａ】



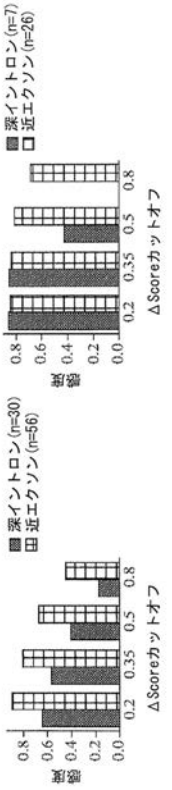
【図４７Ｂ】



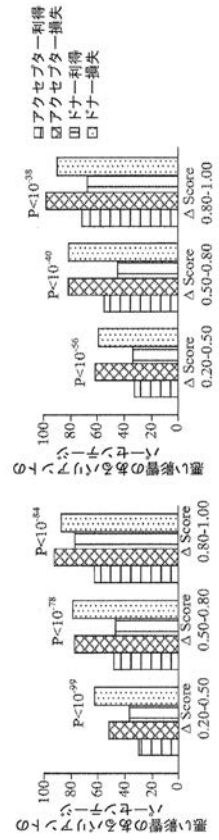
【図４８Ａ】



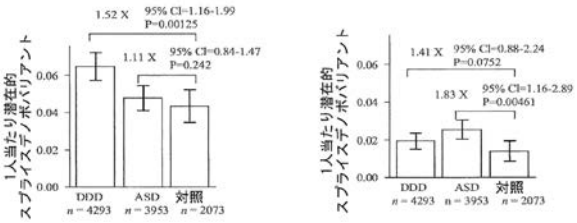
【図４８Ｂ】



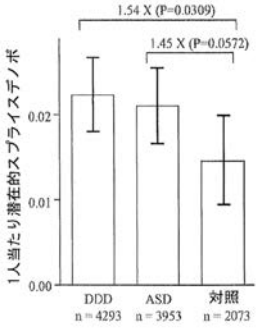
【図４８Ｃ】



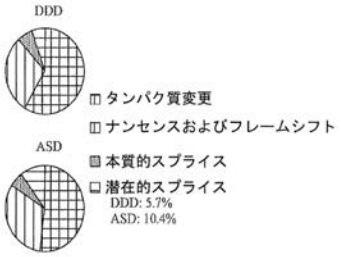
【図４８Ｄ】



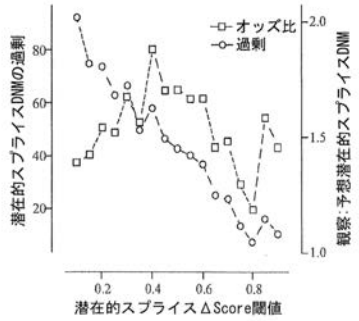
【図４９Ａ】



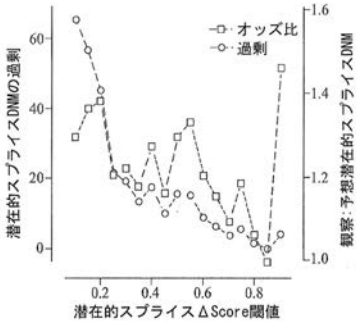
【図４９Ｂ】



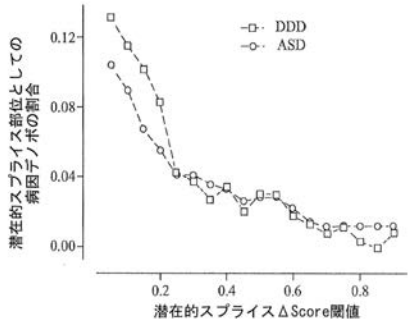
【図４９Ｃ】



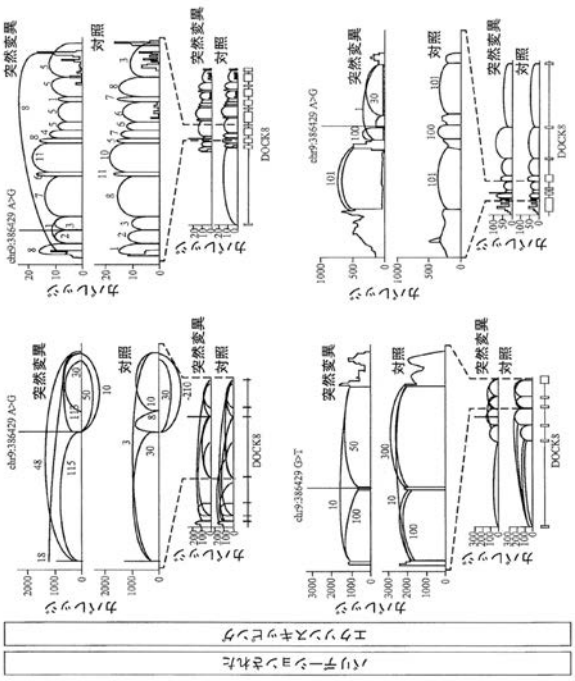
【図５０Ａ】



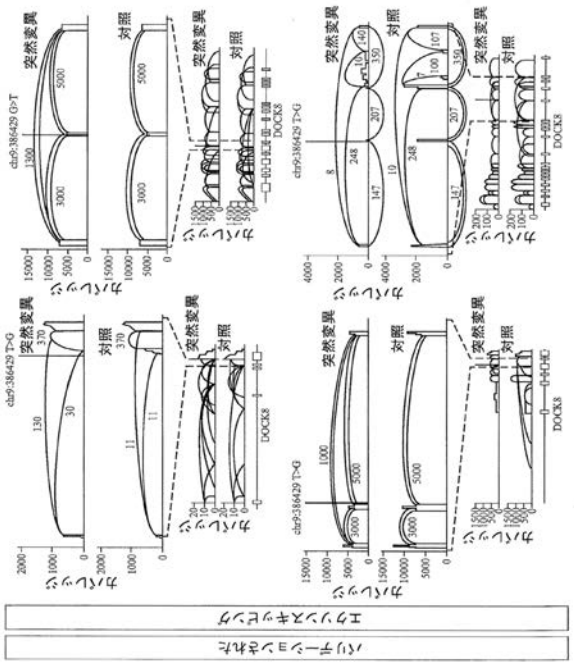
【図５０Ｂ】



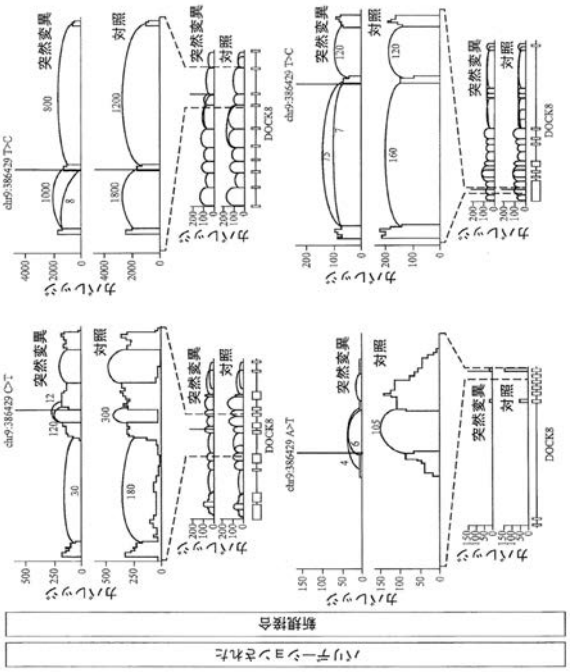
【図５１Ａ】



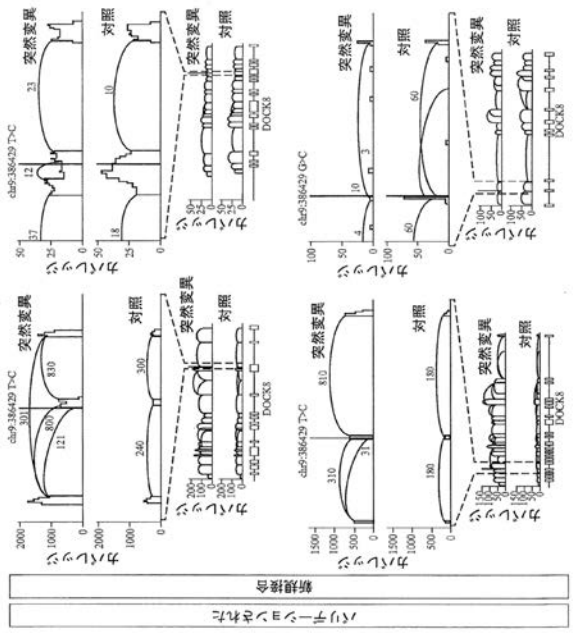
【図５１Ｂ】



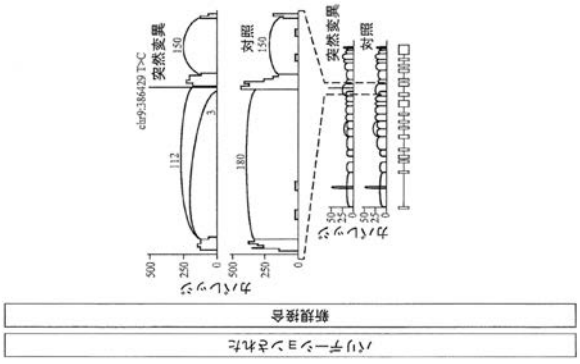
【図５１Ｃ】



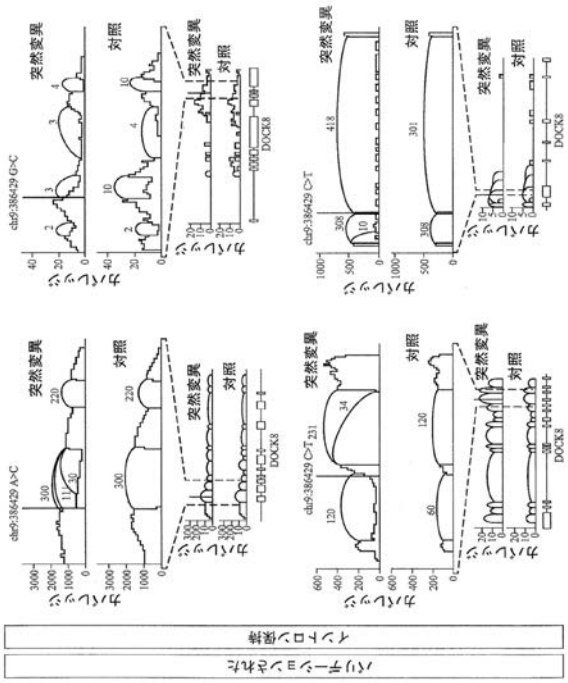
【図５１Ｄ】



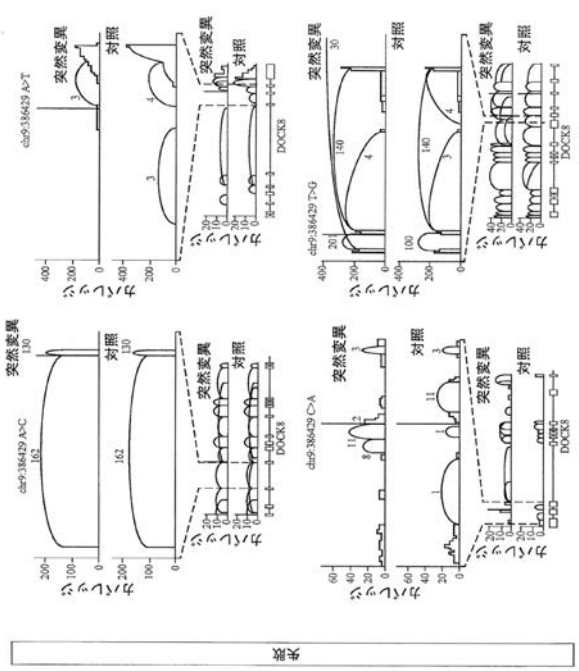
【図５１Ｅ】



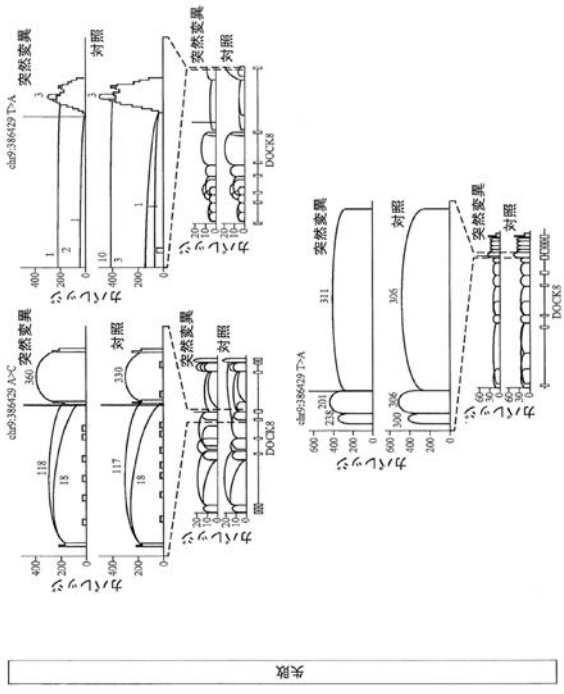
【図５１Ｆ】



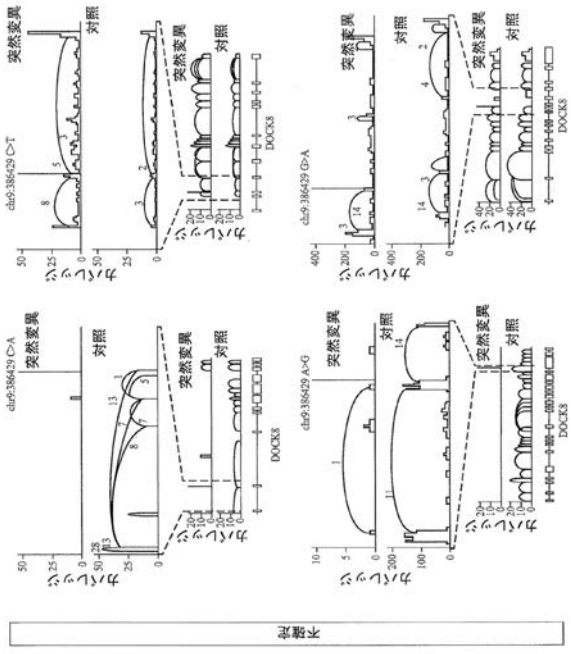
【図５１Ｇ】



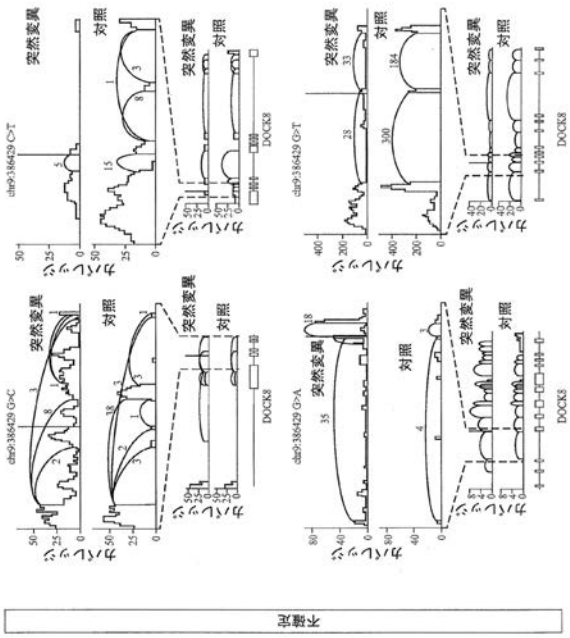
【図５１Ｈ】



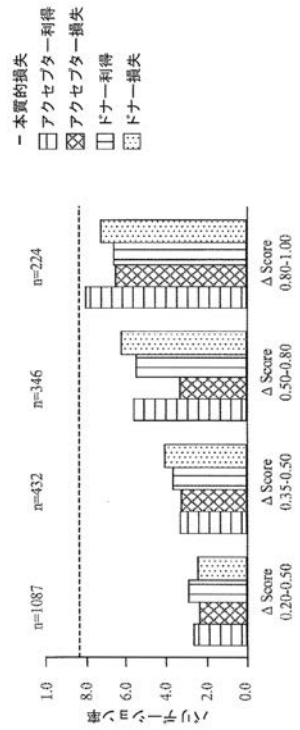
【図５１Ｉ】



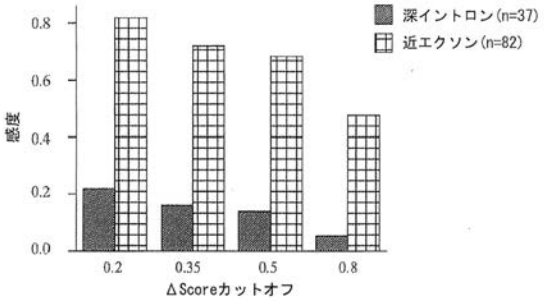
【図５１Ｊ】



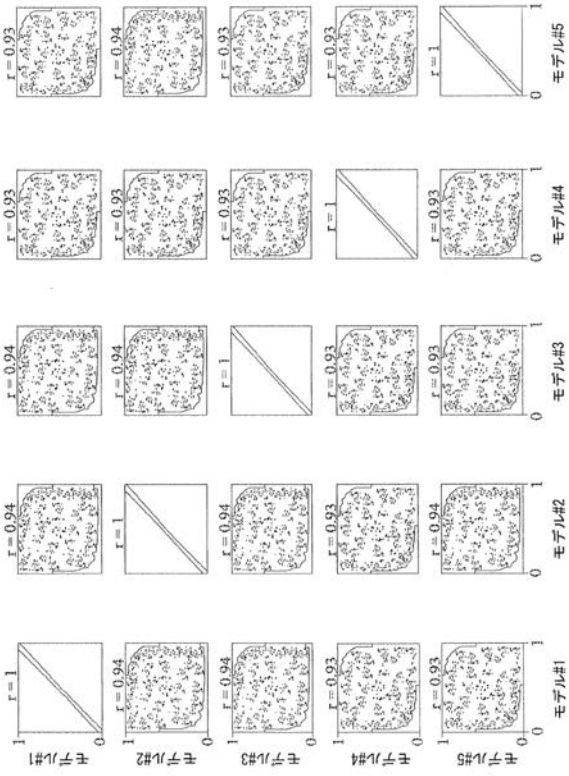
【図５２Ａ】



【図５２Ｂ】



【図５３Ｃ】



【図５３Ａ】

	Top-k精度		精度-再現率AUC	
	アクセプター	ドナー	アクセプター	ドナー
モデル#1	0.9304	0.9353	0.9630	0.9682
モデル#2	0.9297	0.9354	0.9649	0.9695
モデル#3	0.9288	0.9373	0.9650	0.9702
モデル#4	0.9222	0.9307	0.9616	0.9659
モデル#5	0.9267	0.9297	0.9640	0.9673

【図５３Ｂ】

	Top-k精度		精度-再現率AUC	
	アクセプター	ドナー	アクセプター	ドナー
1モデル	0.9304	0.9353	0.9630	0.9682
2モデル	0.9376	0.9428	0.9694	0.9742
3モデル	0.9409	0.9472	0.9715	0.9758
4モデル	0.9424	0.9474	0.9727	0.9765
5モデル	0.9429	0.9479	0.9737	0.9771

【図５４Ａ】

A

SpliceNet-10k

エクソン密度	Top-k精度 アクセプター	Top-k精度 ドナー	精度-再現率AUC アクセプター	精度-再現率AUC ドナー
1エクソン	0.9165	0.9292	0.9592	0.9659
2エクソン	0.9307	0.9357	0.9645	0.9674
3エクソン	0.9519	0.9512	0.9816	0.9822
4エクソン	0.9556	0.9540	0.9833	0.9845
≧5エクソン	0.9579	0.9665	0.9826	0.9872

【図５４Ｂ】

B

MaxEntScan

エクソン密度	Top-k精度 アクセプター	Top-k精度 ドナー	精度-再現率AUC アクセプター	精度-再現率AUC ドナー
1エクソン	0.1834	0.2487	0.1140	0.1660
2エクソン	0.2180	0.3124	0.1512	0.2375
3エクソン	0.2616	0.3348	0.1964	0.2773
4エクソン	0.2950	0.3820	0.2307	0.3347
≧5エクソン	0.3583	0.4652	0.3103	0.4314

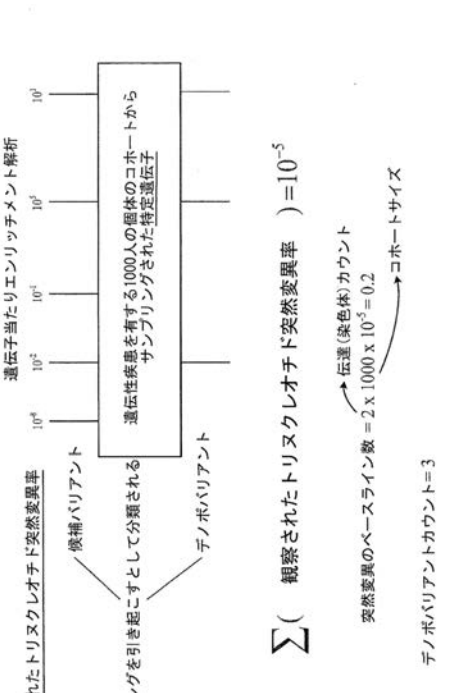
【図５５】

組織	バリアント個体	対照個体
皮膚-日光露光(下肢)	T6MO	OXRK
筋肉-骨格、形質転換線維芽細胞	WHSB	WK11
動脈-腰骨、肺	XOT4, S341, POMQ	WFON
全血	TMMY	T8EM

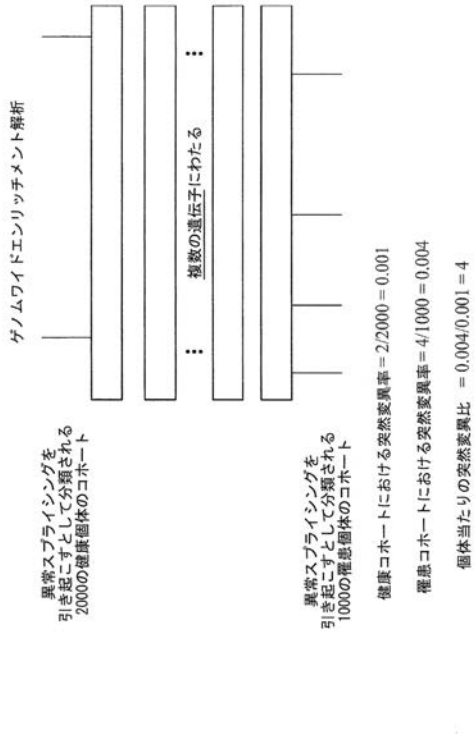
【図５６】

利得 SpliceNet-10k	利得 GeneSplicer	利得 NNSplice	利得 MaxEntScan	損失 SpliceNet-10k	損失 GeneSplicer	損失 NNSplice	損失 MaxEntScan
0.1	103.5145	0.9024	0.901	0.1	1.5183	0.1063	0
0.2	105.6838	0.9288	0.9171	0.2	2.6295	0.2315	0.0068
0.3	107.1061	0.9542	0.9419	0.3	3.4707	0.3554	0.0312
0.4	108.0655	0.967	0.9582	0.4	4.0856	0.4625	0.0646
0.5	108.9011	0.9739	0.9694	0.5	4.6085	0.5349	0.1003
0.6	109.6138	0.9781	0.9759	0.6	5.2064	0.6345	0.1624
0.7	110.388	0.9817	0.9826	0.7	6.4466	0.7573	0.2838
0.8	111.2775	0.9839	0.9875	0.8	8.1053	0.8202	0.4271
0.9	112.6826	0.9858	0.9915	0.9	10.2586	0.8529	0.641

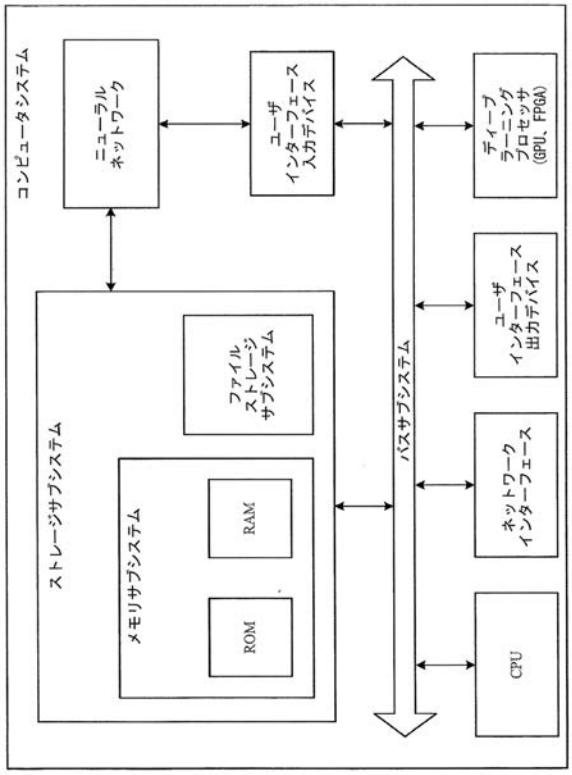
【図５７】



【図５８】


ゲノムワイドエンリッチメント解析
異常スプライシングを
引き起こすとして分類される
2000の健康個体のコホート
複数の遺伝子にわたる
異常スプライシングを
引き起こすとして分類される
1000の罹患個体のコホート
健康コホートにおける突然変異率 = 2/2000 = 0.001
罹患コホートにおける突然変異率 = 4/1000 = 0.004
個体当たりの突然変異比 = 0.004/0.001 = 4

【図５９】


コンピュータシステム
ニューラル
ネットワーク
ユーザ
インターフェース
入力デバイス
ディープ
ラーニング
プロセッサ
(GPU、FPGA)
ストレージサブシステム
メモリサブシステム
ROM
RAM
ファイル
ストレージ
サブシステム
バスサブシステム
ユーザ
インターフェース
出力デバイス
ネットワーク
インターフェース
CPU

フロントページの続き

(31)優先権主張番号 62/573,135
(32)優先日 平成29年10月16日(2017.10.16)
(33)優先権主張国・地域又は機関
米国(US)
(31)優先権主張番号 62/726,158
(32)優先日 平成30年8月31日(2018.8.31)
(33)優先権主張国・地域又は機関
米国(US)

(72)発明者 カイーハウ・ファー
アメリカ合衆国・カリフォルニア・９２１２２・サン・ディエゴ・イルミナ・ウェイ・５２００
(72)発明者 ソフィア・キリアゾポウロウ・パナジオトポウロウ
アメリカ合衆国・カリフォルニア・９２１２２・サン・ディエゴ・イルミナ・ウェイ・５２００
(72)発明者 ジェレミー・フランシス・マクレー
アメリカ合衆国・カリフォルニア・９２１２２・サン・ディエゴ・イルミナ・ウェイ・５２００

【外国語明細書】
2021007035000001.pdf