(19)日本国特許庁(JP)　(12)公表特許公報(A)　(11)公表番号
特表2025-502816
(P2025-502816A)
(43)公表日　令和7年1月28日(2025.1.28)

(51)国際特許分類			FI			テーマコード（参考）
Ｇ１６Ｂ	*40/00*	*(2019.01)*	Ｇ１６Ｂ	40/00		４Ｂ０２９
Ｃ１２Ｍ	*1/00*	*(2006.01)*	Ｃ１２Ｍ	1/00	Ａ	４Ｂ０６３
Ｃ１２Ｍ	*1/34*	*(2006.01)*	Ｃ１２Ｍ	1/34	Ｚ	
Ｃ１２Ｑ	*1/68*	*(2018.01)*	Ｃ１２Ｑ	1/68		

審査請求　未請求　予備審査請求　未請求　（全68頁）

(21)出願番号　特願2024-539696(P2024-539696)
(86)(22)出願日　令和4年12月28日(2022.12.28)
(85)翻訳文提出日　令和6年6月28日(2024.6.28)
(86)国際出願番号　PCT/US2022/054194
(87)国際公開番号　WO2023/129621
(87)国際公開日　令和5年7月6日(2023.7.6)
(31)優先権主張番号　63/294,813
(32)優先日　令和3年12月29日(2021.12.29)
(33)優先権主張国・地域又は機関
米国(US)
(31)優先権主張番号　63/294,816
(32)優先日　令和3年12月29日(2021.12.29)
(33)優先権主張国・地域又は機関
米国(US)
(31)優先権主張番号　63/294,820
最終頁に続く

(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　ペトコ・プラメノフ・フィジエフ
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・イルミナ・インコーポ
最終頁に続く

(54)【発明の名称】　希少バリアント多遺伝子リスクスコア

(57)【要約】
指定された表現型を有する複数の関連遺伝子についての加重総和スコア（１６２）を計算することによって、表現型と関連付けられた遺伝子の関連性の強度、及び表現型応答に対する希少バリアントの寄与を定量化するコンピュータ実装方法であって、総和スコア（１６２）は、関連遺伝子の各々のキャリア状態と、関連遺伝子における１つ以上の希少病原性バリアントの存在に対する表現型応答との間のコホート（１０４）における結果的な非ランダムな関連性を特定することを特定する、コンピュータ実装方法。それぞれの有効強度スコアが、遺伝子ごとの解像度におけるそれぞれの総和スコア（１６２）に基づいて、関連遺伝子から選択された遺伝子についての結果的な非ランダムな関連性について決定される（１４２）。

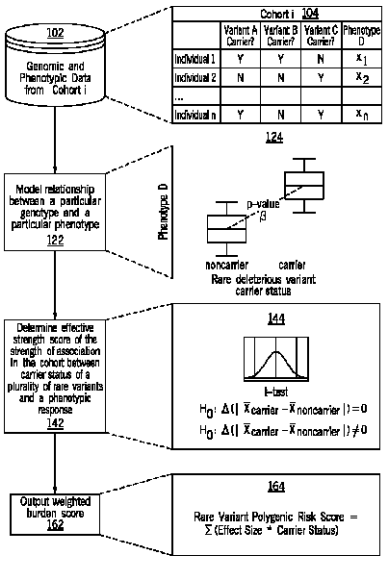


Figure 1A

【特許請求の範囲】
【請求項１】
表現型に対する希少病原性バリアントの効果を決定するコンピュータ実装方法であって、
特定の表現型と関連付けられた特定の遺伝子について：
前記特定の遺伝子における希少病原性バリアントの特定の群を特定することと、
前記希少病原性バリアントの特定の群からの少なくとも１つの希少病原性バリアントを保有する個体のコホートにおけるキャリアを特定することと、
前記希少病原性バリアントの特定の群からのいかなる希少病原性バリアントも保有しない前記個体のコホートにおける非キャリアを特定することと、
総和検定を使用して、前記キャリアを前記非キャリアから分離するキャリア状態に依存して、前記特定の表現型に対する前記希少病原性バリアントの特定の群の効果量を決定することと、を含む、コンピュータ実装方法。
【請求項２】
前記特定の表現型が、定量的バイオマーカー表現型である、請求項１に記載のコンピュータ実装方法。
【請求項３】
前記希少病原性バリアントの特定の群の前記効果量が、前記総和検定の線形回帰成分に対する両側ｔ検定を使用して前記定量的バイオマーカー表現型について決定される、請求項２に記載のコンピュータ実装方法。
【請求項４】
前記両側ｔ検定が、前記キャリア及び前記非キャリアの平均表現型測定値間の差を決定する、請求項３に記載のコンピュータ実装方法。
【請求項５】
前記両側ｔ検定が、２つのｐ値を生成する、請求項４に記載のコンピュータ実装方法。
【請求項６】
前記表現型測定値が、薬物使用量補正される、請求項４に記載のコンピュータ実装方法。
【請求項７】
前記表現型測定値が、共変量補正され、正規化される、請求項４に記載のコンピュータ実装方法。
【請求項８】
前記表現型測定値が、一般的バリアント補正される、請求項４に記載のコンピュータ実装方法。
【請求項９】
前記特定の表現型が、カテゴリ臨床診断表現型である、請求項１に記載のコンピュータ実装方法。
【請求項１０】
前記希少病原性バリアントの特定の群の前記効果量が、前記キャリア状態についてのベータ係数として前記カテゴリ臨床診断表現型に対して決定される、請求項９に記載のコンピュータ実装方法。
【請求項１１】
前記ベータ係数が、前記総和検定のロジスティック回帰成分を使用して決定される、請求項１０に記載のコンピュータ実装方法。
【請求項１２】
前記ロジスティック回帰成分が、前記キャリア状態及び複数の共変量から臨床診断ラベルを予測するようにフィッティングされる、請求項１１に記載のコンピュータ実装方法。
【請求項１３】
前記ロジスティック回帰成分が、前記キャリア状態をバイナリインジケータ変数として符号化する、請求項１２に記載のコンピュータ実装方法。

【請求項１４】
前記ロジスティック回帰成分が、前記複数の共変量を除去して回帰分析する、請求項１２に記載のコンピュータ実装方法。
【請求項１５】
前記複数の共変量が、年齢、性別、遺伝的主成分、民族性、一般的バリアント、及びバイアス項を含む、請求項１２に記載のコンピュータ実装方法。
【請求項１６】
前記ロジスティック回帰成分が、ｐ値を生成する、請求項１２に記載のコンピュータ実装方法。
【請求項１７】
前記希少病原性バリアントの特定の群が、対立遺伝子頻度に適用される異なる希少性閾値を使用して、異なる総和検定について特定される、請求項１に記載のコンピュータ実装方法。
【請求項１８】
前記希少病原性バリアントの特定の群が、０．００１未満である対立遺伝子頻度を有する、請求項１７に記載のコンピュータ実装方法。
【請求項１９】
前記希少病原性バリアントの特定の群が、病原性スコアに適用される異なる病原性閾値を使用して、異なる総和検定に対して特定される、請求項１に記載のコンピュータ実装方法。
【請求項２０】
前記病原性スコアが、畳み込みニューラルネットワークによって生成される、請求項１９に記載のコンピュータ実装方法。
【請求項２１】
前記異なる病原性閾値が、異なるパーセンタイル閾値である、請求項１９に記載のコンピュータ実装方法。
【請求項２２】
それぞれの総和検定を使用して、前記特定の表現型と関連付けられたそれぞれの遺伝子における希少病原性バリアントのそれぞれの群のそれぞれの効果量を決定することを更に含む、請求項１に記載のコンピュータ実装方法。
【請求項２３】
前記それぞれの効果量の加重和に基づいて、前記特定の表現型についての、及び前記個体のコホート内の特定の個体についての希少バリアント多遺伝子リスクスコアを生成することを更に含む、請求項２２に記載のコンピュータ実装方法。
【請求項２４】
前記それぞれの効果量の前記加重和が、前記それぞれの遺伝子にわたる前記特定の個体の前記キャリア状態によって重み付けされる、請求項２３に記載のコンピュータ実装方法。
【請求項２５】
前記特定の個体は、前記特定の個体が前記希少病原性バリアントの特定の群からの少なくとも１つの希少病原性バリアントを保有する場合、前記特定の遺伝子についてのキャリアの前記キャリア状態を有する、請求項２４に記載のコンピュータ実装方法。
【請求項２６】
前記特定の個体が前記キャリアの前記キャリア状態を有する場合、１の重みが使用される、請求項２５に記載のコンピュータ実装方法。
【請求項２７】
前記特定の個体は、前記特定の個体が前記希少病原性バリアントの特定の群からのいかなる希少病原性バリアントも保有しない場合、前記特定の遺伝子についての非キャリアの前記キャリア状態を有する、請求項２４に記載のコンピュータ実装方法。
【請求項２８】

前記特定の個体が前記非キャリアの前記キャリア状態を有する場合、ゼロの重みが使用される、請求項２７に記載のコンピュータ実装方法。
【請求項２９】
前記希少病原性バリアントの特定の群における希少病原性バリアントが、機能喪失バリアントである、請求項１に記載のコンピュータ実装方法。
【請求項３０】
前記希少病原性バリアントが、ミスセンスバリアントである、請求項２９に記載のコンピュータ実装方法。
【請求項３１】
別個の総和検定を使用して、前記表現型に対する前記機能喪失バリアント及び前記ミスセンスバリアントの効果を別個に決定することを更に含む、請求項３０に記載のコンピュータ実装方法。
【請求項３２】
線形回帰成分及びロジスティック回帰成分の訓練された係数を使用して、検定データについての希少バリアント多遺伝子リスクスコアを生成することを更に含む、請求項１に記載のコンピュータ実装方法。
【請求項３３】
前記検定データが、外れ値表現型測定値を有する個体を含む、請求項３２に記載のコンピュータ実装方法。
【請求項３４】
表現型に対する希少病原性バリアントの効果を決定するコンピュータ実装方法であって、
総和検定を使用して、特定の表現型と関連付けられたそれぞれの遺伝子における希少病原性バリアントのそれぞれの群のそれぞれの効果量を決定することと、
前記それぞれの効果量の加重和に基づいて、前記特定の表現型についての、及び個体のコホート内の特定の個体についての希少バリアント多遺伝子リスクスコアを生成することと、を含み、
前記それぞれの効果量の前記加重和が、前記それぞれの遺伝子にわたる前記特定の個体のキャリア状態によって重み付けされる、コンピュータ実装方法。
【請求項３５】
メモリに結合された１つ以上のプロセッサを含むシステムであって、前記メモリには、表現型に対する希少病原性バリアントの効果を決定するためのコンピュータ命令がロードされ、前記命令が、前記プロセッサ上で実行されると、
特定の表現型と関連付けられた特定の遺伝子について：
前記特定の遺伝子における希少病原性バリアントの特定の群を特定することと、
前記希少病原性バリアントの特定の群からの少なくとも１つの希少病原性バリアントを保有する個体のコホートにおけるキャリアを特定することと、
前記希少病原性バリアントの特定の群からのいかなる希少病原性バリアントも保有しない前記個体のコホートにおける非キャリアを特定することと、
総和検定を使用して、前記キャリアを前記非キャリアから分離するキャリア状態に依存して、前記特定の表現型に対する前記希少病原性バリアントの特定の群の効果量を決定することと、を含むアクションを実装する、システム。
【請求項３６】
表現型に対する希少病原性バリアントの効果を決定するためのコンピュータプログラム命令を記憶している非一時的コンピュータ可読記憶媒体であって、前記命令が、プロセッサ上で実行されると、
特定の表現型と関連付けられた特定の遺伝子について：
前記特定の遺伝子における希少病原性バリアントの特定の群を特定することと、
前記希少病原性バリアントの特定の群からの少なくとも１つの希少病原性バリアントを保有する個体のコホートにおけるキャリアを特定することと、

前記希少病原性バリアントの特定の群からのいかなる希少病原性バリアントも保有しない前記個体のコホートにおける非キャリアを特定することと、
総和検定を使用して、前記キャリアを前記非キャリアから分離するキャリア状態に依存して、前記特定の表現型に対する前記希少病原性バリアントの特定の群の効果量を決定することと、を含む方法を実装する、非一時的コンピュータ可読記憶媒体。
【発明の詳細な説明】
【技術分野】
【０００１】
（優先権出願）
本出願は、以下の利益及び優先権を主張する。
２０２１年１２月２９日に出願された「ＰＥＲＩＯＤＩＣ ＭＡＳＫ ＰＡＴＴＥＲＮ ＦＯＲ ＲＥＶＥＬＡＴＩＯＮ ＬＡＮＧＵＡＧＥ ＭＯＤＥＬＳ」と題する米国仮特許出願第６３／２９４，８１３号（代理人整理番号ＩＬＬＭ１０６３－１／ＩＰ－２２９６－ＰＲＶ）、
２０２１年１２月２９日に出願された「ＣＬＡＳＳＩＦＹＩＮＧ ＭＩＬＬＩＯＮＳ ＯＦ ＶＡＲＩＡＮＴＳ ＯＦ ＵＮＣＥＲＴＡＩＮ ＳＩＧＮＩＦＩＣＡＮＣＥ ＵＳＩＮＧ ＰＲＩＭＡＴＥ ＳＥＱＵＥＮＣＩＮＧ ＡＮＤ ＤＥＥＰ ＬＥＡＲＮＩＮＧ」と題する米国仮特許出願第６３／２９４，８１６号（代理人整理番号ＩＬＬＭ１０６４－１／ＩＰ－２２９７－ＰＲＶ）、
２０２１年１２月２９日に出願された「ＩＤＥＮＴＩＦＹＩＮＧ ＧＥＮＥＳ ＷＩＴＨ ＤＩＦＦＥＲＥＮＴＩＡＬ ＳＥＬＥＣＴＩＶＥ ＣＯＮＳＴＲＡＩＮＴ ＢＥＴＷＥＥＮ ＨＵＭＡＮＳ ＡＮＤ ＮＯＮ－ＨＵＭＡＮ ＰＲＩＭＡＴＥＳ」と題する米国仮特許出願第６３／２９４，８２０号（代理人整理番号ＩＬＬＭ１０６５－１／ＩＰ－２２９８－ＰＲＶ）、
２０２１年１２月２９日に出願された「ＤＥＥＰ ＬＥＡＲＮＩＮＧ ＮＥＴＷＯＲＫ ＦＯＲ ＥＶＯＬＵＴＩＯＮＡＲＹ ＣＯＮＳＥＲＶＡＴＩＯＮ」と題する米国非仮特許出願第６３／２９４，８２７号（代理人整理番号ＩＬＬＭ１０６６－１／ＩＰ－２２９９－ＰＲＶ）、
２０２１年１２月２９日に出願された「ＩＮＴＥＲ－ＭＯＤＥＬ ＰＲＥＤＩＣＴＩＯＮ ＳＣＯＲＥ ＲＥＣＡＬＩＢＲＡＴＩＯＮ」と題する米国仮特許出願第６３／２９４，８２８号（代理人整理番号ＩＬＬＭ１０６７－１／ＩＰ－２３０１－ＰＲＶ）、
２０２１年１２月２９日に出願された「ＳＰＥＣＩＥＳ－ＤＩＦＦＥＲＥＮＴＩＡＢＬＥ ＥＶＯＬＵＴＩＯＮＡＲＹ ＰＲＯＦＩＬＥＳ」と題する米国仮特許出願第６３／２９４，８３０号（代理人整理番号ＩＬＬＭ１０６８－１／ＩＰ－２３０２－ＰＲＶ）、
２０２２年６月１０日に出願された「ＯＰＴＩＭＩＺＥＤ ＢＵＲＤＥＮ ＴＥＳＴ ＢＡＳＥＤ ＯＮ ＮＥＳＴＥＤ Ｔ－ＴＥＳＴＳ ＴＨＡＴ ＭＡＸＩＭＩＺＥ ＳＥＰＡＲＡＴＩＯＮ ＢＥＴＷＥＥＮ ＣＡＲＲＩＥＲＳ ＡＮＤ ＮＯＮ－ＣＡＲＲＩＥＲＳ」と題する米国仮特許出願第６３／３５１，２８３号（代理人整理番号ＩＬＬＭ１０７０－１／ＩＰ－２３６８－ＰＲＶ）、
２０２２年６月１０日に出願された「ＲＡＲＥ ＶＡＲＩＡＮＴ ＰＯＬＹＧＥＮＩＣ ＲＩＳＫ ＳＣＯＲＥＳ」と題する米国仮特許出願第６３／３５１，２９９号（代理人整理番号ＩＬＬＭ１０７１－１／ＩＰ－２３７８－ＰＲＶ）、及び
２０２２年６月１０日に出願された「ＣＯＶＡＲＩＡＴＥ ＣＯＲＲＥＣＴＩＯＮ ＩＮＣＬＵＤＩＮＧ ＤＲＵＧ ＵＳＥ ＦＲＯＭ ＴＥＭＰＯＲＡＬ ＤＡＴＡ」と題する米国仮特許出願第６３／３５１，３１７号（代理人整理番号ＩＬＬＭ１０７３－１／ＩＰ－２３８７－ＰＲＶ）。
【０００２】
優先権出願は、本明細書に完全に記載されているかのように参照によって組み込まれる。
【０００３】

（発明の分野）
開示される技術は、人工知能型コンピュータ及びデジタルデータ処理システム、並びに知能（すなわち、知識ベースのシステム、推論システム、及び知識取得システム）を模倣するための対応するデータ処理方法及び製品に関し、不確実性を伴う推論のためのシステム（例えば、ファジー論理システム）、適応システム、機械学習システム、及び人工ニューラルネットワークを含む。具体的には、開示される技術は、順序付きデータを分析するための深層畳み込みニューラルネットワークの使用に関する。

【０００４】
（関連出願）
本出願は、２０２２年１０月１８日に出願された「ＯＰＴＩＭＩＺＥＤ ＢＵＲＤＥＮ ＴＥＳＴ ＢＡＳＥＤ ＯＮ ＮＥＳＴＥＤ Ｔ－ＴＥＳＴＳ ＴＨＡＴ ＭＡＸＩＭＩＺＥ ＳＥＰＡＲＡＴＩＯＮ ＢＥＴＷＥＥＮ ＣＡＲＲＩＥＲＳ ＡＮＤ ＮＯＮ－ＣＡＲＲＩＥＲＳ」と題する米国非仮特許出願第１７／９６８，２８５号（代理人整理番号ＩＬＬＭ１０７０－２／ＩＰ－２３６８－ＵＳ）に関連する。関連出願は、全ての目的のために参照により本明細書に組み込まれる。

【０００５】
本出願は、同時に出願された「ＣＯ ＶＡＲＩＡＴＥ ＣＯＲＲＥＣＴＩＯＮ ＩＮＣＬＵＤＩＮＧ ＤＲＵＧ ＵＳＥ ＦＲＯＭ ＴＥＭＰＯＲＡＬ ＤＡＴＡ」と題する米国非仮特許出願（代理人整理番号ＩＬＬＭ１０７３－２／ＩＰ－２３８７－ＵＳ）に関する。関連出願は、全ての目的のために参照により本明細書に組み込まれる。

【０００６】
（参照による組み込み）
以下は、本明細書に完全に記載されているかのように、全ての目的のために参照により組み込まれる。

Ｓｕｎｄａｒａｍ，Ｌ．ｅｔ ａｌ．Ｐｒｅｄｉｃｔｉｎｇ ｔｈｅ ｃｌｉｎｉｃａｌ ｉｍｐａｃｔ ｏｆ ｈｕｍａｎ ｍｕｔａｔｉｏｎ ｗｉｔｈ ｄｅｅｐ ｎｅｕｒａｌ ｎｅｔｗｏｒｋｓ．Ｎａｔ．Ｇｅｎｅｔ．５０，１１６１－１１７０（２０１８）、

Ｊａｇａｎａｔｈａｎ，Ｋ．ｅｔ ａｌ．Ｐｒｅｄｉｃｔｉｎｇ ｓｐｌｉｃｉｎｇ ｆｒｏｍ ｐｒｉｍａｒｙ ｓｅｑｕｅｎｃｅ ｗｉｔｈ ｄｅｅｐ ｌｅａｒｎｉｎｇ．Ｃｅｌｌ １７６，５３５－５４８（２０１９）、

２０１７年１０月１６日に出願された「ＴＲＡＩＮＩＮＧ Ａ ＤＥＥＰ ＰＡＴＨＯＧＥＮＩＣＩＴＹ ＣＬＡＳＳＩＦＩＥＲ ＵＳＩＮＧ ＬＡＲＧＥ－ＳＣＡＬＥ ＢＥＮＩＧＮ ＴＲＡＩＮＩＮＧ ＤＡＴＡ」と題する米国特許出願第６２／５７３，１４４号（代理人整理番号ＩＬＬＭ１０００－１／ＩＰ－１６１１－ＰＲＶ）、

２０１７年１０月１６日に出願された「ＰＡＴＨＯＧＥＮＩＣＩＴＹ ＣＬＡＳＳＩＦＩＥＲ ＢＡＳＥＤ ＯＮ ＤＥＥＰ ＣＯＮＶＯＬＵＴＩＯＮＡＬ ＮＥＵＲＡＬ ＮＥＴＷＯＲＫＳ（ＣＮＮｓ）」と題する米国特許出願第６２／５７３，１４９号（代理人整理番号ＩＬＬＭ１０００－２／ＩＰ－１６１２－ＰＲＶ）、

２０１７年１０月１６日に出願された「ＤＥＥＰ ＳＥＭＩ－ＳＵＰＥＲＶＩＳＥＤ ＬＥＡＲＮＩＮＧ ＴＨＡＴ ＧＥＮＥＲＡＴＥＳ ＬＡＲＧＥ－ＳＣＡＬＥ ＰＡＴＨＯＧＥＮＩＣ ＴＲＡＩＮＩＮＧ ＤＡＴＡ」と題する米国特許出願第６２／５７３，１５３号（代理人整理番号ＩＬＬＭ１０００－３／ＩＰ－１６１３－ＰＲＶ）、

２０１７年１１月７日に出願された「ＰＡＴＨＯＧＥＮＩＣＩＴＹ ＣＬＡＳＳＩＦＩＣＡＴＩＯＮ ＯＦ ＧＥＮＯＭＩＣ ＤＡＴＡ ＵＳＩＮＧ ＤＥＥＰ ＣＯＮＶＯＬＵＴＩＯＮＡＬ ＮＥＵＲＡＬ ＮＥＴＷＯＲＫＳ（ＣＮＮｓ）」と題する米国特許出願第６２／５８２，８９８号（代理人整理番号ＩＬＬＭ １０００－４／ＩＰ－１６１８－ＰＲＶ）、

２０１８年１０月１５日に出願された「ＤＥＥＰ ＬＥＡＲＮＩＮＧ－ＢＡＳＥＤ ＴＥＣＨＮＩＱＵＥＳ ＦＯＲ ＴＲＡＩＮＩＮＧ ＤＥＥＰ ＣＯＮＶＯＬＵＴＩＯＮＡＬ ＮＥＵＲＡＬ ＮＥＴＷＯＲＫＳ」と題する米国特許出願第１６／１６０，９０３号

（代理人整理番号ＩＬＬＭ１０００－５／ＩＰ－１６１１－ＵＳ）、
２０１８年１０月１５日に出願された「ＤＥＥＰ ＣＯＮＶＯＬＵＴＩＯＮＡＬ ＮＥＵＲＡＬ ＮＥＴＷＯＲＫＳ ＦＯＲ ＶＡＲＩＡＮＴ ＣＬＡＳＳＩＦＩＣＡＴＩＯＮ」と題する米国特許出願第１６／１６０，９８６号（代理人整理番号ＩＬＬＭ１０００－６／ＩＰ－１６１２－ＵＳ）、
２０１８年１０月１５日に出願された「ＳＥＭＩ－ＳＵＰＥＲＶＩＳＥＤ ＬＥＡＲＮＩＮＧ ＦＯＲ ＴＲＡＩＮＩＮＧ ＡＮ ＥＮＳＥＭＢＬＥ ＯＦ ＤＥＥＰ ＣＯＮＶＯＬＵＴＩＯＮＡＬ ＮＥＵＲＡＬ ＮＥＴＷＯＲＫＳ」と題する米国特許出願第１６／１６０，９６８号（代理人整理番号ＩＬＬＭ１０００－７／ＩＰ－１６１３－ＵＳ）、
２０１８年１０月１５日に出願された「ＤＥＥＰ ＬＥＡＲＮＩＮＧ－ＢＡＳＥＤ ＳＰＬＩＣＥ ＳＩＴＥ ＣＬＡＳＳＩＦＩＣＡＴＩＯＮ」と題する米国特許出願第１６／１６０，９７８号（代理人整理番号ＩＬＬＭ１００１－４／ＩＰ－１６８０－ＵＳ）、
２０１９年５月８日に出願された「ＤＥＥＰ ＬＥＡＲＮＩＮＧ－ＢＡＳＥＤ ＴＥＣＨＮＩＱＵＥＳ ＦＯＲ ＰＲＥ－ＴＲＡＩＮＩＮＧ ＤＥＥＰ ＣＯＮＶＯＬＵＴＩＯＮＡＬ ＮＥＵＲＡＬ ＮＥＴＷＯＲＫＳ」と題する米国特許出願第１６／４０７，１４９号（代理人整理番号ＩＬＬＭ１０１０－１／ＩＰ－１７３４－ＵＳ）、
２０２１年４月１５日に出願された「ＤＥＥＰ ＣＯＮＶＯＬＵＴＩＯＮＡＬ ＮＥＵＲＡＬ ＮＥＴＷＯＲＫＳ ＴＯ ＰＲＥＤＩＣＴ ＶＡＲＩＡＮＴ ＰＡＴＨＯＧＥＮＩＣＩＴＹ ＵＳＩＮＧ ＴＨＲＥＥ－ＤＩＭＥＮＳＩＯＮＡＬ（３Ｄ）ＰＲＯＴＥＩＮ ＳＴＲＵＣＴＵＲＥＳ」と題する米国特許出願第１７／２３２，０５６号（代理人整理番号ＩＬＬＭ１０３７－２／ＩＰ－２０５１－ＵＳ）、
２０２１年４月１５日に出願された「ＭＵＬＴＩ－ＣＨＡＮＮＥＬ ＰＲＯＴＥＩＮ ＶＯＸＥＬＩＺＡＴＩＯＮ ＴＯ ＰＲＥＤＩＣＴ ＶＡＲＩＡＮＴ ＰＡＴＨＯＧＥＮＩＣＩＴＹ ＵＳＩＮＧ ＤＥＥＰ ＣＯＮＶＯＬＵＴＩＯＮＡＬ ＮＥＵＲＡＬ ＮＥＴＷＯＲＫＳ」と題する米国特許出願第６３／１７５，４９５号（代理人整理番号ＩＬＬＭ１０４７－１／ＩＰ－２１４２－ＰＲＶ）、
２０２１年４月１６日に出願された「ＥＦＦＩＣＩＥＮＴ ＶＯＸＥＬＩＺＡＴＩＯＮ ＦＯＲ ＤＥＥＰ ＬＥＡＲＮＩＮＧ」と題する米国特許出願第６３／１７５，７６７号（代理人整理番号ＩＬＬＭ１０４８－１／ＩＰ－２１４３－ＰＲＶ）、
２０２１年９月７日に出願された「ＡＲＴＩＦＩＣＩＡＬ ＩＮＴＥＬＬＩＧＥＮＣＥ－ＢＡＳＥＤ ＡＮＡＬＹＳＩＳ ＯＦ ＰＲＯＴＥＩＮ ＴＨＲＥＥ－ＤＩＭＥＮＳＩＯＮＡＬ（３Ｄ）ＳＴＲＵＣＴＵＲＥＳ」と題する米国特許出願第１７／４６８，４１１号（代理人整理番号ＩＬＬＭ１０３７－３／ＩＰ－２０５１Ａ－ＵＳ）、
２０２１年１０月６日に出願された「ＰＲＯＴＥＩＮ ＳＴＲＵＣＴＵＲＥ－ＢＡＳＥＤ ＰＲＯＴＥＩＮ ＬＡＮＧＵＡＧＥ ＭＯＤＥＬＳ」と題する米国仮特許出願第６３／２５３，１２２号（代理人整理番号ＩＬＬＭ１０５０－１／ＩＰ－２１６４－ＰＲＶ）、
２０２１年１１月１９日に出願された「ＰＲＥＤＩＣＴＩＮＧ ＶＡＲＩＡＮＴ ＰＡＴＨＯＧＥＮＩＣＩＴＹ ＦＲＯＭ ＥＶＯＬＵＴＩＯＮＡＲＹ ＣＯＮＳＥＲＶＡＴＩＯＮ ＵＳＩＮＧ ＴＨＲＥＥ－ＤＩＭＥＮＳＩＯＮＡＬ（３Ｄ）ＰＲＯＴＥＩＮ ＳＴＲＵＣＴＵＲＥ ＶＯＸＥＬＳ」と題する米国仮特許出願第６３／２８１，５７９号（代理人整理番号ＩＬＬＭ１０６０－１／ＩＰ－２２７０－ＰＲＶ）、
２０２１年１１月１９日に出願された「ＣＯＭＢＩＮＥＤ ＡＮＤ ＴＲＡＮＳＦＥＲ ＬＥＡＲＮＩＮＧ ＯＦ Ａ ＶＡＲＩＡＮＴ ＰＡＴＨＯＧＥＮＩＣＩＴＹ ＰＲＥＤＩＣＴＯＲ ＵＳＩＮＧ ＧＡＰＥＤ ＡＮＤ ＮＯＮ－ＧＡＰＥＤ ＰＲＯＴＥＩＮ ＳＡＭＰＬＥＳ」と題する米国仮特許出願第６３／２８１，５９２号（代理人整理番号ＩＬＬＭ１０６１－１／ＩＰ－２２７１－ＰＲＶ）、
２０２１年１２月２９日に出願された「ＰＥＲＩＯＤＩＣ ＭＡＳＫ ＰＡＴＴＥＲＮ ＦＯＲ ＲＥＶＥＬＡＴＩＯＮ ＬＡＮＧＵＡＧＥ ＭＯＤＥＬＳ」と題する米国仮特許出願第６３／２９４，８１３号（代理人整理番号ＩＬＬＭ１０６３－１／ＩＰ－２２９６

－ＰＲＶ）、
２０２１年１２月２９日に出願された「ＣＬＡＳＳＩＦＹＩＮＧ ＭＩＬＬＩＯＮＳ ＯＦ ＶＡＲＩＡＮＴＳ ＯＦ ＵＮＣＥＲＴＡＩＮ ＳＩＧＮＩＦＩＣＡＮＣＥ ＵＳＩＮＧ ＰＲＩＭＡＴＥ ＳＥＱＵＥＮＣＩＮＧ ＡＮＤ ＤＥＥＰ ＬＥＡＲＮＩＮＧ」と題する米国仮特許出願第６３／２９４，８１６号（代理人整理番号ＩＬＬＭ１０６４－１／ＩＰ－２２９７－ＰＲＶ）、
２０２１年１２月２９日に出願された「ＩＤＥＮＴＩＦＹＩＮＧ ＧＥＮＥＳ ＷＩＴＨ ＤＩＦＦＥＲＥＮＴＩＡＬ ＳＥＬＥＣＴＩＶＥ ＣＯＮＳＴＲＡＩＮＴ ＢＥＴＷＥＥＮ ＨＵＭＡＮＳ ＡＮＤ ＮＯＮ－ＨＵＭＡＮ ＰＲＩＭＡＴＥＳ」と題する米国仮特許出願第６３／２９４，８２０号（代理人整理番号ＩＬＬＭ１０６５－１／ＩＰ－２２９８－ＰＲＶ）、
２０２１年１２月２９日に出願された「ＤＥＥＰ ＬＥＡＲＮＩＮＧ ＮＥＴＷＯＲＫ ＦＯＲ ＥＶＯＬＵＴＩＯＮＡＲＹ ＣＯＮＳＥＲＶＡＴＩＯＮ」と題する米国非仮特許出願第６３／２９４，８２７号（代理人整理番号ＩＬＬＭ１０６６－１／ＩＰ－２２９９－ＰＲＶ）、
２０２１年１２月２９日に出願された「ＩＮＴＥＲ－ＭＯＤＥＬ ＰＲＥＤＩＣＴＩＯＮ ＳＣＯＲＥ ＲＥＣＡＬＩＢＲＡＴＩＯＮ」と題する米国仮特許出願第６３／２９４，８２８号（代理人整理番号ＩＬＬＭ１０６７－１／ＩＰ－２３０１－ＰＲＶ）、及び
２０２１年１２月２９日に出願された「ＳＰＥＣＩＥＳ－ＤＩＦＦＥＲＥＮＴＩＡＢＬＥ ＥＶＯＬＵＴＩＯＮＡＲＹ ＰＲＯＦＩＬＥＳ」と題する米国仮特許出願第６３／２９４，８３０号（代理人整理番号ＩＬＬＭ１０６８－１／ＩＰ－２３０２－ＰＲＶ）。

【背景技術】

【０００７】

本セクションで考察される主題は、単に本セクションにおける言及の結果として、先行技術であると想定されるべきではない。同様に、本セクションで言及した問題、又は背景として提供された主題と関連付けられた問題は、先行技術において以前に認識されていると想定されるべきではない。本セクションの主題は、単に異なるアプローチを表し、それ自体はまた、特許請求される技術の実施態様に対応し得る。

【０００８】

広義のゲノミクスは、機能的ゲノミクスとも呼ばれ、ゲノム配列決定、トランスクリプトームプロファイリング及びプロテオミクスなどのゲノムスケールアッセイを使用することによって生物の全てのゲノムエレメントの機能を特徴付けることを目的とする。ゲノミクスは、データ主導の科学として生じ、予め考えられたモデル及び仮説を試験することによってではなく、ゲノムスケールデータの調査から新規の特性を発見することによって動作する。ゲノミクスの適用には、遺伝子型と表現型との間の関連を見出すこと、患者の層別化のためのバイオマーカーを発見すること、遺伝子の機能を予測すること、並びに生物統計解析を使用して、転写エンハンサー及び一塩基多型（single nucleotide polymorphism、ＳＮＰ）などの生化学的に活性なゲノム領域及び残基を図表化することが含まれる。

【０００９】

ゲノミクスデータは、ペアワイズ相関の視覚的調査のみによってマイニングするには大きすぎ、かつ複雑すぎる。その代わりに、予期しない関係の発見をサポートし、新規な仮説及びモデルを導き出し、予測を行うために、分析ツールが必要とされる。仮定及び領域専門知識がハード符号化されるいくつかのアルゴリズムとは異なり、機械学習アルゴリズムは、データ内のパターンを自動的に検出するように設計される。したがって、機械学習アルゴリズムは、データ駆動型科学、特にゲノミクスに適している。しかしながら、機械学習アルゴリズムの性能は、データがどのように表されるかに、すなわち、各変数（特徴とも呼ばれる）がどのように計算されるかに強く依存し得る。例えば、蛍光顕微鏡画像から腫瘍を悪性又は良性として分類するために、前処理アルゴリズムは、細胞を検出し、細胞型を特定し、各細胞型について細胞カウントのリストを生成することができる。

【００１０】
機械学習モデルは、人手で設計した特徴の例である推定細胞カウントを、腫瘍を分類するための入力特徴として取ることができる。中心的な問題は、分類性能がこれらの特徴の質及び関連性に大きく依存することである。例えば、細胞形態、細胞間の距離又は器官内の局在化等の関連する視覚的特徴は、細胞カウントにおいて捕捉されず、データのこの不完全な表現は、分類精度を低減させ得る。
【００１１】
機械学習のサブ規律である深層学習は、特徴の計算を機械学習モデル自体に埋め込み、エンドツーエンドモデルを生成することによって、この問題に対処する。この結果は、先行する動作の結果を入力として取ることによってますます複雑になる特徴を計算する、連続する基本動作を含む機械学習モデルである深層ニューラルネットワークの開発によって実現されている。ディープニューラルネットワークは、上記の例における細胞形態及び細胞の空間的構成など、複雑度の高い関連する特徴を発見することによって予測精度を改善することができる。深層ニューラルネットワークの構築及び訓練は、特にグラフィカル処理ユニット（graphical processing unit、ＧＰＵ）の使用を通して、データの爆発、アルゴリズムの進歩、及び計算能力の実質的な増加によって可能になった。
【００１２】
教師あり学習の目標は、特徴を入力として取り、いわゆるターゲット変数の予測を返すモデルを得ることである。教師あり学習問題の例は、標準スプライス部位配列の有無、スプライシング分岐点の位置又はイントロン長などのＲＮＡ上の特徴を考慮して、イントロンがスプライシングされるか否か（標的）を予測する問題である。機械学習モデルを訓練することは、そのパラメータを学習することを指し、これは一般に、未知のデータに対して正確な予測を行う目的で、訓練データに対する損失関数を最小化することを含む。
【００１３】
計算生物学における多くの教師あり学習問題について、入力データは、複数の列又は特徴を有する表として表すことができ、その各々は、予測を行うのに潜在的に有用である数値データ又はカテゴリデータを含有する。いくつかの入力データは、表形式の特徴（例えば、温度又は時間）として自然に表されるが、他の入力データは、表の表現に適合させるために、特徴抽出と呼ばれるプロセスを使用して最初に変換される必要がある（例えば、デオキシリボ核酸（deoxyribonucleic acid、ＤＮＡ）配列をｋ－ｍｅｒカウントに変換する）。イントロン－スプライシング予測問題のために、標準スプライス部位配列の有無、スプライシング分岐点の位置及びイントロン長は、表形式で収集された前処理された特徴であることができる。表形式データは、ロジスティック回帰などの単純な線形モデルから、ニューラルネットワーク及び多くの他のものなどのより柔軟な非線形モデルに及ぶ、広範囲の教師あり機械学習モデルの標準である。
【００１４】
ロジスティック回帰は、バイナリ分類器、すなわち、バイナリターゲット変数を予測する教師あり学習モデルである。具体的には、ロジスティック回帰は、活性化関数の一種であるシグモイド関数を使用して［０，１］区間にマッピングされた入力特徴の加重和を計算することによって、陽性クラスの確率を予測する。ロジスティック回帰、又は異なる活性化関数を使用する他の線形分類器のパラメータは、加重和における重みである。線形分類器は、例えばイントロンがスプライシングされたか否かのクラスが入力特徴の加重和で十分に識別できない場合に失敗する。予測性能を改善するために、新しい入力特徴は、新しい方法で既存の特徴を変換又は組み合わせることによって、例えば、累乗又はペアワイズ積を取ることによって、手動で追加されることができる。
【００１５】
ニューラルネットワークは、隠れ層を使用して、これらの非線形特徴変換を自動的に学習する。各隠れ層は、シグモイド関数又はより一般的な正規化線形ユニット（rectified-linear unit、ＲｅＬＵ）などの非線形活性化関数によって変換された出力を有する複数の線形モデルと考えることができる。同時に、これらの層は、入力特徴を関連する複雑

なパターンに構成し、２つのクラスを区別するタスクを容易にする。

【００１６】

深層ニューラルネットワークは、多くの隠れ層を使用し、各ニューロンが先行する層の全てのニューロンから入力を受信するとき、層は全結合されていると言われる。ニューラルネットワークは、一般に、非常に大きなデータセットでモデルを訓練するのに適したアルゴリズムである確率的勾配降下法を使用して訓練する。現代の深層学習フレームワークを使用するニューラルネットワークの実施態様は、異なるアーキテクチャ及びデータセットを用いたラピッドプロトタイピングを可能にする。全結合ニューラルネットワークは、スプライス因子の結合モチーフの存在又は配列保存などの配列特徴から、所与の配列に対してスプライシングされたエクソンの割合を予測すること、潜在的に疾患を引き起こす遺伝子バリアントを優先順位付けすること、並びに所与のゲノム領域において、クロマチンマーク、遺伝子発現及び進化的保存などの特徴を用いてシス調節エレメントを予測することを含む、いくつかのゲノミクス用途に使用することができる。

【００１７】

効果的な予測のためには、空間的及び縦断的データにおける局所依存性を考慮しなければならない。例えば、ＤＮＡ配列又は画像のピクセルのシャッフリングは、情報パターンを激しく乱す。これらの局所依存性は、特徴の順序付けが任意である表形式データとは別に、空間又は縦断的データを設定する。結合領域が、配列決定（ＣｈＩＰ－ｓｅｑ）データが続く、クロマチン免疫沈降における高信頼度結合事象として定義される、特定の転写因子による結合対非結合としてゲノム領域を分類する問題を考慮されたい。転写因子は、配列モチーフを認識することによってＤＮＡに結合する。配列中のｋ－ｍｅｒインスタンスの数又は位置重み行列（position weight matrix、ＰＷＭ）マッチなどの配列由来の特徴に基づく全結合層を、このタスクに使用することができる。ｋ－ｍｅｒ又はＰＷＭインスタンス頻度は、配列内のモチーフをシフトさせることに対してロバストであるため、そのようなモデルは、異なる位置に位置する同じモチーフを有する配列に十分に一般化することができる。しかしながら、それらは、転写因子結合が明確な間隔を有する複数のモチーフの組み合わせに依存するパターンを認識することができない。更に、可能なｋ－ｍｅｒの数は、ｋ－ｍｅｒ長とともに指数関数的に増加し、これは、保存及び過剰適合の両方の課題をもたらす。

【００１８】

畳み込み層は、同じ全結合層が局所的に、例えば６ｂｐウィンドウ内で、全ての配列位置に適用される、全結合層の特別な形態である。このアプローチはまた、例えば、転写因子ＧＡＴＡ１及びＴＡＬ１について、複数のＰＷＭを使用して配列を走査することとみなすことができる。位置にわたって同じモデルパラメータを使用することによって、パラメータの総数は劇的に低減され、ネットワークは、訓練中に見られない位置でモチーフを検出することができる。各畳み込み層は、フィルタと配列との間の一致を量子化するスカラー値を全ての位置において生成することによって、いくつかのフィルタを用いて配列を走査する。全結合ニューラルネットワークにおけるように、非線形活性化関数（一般にＲｅＬＵ）が各層において適用される。次に、プーリング演算が適用され、これは、位置軸にわたって連続するビン内の活性化を集約し、一般に、各チャネルについて最大又は平均活性化を取る。プーリングは、有効配列長を減少させ、信号を粗大化する。後続の畳み込み層は、前の層の出力を構成し、ＧＡＴＡ１モチーフ及びＴＡＬ１モチーフがある距離範囲に存在したかどうかを検出することができる。最後に、畳み込み層の出力は、最終予測タスクを実行するために全結合ニューラルネットワークへの入力として使用することができる。したがって、異なるタイプのニューラルネットワーク層（例えば、全結合層及び畳み込み層）を単一のニューラルネットワーク内で組み合わせることができる。

【００１９】

畳み込みニューラルネットワーク（convolutional neural network、ＣＮＮ）は、ＤＮＡ配列のみに基づいて様々な分子表現型を予測することができる。用途としては、転写因子結合部位の分類、並びにクロマチン特徴、ＤＮＡコンタクトマップ、ＤＮＡメチ

ル化、遺伝子発現、翻訳効率、ＲＢＰ結合、及びマイクロＲＮＡ（microRNA、ｍｉＲＮＡ）標的などの分子表現型の予測が挙げられる。配列から分子表現型を予測することに加えて、畳み込みニューラルネットワークは、人手で設計したバイオインフォマティクスパイプラインによって伝統的に対処されるより技術的なタスクに適用することができる。例えば、畳み込みニューラルネットワークは、ガイドＲＮＡの特異性を予測し、ＣｈＩＰ－ｓｅｑをノイズ除去し、Ｈｉ－Ｃデータ分解能を向上させ、ＤＮＡ配列から実験室起源を予測し、遺伝子バリアントを呼び出すことができる。畳み込みニューラルネットワークはまた、ゲノムにおける長期依存性をモデル化するために使用されてきた。相互作用する調節エレメントは、折り畳まれていない直鎖状ＤＮＡ配列上に離れて位置し得るが、これらのエレメントは、多くの場合、実際の３Ｄクロマチン立体構造において近位である。したがって、線形ＤＮＡ配列からの分子表現型のモデリングは、クロマチンの粗い近似ではあるが、長期依存性を可能にし、モデルがプロモーター－エンハンサーループなどの３Ｄ組織化の態様を暗示的に学習することを可能にすることによって改善することができる。これは、最大３２ｋｂの受容野を有する拡張畳み込みを使用することによって達成される。拡張畳み込みはまた、スプライス部位が１０ｋｂの受容野を使用して配列から予測されることを可能にし、それによって、典型的なヒトイントロンと同じ長さの距離にわたる遺伝子配列の統合を可能にする（Ｊａｇａｎａｔｈａｎ，Ｋ．ｅｔ ａｌ．Ｐｒｅｄｉｃｔｉｎｇ ｓｐｌｉｃｉｎｇ ｆｒｏｍ ｐｒｉｍａｒｙ ｓｅｑｕｅｎｃｅ ｗｉｔｈ ｄｅｅｐ ｌｅａｒｎｉｎｇ．Ｃｅｌｌ １７６，５３５－５４８（２０１９）を参照）。

【００２０】

異なるタイプのニューラルネットワークは、それらのパラメータ共有スキームによって特徴付けることができる。例えば、全結合層はパラメータ共有を有さないが、畳み込み層は、それらの入力の全ての位置において同じフィルタを適用することによって並進不変性を課す。回帰型ニューラルネットワーク（recurrent neural network、ＲＮＮ）は、異なるパラメータ共有方式を実装する、ＤＮＡ配列又は時系列などの順次データを処理するための畳み込みニューラルネットワークの代替である。回帰型ニューラルネットワークは、各配列エレメントに同じ演算を適用する。この演算は、入力として、前の配列エレメントのメモリ及び新しい入力を取る。これはメモリを更新し、任意選択で出力を発し、この出力は後続の層に渡されるか、又はモデル予測として直接使用されるかのいずれかである。各配列エレメントに同じモデルを適用することによって、回帰型ニューラルネットワークは、処理された配列における位置指数に対して不変である。例えば、回帰型ニューラルネットワークは、配列中の位置にかかわらず、ＤＮＡ配列中のオープンリーディングフレームを検出することができる。このタスクは、開始コドンとそれに続くインフレーム停止コドンなどの特定の一連の入力の認識を必要とする。

【００２１】

畳み込みニューラルネットワークに対する回帰型ニューラルネットワークの主な利点は、理論的には、それらがメモリを介して無限に長い配列を通じて情報を引き継ぐことができることである。更に、回帰型ニューラルネットワークは、ｍＲＮＡ配列のような広く変化する長さの配列を自然に処理することができる。しかしながら、様々なトリック（拡張畳み込みなど）と組み合わされた畳み込みニューラルネットワークは、オーディオ合成及び機械翻訳などの配列モデリングタスクに関して、回帰型ニューラルネットワークに匹敵する、又は回帰型ニューラルネットワークよりも更に良好な性能に達することができる。回帰型ニューラルネットワークは、単一細胞ＤＮＡメチル化状態、ＲＢＰ結合、転写因子結合、及びＤＮＡアクセシビリティを予測するために、畳み込みニューラルネットワークの出力を集約することができる。更に、回帰型ニューラルネットワークは逐次演算を適用するので、容易に並列化することができず、したがって、畳み込みニューラルネットワークよりも計算がはるかに遅い。

【００２２】

ヒト遺伝子コードの大部分は全てのヒトに共通であるが、各ヒトは固有の遺伝子コードを有する。いくつかの場合において、ヒト遺伝子コードは、ヒト集団の比較的小さい群の

個体間で共通であり得る、遺伝子バリアントと呼ばれる外れ値を含み得る。例えば、特定のヒトタンパク質は、アミノ酸の特定の配列を含み得るが、そのタンパク質のバリアントは、他の点では同じ特定の配列において１つのアミノ酸が異なり得る。

【００２３】

遺伝子バリアントは病原的であり得、疾患をもたらし得る。そのような遺伝子バリアントのほとんどは、自然淘汰によってゲノムから枯渇しているが、どの遺伝子バリアントが病原性である可能性が高いかを特定する能力は、研究者がこれらの遺伝子バリアントに焦点を当てて、対応する疾患及びそれらの診断、治療、又は治癒の理解を得る助けとなることができる。何百万ものヒト遺伝子バリアントの臨床的解釈は不明のままである。最も頻繁な病原性バリアントのいくつかは、タンパク質のアミノ酸を変化させる単一ヌクレオチドミスセンス変異である。しかし、全てのミスセンス変異が病原性であるわけではない。

【００２４】

生物学的配列から分子表現型を直接予測することができるモデルは、遺伝的バリエーションと表現型バリエーションとの間の関連を調べるためのインシリコ摂動ツールとして使用することができ、定量的形質遺伝子座特定及びバリアント優先順位付けのための新しい方法として出現した。これらのアプローチは、複雑な表現型のゲノムワイド関連研究によって特定されたバリアントの大部分が非コードであり、それがそれらの効果及び表現型への寄与を推定することを困難にすることを考慮すると、非常に重要である。更に、連鎖不平衡は、共遺伝されるバリアントのブロックをもたらし、これは、個々の原因バリアントを正確に特定することを困難にする。したがって、そのようなバリアントの影響を評価するための照合ツールとして使用することができる配列に基づく深層学習モデルは、複雑な表現型の潜在的なドライバーを見出すための有望なアプローチを提供する。一例としては、転写因子結合、クロマチンアクセシビリティ又は遺伝子発現予測に関して、２つのバリアント間の差異から間接的に非コード単一ヌクレオチドバリアント及び短い挿入又は欠失（インデル）の効果を予測することが挙げられる。別の例としては、配列又はスプライシングに対する遺伝子バリアントの定量的効果から新規スプライス部位生成を予測することが挙げられる。

【００２５】

タンパク質配列及び配列保存データからミスセンスバリアントの病原性を予測するために、バリアント効果予測のためのエンドツーエンド深層学習アプローチが適用される（Ｓｕｎｄａｒａｍ，Ｌ．ｅｔ ａｌ．Ｐｒｅｄｉｃｔｉｎｇ ｔｈｅ ｃｌｉｎｉｃａｌ ｉｍｐａｃｔ ｏｆ ｈｕｍａｎ ｍｕｔａｔｉｏｎ ｗｉｔｈ ｄｅｅｐ ｎｅｕｒａｌ ｎｅｔｗｏｒｋｓ．Ｎａｔ．Ｇｅｎｅｔ．５０，１１６１－１１７０（２０１８）を参照。本明細書では「ＰｒｉｍａｔｅＡＩ」と称される）。ＰｒｉｍａｔｅＡＩは、異種間情報を使用するデータ増強を用いて既知の病原性のバリアントで訓練した深層ニューラルネットワークを使用する。特に、ＰｒｉｍａｔｅＡＩは、野生型及び変異タンパク質の配列を使用して、差異を比較し、訓練した深層ニューラルネットワークを使用して変異の病原性を決定する。病原性予測のためにタンパク質配列を利用するこのようなアプローチは、真円度問題及び以前の知識への過剰適合を回避することができるので、有望である。しかしながら、深層ニューラルネットワークを効果的に訓練するのに十分な数のデータと比較して、ＣｌｉｎＶａｒにおいて利用可能な臨床データの数は比較的少ない。このデータ不足を克服するために、ＰｒｉｍａｔｅＡＩは、一般的なヒトバリアント及び霊長類由来のバリアントを良性データとして使用し、トリヌクレオチド文脈に基づいてシミュレートされたバリアントをラベルなしデータとして使用した。

【００２６】

ＰｒｉｍａｔｅＡＩは、配列アラインメントで直接訓練した場合、従来の方法よりも性能が優れている。ＰｒｉｍａｔｅＡＩは、重要なタンパク質ドメイン、保存されたアミノ酸位置、及び配列依存性を、約１２０，０００のヒトサンプルからなる訓練データから直接学習する。ＰｒｉｍａｔｅＡＩは、候補の発達障害遺伝子における良性及び病原性のデノボ変異を区別すること、及びＣｌｉｎＶａｒにおける事前知識を再現することにおいて

、他のバリアント病原性予測ツールの性能を実質的に上回る。これらの結果は、ＰｒｉｍａｔｅＡＩが、臨床報告の事前知識への依存を減らすことができるバリアント分類ツールのための重要な前進であることを示唆する。

【００２７】

タンパク質生物学の中心は、構造エレメントが観察された機能をどのように生じさせるかの理解である。過度のタンパク質構造データは、構造－機能関係を支配する規則を系統的に導出するための計算方法の開発を可能にする。しかしながら、これらの方法の性能は、タンパク質構造表現の選択に決定的に依存する。

【００２８】

タンパク質部位は、タンパク質構造内の微小環境であり、それらの構造的又は機能的役割によって区別される。部位は、三次元（３Ｄ）位置と、構造又は機能が存在するこの位置の周りの局所近傍とによって定義することができる。合理的なタンパク質工学の中心は、アミノ酸の構造的配置がどのようにしてタンパク質部位内に機能的特徴を作り出すかの理解である。タンパク質中の個々のアミノ酸の構造的及び機能的役割の決定は、タンパク質機能の操作及び改変の助けとなるための情報を提供する。機能的又は構造的に重要なアミノ酸を特定することにより、標的タンパク質の機能特性を改変するための部位特異的変異誘発などの集中的な工学的努力が可能になる。あるいは、この知識は、所望の機能を無効にする工学設計を回避する助けとなることができる。

【００２９】

構造は配列よりもはるかに保存されていることが確立されているので、タンパク質構造データの増加は、データ駆動型アプローチを使用して構造－機能関係を支配する基礎パターンを系統的に研究する機会を提供する。任意の計算タンパク質分析の基本的な態様は、タンパク質構造情報がどのように表されるかである。機械学習方法の性能は、使用される機械学習アルゴリズムよりもデータ表現の選択に依存することが多い。良好な表現は、最も重要な情報を効率的に捕捉するが、不良な表現は、基礎となるパターンのないノイズの多い分布を生成する。

【００３０】

過度のタンパク質構造及び深層学習アルゴリズムの最近の成功は、タンパク質構造のタスク特異的表現を自動的に抽出するためのツールを開発する機会を提供する。

【００３１】

ヒトゲノムにおける７０，０００，０００を超える考えられるミスセンスバリアントのうち、大多数は未知の臨床的重要性を有し、わずか約０．１％が臨床バリアントデータベースにおいてアノテートされている。一般的な疾患における希少浸透性バリアントの役割を理解すべき機会が生じる。良性の結果を有するものからの有害なバリアントの正確な区別は、精密医療及び標的薬物開発の両方の利益をもたらし得る。

【図面の簡単な説明】

【００３２】

本特許又は出願ファイルは、カラーで作成された少なくとも１つの図面を含む。カラー図面を有するこの特許又は特許出願公開のコピーは、要求があり、必要な料金が支払われれば、庁によって提供される。カラー図面はまた、補足コンテンツタブを介してＰＡＩＲで利用可能であり得る。

【００３３】

図面では、同様の参照文字は、概して、異なる図全体を通して同様の部分を指す。また、図面は必ずしも縮尺通りではなく、その代わりに、開示された技術の原理を例示することを強調している。以下の説明において、開示された技術の様々な実施態様は、以下の図面を参照して説明される。

【図１Ａ】加重希少バリアントＰＲＳを決定する方法を例示する概略図である。

【図１Ｂ】特定の複数の遺伝子及び特定の表現型についての希少バリアント多遺伝子リスクスコアの例示的な計算を例示する。

【図２】開示される技術を実装するために使用され得る例示的なコンピュータシステムを

示す。
【図３】遺伝子バリアントの例を例示する。
【図４】遺伝子型の変化に応答した表現型効果を例示する例である。
【図５】心血管疾患患者Ｘに対応する複数の表現型を例示する。
【図６】一般的バリアント対希少バリアントについてのゲノムワイド関連解析をグラフで対比している。
【図７】それぞれ、遺伝子解像度における特定の個々の一般的バリアント又は特定の集約された希少バリアントに対する遺伝的関連性検定を例示する。
【図８】希少バリアント総和検定を最適化するための方法の概略図である。
【図９】バリアントの病原性を決定するためのシステムのプロセスを例示するフロー図である。
【図１０】開示される技術の一実施態様による、病原性分類器の例示的な処理アーキテクチャを示す。
【図１１】開示される技術を実装するために使用され得る例示的なコンピュータシステムを示す。
【図１２】最終的な病原性スコアを決定する一実施態様を例示する。
【図１３】各遺伝子内の多重検定を補正するためのプロセスのフロー図である。
【図１４】共変量について表現型値を補正する方法の概略図である。
【図１５】複数の表現型に対する複数の薬物の使用に応答して表現型シフトを予測することを表す概略図である。
【図１６】特定のコホートについての薬物使用パターン及び表現型データを得るための実験設定を表す概略図である。
【図１７】心血管疾患患者Ｘに対応する複数の表現型を例示する。
【図１８】異なるタイプの総和検定ごとに特定された有意な遺伝子－表現型対の総数を定量化するグラフである。
【図１９－１】病原性分類器ＰｒｉｍａｔｅＡＩ－３Ｄによって特定された疾患重症度及び発症の年齢に影響する希少有害バリアントを例示するグラフの集合である。
【図１９－２】病原性分類器ＰｒｉｍａｔｅＡＩ－３Ｄによって特定された疾患重症度及び発症の年齢に影響する希少有害バリアントを例示するグラフの集合である。
【図１９－３】病原性分類器ＰｒｉｍａｔｅＡＩ－３Ｄによって特定された疾患重症度及び発症の年齢に影響する希少有害バリアントを例示するグラフの集合である。
【図２０】異なる病原性スコアと表現型値との平均絶対スピアマン相関のグラフである。
【図２１】希少有害バリアントと一般的ゲノムワイド関連解析バリアントとのヒートマップ比較である。
【図２２】希少有害バリアントと一般的ゲノムワイド関連解析バリアントとの更なる比較である。
【図２３－１】ＵＫ　Ｂｉｏｂａｎｋコホートにおける全個体にわたるコレステロール経路及び総コレステロール分布を例示する。
【図２３－２】ＵＫ　Ｂｉｏｂａｎｋコホートにおける全個体にわたるコレステロール経路及び総コレステロール分布を例示する。
【図２４】希少バリアントＰＲＳ性能の測定を含む。
【図２５】ＰＲＳ外れ値の濃縮のグラフである。
【図２６】定量的表現型についてのＰＲＳ外れ値のグラフである。
【図２７】２つの別個のコホートからの正規化された総コレステロール分布のグラフを含む。
【図２８】２つの別個のコホート間で希少バリアントＰＲＳ外れ値及び表現型を比較するグラフを含む。
【図２９】民族性による性能結果を例示するグラフを含む。
【図３０】正常なコレステロールレベルに対して使用される一般的ＰＲＳバリアント及び希少ＰＲＳ遺伝子に対する効果量及び頻度を比較する表である。

【図３１】３４個の遺伝子－表現型対のセットに対する異なる病原性スコアリング方法によって説明される表現型分散の平均割合のグラフであり、３４個の遺伝子－表現型対のセットは、希少ミスセンス及びＬｏＦバリアントについてのそれらの濃縮に基づいて選択された。
【図３２】定量的表現型と臨床表現型との間の全てのペアワイズ比較についての、ＧＷＡＳ遺伝子における希少バリアントの濃縮のヒートマップである。
【図３３】同じ遺伝子についてのシングルトンＬｏＦバリアントの平均効果量と最も有意なＧＷＡＳバリアントの効果量との間の比の絶対値の分布のグラフを含む。
【図３４】５％の偽発見率で有意であった遺伝子－表現型対からの遺伝子のいずれかにおける個体当たりのバリアントの数を示す。
【図３５】訓練データ対検定データ分割における効果量の比較を示す。
【図３６】一般的バリアントＰＲＳサブセットと希少バリアントＰＲＳサブセットとを比較するグラフを含む。
【発明を実施するための形態】
【００３４】
以下の考察は、開示される技術を当業者が作製及び使用することを可能にするために提示され、特定の用途及びその要件に関連して提供される。開示される実施態様に対する種々の修正は、当業者には容易に明らかとなり、本明細書で定義される一般原理は、開示される技術の趣旨及び範囲から逸脱することなく、その他の実施態様及び用途に適用され得る。したがって、開示される技術は、示される実施態様に限定されることを意図するものではなく、本明細書に開示される原理及び特徴と一致する最も広い範囲を与えられるものである。
【００３５】
様々な実施態様の詳細な説明は、添付の図面と併せて読むと、より良く理解することができる。図が様々な実施態様の機能ブロックの図を示す限りにおいて、機能ブロックは、必ずしもハードウェア回路間の分割を示すものではない。したがって、例えば、機能ブロック（例えば、モジュール、プロセッサ、又はメモリ）のうちの１つ以上は、単一のハードウェア（例えば、汎用信号プロセッサ又はランダムアクセスメモリのブロック、ハードディスクなど）又は複数のハードウェアに実装されてもよい。同様に、プログラムは、スタンドアロンプログラムであってもよく、オペレーティングシステム内のサブルーチンとして組み込まれてもよく、インストールされたソフトウェアパッケージ内の機能であってもよい、など。様々な実施態様は、図面に示された配置及び手段に限定されないことを理解されたい。
【００３６】
モジュールとして指定された図の処理エンジン及びデータベースは、ハードウェア又はソフトウェアで実装することができ、図に示されるように、正確に同じブロックで分割される必要はない。いくつかのモジュールは、異なるプロセッサ、コンピュータ若しくはサーバ上に実装されてもよく、又は多数の異なるプロセッサ、コンピュータ若しくはサーバの中で広がることもできる。加えて、モジュールの一部は、達成される機能に影響を及ぼすことなく、図に示されるものとは並行して、又は異なる順序で操作され得ることが理解されるであろう。図のモジュールはまた、方法におけるフローチャートステップと考えることができる。また、モジュールは、必ずしもメモリ内に隣接して配置された全てのコードを有する必要はない。コードのいくつかの部分は、他のモジュール又は他の機能からのコードが間に配置された状態で、コードの他の部分から分離することができる。
【００３７】
開示される技術は、希少バリアントについての多遺伝子リスクスコアリングの質を改善するために使用され得る。開示される技術は、重度の早期発症疾患のリスクが最も高く、かつ臨床的介入から最大の利益を受ける、表現型スペクトルの極端な患者を特定するために使用され得る。これらの開示を提示する研究者らは、重度の早期発症疾患と関連付けられた外れ値表現型が、小さい効果量を有する多くの一般的バリアントの集合的アクション

によるよりも、希少浸透性バリアントによってより良好に説明されることを示す。多くの世代にわたる自然淘汰によって、主に有害な結果が取り除かれた一般的バリアントとは異なり、希少バリアントは、大部分がフィルタリングされておらず、複雑な形質及び疾患において高度に浸透性の効果を発揮する潜在性を保持している。開示される技術は、表現型応答に対する希少な病原性バリアントの寄与の定量化のための新規の加重和モデルを含む。

【００３８】

序論

開示される技術は、複数の表現型関連遺伝子からの希少有害バリアントの加重和に基づいて、相補的な希少バリアント多遺伝子リスクスコアモデルを開発する。ゲノムワイド関連解析（genome-wide association study、ＧＷＡＳ）は、遺伝子変異又はバリアントを、遅発性発達障害（delayed development disorder、ＤＤＤ）などの遺伝子疾患及び複雑な形質の表現型と関連付けるためのツールである。ＧＷＡＳツールは、非常に希少なバリアントよりも、一般的又は半一般的バリアントで良好に機能する。例えば、ＵＫ　Ｂｉｏｂａｎｋコホートは、２００，６４３個のエクソーム及び関連する患者データを含む。有害であると考えられる希少バリアントのほぼ半分が、データセット内で１人の個体のみに現れた。従来の統計解析及び総和検定は、大きいコホートにおいて希少バリアントの事例が１つしか存在しない場合に破綻する。

【００３９】

個体の遺伝子型は、個体の表現型に寄与し（表現型は、個体の１つ以上の観察可能な物理的特性として定義される）、したがって、遺伝子型と表現型との間には、特定のバリアント又はバリアント群を有する個体における物理的形質、異常、又は疾患の存在の予測を補助するために使用され得る統計的相関が存在する。特定の遺伝的障害は、単一遺伝子のバリアント（すなわち、単一遺伝子性遺伝的障害）によって引き起こされ得るが、複数の遺伝子のバリアント（すなわち、多遺伝子性遺伝的障害）によって引き起こされる遺伝的障害がより一般的である。したがって、重度の疾患を引き起こす可能性が高い希少有害バリアントについて多遺伝子リスクスコアを決定するためのロバストな方法が必要とされている。

【００４０】

一般集団からの健常な個体の最近の大規模なゲノム及びエクソーム配列決定研究は、平均的なヒトが最近の変異を通じて生じた数十の潜在的に有害な希少バリアントを保有することを明らかにした。多くの世代にわたる自然淘汰によって、主に有害な結果が取り除かれた一般的バリアントとは異なり、希少バリアントは、大部分がフィルタリングされておらず、複雑な形質及び疾患において高度に浸透性の効果を発揮する潜在性を保持している。２００，６４３個のＵＫ　Ｂｉｏｂａｎｋエクソームの公的な公開は、バリアント病原性予測の正確度の急速な進歩と一緒に、一般的なヒト疾患及び複雑な形質の包括的なセットに対する希少浸透性バリアントの影響を調べる機会を生み出すとともに、一般集団のための個人ゲノム配列決定の潜在的有用性への見通しを提供する。

【００４１】

最近の指数関数的なヒト集団の成長は、ＧＷＡＳにおいて特定された一般的バリアントとは対照的に、有害な結果を有するものを取り除くための自然淘汰のための十分な時間を提供することなく、ランダムに生じる変異を介して多数の希少バリアントを作り出した。本発明者らは、軽度の臨床的リスクと関連付けられた一般的バリアントを含有する各ＧＷＡＳ遺伝子座において、はるかに重大な重症度を有する希少有害バリアントも存在するはずであり、バリアントの重症度が集団におけるその頻度に反比例する対立遺伝子系列を形成することを検定することに着手した。

【００４２】

希少有害バリアントの高浸透性にもかかわらず、大多数の個体が所与の表現型について希少有害バリアントを保有しないため、それらの希少性は、表現型分散を説明する能力を集団の小さい割合に限定する。したがって、開示される技術に説明されている希少バリア

ントＰＲＳの一実施態様は、母集団全体にわたって説明される分散に関して、一般的バリアントＰＲＳの約２０分の１のみ実施する。しかしながら、極端な表現型の個体を考慮すると、この傾向が逆転した：外れ値表現型（ｚスコア≧３）を有する個体は、一般的バリアントＰＲＳの１．８倍と比較して、ベースライン集団よりも、１又は９９パーセンタイルの希少バリアントＰＲＳスコアを有する可能性が３倍高かった。

【００４３】

一般的バリアントＰＲＳモデルの臨床的採用に対する顕著な障壁は、異なる祖先を有する集団間におけるそれらの限定された一般化可能性であった。これらの問題は、連鎖不平衡に起因する一般的バリアントＰＲＳモデルにおける予測因子として、疾患に関連するが原因ではないバリアントの組み込みに由来する。原因バリアントの影響は、コホート間で一般化されると予想され得るが、真の原因バリアントとモデルにおいて使用されるバリアントとの間の相関が保たれることになるという保証はなく、むしろ、これらの相関は、集団祖先における差異並びに遺伝子型判定及びインピュテーションにおける技術的アーチファクトによって影響される。比較すると、希少バリアントＰＲＳモデルは、希少有害バリアントを予測因子として直接使用し、母集団における希少及び最近の歴史を考慮すると、一般的バリアントについての相関から因果関係を区別することを困難にする連鎖不平衡問題によって大きく影響されない。むしろ、希少バリアント多遺伝子リスクを予測するための課題は、意義不明のバリアント（variants of uncertain significance、ＶＵＳ）の効果の正確な解釈にあり、これは、有意な希少バリアント関連性を有する遺伝子を特定すること、及び所与の臨床表現型についての希少有害バリアントの効果量を推定することの両方にとって重要である。希少バリアントＰＲＳ遺伝子に出現する希少有害バリアントのほぼ半分が、２００，６４３個のＵＫ　Ｂｉｏｂａｎｋエクソームコホート全体において１個体のみに見られるため、この問題は、従来の統計解析にとって扱い難いようであり得るが、深層学習、ハイスループット実験アッセイ、及び密接に関連する霊長類種からのバリアント情報を採用する最近の進歩は、各々、ゲノムワイドスケールでＶＵＳ問題を解決することに向けた進歩を示している。希少バリアントＰＲＳ遺伝子の根底にある例外的な対立遺伝子不均一性は、バリアント効果予測の観点から課題であり得るが、逆説的に、それはまた、異なるコホートにわたる希少バリアントＰＲＳモデルのロバスト性及び移植性に対する鍵であり、遺伝子当たり数千の固有の希少有害バリアントにわたるシグナルを統合することによって、希少バリアントＰＲＳは、データ中に存在し得る任意のバリアント特異的アーチファクトの効果を平滑化する所望の特性を達成する。

【００４４】

希少及び一般的遺伝的バリアントが一般的な疾患及び複雑な形質のリスクに寄与する程度は、ヒト遺伝学の分野において数十年にわたって議論されてきた。開示される技術は、ＧＷＡＳ遺伝子座の大きい割合における希少浸透性バリアントの存在を示し、全集団分散並びに重度の早期発症疾患のリスクが最も高い外れ値個体に対する希少及び一般的バリアントの相対的寄与を定量化することによって、これらの観点を調和させることを助ける。

【００４５】

精密医療の観点から、ＵＫ　Ｂｉｏｂａｎｋコホートは、概して、ＵＫ集団からの成人の代表的な横断であり、希少浸透性バリアントの総和及び一般集団におけるそれらの効果を特徴付ける固有の機会も提示する。開示される技術の１つの実施態様において研究された９０個の臨床的及び定量的表現型のうちの１つ以上について有意であった５００個の遺伝子にわたって、個体の８６％が、少なくとも１つの希少有害バリアントを保有し、個体当たり平均２．０３個の希少有害バリアントを有した。開示される技術のいくつかの実施態様は、これらの希少有害バリアントが高い浸透効果を有し、同じ遺伝子座に存在する一般的ＧＷＡＳバリアントよりも平均して１０倍大きい効果量に寄与することを示す。平均して、個体の５．２％が、所定の表現型について希少有害バリアントを保有し、個体の０．４％が、所定の遺伝子について希少浸透性バリアントを保有した。

【００４６】

開示される技術は、検出力を最大化し、良性ミスセンス変数をフィルタ除去するために

、希少バリアント総和検定アプローチを補強する。開示される技術のいくつかの実施態様では、総和検定検出力は、総和検定の有意性を最大化する最大対立遺伝子カウントに対するカットオフを見つけるために、観察されたバリアントの対立遺伝子カウントにわたって第１のグリッド検索を実施することによって最大化され、良性ミスセンスバリアントは、総和検定の有意性を最大化する病原性スコアの最適閾値を見つけるために、病原性スコア（例えば、ＰｒｉｍａｔｅＡＩ－３Ｄスコア）閾値に対する複数のカットオフにわたって第２のグリッド検索を実施することによってフィルタ除去される。

【００４７】

開示される技術は、複数の表現型関連遺伝子からの希少有害バリアントの加重和に基づいて、希少有害バリアント多遺伝子リスクスコアモデルを開発する。個々のゲノムワイド関連研究バリアントは、単一遺伝子障害と比較して、複数の遺伝子におけるバリアントによって引き起こされる臨床的実行可能性及び遺伝的障害に対して過度に軽度である効果を与える傾向がある。したがって、複数のバリアントからのシグナルを組み合わせる多遺伝子リスクスコアモデルは、疾患のリスクが高い患者を予測するためにより有効であることが多い。しかしながら、既存の多遺伝子リスクスコアモデルは、主に、一般的バリアントに対して最適化されており、意義不明のバリアントを解釈する困難さ、低検出力研究による課題、及び希少バリアントのサンプルサイズが小さいことに起因する不正確な効果量推定に起因して、希少バリアントを除外する。開示される技術は、個体が特定の遺伝子内に少なくとも１つの希少有害バリアントを保有するかどうかに基づいて、遺伝子解像度ベースでリスクを集約することによって、希少バリアントを含めるために、先行する多遺伝子リスクスコアモデルを改善する。

【００４８】

開示される希少多遺伝子リスクスコアは、関連性の強度が表現型と関連付けられた遺伝子及び表現型応答に対する希少バリアントの寄与を定量化する総和検定として計算される。個体のコホートについて、バリアントキャリア状態及び特定の測定された表現型値を表すデータが得られる。総和スコアは、得られたデータを使用して関連付けられた遺伝子の各々について計算され、総和スコアは、関連付けられた遺伝子の各々のキャリア状態と、１つ以上の希少病原性バリアントの関連付けられた遺伝子における存在に対する表現型応答との間のコホートにおける、結果的な非ランダムな関連性を特定する。キャリア状態は、特定の遺伝子における１つ以上の希少病原性バリアントの存在又は非存在によって決定されるブール変数である。特定の希少バリアントの病原性は、遺伝子がタンパク質として発現される場合の機能に対する特定の希少バリアントの予測される影響によって決定され、予測される影響は、畳み込みニューラルネットワーク病原性分類器によって病原性スコアとして決定される。特定の希少バリアントの希少性は、集団における特定の希少バリアントの所定の閾値を下回る出現によって決定される（すなわち、最大対立遺伝子カウント）。

【００４９】

関連性の強度は、それぞれの遺伝子についての希少病原性バリアントのキャリア状態に対する有効強度スコア、及びそれぞれの遺伝子についての遺伝子ごとの解像度における希少病原性バリアントのキャリア状態に対するそれぞれの表現型応答として定量化される。表現型応答に対する関連付けられた遺伝子にわたるコホート内の対象のキャリア状態の寄与についてのスコアは、有効強度スコアに基づいて決定され得る。開示される技術のいくつかの実施態様では、特定の表現型は、定量的バイオマーカー表現型であり、遺伝子についての結果的な非ランダムな関連性に対する有効強度スコアは、総和スコアの線形回帰成分に対する両側ｔ検定を使用して、定量的バイオマーカー表現型に対して決定される。両側ｔ検定は、キャリアと非キャリアとの平均表現型測定値間の差が所定の有意水準で有意であるか否かを決定するために、ｐ値を生成する。

【００５０】

開示される技術の他の実施態様では、指定された表現型は、カテゴリ臨床診断表現型であり、遺伝子についての結果的な非ランダムな関連性についての有効強度スコアは、キャ

リア状態についてのベータ係数としてカテゴリ臨床診断表現型について決定される。ベータ係数は、総和スコアのロジスティック回帰成分を使用して決定され、ロジスティック回帰成分は、キャリア状態及び複数の共変量に対応するバイナリインジケータ変数から臨床診断ラベルを予測するように適合される。

【００５１】

開示される技術のいくつかの実施態様では、特定の遺伝子及び特定の表現型に対する各総和検定は、特定の遺伝子について最適化されるとともに、特定の表現型値について補正される。開示される技術のいくつかの実施態様は、特定の遺伝子についての最適化又は特定の表現型値についての補正のいずれかを伴い得るが、他方を伴わない。

【００５２】

開示される技術の更なる実施態様は、キャリアと非キャリアとの間の分離を最大化する入れ子型ｔ検定を使用することによって、希少バリアント総和検定を最適化し、総和検定パラメータは、特定の遺伝子に関して最適化される。総和検定検出力は、総和検定の有意性を最大化する最大対立遺伝子カウントに対するカットオフを見つけるために、観察されたバリアントの対立遺伝子カウントにわたって第１のグリッド検索を実施することによって最大化され、良性ミスセンスバリアントは、総和検定の有意性を最大化する病原性スコアの最適閾値を見つけるために、病原性スコア（例えば、ＰｒｉｍａｔｅＡＩ－３Ｄスコア）閾値に対する複数のカットオフにわたって第２のグリッド検索を実施することによってフィルタ除去される。

【００５３】

開示される技術のいくつかの実施態様では、グリッド検索手順は、複数の対立遺伝子カウント及び複数の病原性スコア閾値を検索することを伴い、グリッド検索は、複数の対立遺伝子カウント及び複数の病原性スコア閾値からの対立遺伝子カウント及び病原性スコア閾値の複数の組み合わせの生成と、対立遺伝子カウント及び病原性スコア閾値の複数の組み合わせに対応する希少病原性バリアントの複数の群の特定と、個体のコホートにおける希少病原性バリアントの特定の群のキャリアを個体のコホートにおける希少病原性バリアントの特定の群の非キャリアから分離するキャリア状態に依存した希少病原性バリアントの複数の群の総和検定と、対立遺伝子カウント及び病原性スコア閾値の複数の組み合わせに対応する複数の効果量及びｐ値の決定と、を含む。

【００５４】

最も有意なｐ値を有する対立遺伝子カウント及び病原性スコア閾値の特定の組み合わせが、特定の遺伝子についての最適な組み合わせとして使用されるように選択される。それぞれの遺伝子は、各それぞれの遺伝子のキャリアと非キャリアとの間の最適な分離をもたらす同様の最大対立遺伝子カウント閾値及び／又は最小病原性スコア閾値を有してもよく、又は有していなくてもよい。複数の病原性スコア閾値における病原性スコア閾値が、希少病原性バリアントの群における希少病原性バリアントについて決定された病原性スコアの病原性スコア分位数に対応する。開示される技術のいくつかの実施態様では、病原性スコアは、ＰｒｉｍａｔｅＡＩ－３Ｄなどの畳み込みニューラルネットワーク病原性分類器によって生成される。開示される技術の他の実施態様では、広範囲の追加のＡＩ、機械学習、及び深層学習モデルが、病原性スコアを生成するために採用される。

【００５５】

上記のグリッド検索手順は、生成された偽発見率補正ｐ値に対する多重検定のための補正を必要とする。開示される技術のいくつかの実施態様では、Ｂｅｎｊａｍｉｎｉ－Ｈｏｃｈｂｅｒｇ手順を使用して多重検定補正が実施される。開示される技術の他の実施態様では、適応的並べ替え検定手順を使用して多重検定補正が実施される。

【００５６】

開示される技術の更なる実施態様は、時間的に分散された検出点にわたる薬物使用に共変量補正を適用する。総和検定入力データは、測定された表現型値に関して補正される。表現型応答値は、複数の共変量について共変量補正され、複数の薬物について薬物使用補正される。開示される技術のいくつかの実施態様では、複数の交絡因子を有する個体のコ

ホートの複数の表現型に対する複数の薬物に応答した表現型シフトが、共変量補正及び薬物使用補正された表現型測定値を生成するために予測される。

【００５７】

開示される技術のいくつかの実施態様では、表現型シフトは、個体のコホートについての表現型ごとの複数の交絡因子（例えば、年齢、性別、遺伝的主成分、食事、及び喫煙状態）を有する個体のコホートの複数の表現型に対する複数の薬物の使用、並びに複数の表現型に対する表現型測定値、複数の交絡因子についての共変量測定値、及び２つの別個の時点におけるコホート内の各個体に対する複数の薬物についての薬物使用パターンに応答して予測される。

【００５８】

表現型測定値は、第１の回帰モデルをフィッティングすることによって共変量測定値を除去して回帰分析することによって、共変量測定値に基づいて第１及び第２の時点について共変量補正され、それによって、第１及び第２の時点について共変量補正された表現型測定値を生成する。デルタは、第１及び第２の時点についての共変量補正された表現型測定値間の差に基づいて決定される。薬物使用パターンの各々について、共変量補正された表現型測定値に対する複数の薬物の使用に応答した表現型シフトを予測するためにデルタを使用する第２の回帰モデルがフィッティングされる。

【００５９】

第２の回帰モデルは、薬物使用パターンの各々についての表現型シフトを連続的かつ累積的に含めることによって、デルタを繰り返し予測する前進型選択段階的回帰モデルである。第２の回帰モデルは、薬物使用パターンの各々についてバイナリインジケータ独立変数を有し、薬物使用パターンは、第１及び第２の時点で薬物を服用しないこと、第１及び第２の時点の間で薬物を服用し始めること、第１及び第２の時点の間で薬物を服用することを停止すること、並びに第１及び第２の時点で薬物を服用することを含む。開示される技術のいくつかの実施態様では、共変量補正、デルタ決定、表現型シフト予測、及び薬物使用補正は、複数の薬物中の薬物について薬物ごとに実行される。開示される技術の他の実施態様では、共変量補正、デルタ決定、表現型シフト予測、及び薬物使用補正は、複数の薬物カテゴリ内の薬物カテゴリについて、薬物カテゴリごとに実行される（例えば、スタチン、ＮＳＡＩＤ、オピオイドなど）。

【００６０】

第２の回帰モデルは、個体のコホート内の個体について個体ごとに第１の時点と第２の時点との間で経過した時間に応答した表現型シフト、及び第１の時点と第２の時点との間の平均への回帰に応答した表現型シフトを更にモデル化する。第２の回帰モデルは、複数の薬物における関連薬物のセットに対して共同でフィッティングされる。

【００６１】

開示される技術のいくつかの実施態様では、薬物使用補正は、第３の回帰モデルをフィッティングすることによって実装され、第１の時点と第２の時点との間で薬物を服用し始める第１の薬物使用パターンについての第１のバイナリインジケータ独立変数、第１及び第２の時点で薬物を服用しない第２の薬物使用パターンについての第２のバイナリインジケータ独立変数、並びに個体が第１の時点で特定の薬物を服用していたかどうかを符号化する薬物特異的バイナリインジケータ独立変数に基づいて、第１の時点についての表現型測定値を薬物使用補正することを含む。

【００６２】

第４の回帰モデルがフィッティングされ、第４の回帰モデルは、第１の時点と第２の時点との間で薬物を服用することを停止する第３の薬物使用パターンについての第３のバイナリインジケータ独立変数、第１の時点及び第２の時点で薬物を服用する第４の薬物使用パターンについての第４のバイナリインジケータ独立変数、個体が第２の時点で特定の薬物を服用していたかどうかを符号化する薬物特異的バイナリインジケータ独立変数に基づいて、第２の時点についての表現型測定値を薬物使用補正することを含む。

【００６３】

ランクに基づく逆正規変換が、第１及び第２の時点についての薬物使用補正された表現型測定値に適用されて、第１及び第２の時点についての正規化された薬物使用補正された表現型測定値を生成し得る。次いで、正規化された薬物使用補正された表現型測定値が、第１及び第２の時点について共変量補正され、第１及び第２の時点について共変量補正され正規化された薬物使用補正された表現型測定値を生成する。共変量補正され正規化された薬物使用補正された表現型測定値は、希少バリアント多遺伝子リスクスコアを生成するために使用され得、関心対象の表現型についての測定値は、共変量及び薬物使用パターンに応答した表現型シフトに対して補正される。

【００６４】

開示される技術のいくつかの実施態様では、構築された回帰モデルは、望ましくない副作用及び所望の標的効果などの薬物－表現型関連性を検出するために使用され得る。

【００６５】

希少バリアント多遺伝子リスクスコア（Polygenic Risk Score、ＰＲＳ）

図１Ａは、加重希少バリアントＰＲＳを決定する方法を例示する概略図である。データベース１０２は、コホートｉに属する個体の群に対応するゲノム及び表現型データを含む。コホートｉ１０４は、遺伝子配列決定及び複数の表現型測定値についての検定を受けたＮ人の個体を含む。コホートｉ１０４内の個体についての配列決定データは、複数のバリアントキャリア状態に対応するデータを含む。特定の遺伝子は、３つの公知の考えられる希少有害バリアント（例えば、バリアントＡ、バリアントＢ、及びバリアントＣ）を保有し得、そのうち、個体は、キャリア（すなわち、個体がそれぞれのバリアントを保有し、したがって、それらがそのバリアントのキャリアである）又は非キャリア（すなわち、個体がそれぞれのバリアントを保有せず、したがって、それらがそのバリアントのキャリアではない）であり得る。当業者は、バリアントＡ、バリアントＢ、及びバリアントＣが例示的な例として与えられており、遺伝子が任意の数の希少有害バリアントを保有し得ることを認識するであろう。開示される技術は、バリアントを、集団内のバリアント出現率が所定の閾値を下回る場合、「希少」と定義し、遺伝子が所定の閾値を上回るタンパク質として発現されるときに、バリアントの病原性が影響に対して測定された影響を有すると予測される場合、「有害」と定義する。希少性及び病原性の両方についてのそれぞれの所定の閾値は、遺伝子特異的である（すなわち、各個々の遺伝子は、別の遺伝子と異なり得る、個々に決定された希少性閾値及び病原性閾値に対応することになる）。

【００６６】

コホートｉ１０４内の各個体についてのゲノム及びエクソーム配列決定から得られたゲノムデータに加えて、表現型データもまた、各個体に対して利用可能である（すなわち、個体Ｎは、表現型Ｄについてx_nの測定された値を有する）。特定の遺伝子型と特定の表現型との間の関係についてモデル１２２を構築することによって、特定の表現型に対する特定の遺伝子型の効果量を測定することが可能である。箱ひげ図１２４は、表現型Ｄに対する１つ以上の特定の希少有害バリアントのキャリア状態の影響が、ｐ値（すなわち、キャリアと非キャリアとの間の平均表現型値を比較した場合の有意性の程度）及びβ係数（すなわち、従属変数及び独立変数の分散が１に等しくなるように基礎データが標準化されている、キャリアの平均表現型値と非キャリアの平均表現型値とを結ぶ回帰直線から得られる重み推定値）の両方の形態で測定される、代表的な遺伝的関連性検定を例示する。

【００６７】

特定の表現型が定量的バイオマーカー表現型である本技術のいくつかの実施態様では、関連性の強度は、両側ｔ検定によって決定されるｐ値として定量化される。特定の表現型が定性的表現型（例えば、カテゴリ臨床診断表現型）である、開示される技術の他の実施態様では、関連性の強度は、ロジスティック回帰のβ係数として定量化される。

【００６８】

開示される技術では、キャリア状態は、バリアント解像度ベースではなく遺伝子解像度ベースで定義される。特定の遺伝子について、個体は、個体が特定の遺伝子内に少なくとも１つの希少有害バリアントを保有する場合、キャリアとして定義される（すなわち、２

つの希少有害バリアントに関してキャリアである個体は、１つの希少有害バリアントに関してキャリアである個体と区別されないが、ゼロの希少有害バリアントに関してキャリアである個体は、非キャリアとして区別される）。決定１４２は、複数の希少バリアントのキャリア状態と表現型応答との間のコホートｉ１０４における関連性の強度の有効強度スコアについて行われ得る。

【００６９】

グラフ１４４は、複数の希少バリアントのキャリア状態と表現型応答との間のコホートｉ１０４における関連性の強度の有効強度スコアを決定するためのｔ検定を含む。ｔ検定の帰無仮説は、コホートｉ内のキャリアのサンプル平均とコホートｉ１０４内の非キャリアのサンプル平均との間の差の絶対値がゼロに等しいことを述べている。ｔ検定の対立仮説は、コホートｉ１０４内のキャリアのサンプル平均とコホートｉ１０４内の非キャリアのサンプル平均との間の差の絶対値がゼロに等しくないことを述べている。帰無仮説を受け入れるか拒絶するかの決定は、特定の有意性レベルα及びαに対応する結果として生じるｐ値によって駆動される。それぞれのｔ検定は、特定の共有される表現型応答についての複数の遺伝子特異的有効強度スコア（チャート１２４に表されるように、ｐ値又はβ係数によって測定される）を得るために、複数のそれぞれの遺伝子について実施され得る。

【００７０】

加重総和検定１６２は、複数のそれぞれの遺伝子特異的有効強度スコアから生成され得る。希少バリアントＰＲＳとして説明されるこの計算は、方程式１６４に示され、希少バリアントＰＲＳは、複数の遺伝子についての特定の遺伝子についての効果量及びキャリア状態の積の合計に等しい。

【００７１】

ここで、考察は、複数のそれぞれの遺伝子特異的有効強度スコアを使用して、特定の個体についての希少バリアントＰＲＳ、特定の個体について配列決定された特定の複数のそれぞれの遺伝子、及び関心対象の特定の表現型を生成する例を提供する。

【００７２】

図１Ｂは、特定の複数の遺伝子及び特定の表現型についての希少バリアント多遺伝子リスクスコアの例示的な計算を例示する。ヘッダ行１０８は、特定の個体に関する遺伝子、効果量、及びキャリア状態に対応する列ラベルの群を含む。遺伝子列内に含まれる遺伝子は、遺伝子Ａ（行１１８）、遺伝子Ｂ（行１２８）、遺伝子Ｃ（行１３８）、遺伝子Ｄ（行１４８）、及び遺伝子Ｅ（行１５８）であり、遺伝子Ａ～Ｅは、特定の表現型Ｙに関与する遺伝子であると仮定され得る。各遺伝子についての効果量に関して、効果量Ａが遺伝子Ａ（行１１８）に対応し、効果量Ｂが遺伝子Ｂ（行１２８）に対応し、効果量Ｃが遺伝子Ｃ（行１３８）に対応し、効果量Ｄが遺伝子Ｄ（行１４８）に対応し、効果量Ｅが遺伝子Ｅ（行１５８）に対応する。キャリア状態は、特定の遺伝子内の少なくとも１つの特定された希少有害バリアントの存在を説明するブール変数である（例えば、コホートｉ１０４において検出されたバリアントＡ、Ｂ、及びＣが、遺伝子Ａ（行１１８）に特有であると仮定し、したがって、コホートｉ１０４についての表に列挙された各個体が、遺伝子Ａについてのキャリアとみなされることになる）。ゲノムが図１Ｂで特徴付けられている個体は、それぞれ、遺伝子Ａ（行１１８）、遺伝子Ｃ（行１３８）、遺伝子Ｄ（行１４８）、及び遺伝子Ｅ（行１５８）における少なくとも１つの希少有害バリアントに対するキャリアである。個体は、遺伝子Ｂ（行１２８）における任意の希少有害バリアントに対するキャリアではない。

【００７３】

加重総和スコア１６２を出力するために、方程式１６４が適用される。特定の個体についての方程式１６４は、方程式１８２と等価であるとみなされ得、特定の個体は、遺伝子Ａ（行１１８）に対応する効果量Ａ、遺伝子Ｃ（行１３８）に対応する効果量Ｃ、遺伝子Ｄ（行１４８）に対応する効果量Ｄ、及び遺伝子Ｅ（行１５８）に対応する効果量Ｅの合計と等価である特定の表現型Ｙについての希少バリアントＰＲＳを有する。特定の個体が少なくとも１つの希少有害バリアントに対するキャリアであるそれぞれの遺伝子について

、キャリア状態は、１に等しい（すなわち、効果量＊１＝効果量）。遺伝子Ｂ（行１２８）に対応する効果量Ｂは、特定の個体が少なくとも１つの希少有害バリアントに対するキャリアではないため、方程式１８２には示されず、したがって、キャリア状態は、ゼロに等しく（すなわち、効果量＊０＝０）、遺伝子Ｂ（行１２８）に対応する効果量Ｂは、方程式１８２に存在しない。

【００７４】

図２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例示的コンピュータシステム２００を示す。コンピュータシステム２００は、バスサブシステム２４２を介して多数の周辺デバイスと通信する少なくとも１つの中央処理ユニット（central processing unit、ＣＰＵ）２４４を含む。これらの周辺デバイスは、例えば、メモリデバイス及びファイル記憶サブシステム２３６を含む記憶サブシステム２１０、ユーザインターフェース入力デバイス２３８、ユーザインターフェース出力デバイス２４８、並びにネットワークインターフェースサブシステム２４６を含むことができる。入力デバイス及び出力デバイスは、コンピュータシステム２００とのユーザ対話を可能にする。ネットワークインターフェースサブシステム２４６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。

【００７５】

一実施態様では、希少バリアントＰＲＳモデル２４０は、記憶サブシステム２１０及びユーザインターフェース入力デバイス２３８に通信可能にリンクされる。

【００７６】

ユーザインターフェース入力デバイス２３８は、キーボード、マウス、トラックボール、タッチパッド、又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、音声認識システム及びマイクロフォンなどのオーディオ入力デバイス、並びに他のタイプの入力デバイスを含むことができる。一般に、用語「入力デバイス」の使用は、コンピュータシステム２００に情報を入力するための全ての可能なタイプのデバイス及び方式を含むことを意図している。

【００７７】

ユーザインターフェース出力デバイス２４８は、ディスプレイサブシステム、プリンタ、ファックス装置、又はオーディオ出力デバイスなどの非視覚ディスプレイを含むことができる。ディスプレイサブシステムは、ＬＥＤディスプレイ、陰極線管（cathode ray tube、ＣＲＴ）、液晶ディスプレイ（liquid crystal display、ＬＣＤ）などのフラットパネルデバイス、投影デバイス、又は可視画像を作成するための何らかの他の機構を含むことができる。ディスプレイサブシステムはまた、オーディオ出力デバイスなどの非視覚ディスプレイを提供することができる。一般に、用語「出力デバイス」の使用は、コンピュータシステム２００からユーザ又は別のマシン若しくはコンピュータシステムに情報を出力するための、全ての可能なタイプのデバイス及び方式を含むことを意図している。

【００７８】

記憶サブシステム２１０は、本明細書に説明されるモジュール及び方法のうちのいくつか又は全ての機能を提供するプログラミング及びデータ構築物を記憶する。これらのソフトウェアモジュールは、一般にプロセッサ２４９によって実行される。

【００７９】

プロセッサ２４９は、グラフィック処理ユニット（ＧＰＵ）、フィールドプログラマブルゲートアレイ（field-programmable gate array、ＦＰＧＡ）、特定用途向け集積回路（application-specific integrated circuit、ＡＳＩＣ）、及び／又は粗粒化再構成可能構造（coarse-grained reconfigurable architecture、ＣＧＲＡ）であることができる。プロセッサ２４９は、Ｇｏｏｇｌｅ　Ｃｌｏｕｄ　Ｐｌａｔｆｏｒｍ（商標）、Ｘｉｌｉｎｘ（商標）及びＣｉｒｒａｓｃａｌｅ（商標）などの深層学習クラウドプラットフォームによってホスティングすることができる。プロセッサ２４９の例は、

ＧｏｏｇｌｅのＴｅｎｓｏｒ　Ｐｒｏｃｅｓｓｉｎｇ　Ｕｎｉｔ（ＴＰＵ）（商標）、ＧＸ４　Ｒａｃｋｍｏｕｎｔ　Ｓｅｒｉｅｓ（商標）、ＧＸ２　Ｒａｃｋｍｏｕｎｔ　Ｓｅｒｉｅｓ（商標）のようなラックマウントソリューション、ＮＶＩＤＩＡ　ＤＧＸ－１（商標）、Ｍｉｃｒｏｓｏｆｔ’　Ｓｔｒａｔｉｘ　Ｖ　ＦＰＧＡ（商標）、ＧｒａｐｈｃｏｒｅのＩｎｔｅｌｌｉｇｅｎｔ　Ｐｒｏｃｅｓｓｏｒ　Ｕｎｉｔ（ＩＰＵ）（商標）、Ｓｎａｐｄｒａｇｏｎ　ｐｒｏｃｅｓｓｏｒｓ（商標）を有するＱｕａｌｃｏｍｍのＺｅｒｏｔｈ　Ｐｌａｔｆｏｒｍ（商標）、ＮＶＩＤＩＡのＶｏｌｔａ（商標）、ＮＶＩＤＩＡのＤＲＩＶＥ　ＰＸ（商標）、ＮＶＩＤＩＡのＪＥＴＳＯＮ　ＴＸ１／ＴＸ２　ＭＯＤＵＬＥ（商標）、ＩｎｔｅｌのＮｉｒｖａｎａ（商標）、Ｍｏｖｉｄｉｕｓ　ＶＰＵ（商標）、Ｆｕｊｉｔｓｕ　ＤＰＩ（商標）、ＡＲＭのＤｙｎａｍｉｃＩＱ（商標）、ＩＢＭ　ＴｒｕｅＮｏｒｔｈ（商標）、Ｔｅｓｔａ　Ｖ１００ｓ（商標）を有するＬａｍｂｄａ　ＧＰＵ　Ｓｅｒｖｅｒ、及び他のものを含む。

【００８０】

記憶サブシステム２１０で使用されるメモリサブシステム２２２は、プログラム実行中に命令及びデータを記憶するためのメインランダムアクセスメモリ（random access memory、ＲＡＭ）２３２と、固定命令が記憶された読み取り専用メモリ（read only memory、ＲＯＭ）２３４とを含む多数のメモリを含むことができる。ファイル記憶サブシステム２３６は、プログラム及びデータファイルのための永続的な記憶装置を提供することができ、ハードディスクドライブ、関連する取り外し可能な媒体、ＣＤ－ＲＯＭドライブ、光学ドライブ、又は取り外し可能な媒体カートリッジを含むことができる。特定の実施態様の機能を実施するモジュールは、記憶サブシステム２１０内のファイル記憶サブシステム２３６によって、又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することができる。

【００８１】

バスサブシステム２４２は、コンピュータシステム２００の様々な構成要素及びサブシステムを、意図されるように互いに通信させるための機構を提供する。バスサブシステム２４２は、単一のバスとして概略的に示されているが、バスサブシステムの代替の実施態様は、複数のバスを使用することができる。

【００８２】

コンピュータシステム２００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、緩くネットワーク化されたコンピュータの緩く分散したセット、又は任意の他のデータ処理システム若しくはユーザデバイスを含む様々なタイプのものであり得る。コンピュータ及びネットワークは絶え間なく変化する性質のものであるため、図２に示されるコンピュータシステム２００の説明は、本発明の好ましい実施態様を例示する目的のための特定の実施例としてのみ意図される。コンピュータシステム２００の多くの他の構成は、図２に示されるコンピュータシステムよりも多くの又は少ない構成要素を有することができる。

【００８３】

代替的な希少有害多遺伝子リスクスコアモデル

一実施態様では、開示される技術は、上記で論じられたモデルとは異なる希少有害多遺伝子リスクスコアモデルを使用し得る。遺伝子中のバリアントに割り当てられた重みが、その遺伝子中の全ての希少有害バリアントにわたって観察された効果量に等しかった、すなわち、遺伝子中の全てのバリアントについて同じである、希少有害バリアントの加重和を使用することに代えて、開示される技術は、各有意な遺伝子内のバリアントごとの重みを推定し得る。バリアントごとの重みは、バリアント病原性スコア（例えば、ＰｒｉｍａｔｅＡＩ－３Ｄ）及び個体の形質値に対するｌｏｇ１０変換された対立遺伝子頻度の線形回帰を介して、有意な遺伝子について希少有害バリアントを保有する個体から推定され得る。これは、いくつかの実施態様では、説明される分散に関して１８％だけ性能を改善し得る。

【００８４】
遺伝子型
図１に説明される多遺伝子リスクスコアを計算する方法は、遺伝子型判定アレイ（すなわち、特定の個体の全ゲノム配列決定）、及びエクソーム配列決定（すなわち、特定の個体のゲノムのタンパク質コーディング領域の配列決定）によって決定される、個体の遺伝子型と関連付けられたデータを伴う。ここで、考察は、ゲノムデータ、それらの関連する特徴、及び特定の個体についての遺伝子型と表現型との間の関係の詳細に移る。
【００８５】
図３は、遺伝子バリアントの例を例示する。例示的な参照遺伝子配列Ａ　３０２は、Ｎの長さを保有し、各配列位置は、チミン、アデニン、グアニン、又はシトシンのいずれかのヌクレオチド塩基を含む。参照遺伝子配列Ａ　３０２は、特定の遺伝子のセグメントについての天然の、又は正常な遺伝子配列とみなされ得る。参照遺伝子配列Ａ　３０２は、２つの例示的なバリアント配列１Ａ　３２２及び１Ｂ　３４２と比較され、バリアント配列は、５番目の塩基位置にそれぞれの単一ヌクレオチドバリアントを保有するが、それ以外は参照配列と同一の組成を保有する。例えば、単一ヌクレオチド置換は、参照遺伝子配列Ａ　３０２におけるシトシン３０４と比較して、バリアント１Ａ　３２２におけるアデニン３２４及びバリアント１Ｂ　３４２におけるチミン３４４として示される。参照遺伝子配列Ａ　３０２、バリアント１Ａ　３２２、及びバリアント１Ｂ　３４２が転写され、タンパク質のアミノ酸単位に翻訳される場合、単一ヌクレオチドバリアント３２４及び３４４は、それぞれ、天然タンパク質中に存在するアミノ酸と比較して、少なくとも１つの類似又は異なるアミノ酸をもたらし得る。バリアント１Ａ　３２２又はバリアント１Ｂ　３４２が、それぞれ、変換されたタンパク質において少なくとも１つの変化したアミノ酸をもたらす場合、それぞれのバリアントはまた、それぞれのバリアントを保有する個体について変化した表現型をもたらし得る。
【００８６】
図４は、遺伝子型の変化に応答した表現型効果を例示する例である。例示的な参照遺伝子配列Ａ　４０２は、Ｎの長さを保有し、各配列位置は、チミン、アデニン、グアニン、又はシトシンのいずれかのヌクレオチド塩基を含む。参照遺伝子配列Ａ　４０２は、特定の遺伝子のセグメントについての天然の、又は正常な遺伝子配列とみなされ得る。参照遺伝子配列Ａ　４０２は、３つの例示的なバリアント配列１Ａ　４２２、１Ｂ　４４２、及び１Ｃ　４６２と比較され、バリアント配列は、５番目の塩基位置にそれぞれの単一ヌクレオチドバリアントを保有するが、それ以外は参照配列と同一の組成を有する。例えば、単一ヌクレオチド置換は、参照遺伝子配列Ａ　４０２におけるシトシン４０４と比較して、バリアント１Ａ　４２２におけるアデニン４０４、バリアント１Ｂ　４４２におけるチミン４４４、及びバリアント１Ｃ　４６２におけるグアニン４６４として示される。
【００８７】
参照遺伝子配列Ａ　４０２は、参照遺伝子配列Ａ　４０２を含む特定の遺伝子の天然タンパク質組成物であるタンパク質Ａ　４０６をもたらす。天然タンパク質Ａ　４０６を保有する個体は、健常な表現型である表現型Ａ　４０８を提示することになる。バリアント１Ａ　４２２は、変異タンパク質１Ａ　４２６をもたらし、変異タンパク質１Ａ　４２６は、天然タンパク質Ａ　４０６とは異なるタンパク質構造及び機能をもたらすミスセンス変異を含む。ミスセンスタンパク質１Ａ　４０６を保有する個体は、疾患表現型である表現型１Ａ　４２８を提示することになる。バリアント１Ｂ　４４２は、変異タンパク質１Ｂ　４４６をもたらし、変異タンパク質１Ｂ　４４６は、天然タンパク質Ａ　４０６と比較してタンパク質構造及び機能に変化をもたらさない同義変異を含む（すなわち、バリアント１Ｂ　４４２の５位における単一ヌクレオチドバリアントに応答してアミノ酸変化が起こらない）。同義タンパク質１Ｂ　４４６を保有する個体は、表現型１Ｂ　４４８、表現型Ａ　４０８に類似する健常な表現型を提示することになる。バリアント１Ｃ　４６２は、変異タンパク質１Ｃ　４６６をもたらし、変異タンパク質１Ｃ　４６６は、天然タンパク質Ａ　４０６と比較して、切断された非機能的タンパク質構造及び機能をもたらすナ

ンセンス変異を含む。ナンセンスタンパク質１Ｃ　４６６を保有する個体は、表現型１Ｃ　４６８を提示することになり、生存不能な胚又は有意に短縮された寿命をもたらす可能性が高い。

【００８８】

当業者であれば、バリアント１Ａ　４２２、１Ｂ　４４２、及び１Ｃ　４６２が簡略化された例として列挙されており、潜在的な表現型表現が、限られた数の別個の表現ではなく広いスペクトルに及ぶことを認識するであろう。更に、当業者はまた、多くの表現型応答が、単一遺伝子内の複数のバリアント、又は組み合わせられた多遺伝子バリアント効果に起因して生じることを認識するであろう。開示される技術の多くの実施態様は、個体の生活の質及び／又は平均余命に実質的な損害を引き起こすことが知られている重度の遺伝的障害と関連付けられた特定の表現型応答についての多遺伝子リスクスコアリングに具体的に対処する。

【００８９】

表現型

次に、表現型表現について更に詳細に考察する。図４は、広範な生物レベルにおける複数の表現型の例を例示するが、図５は、ここで、定量的バイオマーカー値又はカテゴリ臨床診断のいずれかとして説明されることができる標的化された測定可能なレベルにおける複数の表現型の例を例示する。

【００９０】

本出願は、「定量的バイオマーカー値」及び「定量的表現型」を互換的に使用することに留意されたい。本出願はまた、「カテゴリ臨床診断」及び「定性的表現型」も互換的に使用することに留意されたい。

【００９１】

図５は、心血管疾患患者Ｘ　５００に対応する複数の表現型を例示する。心血管疾患患者Ｘ　５００は、患者Ｘのゲノム内の複数の遺伝的バリアントに応答した複数の観察可能な表現型によって説明され得る。これらの表現型は、定性的であってもよく、又は定量的であってもよい。定性的表現型は、人口統計出力５０２（例えば、祖先、生物学的性別など）、カテゴリバイオマーカー５０４（例えば、血液中の特定のタンパク質の存在についてのバイナリ変数、決定閾値によって決定される血液中の患者の代謝産物レベルのバイナリ分類、又は特定の組織学的形態の存在についてのマルチクラス変数）、臨床診断５０６（例えば、心血管疾患診断についてのバイナリ変数、特定のがん診断についてのバイナリ変数など）、又は疾患重症度の尺度（例えば、腫瘍転移の進行段階、難治性セリアック病１型対２型などの疾患サブタイプの分類など）を含み得る。

【００９２】

定量的表現型は、正確な血液又は尿バイオマーカー値５２２（例えば、クレアチン、コレステロール、低密度リポタンパク質（low-density lipoprotein、ＬＤＬ）、トリグリセリド、グルコースなど）、ボディマス測定値５４２（例えば、身長、体重、体脂肪率、ボディマス指数など）、又はバイタルサイン（例えば、安静時心拍数、収縮期及び拡張期血圧、呼吸数など）を含み得る。当業者は、心血管疾患患者Ｘ　５００について列挙された定量的及び定性的表現型が非限定的な例であり、個体の健康及び組成を説明するために採用され得る無限数の観察可能な値が存在することを認識するであろう。

【００９３】

遺伝子－表現型関連性検定

これまでの考察は、特定の個体を説明するために使用されるゲノム及び表現型のデータについての定義及び文脈を包含する。図１は、特定の表現型についての遺伝子解像度レベルにおける特定の希少有害バリアントキャリア状態についての遺伝的関連性を決定する方法を高レベルで説明する。ここで、ゲノムワイド関連解析の導入、並びに一般的及び希少バリアントの両方についてのそれらの実装を介して、より詳細にこの方法を再考する。希少バリアントについての従来のゲノムワイド関連解析とは対照的に、考察は、少なくとも１つの特定の希少有害バリアントを含む複数の特定の遺伝子についての複数の希少バリア

ント総和検定が、複数の特定の遺伝子及び特定の表現型についての希少バリアント多遺伝子リスクスコアを決定するために実装される、改善された有用性を有する新規の方法論に向けられる。この考察は、希少有害バリアントが重度の早発性遺伝子疾患に関与する統計的可能性に起因して、特に希少有害バリアントに焦点を当てている。

【００９４】

図６は、一般的バリアント対希少バリアントについてのゲノムワイド関連解析をグラフで対比している。グラフ６０２及び６０４は、それぞれ、コレステロールレベル及び一般的バリアント又は希少バリアントについてのゲノムワイド関連性を説明する。グラフ６０２及び６０４の両方について、交互の色は、連続する染色体を示す。グラフ６０２は、データ点として個々の一般的バリアントを示すのに対して、グラフ６０４は、個々の遺伝子（すなわち、モデル１２２に示されるように単一キャリア状態に凝縮された特定の遺伝子についての全ての希少バリアント）を示す。グラフ６０２に示される一般的バリアントゲノムワイド関連解析は、コレステロールレベルと強く関連付けられるゲノム内の多くの場所を見つける。互いに近い一般的バリアントが頻繁に相関するという事実に起因して、関連性は、多くの関連するバリアントを有するピークとして生じる。結果として、特定の関連する遺伝子座内の任意の原因バリアントを検出することは、かなりの努力を必要とし得る。

【００９５】

グラフ６０４は、ヒトゲノム中の１９，５００個のタンパク質コード遺伝子についての希少バリアントゲノムワイド関連解析結果を示しており、少数の希少バリアントが強く関連付けられている。それらの希少性に起因して、希少バリアントゲノムワイド関連解析は、グラフ６０２に示される一般的バリアントゲノムワイド関連解析と比較して、あまり強力ではない。希少バリアントは、典型的には、互いに相関せず、したがって、効果は近くの遺伝子には及ばないことになる。したがって、有意な遺伝子は、グラフ６０２内のクラスタではなく単離されているように見える。タンパク質配列を変化させるバリアントは、特異的に検定され、これは、原因バリアントが、希少バリアントゲノムワイド関連解析によって特定された有意な遺伝子について特定され得ることを意味する。

【００９６】

図６は、ゲノムワイドレベルにおける複数の遺伝的関連性を例示するが、ここでの考察は、特定の個々のバリアントについての遺伝的関連性検定（すなわち、特定の個々の一般的バリアント対立遺伝子用量と、特定の個々の一般的バリアント対立遺伝子用量値の関連する遺伝子における存在に対する表現型応答との結果的な非ランダムな関連性）、又はバリアントの特定の集約についての遺伝的関連性（すなわち、関連する遺伝子の各々のキャリア状態と、１つ以上の希少病原性バリアントの関連する遺伝子における存在に対する表現型応答との間のコホートにおける結果的な非ランダムな関連性）に移る。

【００９７】

図７は、それぞれ、遺伝子解像度における特定の個々の一般的バリアント又は特定の集約された希少バリアントに対する遺伝的関連性検定を例示する。遺伝子型判定アレイからの単一一般的バリアント遺伝的関連性検定７０２については、各個々のバリアントが、個々の一般的バリアントについてのマイナー対立遺伝子の数（ボックスプロット７０４、７０６、及び７０８は、それぞれ、ゼロ、１つ、又は２つのマイナー対立遺伝子についての表現型値分布に対応する）と、関心対象の表現型（例えば、コレステロール）との間の関係を求めるために評価され得る。線形回帰モデルが、特定のバリアント及び表現型応答についての関連性の有効強度スコアを測定するｐ値及びβ係数を得るために、各マイナー対立遺伝子カウント値についての平均値間で構築され得る。

【００９８】

対照的に、エクソーム配列決定から得られた希少バリアントについて、個々のバリアントごとの検定は、各希少バリアントについての発生率が低すぎるため、有用ではない。希少バリアントの遺伝的関連性を検定するために、総和検定７２２が、モデル１２２において示されているように、特定の遺伝子内の集約されたバリアントに対して実施される。単

一一般的バリアント遺伝的関連性検定７０２と比較して、集約された希少バリアント総和検定７２２は、遺伝子内の少なくとも１つの希少バリアントについてのキャリア状態（ボックスプロット７２４及び７２６は、それぞれ、非キャリア及びキャリアについての表現型値分布に対応する）と、関心対象の表現型との間の関連性の強度を測定する。希少性についてフィルタリングすることに加えて、病原性測定閾値によって決定される有害バリアントのみが含まれる。

【００９９】

開示される技術のいくつかの実施態様では、関心対象の表現型応答は、カテゴリ臨床診断である。したがって、遺伝子についての結果的な非ランダムな関連性についての有効強度スコアは、キャリア状態についてのベータ係数としてカテゴリ臨床診断表現型について決定され、ベータ係数は、総和スコアのロジスティック回帰成分を使用して決定される。

【０１００】

希少バリアント総和検定の最適化

ここで、個々に決定された希少性閾値（すなわち、最大対立遺伝子カウント）及び特定の遺伝子についての病原性スコア閾値の概念を再考する。上記の図は、特定の表現型上の特定の遺伝子についての特定の複数の希少有害バリアントの遺伝的関連性検定を説明しており、最大希少閾値及び最小病原性スコア閾値は、特定の遺伝子に特異的であり、異なる特定の遺伝子についての最大希少閾値及び最小病原性スコア閾値とは異なり得る。

【０１０１】

図８に関して以下で説明される実施態様では、「最適化」は、標的目標が、特定の遺伝子に関する最も有意なｔ検定統計量をもたらすパラメータの組み合わせを特定することであるため、希少性に関する最大閾値及び病原性に関する最小スコアを特定することを指す。しかしながら、他の実施態様では、最適化のための標的目標は、遺伝的及び医学的コンテキストに起因して異なり得る。当業者は、検定統計量（ｐ値など）の解釈が文脈依存的であり、したがって、有意性と関連付けられた目標値又は値の範囲が、サンプルサイズ、統計的検出力、遺伝子ごと又はバリアントごとの効果量の変動、誤差分散、真の効果のベースレート、及び考えられる総和検定結果と関連付けられたペイオフ計算などの、様々なパラメータによって影響されることを認識するであろう。

【０１０２】

本出願は、「希少性閾値」、「最大対立遺伝子カウント」、「対立遺伝子カウント閾値」、及び「ＡＣ」という用語を互換的に使用することに留意されたい。本出願はまた、「病原性閾値」、「病原性スコア閾値」、「最小病原性スコア閾値」、及び「ＰＳＴ」という用語を互換的に使用することに留意されたい。

【０１０３】

次に、考察は、最も有意なｐ値を有する対立遺伝子カウントと病原性スコア閾値との特定の組み合わせを決定し、その特定の組み合わせを最適な組み合わせとして使用するために開示される技術に移る。対立遺伝子カウント及び病原性スコア閾値の複数の組み合わせ内の対立遺伝子カウント及び病原性スコア閾値の各組み合わせについての各ｐ値は、多重検定（すなわち、対立遺伝子カウント及び病原性スコア閾値の複数の組み合わせにわたるグリッド検索内で実施される特定の表現型上の特定の遺伝子についての特定の複数の希少有害バリアントの遺伝的関連性検定についてのｔ検定）を考慮するために補正される。

【０１０４】

図８は、希少バリアント総和検定を最適化するための方法の概略図である。データベース８０２は、コホートｉに属する個体の群に対応するゲノム及び表現型データを含む。コホートｉ８０４は、遺伝子配列決定及び複数の表現型測定値についての検定を受けたＮ人の個体を含む。コホートｉ８０４内の個体についての配列決定データは、複数のバリアントキャリア状態に対応するデータを含む。特定の遺伝子は、３つの公知の考えられる希少有害バリアント（例えば、バリアントＡ、バリアントＢ、及びバリアントＣ）を保有し得、そのうち、個体は、キャリア（すなわち、個体がそれぞれのバリアントを保有し、したがって、それらがそのバリアントのキャリアである）又は非キャリア（すなわち、個体が

それぞれのバリアントを保有せず、したがって、それらがそのバリアントのキャリアではない）であり得る。当業者は、バリアントＡ、バリアントＢ、及びバリアントＣが例示的な例として与えられており、遺伝子が任意の数の希少有害バリアントを保有し得ることを認識するであろう。開示される技術は、バリアントを、集団内のバリアント出現率が所定の閾値を下回る場合、「希少」と定義し、遺伝子が所定の閾値を上回るタンパク質として発現されるときに、バリアントの病原性が影響に対して測定された影響を有すると予測される場合、「有害」と定義する。希少性及び病原性の両方についてのそれぞれの所定の閾値は、遺伝子特異的である（すなわち、各個々の遺伝子は、別の遺伝子と異なり得る、個々に決定された希少性閾値及び病原性閾値に対応することになる）。

【０１０５】

コホートｉ８０４内の各個体についてのゲノム及びエクソーム配列決定から得られたゲノムデータに加えて、表現型データもまた、各個体に対して利用可能である（すなわち、個体Ｎは、表現型Ｄについてx_nの測定された値を有する）。

【０１０６】

希少性及び病原性についての最適な閾値を決定するために、グリッド検索８２２は、全ての考えられる病原性スコア閾値（pathogenicity score threshold、ＰＳＴ）及び最大対立遺伝子カウント閾値（allele count threshold、ＡＣ）を含む空間にわたって検索するように実施される。グリッド８２２内では、各個々の検定（ＰＳＴ値ｍ及びＡＣ値Ｎに対応する）は、ｔ検定８２４である。各ｔ検定８２４について、帰無仮説は、バリアント群のキャリアである個体の表現型値と、バリアント群のキャリアではない個体の表現型値との差が統計的に有意ではないと述べている。対立仮説は、バリアント群のキャリアである個体の表現型値と、バリアント群のキャリアではない個体の表現型値との差が統計的に有意であると述べている。ｔ検定は、不均一な有意水準で各ｔ検定に対するｐ値を得るために、ＰＳＴ及びＡＣ値の各考えられる組み合わせに対してグリッド全体にわたって反復的に実施される。複数のＰＳＴ値及びＡＣ値の組み合わせ並びに関連するｐ値８４２内で、サブコホート（PST_m、AC_n）８４４は、指定されたＰＳＴ値ｍ及びＡＣ値Ｎを満たすＮ人の個体についてのゲノム及び表現型のデータを含む（すなわち、バリアントＢが特定のｔ検定についての病原性又は対立遺伝子カウントについての指定された閾値を満たさない場合、バリアントＢについてのキャリア状態は、集約されたバイナリキャリア状態変数内で考慮されない）。

【０１０７】

最も有意なｐ値に対応するＰＳＴ及びＡＣの組み合わせが、特定の遺伝子に対する最適なＰＳＴ及びＡＣの組み合わせ８６２として出力される。開示される技術のいくつかの実施態様では、決定された最適なＰＳＴ及びＡＣの組み合わせ８６２における最適化に続いて、希少バリアント総和検定が実施される。したがって、最適化された総和検定を実施する方法は、特定の表現型に対する特定の遺伝子における希少病原性バリアントの総和検定効果の有意性を最大化する最大対立遺伝子カウントと最小病原性スコア閾値との最適な組み合わせが特定の遺伝子についての総和検定のためのパラメータとして採用されるように、特定の遺伝子についての全ての最適化された総和検定が開始されるように実施される。

【０１０８】

タンパク質構造に基づく病原性の決定

したがって、ここまでの考察は、個体が各遺伝子において希少有害バリアントを保有するかどうかに基づいて、遺伝子にわたるリスクを集約するように構成された希少バリアント多遺伝子リスクスコアモデルを説明してきた。図１、図６、図７、及び図８に説明される希少バリアント多遺伝子リスクスコアモデルは、特定の表現型に対する特定の遺伝子における希少病原性バリアントの総和検定効果の有意性を最大化する、最大対立遺伝子カウント及び最小病原性スコア閾値の最適な組み合わせで実施される。

【０１０９】

ここで、考察は、タンパク質構造に基づく病原性決定のためのシステムの説明に移り、開示される技術の一実施態様は、バリアントの病原性を決定するように構成された畳み込

みニューラルネットワーク病原性分類器を実装する。特定の遺伝子についての特定のバリアントに対応する病原性スコア出力は、特定のｔ検定８２４における特定の遺伝子についての病原性スコア閾値として実装され得る。複数のｍ個の病原性スコア閾値は、最適病原性スコア閾値８６２などの最適最小病原性スコア閾値を決定するために、グリッド検索８２２内で検定され得る。

【０１１０】

図９は、バリアントの病原性を決定するためのシステムのプロセスを例示するフロー図９００である。ステップ９０２において、システムの配列アクセサ９０４は、参照アミノ酸配列及び代替アミノ酸配列にアクセスする。９１２において、システムの３Ｄ構造生成器９１４は、参照アミノ酸配列の３Ｄタンパク質構造を生成する。いくつかの実施態様では、３Ｄタンパク質構造は、ヒトタンパク質の相同性モデルである。一実施態様では、いわゆるＳｗｉｓｓＭｏｄｅｌ相同性モデリングパイプラインが、予測されたヒトタンパク質構造の公開リポジトリを提供する。別の実施態様では、いわゆるＨＨｐｒｅｄ相同性モデリングは、鋳型構造から標的タンパク質の構造を予測するためにＭｏｄｅｌｌｅｒと呼ばれるツールを使用する。

【０１１１】

タンパク質は、原子の集合及び３Ｄ空間におけるそれらの座標によって表される。アミノ酸は、炭素原子、酸素（Ｏ）原子、窒素（Ｎ）原子、及び水素（Ｈ）原子などの様々な原子を有することができる。原子は、側鎖原子及び骨格原子として更に分類することができる。骨格炭素原子は、アルファ炭素（Ｃα）原子及びβ炭素（Ｃβ）原子を含むことができる。

【０１１２】

ステップ９２２において、システムの座標分類器９２４は、アミノ酸ベースで３Ｄタンパク質構造の３Ｄ原子座標を分類する。一実施態様では、アミノ酸ごとの分類は、３Ｄ原子座標を２１個のアミノ酸カテゴリ（停止又はギャップアミノ酸カテゴリを含む）に帰属させることを含む。一例では、アルファ炭素原子のアミノ酸ごとの分類は、２１個のアミノ酸カテゴリの各々の下にアルファ炭素原子をそれぞれ列挙することができる。別の例では、β炭素原子のアミノ酸ごとの分類は、２１個のアミノ酸カテゴリの各々の下にβ炭素原子をそれぞれ列挙することができる。

【０１１３】

更に別の例では、酸素原子のアミノ酸ごとの分類は、２１個のアミノ酸カテゴリの各々の下に酸素原子をそれぞれ列挙することができる。更に別の例では、窒素原子のアミノ酸ごとの分類は、２１個のアミノ酸カテゴリの各々の下に窒素原子をそれぞれ列挙することができる。更に別の例では、水素原子のアミノ酸ごとの分類は、２１個のアミノ酸カテゴリの各々の下に水素原子をそれぞれ列挙することができる。

【０１１４】

当業者は、種々の実施において、アミノ酸ごとの分類が、２１個のアミノ酸カテゴリのサブセット及び異なる原子エレメントのサブセットを含むことができることを理解するであろう。

【０１１５】

本出願は、非限定的な例としてボクセル及びボクセル化を使用する。当業者であれば、様々な実施態様において、様々な次元の特徴、ベクトル、アレイなどの、データを配置及び処理する異なるフォーマットが、代替として、又は組み合わせとして使用され得ることを理解するであろう。

【０１１６】

ステップ９３２において、システムのボクセルグリッド生成器９３４は、ボクセルグリッドをインスタンス化する。ボクセルグリッドは、任意の解像度、例えば、３×３×３、５×５×５、７×７×７などを有することができる。ボクセルグリッド内のボクセルは、任意のサイズ、例えば、各辺に１オングストローム（Å）、各辺に２Å、各辺に３Åなどであることができる。当業者は、ボクセルが立方体であるので、これらの例示的な寸法が

立方体寸法を指すことを理解するであろう。また、当業者は、これらの例示的な寸法が非限定的であり、ボクセルが任意の立方体寸法を有することができることを理解するであろう。

【０１１７】

ステップ９４２において、システムのボクセルグリッドセンタラ９４４は、アミノ酸レベルで標的バリアントを経験する参照アミノ酸にボクセルグリッドを中心とする。一実施態様では、ボクセルグリッドは、標的バリアントを経験する参照アミノ酸の特定の原子の原子座標、例えば、標的バリアントを経験する参照アミノ酸のアルファ炭素原子の３Ｄ原子座標に中心付けられる。

【０１１８】

距離チャネル

ボクセルグリッド内のボクセルは、複数のチャネル（又は特徴）を有することができる。一実施態様では、ボクセルグリッド内のボクセルは、複数の距離チャネル（例えば、それぞれ、２１個のアミノ酸カテゴリ（停止又はギャップアミノ酸カテゴリを含む）に対する２１個の距離チャネル）を有する。ステップ９５２において、システムの距離チャネル生成器９５４は、ボクセルグリッド内のボクセルに対するアミノ酸ごとの距離チャネルを生成する。距離チャネルは、２１個のアミノ酸カテゴリの各々について独立して生成される。

【０１１９】

例えば、アラニン（Ａ）アミノ酸カテゴリを考慮されたい。更に、例えば、ボクセルグリッドが３×３×３のサイズであり、２７個のボクセルを有することを考慮されたい。次いで、一実施態様では、アラニン距離チャネルは、ボクセルグリッド内の２７個のボクセルに対する２７個の距離値をそれぞれ含む。アラニン距離チャネルにおける２７個の距離値は、ボクセルグリッドにおける２７個のボクセルのそれぞれの中心から、アラニンアミノ酸カテゴリにおけるそれぞれの最も近い原子まで測定される。

【０１２０】

一例では、アラニンアミノ酸カテゴリは、アルファ炭素原子のみを含み、したがって、最も近い原子は、それぞれボクセルグリッド内の２７個のボクセルに最も近接するアラニンアルファ炭素原子である。別の例では、アラニンアミノ酸カテゴリは、β炭素原子のみを含み、したがって、最も近い原子は、それぞれボクセルグリッド内の２７個のボクセルに最も近接するアラニンβ炭素原子である。

【０１２１】

更に別の例では、アラニンアミノ酸カテゴリは酸素原子のみを含み、したがって、最も近い原子は、それぞれボクセルグリッド内の２７個のボクセルに最も近接するアラニン酸素原子である。更に別の例では、アラニンアミノ酸カテゴリは窒素原子のみを含み、したがって、最も近い原子は、それぞれボクセルグリッド内の２７個のボクセルに最も近接するアラニン窒素原子である。更に別の例では、アラニンアミノ酸カテゴリは水素原子のみを含み、したがって、最も近い原子は、それぞれボクセルグリッド内の２７個のボクセルに最も近接するアラニン水素原子である。

【０１２２】

アラニン距離チャネルと同様に、距離チャネル生成器９５４は、残りのアミノ酸カテゴリの各々について距離チャネル（すなわち、ボクセルごとの距離値のセット）を生成する。他の実施態様では、距離チャネル生成器９５４は、２１個のアミノ酸カテゴリのサブセットについてのみ距離チャネルを生成する。

【０１２３】

他の実施態様では、最も近い原子の選択は、特定の原子タイプに限定されない。すなわち、対象アミノ酸カテゴリ内で、特定のボクセルへの最も近い原子が、最も近い原子の原子エレメントに関係なく選択され、特定のボクセルの距離値が、対象アミノ酸カテゴリの距離チャネルに含めるために計算される。

【０１２４】

更に他の実施態様では、距離チャネルは、原子エレメントベースで生成される。アミノ酸カテゴリについて距離チャネルを有する代わりに、又はそれに加えて、原子が属するアミノ酸に関係なく、原子エレメントカテゴリについて距離値を生成することができる。例えば、参照アミノ酸配列中のアミノ酸の原子が、７つの原子エレメント、炭素、酸素、窒素、水素、カルシウム、ヨウ素及び硫黄に及ぶことを考慮されたい。次いで、ボクセルグリッド内のボクセルは、７つの距離チャネルを有するように構成され、その結果、７つの距離チャネルの各々は、対応する原子エレメントカテゴリ内のみの最も近い原子への距離を指定する２７個のボクセルごとの距離値を有する。他の実施態様では、７つの原子エレメントのサブセットのみについての距離チャネルを生成することができる。更に他の実施態様では、原子エレメントカテゴリ及び距離チャネル生成は、同じ原子エレメント、例えば、アルファ炭素（Ｃα）原子及びβ炭素（Ｃβ）原子の変形形態に更に階層化することができる。

【０１２５】

更に他の実施態様では、距離チャネルは、原子タイプベースで生成することができ、例えば、側鎖原子のみについての距離チャネル及び骨格原子のみについての距離チャネルである。

【０１２６】

最も近い原子は、ボクセル中心から所定の最大走査半径（例えば、６オングストローム（Å））内で検索することができる。また、複数の原子が、ボクセルグリッド内の同じボクセルに最も近くてもよい。

【０１２７】

距離は、ボクセル中心の３Ｄ座標と原子の３Ｄ原子座標との間で計算される。また、距離チャネルは、同じ位置に中心付けられる（例えば、標的バリアントを経験する参照アミノ酸のアルファ炭素原子の３Ｄ原子座標に中心付けられる）ボクセルグリッドを用いて生成される。

【０１２８】

距離は、ユークリッド距離であることができる。また、距離は、原子サイズ（又は原子の影響）によって（例えば、問題の原子のレナード－ジョーンズポテンシャル及び／又はファンデルワールス原子半径を使用することによって）パラメータ化することができる。また、距離値は、最大走査半径によって、又は対象アミノ酸カテゴリ若しくは対象原子エレメントカテゴリ若しくは対象原子タイプカテゴリ内の最も遠い最も近い原子の最大観察距離値によって正規化することができる。いくつかの実施態様では、ボクセルと原子との間の距離は、ボクセル及び原子の極座標に基づいて計算される。極座標は、ボクセルと原子との間の角度によってパラメータ化される。一実施態様では、この角度情報は、ボクセルの角度チャネルを生成するために（すなわち、距離チャネルとは無関係に）使用される。いくつかの実施態様では、最も近い原子と隣接原子（例えば、骨格原子）との間の角度は、ボクセルを用いて符号化される特徴として使用されることができる。

【０１２９】

参照対立遺伝子及び代替対立遺伝子チャネル

ボクセルグリッド内のボクセルはまた、参照対立遺伝子及び代替対立遺伝子チャネルを有することができる。ステップ９６２において、システムのワンホットエンコーダ９６４は、参照アミノ酸配列内の参照アミノ酸の参照ワンホット符号化と、代替アミノ酸配列内の代替アミノ酸の代替ワンホット符号化とを生成する。参照アミノ酸は標的バリアントを経験する。代替アミノ酸は標的バリアントである。参照アミノ酸及び代替アミノ酸は、参照アミノ酸配列及び代替アミノ酸配列においてそれぞれ同じ位置に位置する。参照アミノ酸配列及び代替アミノ酸配列は、１つの例外を除いて、同じ位置ごとのアミノ酸組成を有する。例外は、参照アミノ酸配列中の参照アミノ酸及び代替アミノ酸配列中の代替アミノ酸を有する位置である。

【０１３０】

ステップ９７２において、システムの連結器９７４は、アミノ酸ごとの距離チャネルと

参照及び代替ワンホット符号化とを連結する。別の実施態様では、連結器９７４は、原子エレメントごとの距離チャネルと、参照ワンホット符号化及び代替ワンホット符号化とを連結する。更に別の実施態様では、連結器９７４は、原子タイプごとの距離チャネルと、参照ワンホット符号化及び代替ワンホット符号化とを連結する。

【０１３１】

ステップ９８２において、システムのランタイムロジック９８４は、連結されたアミノ酸ごとの／原子エレメントごとの／原子タイプごとの距離チャネル並びに参照及び代替ワンホット符号化を病原性分類器（病原性決定エンジン）を介して処理して、標的バリアントの病原性を決定し、これは次に、アミノ酸レベルで標的バリアントを生成する基礎となるヌクレオチドバリアントの病原性決定として推測される。病原性分類器は、良性及び病原性バリアントのラベル付きデータセットを使用して、例えば、誤差逆伝播アルゴリズムを使用して学習する。良性及び病原性バリアントのラベル付きデータセット、並びに病原性分類器の例示的なアーキテクチャ及び学習に関する更なる詳細は、共有に係る米国特許出願第１６／１６０，９０３号、同第１６／１６０，９８６号、同第１６／１６０，９６８号、及び同第１６／４０７，１４９号に見出すことができる。

【０１３２】

図１０は、開示される技術の一実施態様による、病原性分類器９００の例示的な処理アーキテクチャ１０００を示す。処理アーキテクチャ１０００は、処理モジュール１００６、１０１０、１０１４、１０１８、１０２２、１０１０、１０３０、１０３４、１０３８、及び１０４２のカスケードを含み、その各々は、１Ｄ畳み込み（１×１×１　ＣＯＮＶ）、３Ｄ畳み込み（３×３×３　ＣＯＮＶ）、ＲｅＬＵ非線形性、及びバッチ正規化（batch normalization、ＢＮ）を含むことができる。処理モジュールの他の例は、全結合（fully-connected、ＦＣ）層、ドロップアウト層、平坦化層、並びに良性クラス及び病原性クラスに属する標的バリアントについて指数関数的に正規化されたスコアを生成する最終ソフトマックス層を含む。図１０では、「６４」は、特定の処理モジュールによって適用される畳み込みフィルタの数を示す。図１０では、入力ボクセル１００２のサイズは１５×１５×１５×８である。図１０はまた、処理アーキテクチャ１０００によって生成された中間入力１００４、１００８、１０１２、１０１６、１０２０、１０２４、１０２８、１０３２、１０３６、及び１０４０のそれぞれの体積次元を示す。

【０１３３】

図１１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例示的コンピュータシステム１１００を示す。コンピュータシステム１１００は、バスサブシステム１１４２を介して多数の周辺デバイスと通信する少なくとも１つの中央処理ユニット（ＣＰＵ）１１４４を含む。これらの周辺デバイスは、例えば、メモリデバイス及びファイル記憶サブシステム１１３６を含む記憶サブシステム１１１０、ユーザインターフェース入力デバイス１１３８、ユーザインターフェース出力デバイス１１４８、並びにネットワークインターフェースサブシステム１１４６を含むことができる。入力デバイス及び出力デバイスは、コンピュータシステム１１００とのユーザ対話を可能にする。ネットワークインターフェースサブシステム１１４６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。

【０１３４】

一実施態様では、病原性分類器１０００は、記憶サブシステム１１１０及びユーザインターフェース入力デバイス１１３８に通信可能にリンクされている。

【０１３５】

ユーザインターフェース入力デバイス１１３８は、キーボード、マウス、トラックボール、タッチパッド、又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、音声認識システム及びマイクロフォンなどのオーディオ入力デバイス、並びに他のタイプの入力デバイスを含むことができる。一般に、用語「入力デバイス」の使用は、コンピュータシステム１１００に情報を入

力するための全ての可能なタイプのデバイス及び方式を含むことを意図している。
【０１３６】
ユーザインターフェース出力デバイス１１４８は、ディスプレイサブシステム、プリンタ、ファックス装置、又はオーディオ出力デバイスなどの非視覚ディスプレイを含むことができる。ディスプレイサブシステムは、ＬＥＤディスプレイ、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、投影デバイス、又は可視画像を作成するための何らかの他の機構を含むことができる。ディスプレイサブシステムはまた、オーディオ出力デバイスなどの非視覚ディスプレイを提供することができる。一般に、用語「出力デバイス」の使用は、コンピュータシステム１１００からユーザ又は別のマシン若しくはコンピュータシステムに情報を出力するための、全ての可能なタイプのデバイス及び方式を含むことを意図している。
【０１３７】
記憶サブシステム１１１０は、本明細書に説明されるモジュール及び方法のうちのいくつか又は全ての機能を提供するプログラミング及びデータ構築物を記憶する。これらのソフトウェアモジュールは、一般にプロセッサ１１４８によって実行される。
【０１３８】
プロセッサ１１４８は、グラフィック処理ユニット（ＧＰＵ）、フィールドプログラマブルゲートアレイ（ＦＰＧＡ）、特定用途向け集積回路（ＡＳＩＣ）、及び／又は粗粒化再構成可能構造（ＣＧＲＡ）であることができる。プロセッサ１１４８は、Ｇｏｏｇｌｅ Ｃｌｏｕｄ Ｐｌａｔｆｏｒｍ（商標）、Ｘｉｌｉｎｘ（商標）及びＣｉｒｒａｓｃａｌｅ（商標）などの深層学習クラウドプラットフォームによってホスティングすることができる。プロセッサ１１４８の例は、ＧｏｏｇｌｅのＴｅｎｓｏｒ Ｐｒｏｃｅｓｓｉｎｇ Ｕｎｉｔ（ＴＰＵ）（商標）、ＧＸ４ Ｒａｃｋｍｏｕｎｔ Ｓｅｒｉｅｓ（商標）、ＧＸ１１ Ｒａｃｋｍｏｕｎｔ Ｓｅｒｉｅｓ（商標）のようなラックマウントソリューション、ＮＶＩＤＩＡ ＤＧＸ－１（商標）、Ｍｉｃｒｏｓｏｆｔ’ Ｓｔｒａｔｉｘ Ｖ ＦＰＧＡ（商標）、ＧｒａｐｈｃｏｒｅのＩｎｔｅｌｌｉｇｅｎｔ Ｐｒｏｃｅｓｓｏｒ Ｕｎｉｔ（ＩＰＵ）（商標）、Ｓｎａｐｄｒａｇｏｎ ｐｒｏｃｅｓｓｏｒｓ（商標）を有するＱｕａｌｃｏｍｍのＺｅｒｏｔｈ Ｐｌａｔｆｏｒｍ（商標）、ＮＶＩＤＩＡのＶｏｌｔａ（商標）、ＮＶＩＤＩＡのＤＲＩＶＥ ＰＸ（商標）、ＮＶＩＤＩＡのＪＥＴＳＯＮ ＴＸ１／ＴＸ１１ ＭＯＤＵＬＥ（商標）、ＩｎｔｅｌのＮｉｒｖａｎａ（商標）、Ｍｏｖｉｄｉｕｓ ＶＰＵ（商標）、Ｆｕｊｉｔｓｕ ＤＰＩ（商標）、ＡＲＭのＤｙｎａｍｉｃＩＱ（商標）、ＩＢＭ ＴｒｕｅＮｏｒｔｈ（商標）、Ｔｅｓｔａ Ｖ１００ｓ（商標）を有するＬａｍｂｄａ ＧＰＵ Ｓｅｒｖｅｒ、及び他のものを含む。
【０１３９】
記憶サブシステム１１１０で使用されるメモリサブシステム１１２２は、プログラム実行中に命令及びデータを記憶するためのメインランダムアクセスメモリ（ＲＡＭ）１１３２と、固定命令が記憶された読み取り専用メモリ（ＲＯＭ）１１３４とを含む多数のメモリを含むことができる。ファイル記憶サブシステム１１３６は、プログラム及びデータファイルのための永続的な記憶装置を提供することができ、ハードディスクドライブ、関連する取り外し可能な媒体、ＣＤ－ＲＯＭドライブ、光学ドライブ、又は取り外し可能な媒体カートリッジを含むことができる。特定の実施態様の機能を実施するモジュールは、記憶サブシステム１１１０内のファイル記憶サブシステム１１３６によって、又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することができる。
【０１４０】
バスサブシステム１１４０は、コンピュータシステム１１００の様々な構成要素及びサブシステムを、意図されるように互いに通信させるための機構を提供する。バスサブシステム１１４２は、単一のバスとして概略的に示されているが、バスサブシステムの代替の実施態様は、複数のバスを使用することができる。
【０１４１】
コンピュータシステム１１００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、緩くネットワーク化されたコンピュータの緩く分散したセット、又は任意の他のデータ処理システム若しくはユーザデバイスを含む様々なタイプのものであり得る。コンピュータ及びネットワークは絶え間なく変化する性質のものであるため、図１１に示されるコンピュータシステム１１００の説明は、本発明の好ましい実施態様を例示する目的のための特定の実施例としてのみ意図される。コンピュータシステム１１００の多くの他の構成は、図１１に示されるコンピュータシステムよりも多くの又は少ない構成要素を有することができる。

【０１４２】

図１２は、最終的な病原性スコアを決定する一実施態様１２００を例示する。アクション１２０２において、一実施態様では、病原性分類器１０００は、第１の参照アミノ酸と同じ第１の代替アミノ酸について第１の病原性スコアを生成する。アクション１２１２において、一実施態様では、病原性分類器１０００は、第１の参照アミノ酸とは異なる第２の代替アミノ酸に対する第２の病原性スコアを生成する。アクション１２２２において、一実施態様では、第２の代替アミノ酸についての最終的な病原性スコアは、第２の代替アミノ酸についての第２の病原性スコアである。

【０１４３】

他の代替において、第２の代替アミノ酸についての最終的な病原性スコアは、第１の病原性スコア及び第２の病原性スコアの組み合わせに基づく。１２２２ａにおける第１の代替では、一実施態様では、第２の代替アミノ酸についての最終的な病原性スコアは、第１の病原性スコアと第２の病原性スコアとの合計に対する第２の病原性スコアの比である。１２２２ｂにおける第２の代替では、一実施態様では、第２の代替アミノ酸についての最終的な病原性スコアは、第２の病原性スコアから第１の病原性スコアを減算することによって決定される。

【０１４４】

当業者であれば、他の現在及び将来の人工知能、機械学習、及び深層学習モデル、データセット、及び学習技法を、開示される技術の趣旨から逸脱することなく、開示されるバリアント病原性分類器に組み込むことができることを理解するであろう。

【０１４５】

多重検定のための補正

これまでの議論は、表現型と関連付けられた遺伝子と表現型応答への希少バリアントの寄与との間の関連性の強度が定量化される最適化された総和検定を実施する方法を含む開示される技術の実施態様を説明しており、総和検定は、対立遺伝子カウントと、最も有意なｐ値を有する病原性スコア閾値との特定の組み合わせに対して最適化される。最適化された総和検定は、特定の遺伝子内の少なくとも１つの希少有害バリアントのキャリアと非キャリアとの間の分離を最大化する複数の入れ子型ｔ検定に基づく。開示される技術のこの実施態様では、最適化方法内で実施される多重ｔ検定は、各遺伝子内の多重検定を補正する必要性を保証する。

【０１４６】

次に、考察は、ＰＳＴ値とＡＣ値との最適な組み合わせに対応する最も有意なｐ値の補正の説明に移る。

【０１４７】

図１３は、各遺伝子内の多重検定を補正するためのプロセスのフロー図１３００である。偽発見率（false discovery rate、ＦＤＲ）補正に続いて、ｐ値は、ＰＳＴ及びＡＣの両方についての総和検定最適化を考慮するために多重検定について更に補正される。ＦＤＲ補正されたｐ値１３０２は、値範囲に分離される。開示される技術の一実施態様では、多重検定補正１３００のための以下のアルゴリズムに従う。開示される技術の他の実施態様では、他の形態の多重検定補正が使用され得る。範囲（０，１ｅ－５］内の値について、ＦＤＲ補正されたｐ値１３０２は、Ｂｅｎｊａｍｉｎｉ－Ｈｏｃｈｂｅｒｇ　ＦＤＲ補正１３１０を受ける。

【０１４８】
範囲（０．０１，１］内の値について、ＦＤＲ補正されたｐ値１３０２は、並べ替え検定ステップ１３１２を受ける。ステップ１３１２において、一実施態様では、生成された並べ替えの総数のカウント（Ｎ）がゼロに設定される。他の実施態様では、Ｎは、任意の所定の開始値に設定され得る。他の実施態様では、Ｎは、任意の所定の開始値に設定され得る。ステップ１３１２に続いて、ステップ１３２２は、表現型ラベルの１，０００個の並べ替えの生成を伴い、ＮをＮ＋１０００に等しく設定する。次のステップ１３４２において、データの各並べ替えに対して総和検定が実施され、元の観察されたデータよりも有意なｐ値をもたらす並べ替えの割合ｐがカウントされる。Ｎが１００／ｐよりも大きい場合、ステップ１３６２に進む。ステップ１３６２において、停止し、ｐを出力する。ステップ１３６２における停止点は、Ｎ＜１０，０００を保証し、計算効率を維持する。すなわち、並べ替えの数（Ｎ）は、１０，０００未満に制限され、それによって、計算上実行不可能な、又は計算上非効率的な、又は計算上高価なカウントに達することが防止される。

【０１４９】
範囲（１ｅ－５，０．０１］内の値について、ＦＤＲ補正されたｐ値１３０２は、並べ替え検定ステップ１３１４を受ける。ステップ１３１４において、表現型ラベルの１０，０００個の並べ替えが生成される。ステップ１３１４に続いて、ステップ１３２４は、一般化された極値分布を、並べ替えられたデータからの各それぞれの総和検定に対する検定統計値の絶対値にフィッティングすることを伴う。次のステップ１３４４において、補正されたｐ値は、フィッティングされた分布の曲線下面積から推定される。すなわち、ステップ１３６２では、初期ｐ値が出力されるが、それに対して、ステップ１３４４では、補正されたｐ値が推定される。

【０１５０】
共変量についての表現型値の補正
これまでの考察は、表現型と関連付けられた遺伝子の関連性の強度及び表現型応答に対する希少バリアントの寄与を定量化することを包含しており、希少多遺伝子リスクスコアに寄与する複数の遺伝子の有効強度スコアは、ブールキャリア状態変数によって決定され、希少多遺伝子リスクスコアモデル（すなわち、総和検定）は、対立遺伝子カウントと、最も有意なｐ値を有する病原性スコア閾値との特定の最適な組み合わせについて最適化される。ここで、表現型測定値が共変量補正され、かつ薬物使用補正される、最適化の追加の方法を導入するために、表現型測定値の概念を再考する。

【０１５１】
本出願は、「薬物使用」、「薬物使用パターン」、「薬物カテゴリ使用」、及び「薬物カテゴリ使用パターン」という用語を互換的に使用することに留意されたい。また、本出願は、１つ以上の他の共変量及び関心対象の転帰の両方に関連付けられた共変量を説明するために、「交絡因子」という用語を使用することに留意されたい（例えば、タバコ使用は、心血管疾患の薬物使用パターン及び表現型測定と関連付けられた交絡因子である。当業者は、開示される技術が、腫瘍学的疾患、糖尿病などの、任意の疾患に適用され得ることを理解するであろう。

【０１５２】
図１４は、共変量について表現型値を補正する方法１４００の概略図である。ステップ１４０２において、定量的表現型のデータセットは、非冗長データセットに剪定される。ステップ１４０４において、複数の関心対象の薬物は、薬物カテゴリのセットに群化される。ステップ１４０６において、非冗長定量的表現型データセット内の各表現型は、薬物カテゴリのセット内の各薬物カテゴリについての薬物使用について補正される（例えば、スタチン、血圧薬など）。ステップ１４０８において、各補正された表現型は、逆ランク正規変換される。ステップ１４１０において、補正された表現型は、複数の交絡因子（例えば、年齢、性別、遺伝的主成分、食事、喫煙状態など）について更に補正される。交絡因子に関して、交絡因子は、試験薬物と転帰との間の関係を交絡させる因子である。いく

つかの試験は、試験中の補助的薬物又は精神療法などの補助的治療を可能にし得る。これらは、結果が検定される薬物よりもむしろ補助剤に起因し得るため、交絡因子である。一例として、鎮痛剤の試験は、副木、クリーム、マッサージ、温浴、及びカイロプラクタの操作などの、非薬物鎮痛剤の参加者の使用を無視することが多い。これらの追加の試験治療に対する応答は、試験の結果を著しく混乱させる可能性がある。一般的な交絡因子は、参加者の属性であり、例えば、ボディマス指数、喫煙状態、病気の発症年齢、社会経済的状態、教育的状態、及びサポートネットワークの程度である。人生の出来事もまた、潜在的な交絡因子である。それらは、参加者の気分及び症状レベルに大きな変化を引き起こし得、これは、特に、参加者と評価者との間の関係において、どんなに小さい出来事にも当てはまる。実験群と対照群との間で一致しない全ての既知の交絡因子は、統計分析において考慮され得る。

【０１５３】

ここで、考察は、第１及び第２の時点における個体のコホートの複数の表現型に対する複数の薬物の使用に応答した表現型シフトを予測するように構成された、開示される技術の一実施態様の詳細な説明に移り、コホート内の個体は、２つの時点で測定された並べ替えられた薬物使用パターンに基づいて群化される（例えば、第１及び第２の時点で薬物を服用せず、第１及び第２の時点の間で薬物を服用し始め、第１及び第２の時点の間で薬物を服用することを停止し、第１及び第２の時点で薬物を服用する）。当業者は、開示される技術が、３つ、４つ、５つなどの、任意の数の時点に拡張され得ることを理解するであろう。生物統計学及び集団遺伝学研究において一般的に使用される共変量補正のための他の戦略とは対照的に、単一時点ではなく複数時点の使用は、より低い誤差分散及び表現型シフトに対する特定の交絡因子の影響のより正確な近似をもたらす。

【０１５４】

図１５は、複数の表現型に対する複数の薬物の使用に応答して表現型シフトを予測することを表す概略図である。データベース１５０２は、コホートｊに属する個体群に対応するゲノム及び表現型のデータを含む。ステップ１５２２において、複数の交絡因子についての共変量測定値、複数の表現型についての表現型測定値、及び複数の薬物カテゴリについての薬物使用パターンが、データベース１５０２から２つの時点t_1及びt_2においてアクセスされる。ステップ１５４２、１５６２、１５８２、及び１５９２は、複数の表現型内の各表現型について実施される。

【０１５５】

ステップ１５４２において、表現型測定値は、t_1及びt_2における複数の交絡因子について共変量補正され、第１の回帰モデルをフィッティングすることによって交絡因子補正された値を生成する。ステップ１５６２において、第１の回帰モデルが、交絡因子補正された値に対するデルタδを決定するために使用される。したがって、共変量補正は、第１の回帰モデルをフィッティングすることによって共変量測定値を除去して回帰分析することによって実装される。ステップ１５８２及び１５９２は、複数の薬物カテゴリ内の各薬物カテゴリについて実施される。ステップ１５８２において、δは、第２の回帰モデルをフィッティングすることによって、特定の薬物カテゴリの使用に応答した表現型シフトを予測するために使用される。ステップ１５９２において、交絡因子補正された値は、特定の薬物カテゴリについての薬物使用について更に補正されて、特定の表現型についての交絡因子補正され、薬物カテゴリ使用補正された値を生成する。したがって、表現型シフト予測は、薬物使用パターンの各々について表現型シフトをモデル化する第２の回帰モデルをフィッティングすることによって実装される。第２の回帰モデルは、薬物使用パターンの各々についてバイナリインジケータ変数を有する。

【０１５６】

図１６は、特定のコホートについての薬物使用パターン及び表現型データを得るための実験設定を表す概略図である。データベース１６０２は、コホートｊに属する個体群に対応するゲノム及び表現型のデータを含む。データベース１６０２は、第１の時点t_1において、特定の薬物カテゴリｚを服用していない個体１６２２と、特定の薬物カテゴリｚを

服用している個体１６２４とをそれぞれ含む２つのサブコホートにセグメント化される。特定の薬物カテゴリｚを服用していない個体１６２２を含むサブコホートは、それぞれ、特定の薬物カテゴリｚを継続的に服用しなかった個体１６４２と、第１の時点t_1の後、第２の時点t_2の前に特定の薬物カテゴリｚを服用し始めた個体１６４４とを含む、より小さいサブコホートに更にセグメント化され得る。

【０１５７】

特定の薬物カテゴリｚを服用している個体１６４６を含むサブコホートは、それぞれ、特定の薬物カテゴリｚを服用することを停止した個体１６４６と、第１の時点t_1の後、第２の時点t_2の前に特定の薬物カテゴリｚを服用し続けていた個体１６４８とを含む、より小さいサブコホートに更にセグメント化され得る。

【０１５８】

表現型値傾向グラフ１６６２は、サブコホート１６４２（すなわち、t_1又はt_2において薬物カテゴリｚを服用していなかった患者）についてt_1及びt_2において測定された表現型値の例を例示し、特定の表現型値は、t_1においてｙ軸上で高くなり始め、t_2においてわずかに増加する。表現型値傾向グラフ１６６４は、サブコホート１６４４（すなわち、t_1において薬物カテゴリｚを服用していなかったが、t_2の前に薬物カテゴリｚを服用し始めた患者）についてt_1及びt_2において測定された表現型値の例を例示し、特定の表現型値は、t_1においてｙ軸上で高くなり始め、t_2において減少する。表現型値傾向グラフ１６６６は、サブコホート１６４６（すなわち、t_1において薬物カテゴリｚを服用していたが、t_2の前に薬物カテゴリｚを服用することを停止した患者）についてt_1及びt_2において測定された表現型値の例を例示し、特定の表現型値は、t_1においてｙ軸上で低下し始め、t_2において増加する。表現型値傾向グラフ１６６８は、サブコホート１６４８（すなわち、t_1及びt_2において薬物カテゴリｚを服用していた患者）についてt_1及びt_2において測定された表現型値の例を例示し、特定の表現型値は、t_1においてｙ軸上で低下し始め、t_2において低いままである。

【０１５９】

開示される技術のいくつかの実施態様では、表現型値（すなわち、薬物カテゴリｚに応答した表現型シフト）に対する薬物効果は、モデル１６８２によって学習され得、t_1及びt_2における表現型値に対するデルタ（δ_y）は、薬物カテゴリｚのβ係数を伴う回帰としてモデル化される。このモデルは、帰無仮説が、薬物カテゴリｚについてのβ係数がゼロに等しいと述べ、対立仮説が、薬物カテゴリｚについてのβ係数がゼロに等しくないと述べる場合、有意性について検定され得る。

【０１６０】

開示される技術のいくつかの実施態様では、データベース１６０２からのデータに基づいて構築されるステップ１５４２及び１５８２からの回帰モデルは、以下のプロトコルを介して構築される。

【０１６１】

３４個の薬物カテゴリ（例えば、スタチン、ＮＳＡＩＤ、オピオイドなど）の表現型シフト効果を、５年離れたt_1及びt_2で測定されたそれらの定量的表現型値及び対応する共変量を有した参加者のコホートから推定した。各定量的表現型について、t_1及びt_2における共変量補正された値は、薬物使用を除く全ての共変量（例えば、年齢、性別、遺伝的主成分、食事、喫煙状態など）を除去して回帰分析することによって生成される。２つの時点間の差は、以下のように計算される。

【０１６２】

【数１】

$$\delta = \tilde{Y}_{t2} - \tilde{Y}_{t1}$$

【０１６３】

各薬物カテゴリＸについて、コホート内の全ての個体は、t_1及びt_2におけるそれら

の薬物使用に従って４つの群に分割され（サブコホート１６４２、１６４４、１６４６、及び１６４８に説明されるように）、個体がその群に属するかどうかを符号化するバイナリインジケータ変数が各群に対して導入される。

１）t_1及びt_2の両方の時点で薬物Ｘを服用していなかった個体（インジケータ変数：X_{00}）

２）時点t_1とt_2との間に薬物Ｘの服用を開始した個体（インジケータ変数：X_{01}）

３）時点t_1～t_2の間に薬物Ｘの服用を停止した個体（インジケータ変数：X_{10}）

４）両方の時点t_1及びt_2で薬物Ｘを服用していた個体（インジケータ変数：X_{11}）

【０１６４】

どの薬物が表現型Ｙに対して有意な効果を有するかを決定するために、以下の形態の前進型選択段階的回帰がフィッティングされ、全ての薬物カテゴリにわたって繰り返す。

【０１６５】

【数２】

$$\delta = \beta_{00}X_{00} + \beta_{01}X_{01} + \beta_{10}X_{10} + \beta_{11}X_{11} + \beta_t t + \beta_m\left(\tilde{Y}_{t1} - \text{平均}\left(\tilde{Y}_{t1}\right)\right) + \beta_0$$

【０１６６】

上記の方程式において、項$\beta_t t$は、t_1とt_2との間の経過した時間の効果をモデル化しており（コホート内の各個体についての$t = t_2 - t_1$）、項

【０１６７】

【数３】

$$\beta_m\left(\tilde{Y}_{t1} - \text{平均}\left(\tilde{Y}_{t1}\right)\right)$$

は、t_1とt_2との間の平均への回帰に起因する効果を考慮しており、β_0は、切片である。出発薬物Ｘの効果をモデル化する係数β_{01}のｐ値が10^{-3}を上回って増加したとき、段階的手順が停止される。

【０１６８】

関連薬物のセットＤが決定された後、それらの個々の効果は、以下の回帰をフィッティングすることによって一緒に推定され得る。

【０１６９】

【数４】

$$\delta = \sum_{d \in D}\left(\beta_{00}^{(d)}X_{00}^{(d)} + \beta_{01}^{(d)}X_{01}^{(d)} + \beta_{10}^{(d)}X_{10}^{(d)} + \beta_{11}^{(d)}X_{11}^{(d)}\right) + \beta_t t + \beta_m\left(\tilde{Y}_{t1} - \text{平均}\left(\tilde{Y}_{t1}\right)\right) + \beta_0$$

【０１７０】

【０１７０】

コホート内の全ての個体にわたるt_1における表現型Ｙについての生の値は、以下のように補正され得る。

【０１７１】

【数５】

$$\tilde{Y}_{t1} = Y_{t1} - \sum_{d \in D}\left(\beta_{01}^{(d)} - \beta_{00}^{(d)}\right)X^{(d)}$$

【０１７２】
Ｘ（ｄ）は、個体が初診時ｔ１において薬物Ｘを服用していたかどうかを符号化するバイナリインジケータ変数である。薬物使用について補正した後、値
【０１７３】
【数６】

$$\tilde{Y}_{t1}$$

は、逆ランク正規変換され得、全ての他の共変量が、除去されて回帰分析され得る。共変量補正され正規化され薬物使用補正された表現型測定値が、希少バリアント多遺伝子リスクスコアを生成するために使用され得る。
【０１７４】
開示される技術のいくつかの実施態様では、薬物使用補正は、第３の回帰モデルをフィッティングすることによって実装され、第１の時点と第２の時点との間で薬物を服用し始める第１の薬物使用パターンについての第１のバイナリインジケータ独立変数、第１及び第２の時点で薬物を服用しない第２の薬物使用パターンについての第２のバイナリインジケータ独立変数、並びに個体が第１の時点で特定の薬物を服用していたかどうかを符号化する薬物特異的バイナリインジケータ独立変数に基づいて、第１の時点についての表現型測定値を薬物使用補正することを含む。第４の回帰モデルがフィッティングされ、第４の回帰モデルは、第１の時点と第２の時点との間で薬物を服用することを停止する第３の薬物使用パターンについての第３のバイナリインジケータ独立変数、第１の時点及び第２の時点で薬物を服用する第４の薬物使用パターンについての第４のバイナリインジケータ独立変数、個体が第２の時点で特定の薬物を服用していたかどうかを符号化する薬物特異的バイナリインジケータ独立変数に基づいて、第２の時点についての表現型測定値を薬物使用補正することを含む。
【０１７５】
ランクに基づく逆正規変換が、第１及び第２の時点についての薬物使用補正された表現型測定値に適用されて、第１及び第２の時点についての正規化された薬物使用補正された表現型測定値を生成し得る。次いで、正規化された薬物使用補正された表現型測定値が、第１及び第２の時点について共変量補正され、第１及び第２の時点について共変量補正され正規化された薬物使用補正された表現型測定値を生成する。共変量補正され正規化された薬物使用補正された表現型測定値は、希少バリアント多遺伝子リスクスコアを生成するために使用され得、関心対象の表現型についての測定値は、共変量及び薬物使用パターンに応答した表現型シフトに対して補正される。
【０１７６】
図１７は、心血管疾患患者Ｘ　１７００に対応する複数の表現型を例示する。心血管疾患患者Ｘ　１７００について測定された表現型値は、複数の交絡因子によって影響されることになる。これらの交絡因子の例は、人口統計１７０２（例えば、年齢、性別、民族性など）、心理社会的因子１７２２（メンタルヘルス関連の交絡因子、社会経済的クラスなど）、タバコ及びアルコール消費１７４２、薬物使用１７０４（例えば、違法薬物などの正式な薬物効果分析に含まれない薬物）、ボディマス測定値１７２４（例えば、ボディマス指数、体脂肪率、腹部脂肪密度など）、又は食事及びライフスタイル因子１７４４（例えば、食事制限、食習慣、運動など）を含む。当業者は、これらが非限定的な交絡因子であり、測定された表現型値に影響を及ぼし得る多数の追加の交絡因子が存在することを認識するであろう。
【０１７７】
非自明性及び進歩性の客観的な指標としての性能測定結果
これまでの考察は、特定の表現型についての特定の遺伝子内の希少有害バリアントキャリア状態について構築された多遺伝子リスクスコアモデルを含む希少バリアント総和検定について開示された技術の複数の実施態様を包含しており、モデルパラメータ（すなわち

、病原性スコア閾値及び最大対立遺伝子カウント）は、入れ子型ｔ検定（すなわち、グリッド検索）を介して最適化され、両方の表現型値は、複数の共変量について共変量補正され、複数の薬物使用パターンについて薬物使用補正される。次に、考察は、開示される技術の様々な実施態様の性能結果に移る。

【０１７８】

図１８は、異なるタイプの総和検定ごとに特定された有意な遺伝子－表現型対の総数を定量化するグラフ１８００である。異なる型の総和検定について９０個の表現型にわたって特定された有意な遺伝子－表現型対の総数は、ＰｒｉｍａｔｅＡＩ－３Ｄによって優先順位付けされた希少ＬｏＦ及び有害ミスセンスバリアントについての複合検定が他のアプローチよりも性能が優れていることを示す。陰性対照として、同義バリアントに対して行われた総和検定についての有意な遺伝子型－表現型対の数も示される。

【０１７９】

図１９は、病原性分類器ＰｒｉｍａｔｅＡＩ－３Ｄによって特定された疾患重症度及び発症の年齢に影響する希少有害バリアントを例示するグラフの集合である。グラフ１９０２は、ＬＤＬＲ遺伝子、並びにＬＤＬコレステロール及びリポタンパク質代謝の障害とのその関連性に対応する。ＬＤＬＲ遺伝子における希少ミスセンスバリアントのキャリアについては、ＬＤＬコレステロールレベル（ｙ軸）は、ＰｒｉｍａｔｅＡＩ－３Ｄパーセンタイルスコア（ｘ軸）と正に相関する。以下、ＰｒｉｍａｔｅＡＩ－３Ｄスコアは、遺伝子間の比較を容易にするために各遺伝子内で０から１に正規化されたＰｒｉｍａｔｅＡＩ－３Ｄパーセンタイルスコアを指す。グラフ１９０４は、ＰｒｉｍａｔｅＡＩ－３Ｄスコアが、ＬＤＬＲ遺伝子における希少ミスセンスバリアントのキャリアについて脂質異常症の発症年齢を予測することを例示する。

【０１８０】

グラフ１９２２は、ＬＤＬＲの下方制御因子である、ＰＣＳＫ９遺伝子に対応する。ＬＤＬＲの下方制御因子である、ＰＣＳＫ９におけるミスセンスバリアントのキャリアのＬＤＬコレステロールレベルは、ＰｒｉｍａｔｅＡＩ－３Ｄスコアと負に相関する。グラフ１９２４は、キャリアのＬＤＬコレステロールレベルが、非キャリアと同様の速度で年齢とともに増加するが、平均して同じ年齢群の非キャリアよりも低いことを示す。グラフ１９４２は、ＧＣＫ遺伝子に対応しており、ＧＣＫ遺伝子における希少ミスセンスバリアントのキャリアについて、ＨｂＡ１ｃレベルがＰｒｉｍａｔｅＡＩ－３Ｄと相関することを示している。グラフ１９４４は、キャリアのＨｂＡ１ｃレベルが、非キャリアと同様の速度で年齢とともに増加するが、平均して、キャリアの人生においてより早く前糖尿病閾値に達することを示す。

【０１８１】

図２０は、異なる病原性スコアと表現型値との平均絶対スピアマン相関のグラフ２０００である。グラフ２０００は、同じバリアントのキャリアによる理論上の上限セットと比較した、３４個の遺伝子－表現型対のベンチマークセットにおける表現型値との異なる病原性スコアの平均絶対スピアマン相関を示す。

【０１８２】

図２１は、希少有害バリアントと一般的ゲノムワイド関連解析バリアントとのヒートマップ２１００の比較である。表現型の各対について、ヒートマップは、ｙ軸上の表現型と関連付けられたＧＷＡＳ遺伝子と、ｘ軸上の表現型と関連付けられた希少バリアント遺伝子との間の重複の統計的有意性を示す。

【０１８３】

図２２は、希少有害バリアントと一般的ゲノムワイド関連解析バリアントとの更なる比較である。グラフ２２００Ａは、ＬｏＦ、有害ミスセンス（ＰｒｉｍａｔｅＡＩ－３Ｄ＞０．５）、及び潜在性スプライス（ＳｐｌｉｃｅＡＩ＞０．２）バリアントについてのバリアント効果量とバリアント対立遺伝子頻度との関係を示す。同義及び良性のミスセンス（ＰｒｉｍａｔｅＡＩ－３Ｄ＜０．５）バリアントが陰性対照として示される。ドットサイズは、各ドットにおけるバリアントの数の平方根に比例する。グラフ２２００Ｂは、最

も希少なＬｏＦバリアントの効果量を、同じ遺伝子に対するリードＧＷＡＳバリアントの効果量で除算した分布を含む。高ｐＬＩ（上のプロット）及び低ｐＬＩ（下のプロット）遺伝子の両方のヒストグラムが示される。赤色の垂直線は、各分布の平均を示す。グラフ２２００Ｃは、異なるＧＷＡＳ有意性を有する遺伝子によって層別化された、希少バリアント濃縮（公称ｐ値≦０．０５）を有する明確にマッピングされたＧＷＡＳ遺伝子のパーセンテージを示す。結果は、高ｐＬＩ遺伝子（ｐＬＩ＞０．５）及び低ｐＬＩ遺伝子（ｐＬＩ＜０．５）について別々にプロットされる。高信頼度ＧＷＡＳ遺伝子は、ｐ値＜１０～１００を有するリードバリアントを有し、関連する遺伝子におけるコーディングバリアントと強いＬＤ（r^2＞＝０．９）を有するものとして定義された。破線は、公称ｐ値閾値（ｐ＜０．０５）を偶然に満たすことになる遺伝子の期待されるパーセンテージを示す。

【０１８４】

図２３は、ＵＫ　Ｂｉｏｂａｎｋコホートにおける全個体にわたるコレステロール経路及び総コレステロール分布を例示する。例示２３０２は、コレステロール経路を含み、希少バリアントＰＲＳモデルにおける遺伝子が写真上に重ねられている。各遺伝子について、数字及び矢印は、効果量及び効果の方向を示す。グラフ２３０４は、ＵＫ　Ｂｉｏｂａｎｋコホートにおける全ての個体にわたる総コレステロールの分布を示し、一般的及び希少バリアントの平均効果量が示されている。

【０１８５】

図２４は、希少バリアントＰＲＳ性能の測定を含む。グラフ２４０２は、ＵＫＢＢの９０％をＰＲＳ訓練に使用し、１０％を検定に使用した、一般的バリアントＰＲＳ（ＡＦ＞１％）、希少バリアントＰＲＳ（ＡＦ＜０．１％）、及び７２個の表現型を有する複合ＰＲＳの平均相関を示す。箱ひげ図は、形質間のピアソン相関（Ｒ）の分布を示す。グラフ２４０４は、モデルへの遺伝子又は遺伝子座の包含を決定するための異なる有意性カットオフを使用したＰＲＳ性能（表現型との平均相関）の比較である。データ点サイズは、各ＰＲＳモデルにおいて使用される遺伝子又は遺伝子座の数に対応する。

【０１８６】

図２５は、ＰＲＳ外れ値の濃縮のグラフ２５００である。グラフ２５００は、表現型外れ値である個体内のＰＲＳ外れ値の濃縮を示す。ｘ軸は、表現型外れ値を定義するために使用されるｚスコア閾値を示し、ｙ軸は、ベースライン集団に対する表現型外れ値個体におけるＰＲＳ外れ値の濃縮を示す。

【０１８７】

図２６は、定量的表現型についてのＰＲＳ外れ値のグラフ２６００である。グラフ２６００は、５６個の定量的表現型についてのＰＲＳ外れ値を示す。バーは、９９及び９９．９パーセンタイルの表現型外れ値閾値における個体を説明するための、一般的バリアントＰＲＳ又は希少バリアントＰＲＳのいずれかについての統計的有意性を示す。

【０１８８】

図２７は、２つの別個のコホートからの正規化された総コレステロール分布のグラフ２７０２及び２７０４を含む。グラフ２７０２は、コレステロールＰＲＳについて下位０．５％（低）、０．５～９９．５％（中）、及び上位０．５％（高）群にあるＵＫ　Ｂｉｏｂａｎｋ個体からの正規化された総コレステロール分布を示し、コホートの５０％は、ＰＲＳを訓練するために使用され、残りは検定のために使用される。グラフ２７０４は、コレステロールＰＲＳについて下位０．５％（低）、０．５～９９．５％（中）、及び上位０．５％（高）群にあるＭＧＢ個体からの正規化された総コレステロール分布を示し、コホートの５０％は、ＰＲＳを訓練するために使用され、残りは検定のために使用される。

【０１８９】

図２８は、２つの別個のコホート間で希少バリアントＰＲＳ外れ値及び表現型を比較するグラフ２８０２及び２８０４を含む。グラフ２８０２は、両方のコホートにおいて測定された１７個の定量的表現型についての希少バリアントＰＲＳ外れ値検定を示す。ＰＲＳ及び表現型の外れ値を、集団の上位０．５％及び下位０．５％の個体として定義した。グ

ラフ２８０４は、１７個の定量的表現型について、希少バリアントＰＲＳと表現型との間の相関を示す。

【０１９０】

図２９は、民族性による性能結果を例示するグラフ２９０２及び２９０４を含む。グラフ２９０２は、ＭＧＢコホートにおける低及び高ＰＲＳ群についての正規化されたコレステロール分布を示しており、白人及び非白人個体の両方について結果が示されている。グラフ２９０４は、白人及び非白人個体の両方について、ＭＧＢにおける１１個の表現型の各々についての低ＰＲＳ（＜０．５％）群と高ＰＲＳ（＞９９．５％）群との間の平均ｚスコア距離の比較を示す。

【０１９１】

図３０は、正常なコレステロールレベルに対して使用される一般的ＰＲＳバリアント及び希少ＰＲＳ遺伝子に対する効果量及び頻度を比較する表３０００である。

【０１９２】

図３１は、３４個の遺伝子－表現型対のセットに対する異なる病原性スコアリング方法によって説明される表現型分散の平均割合のグラフ３１００であり、３４個の遺伝子－表現型対のセットは、希少ミスセンス及びＬｏＦバリアントについてのそれらの濃縮に基づいて選択された。説明される表現型分散の割合は、各スコアリング方法からの各スコアと、対応するバリアントのキャリアの表現型値との間の二乗スピアマン相関として計算される。方法は、グラフ３１００において、同じミスセンスバリアントのキャリアの表現型値を使用することによって計算された理論上の上限と、同じ同義バリアントのキャリア及びランダムスコアである、２つの理論上の下限と比較される。

【０１９３】

図３２は、定量的表現型と臨床表現型との間の全てのペアワイズ比較についての、ＧＷＡＳ遺伝子における希少バリアントの濃縮のヒートマップ３２００である。表現型の各対について、ヒートマップ３２００は、１つの表現型（ｙ軸）のＧＷＡＳ遺伝子の最も有意なサブセットの順位付けの統計的有意性を、第２の表現型（ｘ軸）に影響を及ぼす希少有害バリアントに対するそれらの濃縮によって示す。各表現型からの全ての有意なＧＷＡＳバリアントの効果を除去して回帰分析した後、希少有害バリアントについての総和検定が計算される。ヒートマップ２１００は、ヒートマップ３２００の主対角線上に濃縮された表現型のサブセットを示す。

【０１９４】

図３３は、同じ遺伝子についてのシングルトンＬｏＦバリアントの平均効果量と最も有意なＧＷＡＳバリアントの効果量との間の比の絶対値の分布のグラフ３３００を含む。グラフ３３００Ａは、グラフ２２００Ｂからの高ｐＬＩ遺伝子に対する同じ遺伝子についてのシングルトンＬｏＦバリアントの平均効果量と最も有意なＧＷＡＳバリアントの効果量との間の比の絶対値の分布を示す。グラフ３３００Ｂは、グラフ２２００Ｂからの低ｐＬＩ遺伝子に対する同じ遺伝子についてのシングルトンＬｏＦバリアントの平均効果量と最も有意なＧＷＡＳバリアントの効果量との間の比の絶対値の分布を示す。

【０１９５】

図３４は、５％の偽発見率で有意であった遺伝子－表現型対からの遺伝子のいずれかにおける個体当たりのバリアントの数を示す。グラフ３４００は、有意性閾値を通過した１０３１個の遺伝子－表現型対からの５００個の遺伝子を含む。１人当たりのバリアントの平均数は２．０３であり、標準偏差は１．５１である。

【０１９６】

図３５は、訓練データ対検定データ分割における効果量の比較３５００を示す。コホートの５０％を訓練セットとして使用し、その他の５０％を試験セットとして使用して、希少バリアント検定を実施した。グラフ３５００では、訓練セットにおいてｐ値＜10^{-7}を有する遺伝子について、訓練セット対検定セットからの効果量が比較される。

【０１９７】

図３６は、一般的バリアントＰＲＳサブセットと希少バリアントＰＲＳサブセットとを

比較するグラフ３６００を含む。グラフ３６００Ａは、遺伝子座が一般的及び希少検定の両方において有意であることが必要とされた場合の一般的及び希少ＰＲＳ相関を示す。ＰＲＳを９０％で訓練し、残りで検定した。グラフ３６００Ｂは、訓練サンプルの数による、最大値に対する一般的及び希少ＰＲＳ相関を示す。グラフ３６００Ｃは、ＰｒｉｍａｔｅＡＩ－３Ｄを使用した希少ＰＲＳモデル対ＰｒｉｍａｔｅＡＩ－３Ｄなしで構築された希少バリアントＰＲＳモデルから説明される効果量及び分散の比を示す。

【０１９８】

条項

開示される技術、特に、このセクションで開示される条項は、システム、方法、又は製品として実施され得る。実施態様の１つ以上の特徴を、塩基実施態様と組み合わせることができる。相互に排他的でない実施態様は、組み合わせ可能であると教示されている。実施態様の１つ以上の特徴を他の実施態様と組み合わせることができる。本開示は、これらのオプションのユーザを定期的に通知する。これらの選択肢を繰り返す列挙のいくつかの実施態様からの省略は、前述のセクションで教示されている組み合わせを制限するものとして解釈されるべきではない。これらの記載は、以下の実施のそれぞれに参照することにより本明細書に組み込まれる。

【０１９９】

開示される技術、又はその要素の１つ以上の実施態様及び条項は、示された方法ステップを実行するためのコンピュータ使用可能なプログラムコードを備えた非一時的コンピュータ可読記憶媒体を含むコンピュータ製品の形態で実装することができる。更に、開示される技術、又はその要素の１つ以上の実施態様及び条項は、メモリと、メモリに結合され、例示的な方法ステップを実行するように動作する少なくとも１つのプロセッサと、を含む装置の形態で実装することができる。更に、別の態様では、開示される技術又はその要素の１つ以上の実施態様及び条項は、本明細書に記載の方法ステップのうちの１つ以上を実行するための手段の形態で実装することができ、この手段は、（ｉ）ハードウェアモジュール、（ｉｉ）１つ以上のハードウェアプロセッサ上で実行されるソフトウェアモジュール、又は（ｉｉｉ）ハードウェア及びソフトウェアモジュールの組み合わせ、を含むことができ、（ｉ）～（ｉｉｉ）のいずれかが、本明細書に記載の特定の技術を実施し、ソフトウェアモジュールは、コンピュータ可読記憶媒体（又は複数のそのような媒体）に記憶される。

【０２００】

このセクションで説明される条項は、特徴として組み合わせることができる。簡潔性の目的で、特徴の組み合わせは、個別に列挙されず、特徴の各ベースセットで繰り返されない。読者は、このセクションに記載される条項で特定された特徴が、本出願の他のセクションにおける実施態様として特定された基本特徴のセットと容易に組み合わせることができる方法を理解するであろう。これらの条項は、相互排他的、網羅的、又は制限的であることを意味せず、開示される技術は、これらの条項に限定されず、むしろ、特許請求される技術及びその均等物の範囲内の全ての可能な組み合わせ、修正、及び変形を包含する。

【０２０１】

このセクションで説明される条項の他の実施態様は、このセクションに記載される条項のいずれか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むことができる。このセクションに記載される条項の更に別の実施態様は、メモリと、メモリ内に記憶された命令を実行してこのセクションに記載される条項のいずれかを行うように動作可能な１つ以上のプロセッサとを含むシステムを含むことができる。

【０２０２】

本発明者らは、以下の条項を開示する。

【０２０３】

条項セット１

１．表現型と関連付けられた遺伝子の関連性の強度及び表現型応答に対する希少バリア

ントの寄与を定量化するコンピュータ実装方法であって、方法が、
ａ）指定された表現型を呈する対象のコホートにおける変異を表すデータを得ることと、
ｂ）プロセッサにおいて、得られたデータを使用して関連遺伝子の各々についての総和スコアを計算することであって、総和スコアが、関連遺伝子の各々のキャリア状態と、１つ以上の希少病原性バリアントの関連遺伝子における存在に対する表現型応答との間のコホートにおける結果的な非ランダムな関連性を特定し、
ｃ）
キャリア状態が、特定の遺伝子における１つ以上の希少病原性バリアントの存在又は非存在によって決定されるブールであり、
特定の希少バリアントの病原性は、遺伝子がタンパク質として発現される場合、機能に対する特定の希少バリアントの予測される影響によって決定され、
特定の希少バリアントの希少性が、集団における特定の希少バリアントの第１の所定の閾値を下回る出現によって決定される、計算することと、
ｄ）それぞれの有効強度スコアを、遺伝子ごとの解像度におけるそれぞれの総和スコアに基づいて、関連遺伝子から選択された遺伝子についての結果的な非ランダムな関連性について決定することと、を含む、コンピュータ実装方法。
２．有効強度スコアに基づいて、表現型応答に対する関連遺伝子にわたるコホート内の対象のキャリア状態の寄与についてのスコアを決定することを更に含む、条項１に記載のコンピュータ実装方法。
３．指定された表現型が、定量的バイオマーカー表現型である、条項１に記載のコンピュータ実装方法。
４．遺伝子についての結果的な非ランダムな関連性に対する有効強度スコアが、総和スコアの線形回帰成分に対する両側ｔ検定を使用して、定量的バイオマーカー表現型に対して決定される、条項３に記載のコンピュータ実装方法。
５．両側ｔ検定が、キャリア及び非キャリアの平均表現型測定値間の差を決定する、条項４に記載のコンピュータ実装方法。
６．両側ｔ検定が、２つのｐ値を生成する、条項５に記載のコンピュータ実装方法。
７．指定された表現型の測定値が、薬物使用量補正される、条項５に記載のコンピュータ実装方法。
８．指定された表現型の測定値が、共変量補正され、正規化される、条項５に記載のコンピュータ実装方法。
９．指定された表現型の測定値が、一般的バリアント補正される、条項５に記載のコンピュータ実装方法。
１０．指定された表現型が、カテゴリ臨床診断表現型である、条項１に記載のコンピュータ実装方法。
１１．遺伝子についての結果的な非ランダムな関連性についての有効強度スコアが、キャリア状態についてのベータ係数としてカテゴリ臨床診断表現型について決定される、条項１０に記載のコンピュータ実装方法。
１２．ベータ係数が、総和スコアのロジスティック回帰成分を使用して決定される、条項１１に記載のコンピュータ実装方法。
１３．ロジスティック回帰成分が、キャリア状態及び複数の共変量から臨床診断ラベルを予測するようにフィッティングされる、条項１２に記載のコンピュータ実装方法。
１４．ロジスティック回帰成分が、キャリア状態をバイナリインジケータ変数として符号化する、条項１３に記載のコンピュータ実装方法。
１５．ロジスティック回帰成分が、複数の共変量を除去して回帰分析する、条項１３に記載のコンピュータ実装方法。
１６．複数の共変量が、年齢、性別、遺伝的主成分、民族性、一般的バリアント、及びバイアス項を含む、条項１３に記載のコンピュータ実装方法。
１７．ロジスティック回帰成分が、ｐ値を生成する、条項１３に記載のコンピュータ実

装方法。
１８．関連遺伝子についての希少病原性バリアントが、対立遺伝子頻度に適用される異なる希少性閾値を使用して、異なる総和スコアについて特定される、条項１に記載のコンピュータ実装方法。
１９．関連遺伝子についての希少病原性バリアントが、０．００１未満である対立遺伝子頻度を有する、条項１８に記載のコンピュータ実装方法。
２０．関連遺伝子についての希少病原性バリアントが、病原性スコアに適用される異なる病原性閾値を使用して、異なる総和検定に対して特定される、条項１に記載のコンピュータ実装方法。
２１．病原性スコアが、畳み込みニューラルネットワークによって生成される、条項２０に記載のコンピュータ実装方法。
２２．異なる病原性閾値が、異なるパーセンタイル閾値である、条項２０に記載のコンピュータ実装方法。
２３．それぞれの有効強度スコアの加重和に基づいて、指定された表現型についての、及びコホート内の特定の個体についての希少バリアント多遺伝子リスクスコアを決定することを更に含む、条項１に記載のコンピュータ実装方法。
２４．それぞれの有効強度スコアの加重和が、関連遺伝子にわたる特定の個体のキャリア状態によって重み付けされる、条項２３に記載のコンピュータ実装方法。
２５．希少病原性バリアントが、機能喪失バリアントである、条項１に記載のコンピュータ実装方法。
２６．希少病原性バリアントが、ミスセンスバリアントである、条項２５に記載のコンピュータ実装方法。
２７．別個の総和スコアを使用して、指定された表現型に対する機能喪失バリアント及びミスセンスバリアントの有効強度スコアを別個に決定することを更に含む、条項２６に記載のコンピュータ実装方法。
２８．線形回帰成分又はロジスティック回帰成分の訓練された係数を使用して、検定データについての希少バリアント多遺伝子リスクスコアを生成することを更に含む、条項１に記載のコンピュータ実装方法。
２９．検定データが、外れ値表現型測定値を有する個体を含む、条項２８に記載のコンピュータ実装方法。
３０．メモリに結合された１つ以上のプロセッサを含むシステムであって、メモリに、表現型と関連付けられた遺伝子の関連性の強度及び表現型応答に対する希少バリアントの寄与を定量化するためのコンピュータ命令がロードされており、命令が、プロセッサ上で実行されると、
ａ）指定された表現型を呈する対象のコホートにおける変異を表すデータを得ることと、
ｂ）プロセッサにおいて、得られたデータを使用して関連遺伝子の各々についての総和スコアを計算することであって、総和スコアが、関連遺伝子の各々のキャリア状態と、１つ以上の希少病原性バリアントの関連遺伝子における存在に対する表現型応答との間のコホートにおける結果的な非ランダムな関連性を特定し、
ｃ）
キャリア状態が、特定の遺伝子における１つ以上の希少病原性バリアントの存在又は非存在によって決定されるブールであり、
特定の希少バリアントの病原性は、遺伝子がタンパク質として発現される場合、機能に対する特定の希少バリアントの予測される影響によって決定され、
特定の希少バリアントの希少性が、集団における特定の希少バリアントの第１の所定の閾値を下回る出現によって決定される、計算することと、
ｄ）それぞれの有効強度スコアを、遺伝子ごとの解像度におけるそれぞれの総和スコアに基づいて、関連遺伝子から選択された遺伝子についての結果的な非ランダムな関連性について決定することと、を含むアクションを実装する、システム。

３１．有効強度スコアに基づいて、表現型応答に対する関連遺伝子にわたるコホート内の対象のキャリア状態の寄与についてのスコアを決定することを更に含む、条項３０に記載のシステム。
３２．指定された表現型が、定量的バイオマーカー表現型である、条項３０に記載のシステム。
３３．遺伝子についての結果的な非ランダムな関連性に対する有効強度スコアが、総和スコアの線形回帰成分に対する両側ｔ検定を使用して、定量的バイオマーカー表現型に対して決定される、条項３２に記載のシステム。
３４．両側ｔ検定が、キャリア及び非キャリアの平均表現型測定値間の差を決定する、条項３３に記載のシステム。
３５．両側ｔ検定が、２つのｐ値を生成する、条項３４に記載のシステム。
３６．指定された表現型の測定値が、薬物使用量補正される、条項３４に記載のシステム。
３７．指定された表現型の測定値が、共変量補正され、正規化される、条項３４に記載のシステム。
３８．指定された表現型の測定値が、一般的バリアント補正される、条項３４に記載のシステム。
３９．指定された表現型が、カテゴリ臨床診断表現型である、条項３０に記載のシステム。
４０．遺伝子についての結果的な非ランダムな関連性についての有効強度スコアが、キャリア状態についてのベータ係数としてカテゴリ臨床診断表現型について決定される、条項３９に記載のシステム。
４１．ベータ係数が、総和スコアのロジスティック回帰成分を使用して決定される、条項４０に記載のシステム。
４２．ロジスティック回帰成分が、キャリア状態及び複数の共変量から臨床診断ラベルを予測するようにフィッティングされる、条項４１に記載のシステム。
４３．ロジスティック回帰成分が、キャリア状態をバイナリインジケータ変数として符号化する、条項４２に記載のシステム。
４４．ロジスティック回帰成分が、複数の共変量を除去して回帰分析する、条項４２に記載のシステム。
４５．複数の共変量が、年齢、性別、遺伝的主成分、民族性、一般的バリアント、及びバイアス項を含む、条項４２に記載のシステム。
４６．ロジスティック回帰成分が、ｐ値を生成する、条項４２に記載のシステム。
４７．関連遺伝子についての希少病原性バリアントが、対立遺伝子頻度に適用される異なる希少性閾値を使用して、異なる総和スコアについて特定される、条項３０に記載のシステム。
４８．関連遺伝子についての希少病原性バリアントが、０．００１未満である対立遺伝子頻度を有する、条項４７に記載のシステム。
４９．関連遺伝子についての希少病原性バリアントが、病原性スコアに適用される異なる病原性閾値を使用して、異なる総和検定に対して特定される、条項３０に記載のシステム。
５０．病原性スコアが、畳み込みニューラルネットワークによって生成される、条項４９に記載のシステム。
５１．異なる病原性閾値が、異なるパーセンタイル閾値である、条項４９に記載のシステム。
５２．それぞれの有効強度スコアの加重和に基づいて、指定された表現型についての、及びコホート内の特定の個体についての希少バリアント多遺伝子リスクスコアを決定することを更に含む、条項３０に記載のシステム。
５３．それぞれの有効強度スコアの加重和が、関連遺伝子にわたる特定の個体のキャリア状態によって重み付けされる、条項５２に記載のシステム。

５４．希少病原性バリアントが、機能喪失バリアントである、条項３０に記載のシステム。
５５．希少病原性バリアントが、ミスセンスバリアントである、条項５４に記載のシステム。
５６．別個の総和スコアを使用して、指定された表現型に対する機能喪失バリアント及びミスセンスバリアントの有効強度スコアを別個に決定することを更に含む、条項５５に記載のシステム。
５７．線形回帰成分又はロジスティック回帰成分の訓練された係数を使用して、検定データについての希少バリアント多遺伝子リスクスコアを生成することを更に含む、条項３０に記載のシステム。
５８．検定データが、外れ値表現型測定値を有する個体を含む、条項５７に記載のシステム。
５９．表現型と関連付けられた遺伝子の関連性の強度及び表現型応答に対する希少バリアントの寄与を定量化するためのコンピュータプログラム命令を記憶している非一時的コンピュータ可読記憶媒体であって、命令が、プロセッサ上で実行されると、
ａ）指定された表現型を呈する対象のコホートにおける変異を表すデータを得ることと、
ｂ）プロセッサにおいて、得られたデータを使用して関連遺伝子の各々についての総和スコアを計算することであって、総和スコアが、関連遺伝子の各々のキャリア状態と、１つ以上の希少病原性バリアントの関連遺伝子における存在に対する表現型応答との間のコホートにおける結果的な非ランダムな関連性を特定し、
ｃ）
キャリア状態が、特定の遺伝子における１つ以上の希少病原性バリアントの存在又は非存在によって決定されるブールであり、
特定の希少バリアントの病原性は、遺伝子がタンパク質として発現される場合、機能に対する特定の希少バリアントの予測される影響によって決定され、
特定の希少バリアントの希少性が、集団における特定の希少バリアントの第１の所定の閾値を下回る出現によって決定される、計算することと、
ｄ）それぞれの有効強度スコアを、遺伝子ごとの解像度におけるそれぞれの総和スコアに基づいて、関連遺伝子から選択された遺伝子についての結果的な非ランダムな関連性について決定することと、を含む方法を実装する、非一時的コンピュータ可読記憶媒体。
６０．有効強度スコアに基づいて、表現型応答に対する関連遺伝子にわたるコホート内の対象のキャリア状態の寄与についてのスコアを決定することを更に含む、条項５９に記載の非一時的コンピュータ可読記憶媒体。
６１．指定された表現型が、定量的バイオマーカー表現型である、条項５９に記載の非一時的コンピュータ可読記憶媒体。
６２．遺伝子についての結果的な非ランダムな関連性に対する有効強度スコアが、総和スコアの線形回帰成分に対する両側ｔ検定を使用して、定量的バイオマーカー表現型に対して決定される、条項６１に記載の非一時的コンピュータ可読記憶媒体。
６３．両側ｔ検定が、キャリア及び非キャリアの平均表現型測定値間の差を決定する、条項６２に記載の非一時的コンピュータ可読記憶媒体。
６４．両側ｔ検定が、２つのｐ値を生成する、条項６３に記載の非一時的コンピュータ可読記憶媒体。
６５．指定された表現型の測定値が、薬物使用量補正される、条項６３に記載の非一時的コンピュータ可読記憶媒体。
６６．指定された表現型の測定値が、共変量補正され、正規化される、条項６３に記載の非一時的コンピュータ可読記憶媒体。
６７．指定された表現型の測定値が、一般的バリアント補正される、条項６３に記載の非一時的コンピュータ可読記憶媒体。
６８．指定された表現型が、カテゴリ臨床診断表現型である、条項５９に記載の非一時

的コンピュータ可読記憶媒体。
６９．遺伝子についての結果的な非ランダムな関連性についての有効強度スコアが、キャリア状態についてのベータ係数としてカテゴリ臨床診断表現型について決定される、条項６８に記載の非一時的コンピュータ可読記憶媒体。
７０．ベータ係数が、総和スコアのロジスティック回帰成分を使用して決定される、条項６９に記載の非一時的コンピュータ可読記憶媒体。
７１．ロジスティック回帰成分が、キャリア状態及び複数の共変量から臨床診断ラベルを予測するようにフィッティングされる、条項７０に記載の非一時的コンピュータ可読記憶媒体。
７２．ロジスティック回帰成分が、キャリア状態をバイナリインジケータ変数として符号化する、条項７１に記載の非一時的コンピュータ可読記憶媒体。
７３．ロジスティック回帰成分が、複数の共変量を除去して回帰分析する、条項７１に記載の非一時的コンピュータ可読記憶媒体。
７４．複数の共変量が、年齢、性別、遺伝的主成分、民族性、一般的バリアント、及びバイアス項を含む、条項７１に記載の非一時的コンピュータ可読記憶媒体。
７５．ロジスティック回帰成分が、ｐ値を生成する、条項７１に記載の非一時的コンピュータ可読記憶媒体。
７６．関連遺伝子についての希少病原性バリアントが、対立遺伝子頻度に適用される異なる希少性閾値を使用して、異なる総和スコアについて特定される、条項５９に記載の非一時的コンピュータ可読記憶媒体。
７７．関連遺伝子についての希少病原性バリアントが、０．００１未満である対立遺伝子頻度を有する、条項７６に記載の非一時的コンピュータ可読記憶媒体。
７８．関連遺伝子についての希少病原性バリアントが、病原性スコアに適用される異なる病原性閾値を使用して、異なる総和検定に対して特定される、条項５９に記載の非一時的コンピュータ可読記憶媒体。
７９．病原性スコアが、畳み込みニューラルネットワークによって生成される、条項７８に記載の非一時的コンピュータ可読記憶媒体。
８０．異なる病原性閾値が、異なるパーセンタイル閾値である、条項７８に記載の非一時的コンピュータ可読記憶媒体。
８１．それぞれの有効強度スコアの加重和に基づいて、指定された表現型についての、及びコホート内の特定の個体についての希少バリアント多遺伝子リスクスコアを決定することを更に含む、条項５９に記載の非一時的コンピュータ可読記憶媒体。
８２．それぞれの有効強度スコアの加重和が、関連遺伝子にわたる特定の個体のキャリア状態によって重み付けされる、条項８１に記載の非一時的コンピュータ可読記憶媒体。
８３．希少病原性バリアントが、機能喪失バリアントである、条項５９に記載の非一時的コンピュータ可読記憶媒体。
８４．希少病原性バリアントが、ミスセンスバリアントである、条項８３に記載の非一時的コンピュータ可読記憶媒体。
８５．別個の総和スコアを使用して、指定された表現型に対する機能喪失バリアント及びミスセンスバリアントの有効強度スコアを別個に決定することを更に含む、条項８４に記載の非一時的コンピュータ可読記憶媒体。
８６．線形回帰成分又はロジスティック回帰成分の訓練された係数を使用して、検定データについての希少バリアント多遺伝子リスクスコアを生成することを更に含む、条項５９に記載の非一時的コンピュータ可読記憶媒体。
８７．検定データが、外れ値表現型測定値を有する個体を含む、条項８６に記載の非一時的コンピュータ可読記憶媒体。
【０２０４】
条項セット２
１．表現型に対する希少病原性バリアントの効果を決定するコンピュータ実装方法であって、

特定の表現型と関連付けられた特定の遺伝子について：
特定の遺伝子における希少病原性バリアントの特定の群を特定することと、
希少病原性バリアントの特定の群からの少なくとも１つの希少病原性バリアントを保有する個体のコホートにおけるキャリアを特定することと、
希少病原性バリアントの特定の群からのいかなる希少病原性バリアントも保有しない個体のコホートにおける非キャリアを特定することと、
総和検定を使用して、キャリアを非キャリアから分離するキャリア状態に依存して、特定の表現型に対する希少病原性バリアントの特定の群の効果量を決定することと、を含む、コンピュータ実装方法。
２．特定の表現型が、定量的バイオマーカー表現型である、条項１に記載のコンピュータ実装方法。
３．希少病原性バリアントの特定の群の効果量が、総和検定の線形回帰成分に対する両側ｔ検定を使用して定量的バイオマーカー表現型について決定される、条項２に記載のコンピュータ実装方法。
４．両側ｔ検定が、キャリア及び非キャリアの平均表現型測定値間の差を決定する、条項３に記載のコンピュータ実装方法。
５．両側ｔ検定が、２つのｐ値を生成する、条項４に記載のコンピュータ実装方法。
６．表現型測定値が、薬物使用量補正される、条項４に記載のコンピュータ実装方法。
７．表現型測定値が、共変量補正され、正規化される、条項４に記載のコンピュータ実装方法。
８．表現型測定値が、一般的バリアント補正される、条項４に記載のコンピュータ実装方法。
９．特定の表現型が、カテゴリ臨床診断表現型である、条項１に記載のコンピュータ実装方法。
１０．希少病原性バリアントの特定の群の効果量が、キャリア状態についてのベータ係数としてカテゴリ臨床診断表現型に対して決定される、条項９に記載のコンピュータ実装方法。
１１．ベータ係数が、総和検定のロジスティック回帰成分を使用して決定される、条項１０に記載のコンピュータ実装方法。
１２．ロジスティック回帰成分が、キャリア状態及び複数の共変量から臨床診断ラベルを予測するようにフィッティングされる、条項１１に記載のコンピュータ実装方法。
１３．ロジスティック回帰成分が、キャリア状態をバイナリインジケータ変数として符号化する、条項１２に記載のコンピュータ実装方法。
１４．ロジスティック回帰成分が、複数の共変量を除去して回帰分析する、条項１２に記載のコンピュータ実装方法。
１５．複数の共変量が、年齢、性別、遺伝的主成分、民族性、一般的バリアント、及びバイアス項を含む、条項１２に記載のコンピュータ実装方法。
１６．ロジスティック回帰成分が、ｐ値を生成する、条項１２に記載のコンピュータ実装方法。
１７．希少病原性バリアントの特定の群が、対立遺伝子頻度に適用される異なる希少性閾値を使用して、異なる総和検定について特定される、条項１に記載のコンピュータ実装方法。
１８．希少病原性バリアントの特定の群が、０．００１未満である対立遺伝子頻度を有する、条項１７に記載のコンピュータ実装方法。
１９．希少病原性バリアントの特定の群が、病原性スコアに適用される異なる病原性閾値を使用して、異なる総和検定に対して特定される、条項１に記載のコンピュータ実装方法。
２０．病原性スコアが、畳み込みニューラルネットワークによって生成される、条項１９に記載のコンピュータ実装方法。
２１．異なる病原性閾値が、異なるパーセンタイル閾値である、条項１９に記載のコン

ピュータ実装方法。
２２．それぞれの総和検定を使用して、特定の表現型と関連付けられたそれぞれの遺伝子における希少病原性バリアントのそれぞれの群のそれぞれの効果量を決定することを更に含む、条項１に記載のコンピュータ実装方法。
２３．それぞれの効果量の加重和に基づいて、特定の表現型についての、及び個体のコホート内の特定の個体についての希少バリアント多遺伝子リスクスコアを生成することを更に含む、条項２２に記載のコンピュータ実装方法。
２４．それぞれの効果量の加重和が、それぞれの遺伝子にわたる特定の個体のキャリア状態によって重み付けされる、条項２３に記載のコンピュータ実装方法。
２５．特定の個体は、特定の個体が希少病原性バリアントの特定の群からの少なくとも１つの希少病原性バリアントを保有する場合、特定の遺伝子についてのキャリアのキャリア状態を有する、条項２４に記載のコンピュータ実装方法。
２６．特定の個体がキャリアのキャリア状態を有する場合、１の重みが使用される、条項２５に記載のコンピュータ実装方法。
２７．特定の個体は、特定の個体が希少病原性バリアントの特定の群からのいかなる希少病原性バリアントも保有しない場合、特定の遺伝子についての非キャリアのキャリア状態を有する、条項２４に記載のコンピュータ実装方法。
２８．特定の個体が非キャリアのキャリア状態を有する場合、ゼロの重みが使用される、条項２７に記載のコンピュータ実装方法。
２９．希少病原性バリアントの特定の群における希少病原性バリアントが、機能喪失バリアントである、条項１に記載のコンピュータ実装方法。
３０．希少病原性バリアントが、ミスセンスバリアントである、条項２９に記載のコンピュータ実装方法。
３１．別個の総和検定を使用して、表現型に対する機能喪失バリアント及びミスセンスバリアントの効果を別個に決定することを更に含む、条項３０に記載のコンピュータ実装方法。
３２．線形回帰成分及びロジスティック回帰成分の訓練された係数を使用して、検定データについての希少バリアント多遺伝子リスクスコアを生成することを更に含む、条項１に記載のコンピュータ実装方法。
３３．検定データが、外れ値表現型測定値を有する個体を含む、条項３２に記載のコンピュータ実装方法。
３４．表現型に対する希少病原性バリアントの効果を決定するコンピュータ実装方法であって、
総和検定を使用して、特定の表現型と関連付けられたそれぞれの遺伝子における希少病原性バリアントのそれぞれの群のそれぞれの効果量を決定することと、
それぞれの効果量の加重和に基づいて、特定の表現型についての、及び個体のコホート内の特定の個体についての希少バリアント多遺伝子リスクスコアを生成することと、を含み、
それぞれの効果量の加重和が、それぞれの遺伝子にわたる特定の個体のキャリア状態によって重み付けされる、コンピュータ実装方法。
３５．特定の個体は、特定の個体が希少病原性バリアントの特定の群からの少なくとも１つの希少病原性バリアントを保有する場合、特定の遺伝子についてのキャリアのキャリア状態を有する、条項３４に記載のコンピュータ実装方法。
３６．特定の個体がキャリアのキャリア状態を有する場合、１の重みが使用される、条項３５に記載のコンピュータ実装方法。
３７．特定の個体は、特定の個体が希少病原性バリアントの特定の群からのいかなる希少病原性バリアントも保有しない場合、特定の遺伝子についての非キャリアのキャリア状態を有する、条項３５に記載のコンピュータ実装方法。
３８．特定の個体が非キャリアのキャリア状態を有する場合、ゼロの重みが使用される、条項３７に記載のコンピュータ実装方法。

３９．メモリに結合された１つ以上のプロセッサを含むシステムであって、メモリには、表現型に対する希少病原性バリアントの効果を決定するためのコンピュータ命令がロードされ、命令が、プロセッサ上で実行されると、
特定の表現型と関連付けられた特定の遺伝子について：
特定の遺伝子における希少病原性バリアントの特定の群を特定することと、
希少病原性バリアントの特定の群からの少なくとも１つの希少病原性バリアントを保有する個体のコホートにおけるキャリアを特定することと、
希少病原性バリアントの特定の群からのいかなる希少病原性バリアントも保有しない個体のコホートにおける非キャリアを特定することと、
総和検定を使用して、キャリアを非キャリアから分離するキャリア状態に依存して、特定の表現型に対する希少病原性バリアントの特定の群の効果量を決定することと、を含むアクションを実装する、システム。
４０．特定の表現型が、定量的バイオマーカー表現型である、条項３９に記載のシステム。
４１．希少病原性バリアントの特定の群の効果量が、総和検定の線形回帰成分に対する両側ｔ検定を使用して定量的バイオマーカー表現型について決定される、条項４０に記載のシステム。
４２．両側ｔ検定が、キャリア及び非キャリアの平均表現型測定値間の差を決定する、条項４１に記載のシステム。
４３．両側ｔ検定が、２つのｐ値を生成する、条項４２に記載のシステム。
４４．表現型測定値が、薬物使用量補正される、条項４２に記載のシステム。
４５．表現型測定値が、共変量補正され、正規化される、条項４２に記載のシステム。
４６．表現型測定値が、一般的バリアント補正される、条項４２に記載のシステム。
４７．特定の表現型が、カテゴリ臨床診断表現型である、条項３９に記載のシステム。
４８．希少病原性バリアントの特定の群の効果量が、キャリア状態についてのベータ係数としてカテゴリ臨床診断表現型に対して決定される、条項４７に記載のシステム。
４９．ベータ係数が、総和検定のロジスティック回帰成分を使用して決定される、条項４８に記載のシステム。
５０．ロジスティック回帰成分が、キャリア状態及び複数の共変量から臨床診断ラベルを予測するようにフィッティングされる、条項４９に記載のシステム。
５１．ロジスティック回帰成分が、キャリア状態をバイナリインジケータ変数として符号化する、条項５０に記載のシステム。
５２．ロジスティック回帰成分が、複数の共変量を除去して回帰分析する、条項５０に記載のシステム。
５３．複数の共変量が、年齢、性別、遺伝的主成分、民族性、一般的バリアント、及びバイアス項を含む、条項５０に記載のシステム。
５４．ロジスティック回帰成分が、ｐ値を生成する、条項５０に記載のシステム。
５５．希少病原性バリアントの特定の群が、対立遺伝子頻度に適用される異なる希少性閾値を使用して、異なる総和検定について特定される、条項３９に記載のシステム。
５６．希少病原性バリアントの特定の群が、０．００１未満である対立遺伝子頻度を有する、条項５５に記載のシステム。
５７．希少病原性バリアントの特定の群が、病原性スコアに適用される異なる病原性閾値を使用して、異なる総和検定に対して特定される、条項３９に記載のシステム。
５８．病原性スコアが、畳み込みニューラルネットワークによって生成される、条項５７に記載のシステム。
５９．異なる病原性閾値が、異なるパーセンタイル閾値である、条項５７に記載のシステム。
６０．それぞれの総和検定を使用して、特定の表現型と関連付けられたそれぞれの遺伝子における希少病原性バリアントのそれぞれの群のそれぞれの効果量を決定することを含むアクションを更に実装する、条項３９に記載のシステム。

６１．それぞれの効果量の加重和に基づいて、特定の表現型についての、及び個体のコホート内の特定の個体についての希少バリアント多遺伝子リスクスコアを生成することを含むアクションを更に実装する、条項６０に記載のシステム。
６２．それぞれの効果量の加重和が、それぞれの遺伝子にわたる特定の個体のキャリア状態によって重み付けされる、条項６１に記載のシステム。
６３．特定の個体は、特定の個体が希少病原性バリアントの特定の群からの少なくとも１つの希少病原性バリアントを保有する場合、特定の遺伝子についてのキャリアのキャリア状態を有する、条項６２に記載のシステム。
６４．特定の個体がキャリアのキャリア状態を有する場合、１の重みが使用される、条項６３に記載のシステム。
６５．特定の個体は、特定の個体が希少病原性バリアントの特定の群からのいかなる希少病原性バリアントも保有しない場合、特定の遺伝子についての非キャリアのキャリア状態を有する、条項６２に記載のシステム。
６６．特定の個体が非キャリアのキャリア状態を有する場合、ゼロの重みが使用される、条項６５に記載のシステム。
６７．希少病原性バリアントの特定の群における希少病原性バリアントが、機能喪失バリアントである、条項３９に記載のシステム。
６８．希少病原性バリアントが、ミスセンスバリアントである、条項６７に記載のシステム。
６９．別個の総和検定を使用して、表現型に対する機能喪失バリアント及びミスセンスバリアントの効果を別個に決定することを含むアクションを更に実装する、条項６８に記載のシステム。
７０．線形回帰成分及びロジスティック回帰成分の訓練された係数を使用して、検定データについての希少バリアント多遺伝子リスクスコアを生成することを含むアクションを更に実装する、条項３９に記載のシステム。
７１．検定データが、外れ値表現型測定値を有する個体を含む、条項７０に記載のシステム。
７２．表現型に対する希少病原性バリアントの効果を決定するシステムであって、
総和検定を使用して、特定の表現型と関連付けられたそれぞれの遺伝子における希少病原性バリアントのそれぞれの群のそれぞれの効果量を決定することと、
それぞれの効果量の加重和に基づいて、特定の表現型についての、及び個体のコホート内の特定の個体についての希少バリアント多遺伝子リスクスコアを生成することと、を含み、
それぞれの効果量の加重和が、それぞれの遺伝子にわたる特定の個体のキャリア状態によって重み付けされる、システム。
７３．特定の個体は、特定の個体が希少病原性バリアントの特定の群からの少なくとも１つの希少病原性バリアントを保有する場合、特定の遺伝子についてのキャリアのキャリア状態を有する、条項７２に記載のシステム。
７４．特定の個体がキャリアのキャリア状態を有する場合、１の重みが使用される、条項７３に記載のシステム。
７５．特定の個体は、特定の個体が希少病原性バリアントの特定の群からのいかなる希少病原性バリアントも保有しない場合、特定の遺伝子についての非キャリアのキャリア状態を有する、条項７３に記載のシステム。
７６．特定の個体が非キャリアのキャリア状態を有する場合、ゼロの重みが使用される、条項７５に記載のシステム。
７７．表現型に対する希少病原性バリアントの効果を決定するためのコンピュータプログラム命令を記憶している非一時的コンピュータ可読記憶媒体であって、命令が、プロセッサ上で実行されると、
特定の表現型と関連付けられた特定の遺伝子について：
特定の遺伝子における希少病原性バリアントの特定の群を特定することと、

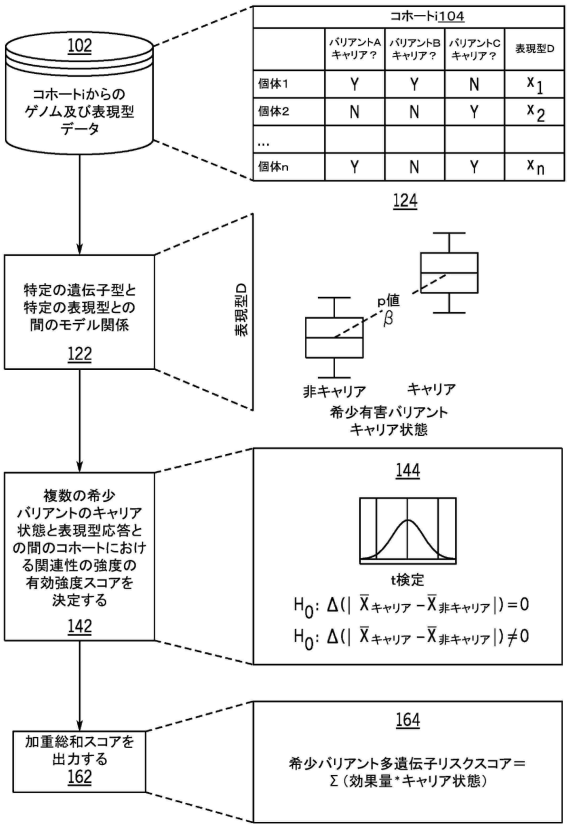
希少病原性バリアントの特定の群からの少なくとも１つの希少病原性バリアントを保有する個体のコホートにおけるキャリアを特定することと、
希少病原性バリアントの特定の群からのいかなる希少病原性バリアントも保有しない個体のコホートにおける非キャリアを特定することと、
総和検定を使用して、キャリアを非キャリアから分離するキャリア状態に依存して、特定の表現型に対する希少病原性バリアントの特定の群の効果量を決定することと、を含む方法を実装する、非一時的コンピュータ可読記憶媒体。
７８．特定の表現型が、定量的バイオマーカー表現型である、条項７７に記載の非一時的コンピュータ可読記憶媒体。
７９．希少病原性バリアントの特定の群の効果量が、総和検定の線形回帰成分に対する両側ｔ検定を使用して定量的バイオマーカー表現型について決定される、条項７８に記載の非一時的コンピュータ可読記憶媒体。
８０．両側ｔ検定が、キャリア及び非キャリアの平均表現型測定値間の差を決定する、条項７９に記載の非一時的コンピュータ可読記憶媒体。
８１．両側ｔ検定が、２つのｐ値を生成する、条項８０に記載の非一時的コンピュータ可読記憶媒体。
８２．表現型測定値が、薬物使用量補正される、条項８０に記載の非一時的コンピュータ可読記憶媒体。
８３．表現型測定値が、共変量補正され、正規化される、条項８０に記載の非一時的コンピュータ可読記憶媒体。
８４．表現型測定値が、一般的バリアント補正される、条項８０に記載の非一時的コンピュータ可読記憶媒体。
８５．特定の表現型が、カテゴリ臨床診断表現型である、条項７７に記載の非一時的コンピュータ可読記憶媒体。
８６．希少病原性バリアントの特定の群の効果量が、キャリア状態についてのベータ係数としてカテゴリ臨床診断表現型に対して決定される、条項８５に記載の非一時的コンピュータ可読記憶媒体。
８７．ベータ係数が、総和検定のロジスティック回帰成分を使用して決定される、条項８６に記載の非一時的コンピュータ可読記憶媒体。
８８．ロジスティック回帰成分が、キャリア状態及び複数の共変量から臨床診断ラベルを予測するようにフィッティングされる、条項８７に記載の非一時的コンピュータ可読記憶媒体。
８９．ロジスティック回帰成分が、キャリア状態をバイナリインジケータ変数として符号化する、条項８８に記載の非一時的コンピュータ可読記憶媒体。
９０．ロジスティック回帰成分が、複数の共変量を除去して回帰分析する、条項８８に記載の非一時的コンピュータ可読記憶媒体。
９１．複数の共変量が、年齢、性別、遺伝的主成分、民族性、一般的バリアント、及びバイアス項を含む、条項８８に記載の非一時的コンピュータ可読記憶媒体。
９２．ロジスティック回帰成分が、ｐ値を生成する、条項８８に記載の非一時的コンピュータ可読記憶媒体。
９３．希少病原性バリアントの特定の群が、対立遺伝子頻度に適用される異なる希少性閾値を使用して、異なる総和検定について特定される、条項７７に記載の非一時的コンピュータ可読記憶媒体。
９４．希少病原性バリアントの特定の群が、０．００１未満である対立遺伝子頻度を有する、条項９３に記載の非一時的コンピュータ可読記憶媒体。
９５．希少病原性バリアントの特定の群が、病原性スコアに適用される異なる病原性閾値を使用して、異なる総和検定に対して特定される、条項７７に記載の非一時的コンピュータ可読記憶媒体。
９６．病原性スコアが、畳み込みニューラルネットワークによって生成される、条項９５に記載の非一時的コンピュータ可読記憶媒体。

９７．異なる病原性閾値が、異なるパーセンタイル閾値である、条項９５に記載の非一時的コンピュータ可読記憶媒体。
９８．それぞれの総和検定を使用して、特定の表現型と関連付けられたそれぞれの遺伝子における希少病原性バリアントのそれぞれの群のそれぞれの効果量を決定することを更に含む方法を実装する、条項７７に記載の非一時的コンピュータ可読記憶媒体。
９９．それぞれの効果量の加重和に基づいて、特定の表現型についての、及び個体のコホート内の特定の個体についての希少バリアント多遺伝子リスクスコアを生成することを更に含む方法を実装する、条項９８に記載の非一時的コンピュータ可読記憶媒体。
１００．それぞれの効果量の加重和が、それぞれの遺伝子にわたる特定の個体のキャリア状態によって重み付けされる、条項９９に記載の非一時的コンピュータ可読記憶媒体。
１０１．特定の個体は、特定の個体が希少病原性バリアントの特定の群からの少なくとも１つの希少病原性バリアントを保有する場合、特定の遺伝子についてのキャリアのキャリア状態を有する、条項１００に記載の非一時的コンピュータ可読記憶媒体。
１０２．特定の個体がキャリアのキャリア状態を有する場合、１の重みが使用される、条項１０１に記載の非一時的コンピュータ可読記憶媒体。
１０３．特定の個体は、特定の個体が希少病原性バリアントの特定の群からのいかなる希少病原性バリアントも保有しない場合、特定の遺伝子についての非キャリアのキャリア状態を有する、条項１００に記載の非一時的コンピュータ可読記憶媒体。
１０４．特定の個体が非キャリアのキャリア状態を有する場合、ゼロの重みが使用される、条項１０３に記載の非一時的コンピュータ可読記憶媒体。
１０５．希少病原性バリアントの特定の群における希少病原性バリアントが、機能喪失バリアントである、条項７７に記載の非一時的コンピュータ可読記憶媒体。
１０６．希少病原性バリアントが、ミスセンスバリアントである、条項１０５に記載の非一時的コンピュータ可読記憶媒体。
１０７．別個の総和検定を使用して、表現型に対する機能喪失バリアント及びミスセンスバリアントの効果を別個に決定することを更に含む方法を実装する、条項１０６に記載の非一時的コンピュータ可読記憶媒体。
１０８．線形回帰成分及びロジスティック回帰成分の訓練された係数を使用して、検定データについての希少バリアント多遺伝子リスクスコアを生成することを更に含む方法を実装する、条項７７に記載の非一時的コンピュータ可読記憶媒体。
１０９．検定データが、外れ値表現型測定値を有する個体を含む、条項１０８に記載の非一時的コンピュータ可読記憶媒体。
１１０．表現型に対する希少病原性バリアントの効果を決定する非一時的コンピュータ可読記憶媒体であって、
総和検定を使用して、特定の表現型と関連付けられたそれぞれの遺伝子における希少病原性バリアントのそれぞれの群のそれぞれの効果量を決定することと、
それぞれの効果量の加重和に基づいて、特定の表現型についての、及び個体のコホート内の特定の個体についての希少バリアント多遺伝子リスクスコアを生成することと、を含み、
それぞれの効果量の加重和が、それぞれの遺伝子にわたる特定の個体のキャリア状態によって重み付けされる、非一時的コンピュータ可読記憶媒体。
１０１．特定の個体は、特定の個体が希少病原性バリアントの特定の群からの少なくとも１つの希少病原性バリアントを保有する場合、特定の遺伝子についてのキャリアのキャリア状態を有する、条項１１０に記載の非一時的コンピュータ可読記憶媒体。
１０２．特定の個体がキャリアのキャリア状態を有する場合、１の重みが使用される、条項１０１に記載の非一時的コンピュータ可読記憶媒体。
１０３．特定の個体は、特定の個体が希少病原性バリアントの特定の群からのいかなる希少病原性バリアントも保有しない場合、特定の遺伝子についての非キャリアのキャリア状態を有する、条項１０１に記載の非一時的コンピュータ可読記憶媒体。
１０４．特定の個体が非キャリアのキャリア状態を有する場合、ゼロの重みが使用され

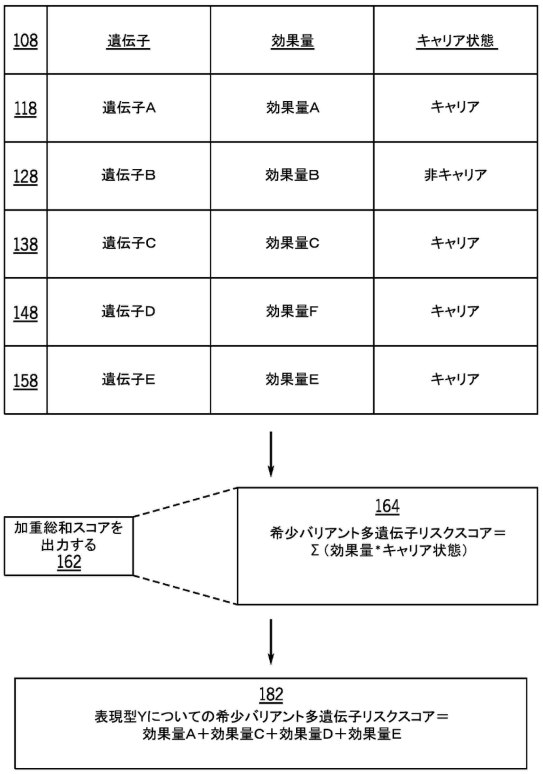
る、条項１０３に記載の非一時的コンピュータ可読記憶媒体。

【図面】

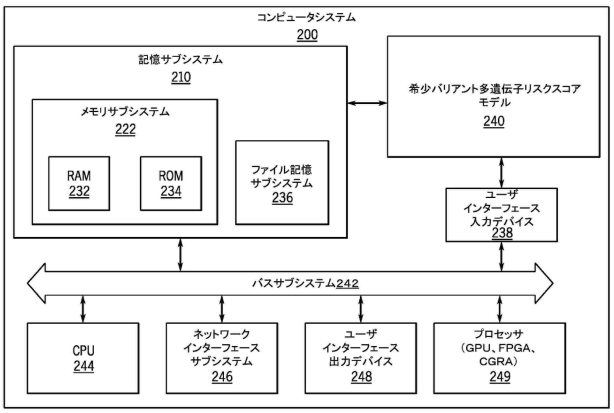
【図１Ａ】



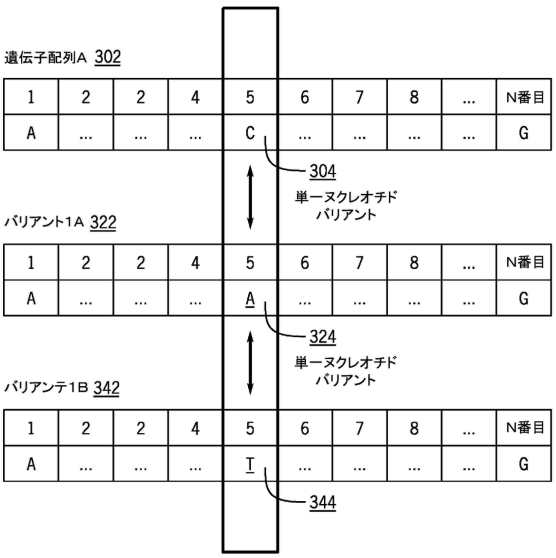
【図１Ｂ】



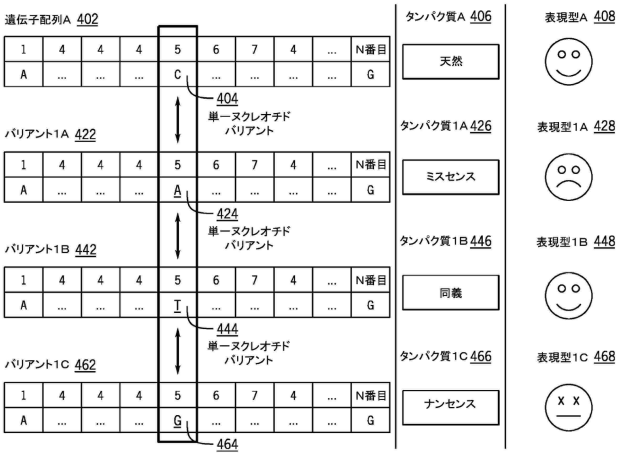
【図２】



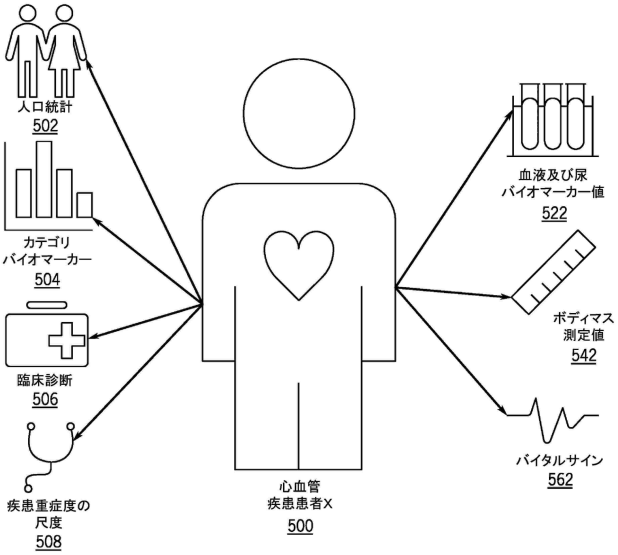
【図３】



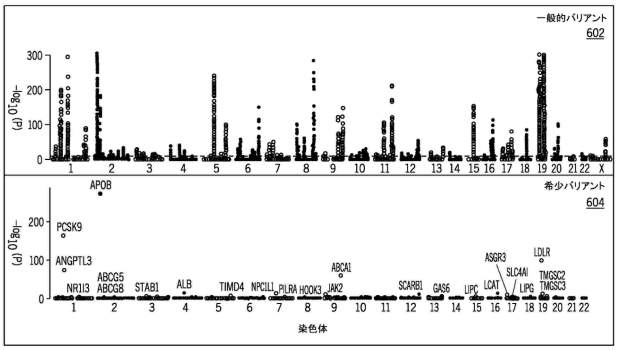
【図４】


遺伝子配列A 402
1 4 4 4 5 6 7 4 ... N番目
A ... ... C ... ... G
404
単一ヌクレオチド
バリアント
バリアント1A 422
A
424
バリアント1B 442
T
444
バリアント1C 462
G
464
タンパク質A 406
天然
タンパク質1A 426
ミスセンス
タンパク質1B 446
同義
タンパク質1C 466
ナンセンス
表現型A 408
表現型1A 428
表現型1B 448
表現型1C 468

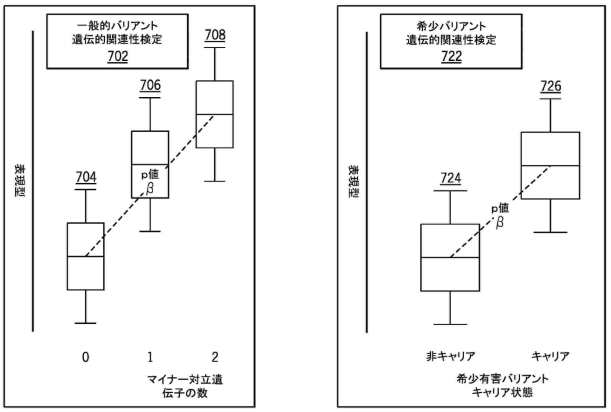
【図５】


人口統計
502
カテゴリ
バイオマーカー
504
臨床診断
506
疾患重症度の
尺度
508
心血管
疾患患者X
500
血液及び尿
バイオマーカー値
522
ボディマス
測定値
542
バイタルサイン
562

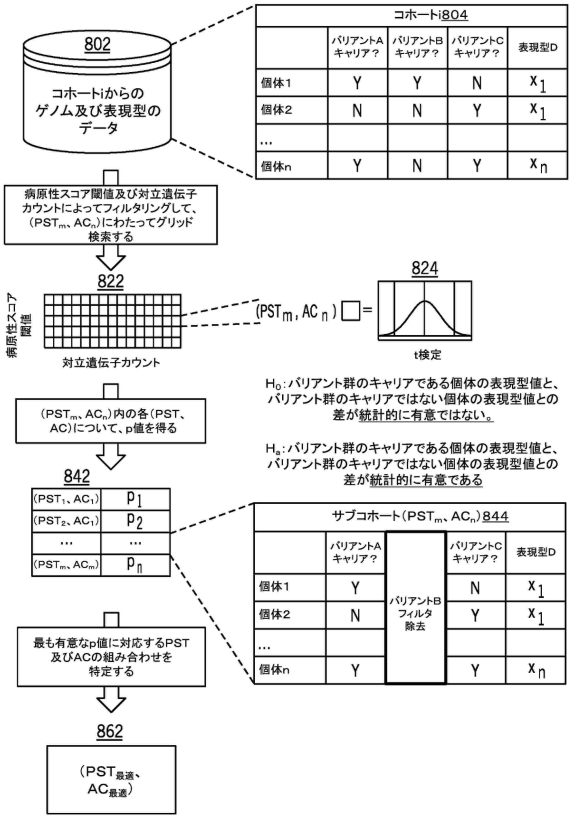
【図６】


一般的バリアント
602
希少バリアント
604
染色体
APOB
PCSK9
ANGPTL3
NR1I3
ABCG5
ABCG8
STAB1
ALB
TIMD4
NPC1L1
PLRA
HOOK3
ABCA1
JAK2
SCARB1
GAS6
LIPC
LCAT
ASGR3
SLC4A1
LIPG
LDLR
TMSC2
TMSC3

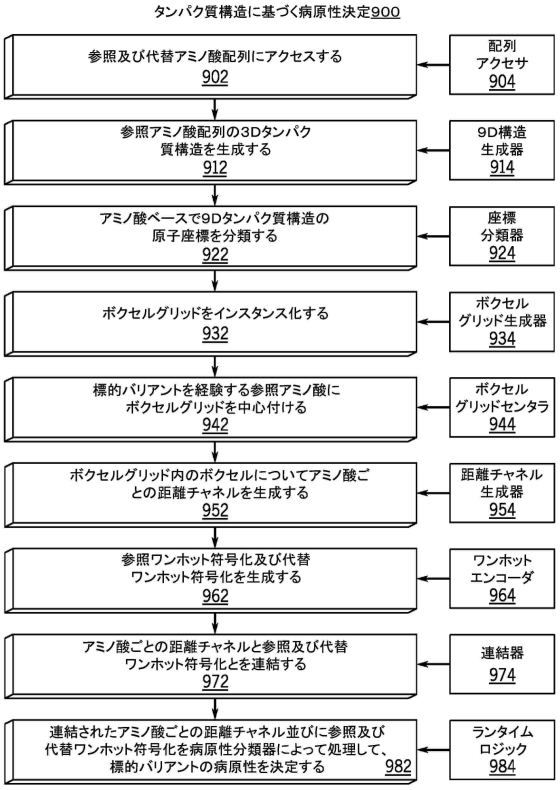
【図７】


一般的バリアント
遺伝的関連性検定
702
704
706
708
p値
β
表現型
0
1
2
マイナー対立遺
伝子の数
希少バリアント
遺伝的関連性検定
722
724
726
非キャリア
キャリア
希少有害バリアント
キャリア状態

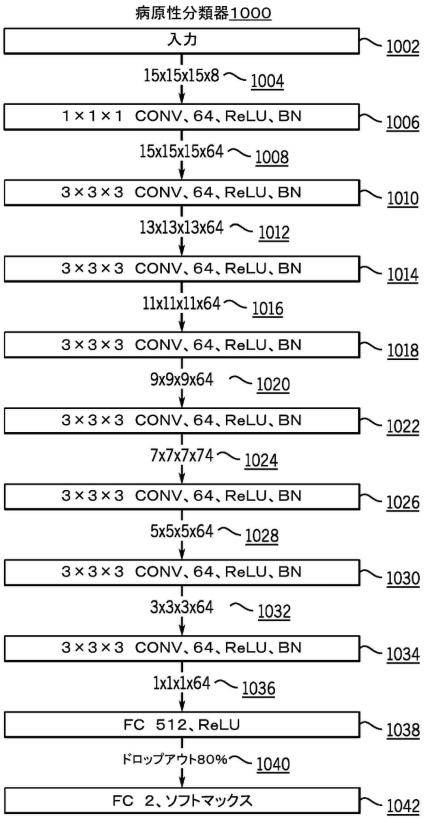
【図８】


802
コホートiからのゲノム及び表現型のデータ
コホートi804
バリアントAキャリア？
バリアントBキャリア？
バリアントCキャリア？
表現型D
個体1 Y Y N x1
個体2 N N Y x1
…
個体n Y N Y xn
病原性スコア閾値及び対立遺伝子カウントによってフィルタリングして、(PSTm、ACn)にわたってグリッド検索する
822
824
病原性スコア閾値
対立遺伝子カウント
(PSTm, ACn) □=
t検定
(PSTm、ACn)内の各(PST、AC)について、p値を得る
H0：バリアント群のキャリアである個体の表現型値と、バリアント群のキャリアではない個体の表現型値との差が統計的に有意ではない。
Ha：バリアント群のキャリアである個体の表現型値と、バリアント群のキャリアではない個体の表現型値との差が統計的に有意である
842
(PST1、AC1) p1
(PST2、AC1) p2
(PSTm、ACn) pn
サブコホート(PSTm、ACn)844
バリアントBフィルタ除去
最も有意なp値に対応するPST及びACの組み合わせを特定する
862
(PST最適、AC最適)

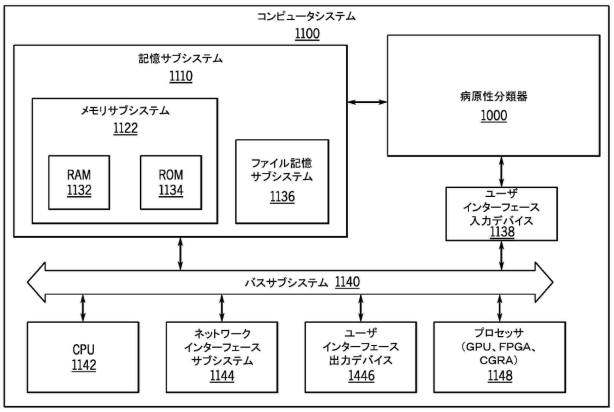
【図９】


タンパク質構造に基づく病原性決定900
参照及び代替アミノ酸配列にアクセスする 902
配列アクセサ 904
参照アミノ酸配列の3Dタンパク質構造を生成する 912
9D構造生成器 914
アミノ酸ベースで9Dタンパク質構造の原子座標を分類する 922
座標分類器 924
ボクセルグリッドをインスタンス化する 932
ボクセルグリッド生成器 934
標的バリアントを経験する参照アミノ酸にボクセルグリッドを中心付ける 942
ボクセルグリッドセンタラ 944
ボクセルグリッド内のボクセルについてアミノ酸ごとの距離チャネルを生成する 952
距離チャネル生成器 954
参照ワンホット符号化及び代替ワンホット符号化を生成する 962
ワンホットエンコーダ 964
アミノ酸ごとの距離チャネルと参照及び代替ワンホット符号化とを連結する 972
連結器 974
連結されたアミノ酸ごとの距離チャネル並びに参照及び代替ワンホット符号化を病原性分類器によって処理して、標的バリアントの病原性を決定する 982
ランタイムロジック 984

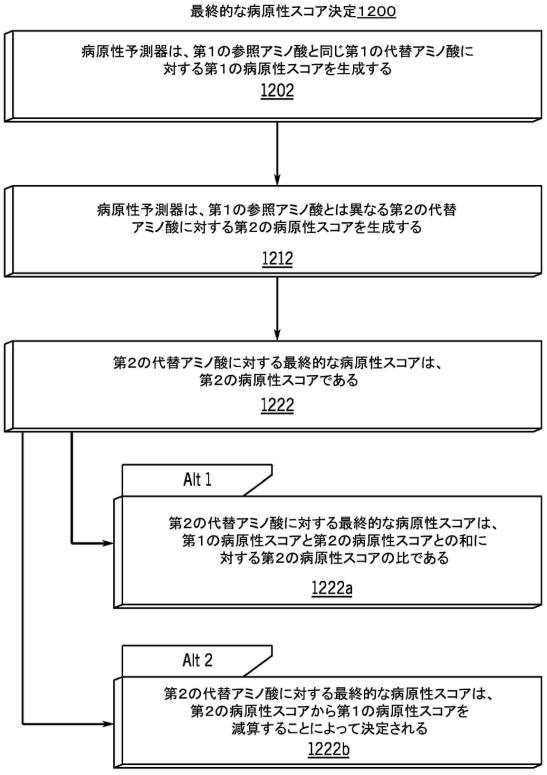
【図１０】


病原性分類器1000
入力 1002
15x15x15x8 1004
1×1×1 CONV、64、ReLU、BN 1006
15x15x15x64 1008
3×3×3 CONV、64、ReLU、BN 1010
13x13x13x64 1012
3×3×3 CONV、64、ReLU、BN 1014
11x11x11x64 1016
3×3×3 CONV、64、ReLU、BN 1018
9x9x9x64 1020
3×3×3 CONV、64、ReLU、BN 1022
7x7x7x74 1024
3×3×3 CONV、64、ReLU、BN 1026
5x5x5x64 1028
3×3×3 CONV、64、ReLU、BN 1030
3x3x3x64 1032
3×3×3 CONV、64、ReLU、BN 1034
1x1x1x64 1036
FC 512、ReLU 1038
ドロップアウト80% 1040
FC 2、ソフトマックス 1042

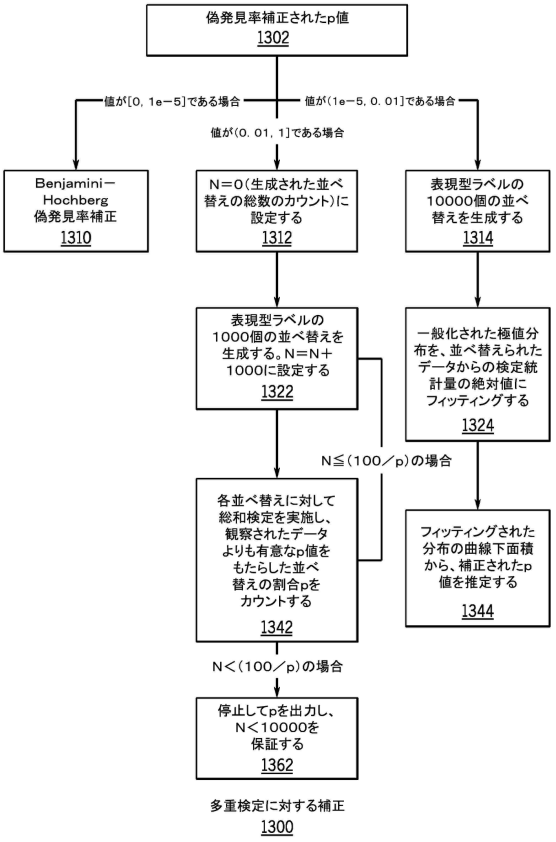
【図１１】


コンピュータシステム 1100
記憶サブシステム 1110
メモリサブシステム 1122
RAM 1132
ROM 1134
ファイル記憶サブシステム 1136
病原性分類器 1000
ユーザインターフェース入力デバイス 1138
バスサブシステム 1140
CPU 1142
ネットワークインターフェースサブシステム 1144
ユーザインターフェース出力デバイス 1446
プロセッサ(GPU、FPGA、CGRA) 1148

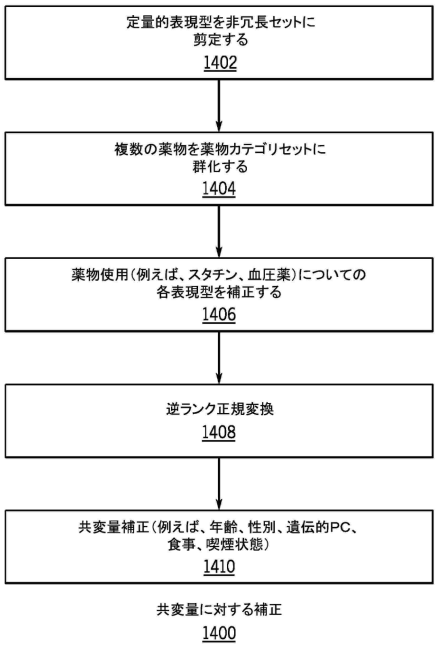
【図１２】



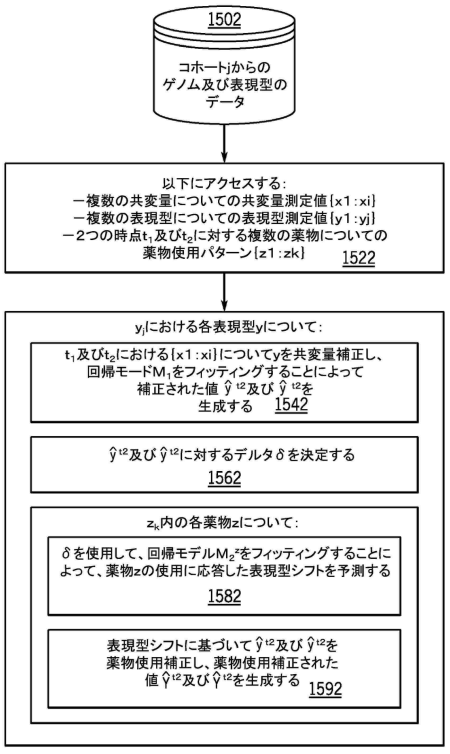
【図１３】



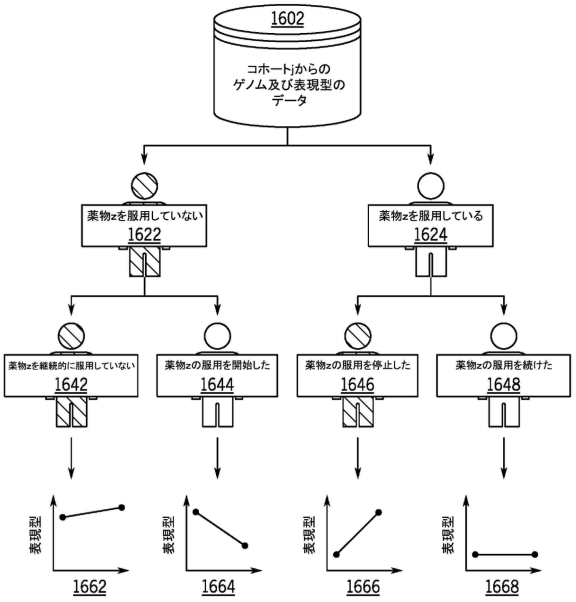
【図１４】



【図１５】



【図１６】



1. 薬物効果のモデルを学習する:

$$\delta_y = y_2 - y_1 = \beta_1 S + \varepsilon$$

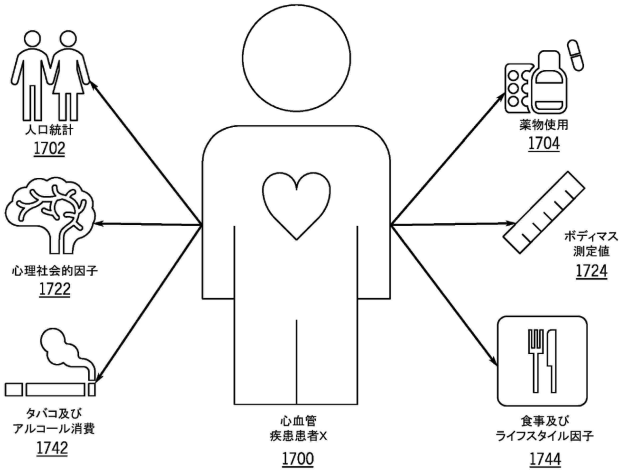
1682

2. 有意性の効果について検定する:

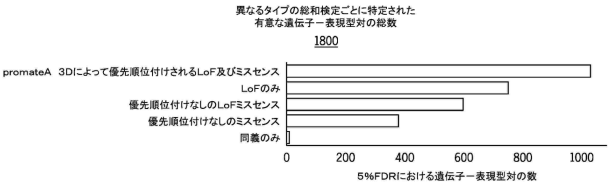
$$\beta_z \neq 0$$

1692

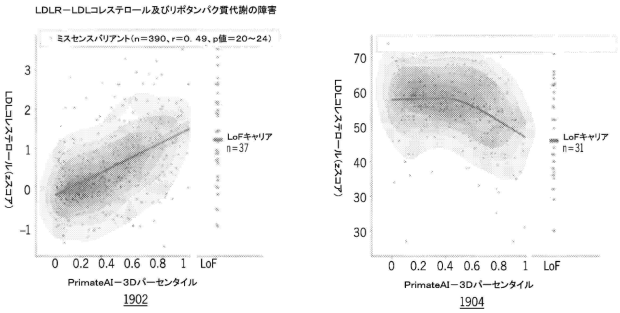
【図１７】



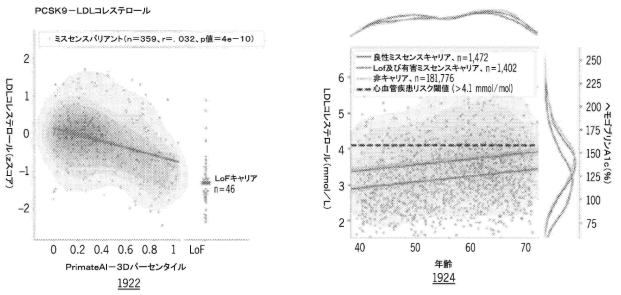
【図１８】



【図１９－１】

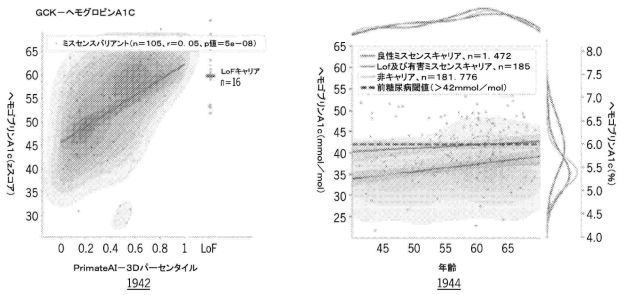


【図１９－２】



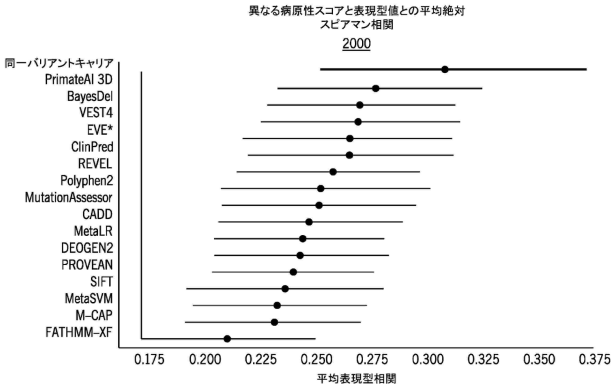
1922 1924

【図１９－３】



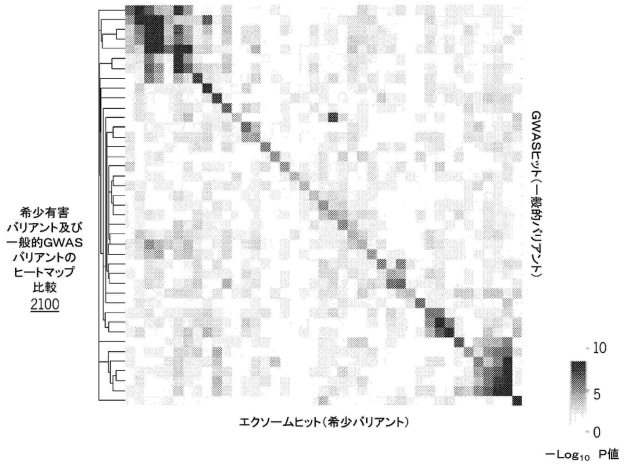
1942 1944

【図２０】



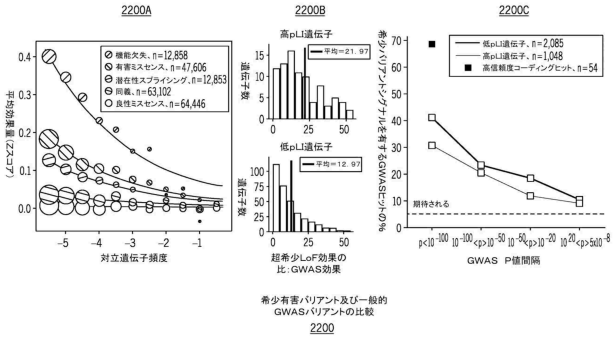
2000

【図２１】



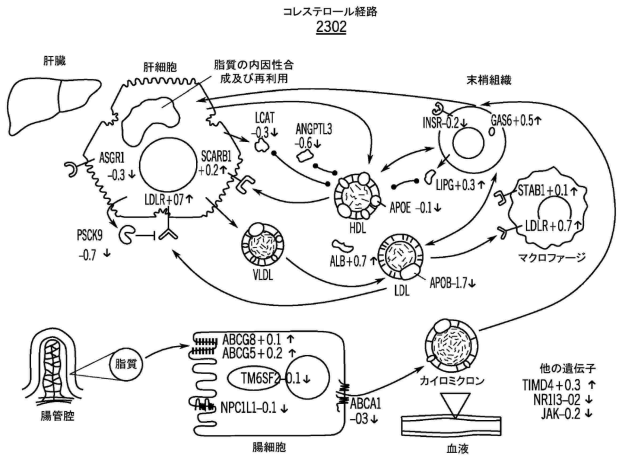
2100

【図２２】



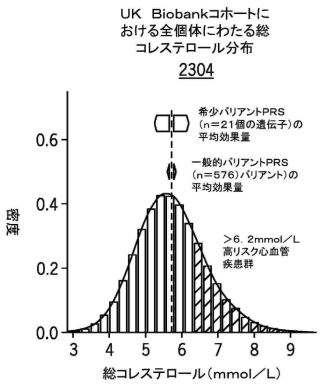
2200

【図２３－１】

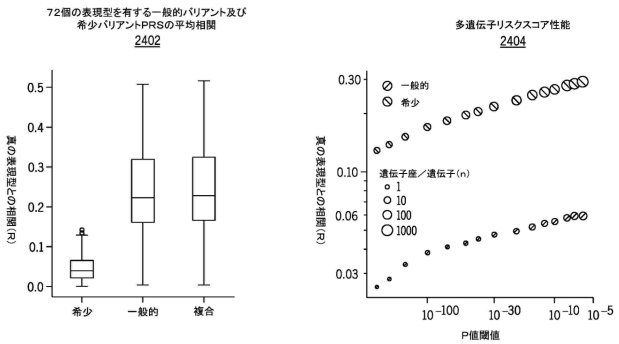


2302

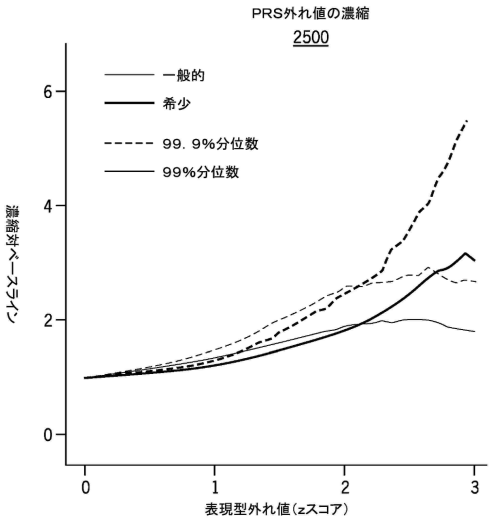
【図２３－２】



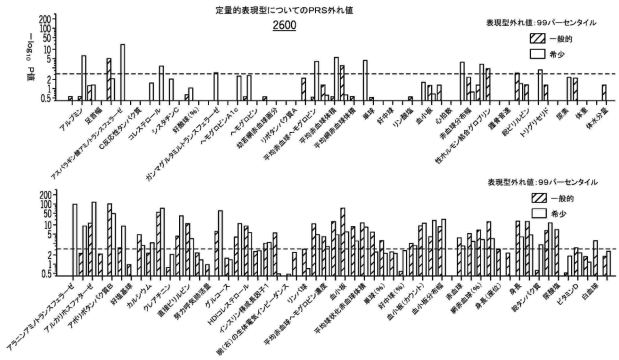
【図２４】



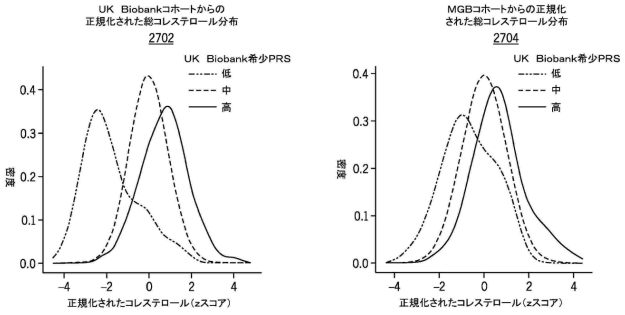
【図２５】



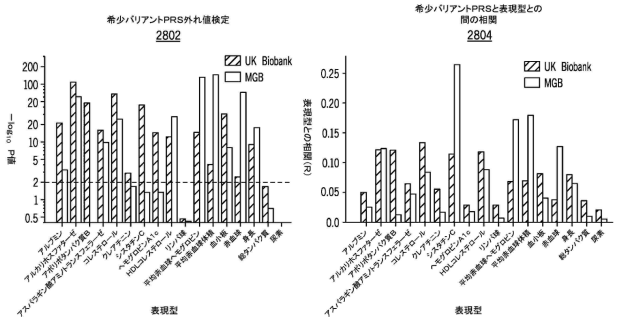
【図２６】



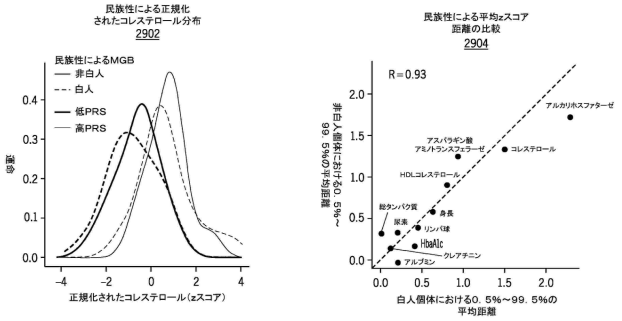
【図２７】



【図２８】



【図２９】



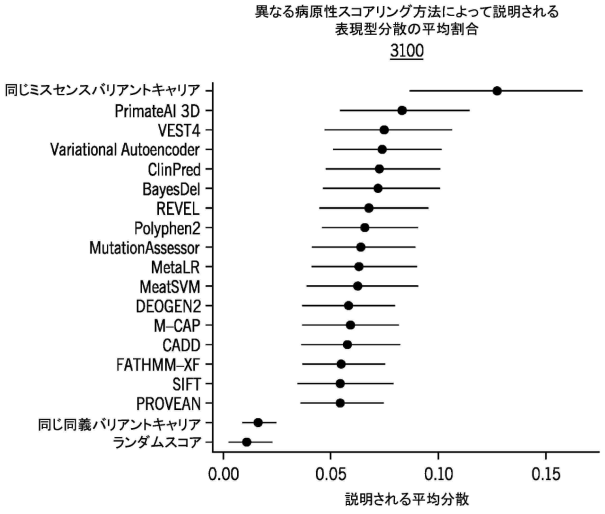
【図３０】

正常なコレステロールレベルに使用される一般的PRSバリアント及び
希少PRS遺伝子に対する効果量及び頻度の比較

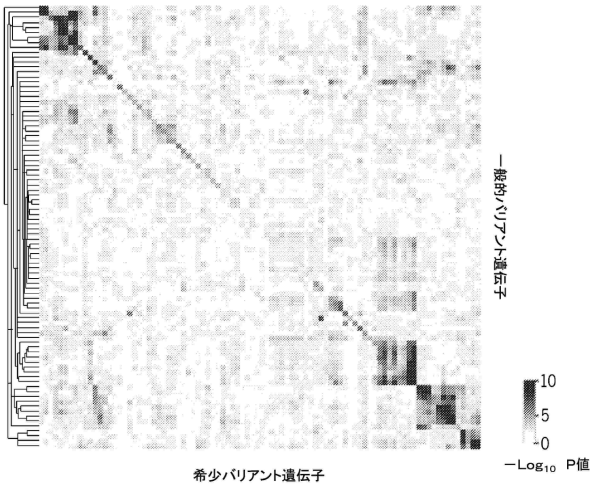
3000

		一般的PRSバリアント						希少PRS遺伝子			
染色体	位置（GRCh37）	メジャー対立遺伝子	マイナー対立遺伝子	対立遺伝子頻度	効果量（zスコア）	P値		遺伝子	集約対立遺伝子頻度	効果量（zスコア）	P値
18	47109955	A	G	0.014	0.198	2.6×10^{-84}		*ABCA1*	0.010	−0.344	1.0×10^{-57}
6	32644523	T	C	0.171	0.045	1.8×10^{-54}		*ABCG5*	0.009	0.156	3.5×10^{-13}
6	160575985	C	T	0.373	−0.044	1.8×10^{-127}		*ABCG8*	0.007	0.139	8.1×10^{-8}
13	114544024	T	C	0.297	0.033	1.4×10^{-67}		*ALB*	0.001	0.709	1.2×10^{-16}
6	29925268	C	A	0.068	0.042	4.9×10^{-20}		*ANGPTL3*	0.004	−0.569	1.4×10^{-67}
8	28843833	G	A	0.041	0.032	3.2×10^{-8}		*APOB*	0.002	−1.686	9.7×10^{-268}
11	2963612	C	T	0.391	−0.014	2.1×10^{-14}		*APOE*	0.007	−0.138	7.5×10^{-8}
19	7827830	G	A	0.152	−0.023	1.1×10^{-15}		*ASGR1*	0.002	−0.330	3.9×10^{-11}
14	94847262	T	A	0.049	0.038	8.3×10^{-13}		*GAS6*	0.001	0.479	4.3×10^{-7}
20	62362116	G	C	0.245	0.018	5.1×10^{-13}		*INSR*	0.003	−0.212	3.0×10^{-8}
17	38175462	C	T	0.368	0.016	1.7×10^{-20}		*JAK2*	0.003	−0.213	2.2×10^{-7}
19	46436564	C	T	0.016	−0.115	1.4×10^{-32}		*LCAT*	0.003	−0.339	7.1×10^{-17}
2	43987888	CTT	C	0.067	−0.042	4.7×10^{-21}		*LDLR*	0.004	0.721	1.7×10^{-107}
9	78730766	A	G	0.423	0.014	1.3×10^{-20}		*LIPG*	0.001	0.347	2.4×10^{-8}
5	156369171	T	C	0.347	−0.046	1.0×10^{-154}		*NPC1L1*	0.015	−0.128	8.4×10^{-14}
9	107589744	C	T	0.153	0.055	3.6×10^{-71}		*NR1I3*	0.002	−0.239	2.7×10^{-7}
		...						*PCSK9*	0.005	−0.714	2.8×10^{-134}
		558個のバリアントは省略した						*SCARB1*	0.008	0.157	6.7×10^{-11}
		...						*STAB1*	0.014	0.095	2.6×10^{-7}
4	104250412	C	T	0.132	−0.018	4.7×10^{-9}		*TIMD4*	0.002	0.309	2.8×10^{-9}
9	110530324	G	A	0.288	0.013	2.4×10^{-10}		*TM6SF2*	0.009	−0.141	1.7×10^{-10}

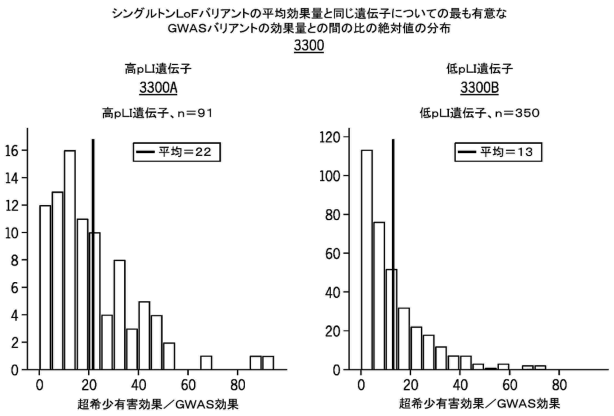
【図３１】



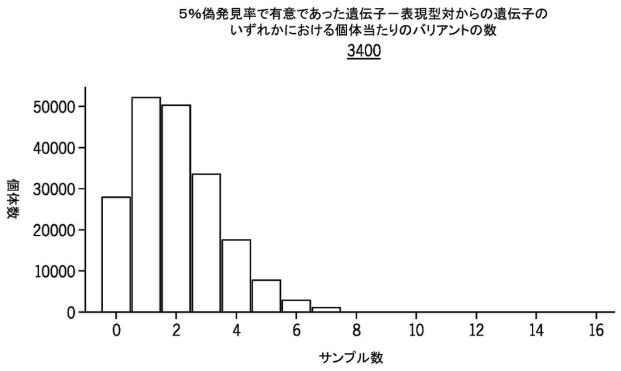
【図３２】



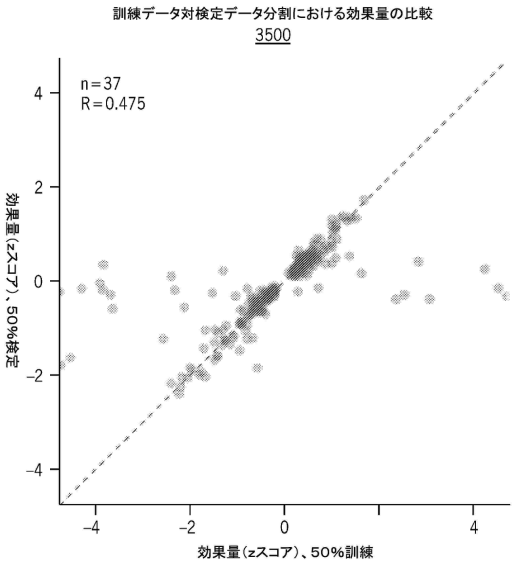
【図３３】



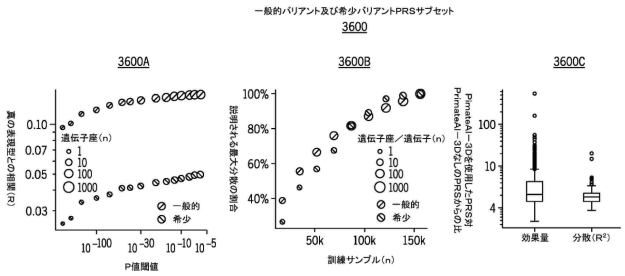
【図３４】



【図３５】



【図３６】



【国際調査報告】

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2022/054194

A. CLASSIFICATION OF SUBJECT MATTER
INV. G16B20/20 G16B30/10 G16B20/40 G16B40/20 G16B5/10
ADD. G16B25/10 G16B15/30

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
G16B

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
EPO-Internal, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	POVYSIL GUNDULA ET AL: "Rare-variant collapsing analyses for complex traits: guidelines and applications", NATURE REVIEWS GENETICS, NATURE PUBLISHING GROUP, GB, vol. 20, no. 12, 11 October 2019 (2019-10-11), pages 747-759, XP036927755, ISSN: 1471-0056, DOI: 10.1038/S41576-019-0177-4 [retrieved on 2019-10-11] the whole document in particular pages 752, 753, 755-757; figures 1,2, box1,2 ----- -/--	1-36

[x] Further documents are listed in the continuation of Box C.
[x] See patent family annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance;; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance;; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
24 April 2023	04/05/2023

Name and mailing address of the ISA/	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016	Rákossy, Z

Form PCT/ISA/210 (second sheet) (April 2005)

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2022/054194

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	WO 2020/242976 A1 (UNIV LELAND STANFORD JUNIOR [US]; PRITCHARD JONATHAN [US] ET AL.) 3 December 2020 (2020-12-03) the whole document in particular paragraph [0022] - paragraph [0028]; figures 5,13 paragraph [0138] - paragraph [0166] -----	1-36
X	FRAZER JONATHAN ET AL: "Disease variant prediction with deep generative models of evolutionary data", NATURE, NATURE PUBLISHING GROUP UK, LONDON, vol. 599, no. 7883, 27 October 2021 (2021-10-27), pages 91-95, XP037645771, ISSN: 0028-0836, DOI: 10.1038/S41586-021-04043-8 [retrieved on 2021-10-27] the whole document in particular page 92 - page 94 -----	1-36

Form PCT/ISA/210 (continuation of second sheet) (April 2005)

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/US2022/054194

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO 2020242976 A1	03-12-2020	NONE	

フロントページの続き

(32)優先日　令和3年12月29日(2021.12.29)
(33)優先権主張国・地域又は機関
米国(US)
(31)優先権主張番号　63/294,827
(32)優先日　令和3年12月29日(2021.12.29)
(33)優先権主張国・地域又は機関
米国(US)
(31)優先権主張番号　63/294,828
(32)優先日　令和3年12月29日(2021.12.29)
(33)優先権主張国・地域又は機関
米国(US)
(31)優先権主張番号　63/294,830
(32)優先日　令和3年12月29日(2021.12.29)
(33)優先権主張国・地域又は機関
米国(US)
(31)優先権主張番号　63/351,283
(32)優先日　令和4年6月10日(2022.6.10)
(33)優先権主張国・地域又は機関
米国(US)
(31)優先権主張番号　63/351,299
(32)優先日　令和4年6月10日(2022.6.10)
(33)優先権主張国・地域又は機関
米国(US)
(31)優先権主張番号　63/351,317
(32)優先日　令和4年6月10日(2022.6.10)
(33)優先権主張国・地域又は機関
米国(US)
(31)優先権主張番号　17/968,723
(32)優先日　令和4年10月18日(2022.10.18)
(33)優先権主張国・地域又は機関
米国(US)
(81)指定国・地域　AP(BW,CV,GH,GM,KE,LR,LS,MW,MZ,NA,RW,SD,SL,ST,SZ,TZ,UG,ZM,ZW),EA(AM,AZ,BY,KG,KZ,RU,TJ,TM),EP(AL,AT,BE,BG,CH,CY,CZ,DE,DK,EE,ES,FI,FR,GB,GR,HR,HU,IE,IS,IT,LT,LU,LV,MC,ME,MK,MT,NL,NO,PL,PT,RO,RS,SE,SI,SK,SM,TR),OA(BF,BJ,CF,CG,CI,CM,GA,GN,GQ,GW,KM,ML,MR,NE,SN,TD,TG),AE,AG,AL,AM,AO,AT,AU,AZ,BA,BB,BG,BH,BN,BR,BW,BY,BZ,CA,CH,CL,CN,CO,CR,CU,CV,CZ,DE,DJ,DK,DM,DO,DZ,EC,EE,EG,ES,FI,GB,GD,GE,GH,GM,GT,HN,HR,HU,ID,IL,IN,IQ,IR,IS,IT,JM,JO,JP,KE,KG,KH,KN,KP,KR,KW,KZ,LA,LC,LK,LR,LS,LU,LY,MA,MD,MG,MK,MN,MW,MX,MY,MZ,NA,NG,NI,NO,NZ,OM,PA,PE,PG,PH,PL,PT,QA,RO,RS,RU,RW,SA,SC,SD,SE,SG,SK,SL,ST,SV,SY,TH,TJ,TM,TN,TR,TT,TZ,UA,UG,US,UZ,VC,VN,WS,ZA,ZM,ZW
レイテッド内
(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・イルミナ・インコーポレイテッド内
(72)発明者　カイ－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・イルミナ・インコーポレイテッド内
Ｆターム（参考）　4B029　AA07 AA23 BB20
4B063　QA08 QQ02 QQ42 QR32 QR72 QS39