(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7635995号
(P7635995)

(45)発行日　令和7年2月26日(2025.2.26)

(24)登録日　令和7年2月17日(2025.2.17)

(51)国際特許分類
Ｇ１６Ｂ　40/00　(2019.01)

ＦＩ
Ｇ１６Ｂ　40/00

請求項の数　17　（全31頁）

(21)出願番号　特願2021-556270(P2021-556270)
(86)(22)出願日　令和2年3月19日(2020.3.19)
(65)公表番号　特表2022-525638(P2022-525638A)
(43)公表日　令和4年5月18日(2022.5.18)
(86)国際出願番号　PCT/US2020/023633
(87)国際公開番号　WO2020/191195
(87)国際公開日　令和2年9月24日(2020.9.24)
審査請求日　令和5年3月20日(2023.3.20)
(31)優先権主張番号　62/820,286
(32)優先日　平成31年3月19日(2019.3.19)
(33)優先権主張国・地域又は機関
米国(US)

(73)特許権者　520446045
センバ 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 ９４０７０，サン カルロス，スイート ４１０，２０１ インダストリアル ロード
(74)代理人　100114775
弁理士　高岡 亮一
(74)代理人　100121511
弁理士　小田 直
(74)代理人　100202751
弁理士　岩堀 明代
(74)代理人　100208580
弁理士　三好 玲奈
(74)代理人　100191086
弁理士　高橋 香元

最終頁に続く

(54)【発明の名称】　非メンデル表現型に対する遺伝学的リスクを決定するための親類の情報の使用

(57)【特許請求の範囲】

【請求項１】

非メンデル表現型リスクスコアを出力するための方法であって、

第１のデータセットから、（ｉ）２つ以上の関心のある非メンデル遺伝子を有する対象に対する遺伝子型データ及び（ｉｉ）前記関心のある非メンデル遺伝子の１つ以上を有する対象の１人以上の血縁者に対する遺伝子型データ及び表現型データを受け取る工程と、

第２のデータセットから、遺伝子型集団データ及び表現型集団データを受け取る工程であって、前記集団が２人以上の血縁者であって、前記２つ以上の関心のある非メンデル遺伝子を有する２人以上の血縁者の１つ以上のセットを含む、工程と、

前記対象に対する前記２つ以上の関心のある非メンデル遺伝子と関連がある前記表現型リスクスコアを、モデルを用いて決定する工程であって、（ｉ）前記モデルは前記第２のデータセットを用いてトレーニングし、（ｉｉ）前記表現型リスクスコアは多遺伝子リスクスコアであり、かつ（ｉｉｉ）前記多遺伝子リスクスコアは、前記対象の前記１人以上の血縁者において表現型が観察されるか否かを示す前記第１のデータセットに基づく、工程と、

前記対象に対する前記２つ以上の関心のある非メンデル遺伝子に対して前記表現型リスクスコアを出力する工程と

を含む、方法。

【請求項２】

前記第２のデータセットが、２人以上の血縁者の複数セットに対する遺伝子型集団データ及び表現型集団データを含む、請求項１に記載の方法。

【請求項３】
前記第１のデータセット中の血縁者が、前記対象の母、父、兄弟、姉妹、息子、娘、祖父、祖母、叔母、叔父、姪、甥及び従兄弟姉妹（ｆｉｒｓｔ　ｃｏｕｓｉｎ）のうち１人以上を含み、
前記第２のデータセットが、前記第１のデータセット中の前記対象と同じ血縁を有する２人以上の対象を含む、請求項１に記載の方法。
【請求項４】
（ｉ）前記血縁者の１人以上が男性の親類である、又は（ｉｉ）前記血縁者の１人以上が女性の親類である、請求項１に記載の方法。
【請求項５】
前記第１のデータセットが、前記対象の複数の血縁者に対するデータを含む、請求項１に記載の方法。
【請求項６】
前記関心のある遺伝子が関心のある遺伝子変異体である、請求項１に記載の方法。
【請求項７】
前記第１のデータセット及び第２のデータセットが、表現型の開始年齢と関連するデータを含む、請求項１に記載の方法。
【請求項８】
プロセッサと、
前記プロセッサにより実行される場合に、前記プロセッサに演算を実行させる命令を格納するための前記プロセッサと結合されるメモリと、
を含むシステムであって、前記演算が以下の処理、
第１のデータセットから、（ｉ）１つ以上の関心のある非メンデル遺伝子を有する対象に対する遺伝子型データ及び（ｉｉ）前記対象の１人以上の血縁者に対する遺伝子型データ及び表現型データであって、前記１人以上の血縁者は前記関心のある非メンデル遺伝子のうち１つ以上を有する、遺伝子型データ及び表現型データを受け取る工程、
第２のデータセットから、遺伝子型集団データ及び表現型集団データを受け取る工程であって、前記集団が、前記１つ以上の関心のある非メンデル遺伝子を有する２人以上の血縁者の１つ以上のセットを含む、工程、
前記対象に対する前記１つ以上の関心のある非メンデル遺伝子と関連がある表現型リスクスコアを、モデルを用いて決定する工程であって、（i）前記モデルは前記第２のデータセットを用いてトレーニングし、（ｉｉ）前記表現型リスクスコアは、前記１つ以上の関心のある非メンデル遺伝子を有する前記１人以上の血縁者において表現型が観察されるか否かを示す前記第１のデータセットに基づく、工程、及び
前記対象に対する前記１つ以上の関心のある非メンデル遺伝子に対して前記表現型リスクスコアを出力する工程
を含む、システム。
【請求項９】
プロセッサにより実行される場合、前記プロセッサに演算を実施させる、そこに格納される命令を有する非一時的な機械可読媒体であって、前記演算が以下の処理、
第１のデータセットから、（ｉ）１つ以上の関心のある非メンデル遺伝子を有する対象に対する遺伝子型データ及び（ｉｉ）前記対象の１人以上の血縁者に対する遺伝子型データ及び表現型データを受け取る工程であって、前記１人以上の血縁者は前記関心のある非メンデル遺伝子のうち１つ以上を有する、工程、
第２のデータセットから、遺伝子型集団データ及び表現型集団データを受け取る工程であって、前記集団が、前記１つ以上の関心のある非メンデル遺伝子を有する２人以上の血縁者の１つ以上のセットを含む、工程、
前記プロセッサによって、前記対象に対する前記１つ以上の関心のある非メンデル遺伝子と関連がある表現型リスクスコアを、モデルを用いて決定する工程であって、（i）前記モデルは前記第２のデータセットを用いてトレーニングし、（ｉｉ）前記表現型リスクスコ

アは、前記１つ以上の関心のある非メンデル遺伝子を有する前記１人以上の血縁者において表現型が観察されるか否かを示す前記第１のデータセットに基づく、工程、及び
前記対象に対する前記１つ以上の関心のある非メンデル遺伝子に対して前記表現型リスクスコアを出力する工程
を含む、非一時的な機械可読媒体。
【請求項１０】
前記第２のデータセットが、２人以上の血縁者の複数セットに対する遺伝子型集団データ及び表現型集団データを含む、請求項９に記載の非一時的な機械可読媒体。
【請求項１１】
前記第１のデータセット中の血縁者が、前記対象の母、父、兄弟、姉妹、息子、娘、祖父、祖母、叔母、叔父、姪、甥及び従兄弟姉妹（ｆｉｒｓｔ　ｃｏｕｓｉｎ）のうち１人以上を含み、
前記第２のデータセットが、前記第１のデータセット中の前記対象と同じ血縁関係を有する２人以上の対象を含む、請求項９に記載の非一時的な機械可読媒体。
【請求項１２】
（ｉ）前記血縁者の１人以上が男性の親類である、又は（ｉｉ）前記血縁者の１人以上が女性の親類である、請求項９に記載の非一時的な機械可読媒体。
【請求項１３】
前記第１のデータセットが前記対象の複数の血縁者に対するデータを含む、請求項９に記載の非一時的な機械可読媒体。
【請求項１４】
前記関心のある遺伝子が関心のある遺伝子変異体である、請求項９に記載の非一時的な機械可読媒体。
【請求項１５】
前記第１のデータセット及び第２のデータセットが、表現型の開始年齢に関連するデータを含む、請求項９に記載の非一時的な機械可読媒体。
【請求項１６】
表現型リスクスコアを出力するための方法であって、
第１のデータセットから、（ｉ）１つ以上の関心のある非メンデル遺伝子を有する対象に対する遺伝子型データ及び（ｉｉ）前記対象の１人以上の血縁者に対する遺伝子型データ及び表現型データを受け取る工程であって、前記１人以上の血縁者は前記関心のある非メンデル遺伝子のうち１つ以上を有する、工程と、
第２のデータセットから、遺伝子型集団データ及び表現型集団データを受け取る工程であって、前記集団が、前記１つ以上の関心のある非メンデル遺伝子を有する２人以上の血縁者の１つ以上のセットを含む、工程と、
前記対象に対する前記１つ以上の関心のある非メンデル遺伝子と関連がある前記表現型リスクスコアを、モデルを用いて決定する工程であって、（i）前記モデルは前記第２のデータセットを用いてトレーニングし、（ｉｉ）前記表現型リスクスコアは、前記１つ以上の関心のある非メンデル遺伝子を有する前記１人以上の血縁者において表現型が観察されるか否かを示す前記第１のデータセットに基づく、工程と、
前記対象に対する前記１つ以上の関心のある非メンデル遺伝子に対して前記表現型リスクスコアを出力する工程と
を含む、方法。
【請求項１７】
前記血縁者の表現型データが与えられたとき、前記対象におけるリスクと比較して、前記対象におけるリスクが１つ以上の関心のある非メンデル遺伝子によりどのように変化させられるかを予測するために、前記第１及び第２のデータセットにおいてモデルをトレーニングすること
を含む、請求項１６に記載の方法。
【発明の詳細な説明】

【技術分野】
【０００１】
関連出願に対する相互参照
本願は、それらの全体において参照により本明細書中で組み込まれる２０１９年３月１９日出願の米国特許仮出願第６２／８２０，２８６号明細書の優先権を主張する。
【０００２】
親類の遺伝学的情報を使用して非メンデル表現型の遺伝学的リスクを決定するための方法が記載される。
【背景技術】
【０００３】
メンデル遺伝子の場合、突然変異遺伝子の０、１又は２のバージョンを対象が受け継ぐか否か及びこの遺伝子が顕性又は潜性遺伝を呈するか否かに依存して、表現型を発現する確率は、大まかに言って０又は１である。メンデル表現型に対して、対象に対するリスクは、明確に定義された方法で家系図及び対象の親類の病歴を分析することによって確立される。非メンデル遺伝子の場合、特定の遺伝子突然変異を有する対象が表現型を発現する確率は絶対的に０又は１ではない。さらに、非メンデル表現型は、一般的には複数の遺伝子により影響を受ける。複数遺伝子の影響は、一般的には多遺伝子リスクモデルにおいて捉えられ、これは不正確になりがちであり、各遺伝子の影響を較正するために集団レベルデータを使用する。当技術分野で、対象が非メンデル表現型に対するリスクを有するか否かを決定するためのより正確な方法、特に家族病歴を組み込み得る方法が必要とされている。
【発明の概要】
【０００４】
対象の親類の疾患又は表現型の状況を使用することによって各対象に対してより正確になる、非メンデル表現型リスクスコアを出力するための方法が提供される。いくつかの態様は、第１のデータセットから、（ｉ）１つ以上の関心のある非メンデル遺伝子を有する対象に対する遺伝子型データ及び（ｉｉ）関心のある非メンデル遺伝子の１つ以上を有する対象の１人以上の血縁者に対する遺伝子型データ及び表現型データを受け取ることを含む。いくつかの態様は、第２のデータセットから、遺伝子型集団データ及び表現型集団データを受け取ることを含み、この集団は２人以上の血縁者の１つ以上のセットを含む。いくつかの態様は、関心のある非メンデル遺伝子の１つ以上と関連する対象においてリスクを決定するために、第１及び第２のデータセットにおいてモデルをトレーニングすることを含む。いくつかの態様は、対象に対する表現型リスクスコアを出力することを含む。
【０００５】
いくつかの態様では、第２のデータセットは、２人以上の血縁者の複数セットに対する遺伝子型集団データ及び表現型集団データを含む。
【０００６】
いくつかの態様では、第１のデータセット中の血縁者は、対象の母、父、兄弟、姉妹、息子、娘、祖父、祖母、叔母、叔父、姪、甥及び従兄弟姉妹（ｆｉｒｓｔ　ｃｏｕｓｉｎ）の１人以上を含む。いくつかの態様では、第２のデータセットは、第１のデータセット中の対象と同じ血縁関係を有する２人以上の対象を含む。
【０００７】
いくつかの態様では、血縁者の１人以上は男性の親類である。いくつかの態様では、血縁者の１人以上は女性の親類である。
【０００８】
いくつかの態様では、第１のデータセットは、対象の複数の血縁者に対するデータを含む。いくつかの態様では、血縁者の１人以上は男性の親類であり、血縁者の１人以上は女性の親類である。
【０００９】
いくつかの態様では、関心のある遺伝子は関心のある遺伝子変異体である。

【００１０】
いくつかの態様では、第１のデータセット及び第２のデータセットは、表現型開始年齢と関連するデータを含む。
【００１１】
プロセッサと；プロセッサにより実行される際にこのプロセッサに対して演算を実施させる命令を格納するための、プロセッサと結合されるメモリと、を含むシステムも提供され、この演算は、第１のデータセットから、（ｉ）１つ以上の関心のある非メンデル遺伝子を有する対象に対する遺伝子型データ及び（ｉｉ）関心のある遺伝子のうち１つ以上を有する前記対象の１人以上の血縁者に対する遺伝子型データ及び表現型データを受け取り；第２のデータセットから、遺伝子型集団データ及び表現型集団データを受け取り、この集団が２人以上の血縁者の１つ以上のセットを含み；関心のある非メンデル遺伝子の１つ以上と関連する対象におけるリスクを決定するために第１及び第２のデータセットにおいてモデルをトレーニングし、対象に対する表現型リスクスコアを出力することを含む。
【００１２】
プロセッサにより実行される際に、このプロセッサに対して演算を実施させる、そこに格納される命令を有する非一時的な機械可読媒体も提供され、この演算は、第１のデータセットから、（ｉ）１つ以上の関心のある非メンデル遺伝子を有する対象に対する遺伝子型データ及び（ｉｉ）関心のある遺伝子のうち１つ以上を有する対象の１人以上の血縁者に対する遺伝子型データ及び表現型データを受け取り；第２のデータセットから、遺伝子型集団データ及び表現型集団データを受け取り、この集団が２人以上の血縁者の１つ以上のセットを含み；関心のある非メンデル遺伝子の１つ以上と関連する対象におけるリスクを決定するために第１及び第２のデータセットにおいてモデルをトレーニングし、対象に対する表現型リスクスコアを出力することを含む。
【００１３】
システム又は非一時的な機械可読媒体に関連するいくつかの態様では、第２のデータセットは、２人以上の血縁者に対する遺伝子型集団データ及び表現型集団データを含む。いくつかの態様では、第１のデータセット中の血縁者は、対象の母、父、兄弟、姉妹、息子、娘、祖父、祖母、叔母、叔父、姪、甥及び従兄弟姉妹（ｆｉｒｓｔ　ｃｏｕｓｉｎ）のうち１人以上を含む。いくつかの態様では、第２のデータセットは、第１のデータセット中の対象と同じ血縁関係を有する２人以上の対象を含む。いくつかの態様では、この血縁者の１人以上は男性の親類である。いくつかの態様では、この血縁者の１人以上は女性の親類である。
【００１４】
システム又は非一時的な機械可読媒体に関連するいくつかの態様では、第１のデータセットは、対象の複数の血縁者に対するデータを含む。いくつかの態様では、この血縁者の１人以上は男性の親類であり、この血縁者の１人以上は女性の親類である。
【００１５】
システム又は非一時的な機械可読媒体に関連するいくつかの態様では、関心のある遺伝子は関心のある遺伝子変異体である。
【００１６】
システム又は非一時的な機械可読媒体に関連するいくつかの態様では、第１のデータセット及び第２のデータセットは、表現型開始年齢と関連するデータを含む。
【００１７】
多遺伝子リスクスコアを出力するための方法も提供され、この方法は、第１のデータセットから、（ｉ）１つ以上の関心のある非メンデル遺伝子を有する対象に対する遺伝子型データ及び（ｉｉ）関心のある非メンデル遺伝子のうち１つ以上を有する対象の１人以上の血縁者に対する遺伝子型データ及び表現型データを受け取り；第２のデータセットから、遺伝子型集団データ及び表現型集団データを受け取り、この集団が２人以上の血縁者の１つ以上のセットを含み；１つ以上の関心のある非メンデル遺伝子に基づき、対象におけるリスクを決定するために第１及び第２のデータセットにおいてモデルをトレーニングし

、対象に対する表現型リスクスコアを出力することを含む。いくつかの態様は、血縁者の表現型データが与えられたとき、対象におけるリスクと比較して、対象におけるリスクが１つ以上の関心のある非メンデル遺伝子によりどのように変化させられるかを予測するために、第１及び第２のデータセットにおいてモデルをトレーニングすることを含む。

【００１８】

表現型リスクスコアに基づいて対象を処置する方法も提供される。

【図面の簡単な説明】

【００１９】

【図１】図１は、平均発生年齢が６０歳である発現される表現型の模擬ヒストグラムを示す。

【図２】図２は、計算装置のブロック図の一例である。

【図３】図３は、３つの遺伝子に適用される方法の態様を例示するシミュレーションの結果であり、この第３の遺伝子の集団内頻度は１．０％であり；図３Ａ及び３Ｂは、関連遺伝子のサブセットのみがモデルにおいて利用可能である対象に対する予測のヒストグラムを示し；図３Ｃは、全ての遺伝学的変数が含まれる対象に対する予測のヒストグラムを示す。

【図４】図４は、３つの遺伝子に適用される方法の態様を例示するシミュレーションの結果であり、この第３の遺伝子の集団内頻度は０．２％であり；図４Ａ及び４Ｂは、関連遺伝子のサブセットのみがモデルにおいて利用可能である対象に対する予測のヒストグラムを示し；図４Ｃは、全ての遺伝学的変数が含まれる対象に対する予測のヒストグラムを示す。

【図５】図５は、３つの遺伝子に適用される方法の態様を例示するシミュレーションの結果であり、この第３の遺伝子の集団内頻度は０．０５％であり；図５Ａ及び５Ｂは、関連遺伝子のサブセットのみがモデルにおいて利用可能である対象に対する予測のヒストグラムを示し；図５Ｃは全ての遺伝学的変数が含まれる対象に対する予測のヒストグラムを示す。

【発明を実施するための形態】

【００２０】

本明細書中で使用される技術及び科学用語は、別段定められない限り、本発明が属する技術分野の当業者により一般的に理解される意味を有する。次の記載及び実施例において言及される物質は、別段示されない限り、市販供給源から入手可能である。

【００２１】

本明細書中で使用される場合、「ａ」、「ａｎ」及び「ｔｈｅ」の単数形は、単数のみを示すことが明らかに述べられない限り、単数及び複数の両方を示す。

【００２２】

「約」という用語は、包含される数が本明細書中で示される正確な数に限定されないことを意味し、本発明の範囲から逸脱することなく、実質的に引用される数の前後の数を指すものとする。本明細書中で使用される場合、「約」は当業者により理解され、それが使用される文脈においてある程度まで変動する。それが使用される文脈が与えられるとき、当技術分野の熟練者にとって明確でない用語の使用がある場合、「約」は、特定の用語のプラス又はマイナス１０％までを意味する。

【００２３】

「血縁者」という用語は、１人以上の共通祖先を有する２人以上の対象を指す。対象の血縁者の非限定例としては、対象の母、父、兄弟、姉妹、息子、娘、祖父、祖母、叔母、叔父、姪、甥及び／又は従兄弟姉妹（ｆｉｒｓｔ　ｃｏｕｓｉｎ）が挙げられる。いくつかの態様では、血縁者は男性である。いくつかの態様では、血縁者は女性である。

【００２４】

「遺伝子」という用語は、ポリペプチドをコードするか又は生物において機能的役割を果たす一続きのＤＮＡ又はＲＮＡに関する。遺伝子は、野生型遺伝子又は野生型遺伝子の変異体若しくは突然変異であり得る。「関心のある遺伝子」は、特定の表現型又は特定の

表現型のリスクと関連することが知られていてもよいし、知られていなくてもよい、遺伝子又は遺伝子の変異体を指す。

【００２５】

「発現」は、ＤＮＡ鋳型から（ｍＲＮＡ又は他のＲＮＡ転写物などへ）ポリヌクレオチドが転写される過程及び／又は転写されたｍＲＮＡが次にペプチド、ポリペプチド又はタンパク質へ翻訳される過程を指す。核酸配列がペプチド、ポリペプチド又はタンパク質をコードする場合、遺伝子発現は、核酸（例えばＤＮＡ又はＲＮＡ、例えばｍＲＮＡなど）及び／又はペプチド、ポリペプチド若しくはタンパク質の産生に関する。従って、「発現レベル」は、試料中の核酸（例えばｍＲＮＡ）又はタンパク質の量を指し得る。

【００２６】

対象が表現型を有するリスクを決定するために遺伝学的情報を使用する、新規及び予想外の方法を記載する。非メンデル遺伝子の場合、対象が表現型を発現する確率は集団データから算出され得る。しかし、対象が、それらの親類と同じ突然変異である遺伝子突然変異を有する場合、及び親類がその表現型を有する場合、対象が表現型を発現する確率は、親類のデータなしで算出した集団リスクを用いる場合よりも正確に算出され得る。

【００２７】

遺伝子選択

関心のある遺伝子は、当技術分野で公知の何らかの手段により同定され得る。例えば，関心のある遺伝子は、対象の個人ゲノムに基づいて選択され得る。いくつかの態様では、関心のある遺伝子は既知の非メンデル遺伝子である。いくつかの態様では、関心のある遺伝子は、関心のある遺伝子変異体である。いくつかの態様では、関心のある遺伝子は独立に、観察される表現型と統計学的に有意に関連付けられていない。いくつかの態様では、関心のある遺伝子は、観察される表現型と関連付けられることが知られている。

【００２８】

データセット選択

リスクを判定するためのデータセットは、当技術分野で公知の何れかの手段により得られ得る。例えば、第１のデータセットは、対象に対する、及びまた対象の１人以上の血縁者に対する、遺伝子型データ及び表現型データを含み得る。遺伝子型データは、１つ以上の関心のある遺伝子に対する発現データを含み得る。表現型データは、疾患の特定の症状又は何れの疾患とも関連づけられない対象の観察可能な特徴を含め、観察可能な特徴又は疾患の形質を含み得る。

【００２９】

第１のデータセットは、対象及び対象の１人以上の血縁者において１つ以上の関心のある遺伝子の発現を検出することによって、用意され得る。いくつかの態様では、対象からの及び対象の１人以上の血縁者からの遺伝子型データ及び／又は表現型データは、複数の供給源から獲得される。

【００３０】

いくつかの態様では、第１のデータセットは、対象及び／又は血縁者の年齢に関する情報をさらに含む。いくつかの態様では、第１のデータセットは、対象及び／又は対象の血縁者における、表現型（例えば疾患若しくは状態又は疾患若しくは状態が関連付けられる特定の症状）の開始年齢に関する情報を含む。

【００３１】

いくつかの態様では、対象は特定の表現型を有する。いくつかの態様では、対象はその表現型を持たない。いくつかの態様では、対象は１つ以上の関心のある遺伝子を保有する。いくつかの態様では、対象は関心のある遺伝子を保有しない。いくつかの態様では、対象の１人以上の血縁者は関心のある遺伝子の１つ以上を保有し、対象でも観察される表現型を示す。いくつかの態様では、対象の血縁者の１人以上が関心のある遺伝子の１つ以上を保有し、対象で観察されない表現型を示す。いくつかの態様では、対象の血縁者の１人以上は、関心のある遺伝子の１つ以上を保有し、対象でも観察される表現型を示す。いくつかの態様では、対象の血縁者の１人以上は、関心のある遺伝子の１つ以上を保有せず、

対象で観察されない表現型を示す。

【００３２】

遺伝子型集団データ及び表現型集団データを有する第２のデータセットが使用され得る。非メンデル遺伝子に対するこのような集団データは、表現型を発現する対象の確率を決定するために使用され得る。いくつかの態様では、集団データは、２人以上の血縁者からのデータを含む。いくつかの態様では、集団データは、２人以上の血縁者の１つ以上のセット、例えば血縁者の、２セット、３セット、４セット、５セット、１０セット以上からのデータを含む。血縁者間の関係は、第１のデータセット中の対象と血縁者との間の関連と同じであり得るか、異なり得るか又は重複し得る。いくつかの態様では、集団データからの２人以上の血縁者は、第１のデータセットに対して使用される対象に対する血縁者ではない。いくつかの態様では、第２のデータセットに対するデータは、１つ以上の公開データベースから収集される。このようなデータベースの非限定例としては、Ｕｎｉｔｅｄ Ｋｉｎｇｄｏｍ（ＵＫ）Ｂｉｏｂａｎｋ；Ｎａｔｉｏｎａｌ Ｃｅｎｔｅｒ ｆｏｒ Ｂｉｏｔｅｃｈｎｏｌｏｇｙ Ｉｎｆｏｒｍａｔｉｏｎ（ＮＣＢＩ）により維持されるＤａｔａｂａｓｅ ｏｆ Ｇｅｎｏｔｙｐｅ ａｎｄ Ｐｈｅｎｏｔｙｐｅ（ｄｂＧａＰ）の一部である様々な遺伝子型－表現型データセット；Ｔｈｅ Ｅｕｒｏｐｅａｎ Ｇｅｎｏｍｅ－ｐｈｅｎｏｍｅ Ａｒｃｈｉｖｅ；ＯＭＩＭ；ＧＷＡＳｄｂ；ＰｈｅＧｅｎｌ；Ｇｅｎｅｔｉｃ Ａｓｓｏｃｉａｔｉｏｎ Ｄａｔａｂａｓｅ（ＧＡＤ）；及びＰｈｅｎｏｍｉｃＤＢが挙げられ得る。

【００３３】

データセットは、様々な組織又は体液の１つ以上からのデータを使用して収集され得る。例えば、第１及び／又は第２のデータセットは独立に、脳組織、心臓組織、肺組織、腎臓組織、肝臓組織、筋肉組織、骨組織、胃組織、腸組織、食道組織及び／又は皮膚組織又はこのような組織のあらゆる組み合わせと関連するデータを含み得る。さらに又は或いは、データセットは、体液、例えば尿、血液、血漿、血清、唾液、精液、痰、脳脊髄液、粘液、汗、硝子体液及び／又は乳汁又はこのような体液のあらゆる組み合わせなどと関連するデータを含み得る。

【００３４】

いくつかの態様では、データセットは、特定の状態及び／又は特定の症状を有する対象からのデータを使用して収集される。いくつかの態様では、データセットは、複数の組織及び／又は複数の体液からの試料を使用して収集される。

【００３５】

表現型リスクスコア

いくつかの態様は、対象に対する表現型リスクスコアを決定することを含む。表現型リスクスコアは、対象が特定の表現型（例えば疾患若しくは状態又は疾患若しくは状態の症状）を発現する見込みを示し得る。多遺伝子リスクスコアは、機械学習（教師あり及び／又は教師なし機械学習アルゴリズムを含む）を使用して決定され得る。いくつかの態様では、多遺伝子リスクスコアは、（例えば対象及び対象の１人以上の血縁者に対する遺伝子型データ及び表現型データを有する）第１のデータセット及び（例えば遺伝子型集団データ及び表現型集団データを有する）第２のデータセットにおいてモデルをトレーニングすることにより計算され得る。いくつかの態様では、トレーニングは、正規化（例えばハウスキーピング遺伝子の発現レベルに対して関心のある遺伝子の転写発現レベルを正規化する）及び／又は標準化段階（例えばゼロ平均に対して転写物の存在量をスケーリングするためにＳＶＭを介する）を含む。

【００３６】

いくつかの態様では、表現型リスクスコアは、オーバーサンプリング又はアンダーサンプリングなど、リサンプリング技術を使用して決定される。いくつかの態様は、ビニング及び／又はバギング技術を使用することを含む。いくつかの態様では、パラメトリック及び／又は非パラメトリック統計学的検定を使用して、対象間の発現の相違を評価する。

【００３７】

いくつかの態様では、表現型のリスクがあるものとして対象を分類するために表現型リスクスコアが使用され得る。分類は、例えばＳＶＭ、ロジスティック回帰分析、ランダムフォレスト、ナイーブベイズ及び／又はアダブーストを使用して行われ得る。いくつかの態様では、表現型リスクスコアは、対象が表現型を発現する確率である。いくつかの態様では、表現型リスクスコアは、特定の年齢までに対象が表現型を発現する確率である。

【００３８】

いくつかの態様では、表現型リスクスコアは、曲線下面積（ＡＵＣ）測定を使用して決定される。例えばＡＵＣは、約０．５超、約０．５５超、約０．６超、約０．６５超、約０．７超、約０．７５超、約０．８超、約０．８５超、約０．９超、約０．９５超、約０．９７超、約０．９８超又は約０．９９超であり得る。

【００３９】

実行システム

本明細書中で記載される方法は、様々なシステム上で実行され得る。例えば、いくつかの態様では、表現型リスクスコアを決定するためのシステムは、メモリに結合される１つ以上のプロセッサを含む。本方法は、１つ以上の電子デバイス上で格納され実行されるコード及びデータを使用して実行され得る。このような電子デバイスは、非一時的なコンピュータ可読記憶媒体などのコンピュータ可読媒体（例えば磁気ディスク；光学ディスク；ランダムアクセスメモリ；読み取り専用メモリ；フラッシュメモリ装置；相変化メモリ）及び一時的コンピュータ可読伝送媒体（例えば電気的、光学的、音響的又は他の形態の伝播シグナル－搬送波、赤外線シグナル、デジタルシグナルなど）を使用して、コード及びデータを（内部に及び／又はネットワーク上の他の電子デバイスで）格納し、通信し得る。

【００４０】

メモリには、表現型リスクスコアを決定するためにモデルをトレーニングするためのコンピュータ命令が搭載され得る。いくつかの態様では、このシステムは、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、スーパーコンピュータ、超並列コンピュータプラットフォーム、テレビ、メインフレーム、サーバファーム、広範囲分散型の一連の緩やかにネットワーク化されたコンピュータ又は何らかの他のデータ処理システム又はユーザーデバイスなど、コンピュータ上で実行される。

【００４１】

本方法は、ハードウェア（例えば回路網、専用論理など）、ファームウェア、ソフトウェア（例えば非一時的コンピュータ可読媒体上で統合）又は両者の組み合わせを含むロジックを処理することによって行われ得る。あらゆる順序で又は平行して、記載される演算が行われ得る。

【００４２】

一般に、プロセッサは、読み取り専用メモリ又はランダムアクセスメモリ又は両方から命令及びデータを受信し得る。コンピュータは一般に、命令に従いアクションを行い得るプロセッサ及び命令及びデータを格納するための１つ以上の記憶装置を含有する。一般に、コンピュータは、そこからデータを受け取るか又はそこにデータを伝達するか又は両方のために、データを格納するための１つ以上の大容量記憶装置、例えば磁気ディスク、光磁気ディスク、光学ディスク又はソリッドステートドライブも含むか又はそれに操作可能に結合される。しかし、コンピュータは、このようなデバイスを有する必要はない。さらに、コンピュータは、別のデバイス、数例挙げると、例えばスマートフォン、携帯型オーディオ又はメディアプレーヤー、ゲームコンソール、全地球測位システム（ＧＰＳ）受信機又はポータブル記憶装置（例えばユニバーサル・シリアル・バス（ＵＳＢ）フラッシュドライブ）に埋め込まれ得る。コンピュータプログラム命令及びデータを格納するのに適切な装置としては、例として半導体記憶装置、例えば、ＥＰＲＯＭ、ＥＥＰＲＯＭ及びフラッシュメモリ装置；磁気ディスク、例えば内部ハードディスク又はリムーバブルディスク；光磁気ディスク；及びＣＤ　ＲＯＭ及びＤＶＤ－ＲＯＭディスクを含む、不揮発性メモリ、媒体及び記憶装置の全ての形態が挙げられる。プロセッサ及びメモリは、特別な目

的の理論回路により供給され得るか又はそれに組み込まれ得る。
【００４３】
１つ以上のコンピュータのシステムは、演算においてアクションを実行させるか又はシステムがアクションを実行するようにするシステム上にインストールされる、ソフトウェア、ファームウェア、ハードウェア又はそれらの組み合わせを有することにより、特定の演算又はアクションを行うために構成され得る。１つ以上のコンピュータプログラムは、データ処理装置により実行される際にその装置がアクションを実施するようにする命令を含むことによって特定の演算又はアクションを行うために構成され得る。
【００４４】
代表的な実行システムを図２で示す。このようなシステムは、本明細書中に記載の演算の１つ以上を実施するために使用され得る。計算装置は、ＬＡＮ、イントラネット、エクストラネット及び／又はインターネットで他の計算装置と連結され得る。計算装置は、クライアントサーバーネットワーク環境でのサーバーマシンの容量において又はピアトゥピアネットワーク環境でのクライアントの容量において動作し得る。
【００４５】
診断及び処置
いくつかの態様では、対象（例えばヒト対象）は、表現型リスクスコアに基づき、状態若しくは疾患を有するものとして、又は状態若しくは疾患を有するリスクがあるものとして診断される。例えば、いくつかの態様では、特定の表現型リスクスコアを有する対象は、状態又は疾患を有すると診断される。いくつかの態様では、特定の表現型リスクスコアを有する対象は、状態若しくは疾患又は１つ以上のその症状を発現するリスクが上昇していると判定される。
【００４６】
いくつかの態様は、状態若しくは疾患又は疾患若しくは状態の１つ以上の症状を有するか又はそのリスクが上昇していると判定された対象を処置することを含む。「処置する（ｔｒｅａｔ）」という用語は、（１）疾患若しくは状態の発症若しくは進行を遅延させるか若しくは予防すること；（２）疾患若しくは状態の症状の進行、増悪若しくは悪化を遅延若しくは停止させること；（３）疾患若しくは状態の症状を寛解させること；又は（４）疾患若しくは状態を治癒させることを目的とする方法又は過程を特徴付けるために本明細書中で使用される。疾患又は状態の開始後に処置が施され得る。或いは、予防的又は予防行為のために、疾患又は状態の発症前に処置が施され得る。この場合、「予防」という用語が使用される。いくつかの態様では、この処置は、その全体において参照により本明細書中に組み込まれるＦＤＡ’ｓ　Ｏｒａｎｇｅ　Ｂｏｏｋの最新バージョンで列挙される薬物製品を投与することを含む。代表的な状態及び処置も、それぞれがその全体において参照により本明細書中に組み込まれるＰＨＹＳＩＣＩＡＮＳ’　ＤＥＳＫ　ＲＥＦＥＲＥＮＣＥ（ＰＲＤ　Ｎｅｔｗｏｒｋ　７１ｓｔ　ｅｄ．２０１６）；及びＴＨＥ　ＭＥＲＣＫ　ＭＡＮＵＡＬ　ＯＦ　ＤＩＡＧＮＯＳＩＳ　ＡＮＤ　ＴＨＥＲＡＰＹ（Ｍｅｒｃｋ　２０ｔｈ　ｅｄ．２０１８）に記載されている。
【００４７】
次の実施例は、本発明を例示するために提供されるが、これらの実施例の具体的な条件又は詳細に本発明が限定されないことを理解されたい。
【００４８】
実施例
実施例１：親類の情報を使用したリスクの精密化
単純化した実施例として、遺伝子ｇ上の可能性のある突然変異ｍを考え、X_{gm}は、突然変異が存在するときに$X_{gm}=1$であり、突然変異がないときに$X_{gm}=0$である、バイナリ指標変数とする。効率性のために、X_{gm}は、交換可能に、突然変異、突然変異の遺伝子座を指すために、及びその遺伝子座に突然変異が存在するか否かの指標として、使用される。突然変異X_{gm}があるサブ集団において、表現型は、$P(X_{gm})=p_{gm}$の確率で生じる（この記号は続く実施例を通じて使用する）。一元配置p_{gm}は試験から測定さ

れ得：

【数１】

$$p_{gm} = \frac{N_{gm,affected}}{N_{gm,affected} + N_{gm,unaffected}}$$

（式中、$N_{gm,affected}$及び$N_{gm,unaffected}$は、それぞれ表現型を有する及び有しない、突然変異X_{gm}がある対象（例えばヒト）の数である）である。

【００４９】

この実施例に対して、X_{gm}の他に１つのみ他の突然変異が表現型に影響を与えることが知られており（例えば突然変異ｎ及び遺伝子ｈ、X_{hn}）、X_{hn}がX_{gm}と連鎖不平衡ではないと仮定されるゲノムにおいて未知の位置にあると仮定する。この実施例に対して、X_{hn}は、X_{gm}及びX_{hn}が突然変異している場合には対象が表現型を発現するが、X_{gm}又はX_{hn}のみが突然変異している場合には対象が表現型を発現しないという、スイッチのように働くと仮定する。母及び子がX_{gm}突然変異しており、母が表現型を有する場合、子のリスクは、p_{gm}としてサブ集団試験に基づいてリスクが決定される場合よりも、正確に予測され得る。この実施例に対して、突然変異X_{hn}は、複数コピーを有する父又は母からこの突然変異を受け取る確率を無視できるほど十分に稀であると仮定する。従って、子が母からX_{hn}突然変異を受け継ぐ５０％の見込みがあるので、子が表現型を発現する見込みはおよそ５０％である。この実施例に対して、表現型について一般的集団リスクが１％前後であり、突然変異X_{gm}がリスクを５０％上昇させる稀な突然変異であり、血縁者からのデータが含まれない突然変異X_{gm}を有する個体の場合、およそ１．５％までリスクを上昇させる、と仮定する。子がX_{gm}突然変異しており、母がX_{gm}突然変異していて、表現型を有することが分かっている場合、子のリスクはここで１．５％ではなく５０％となる。つまり、５０％の中程度のリスク上昇として考えても、X_{hn}がX_{gm}に対するスイッチとして働くという単純化したシナリオが与えられる場合、突然変異及び表現型を有する母の情報の影響は相当なものである。

【００５０】

X_{gm}と相互作用する突然変異全てが表現型に影響を与えること又はそれらの相互作用の機序について分からないというシナリオにおいて、血縁者が同じ突然変異及び関連付けられる表現型を有する場合、対象が表現型を発現する確率を経験的に推定するために、上で概説する概念を適用し得る。これは、特定の縁戚関係及び特定の突然変異又は遺伝子に対して固有のリスクを計算するために遺伝子型－表現型データベースから情報を抽出することを含む。対象は、血縁者ｒ（ｒは母、父、兄弟、姉妹、息子、娘、祖父、祖母、叔母、叔父、姪、甥、従姉妹、従兄弟などであり得る）と突然変異X_{gm}を共有すると仮定する。ここで、対象が、表現型が発現されると思われる年齢よりも前の年齢であると仮定すると、対象の生涯リスクは、対象の現在の年齢の影響について調整せずに考え得る（以下で論じるように別々に組み込まれ得る）。突然変異X_{gm}を有しており、突然変異X_{gm}及び表現型がある親類ｒを有し、及び、その人において表現型が発現するであろう場合は、死亡しているか又は表現型が発現しているであろう年齢であるかの何れかである、データベース$N_{gm,r}$中の者の人数を求める（そうすると完全な生涯リスクを計算し得るようになる）。次に、表現型により影響された$N_{gm,r}$からの人数$N_{gm,r,affected}$を求める。対象が表現型を発現する推定確率は次のとおりである：

【数２】

$$\hat{p}_{gm,r} = \frac{N_{gm,r,affected}}{N_{gm,r}}$$

【００５１】

実施例２－限定データの管理

二項分布の正規近似に対して－小さい数に対して正確な二項を使用し得る－
【数３】

$$\hat{p}_{gm,\ r}$$

の推定値の分散を求める：
【数４】

$$\hat{\sigma}_{gm,r}{}^{2} = \frac{\hat{p}_{gm,r}\left(1-\hat{p}_{gm,r}\right)}{N_{gm,r}}$$

ｐｇｍは、突然変異Ｘｇｍが与えられる場合、親類における情報とは独立に表現型を発現する確率を表す。十分な信頼性で、例えば２標準偏差分、ｐｇｍとは異なる場合、即ち
【数５】

$$\left|p_{gm}-\hat{p}_{gm,r}\right| > 2\hat{\sigma}_{gm,r}$$

である場合、
【数６】

$$\hat{p}_{gm,r}$$

を使用し得、
又は、ｐｇｍの経験的推定値も求められている場合：
【数７】

$$\hat{p}_{gm} = \frac{N_{gm,affected}}{N_{gm}},\ \ \hat{\sigma}_{gm}{}^{2} = \frac{\hat{p}_{gm}\left(1-\hat{p}_{gm}\right)}{N_{gm}}$$

である。次の基準：
【数８】

$$\left|\hat{p}_{gm}-\hat{p}_{gm,r}\right| > 2\sqrt{\hat{\sigma}_{gm}{}^{2}+\hat{\sigma}_{gm,r}{}^{2}}$$

を使用し得る
【００５２】
又は、控えめにする（ｃｏｎｓｅｒｖａｔｉｓｍ）目的でｐｇｍの方向でいくつかの標準偏差で、
【数９】

$$\hat{p}_{gm,r}$$

を調整し得る：例えば２－シグマ調整を使用して、
【数１０】

$$\hat{p}_{gm,r} > p_{gm}$$

である場合、
【数１１】

$$\hat{p}_{gm,r} \rightarrow \max\left(p_{gm,}\,\hat{p}_{gm,r}-2\hat{\sigma}_{gm,r}\right)$$

である。別のアプローチは、データベースを複数のサブデータベースに分割することであり、経験的に各サブデータベースについて、
【数１２】

$$\hat{p}_{gm,r}$$

を計算し、標本分散を算出することによって、
【数１３】

$$\hat{p}_{gm,r}$$

の推定値において分散を上に有界にする(upper-bounding)。
【００５３】
【数１４】

$$\hat{p}_{gm,r}$$

の計算において使用されない試験データベースも使用し得る。例えば、突然変異Ｘｇｍを有する、及び死亡している、試験データ中の全対象を特定し得る。次に、親類情報を組み込む、
【数１５】

$$\hat{p}_{gm,r}$$

がｐｇｍよりも正確な予測を提供するか否かを決定するために、トレーニングデータを使用して、これらの各対象に対して
【数１６】

$$\hat{p}_{gm,r}$$

を算出し、対象が表現型を発現したか否かに対して比較し得る。
【００５４】
実施例３：類似の親類関係の組み合わせ
別のアプローチは、男性及び女性の親類におけるデータを組み合わせることであり、Ｘ染色体上に存在し、Ｙ染色体上に存在しない遺伝子は、表現型の発現での影響が最小であると仮定する。
【００５５】
さらに、関心のある対象と同様の量の遺伝物質を共有する親類からの情報を組み合わせ得る。この場合、ｒは、対象と同量の遺伝情報を共有する親類の各群を表すこととする。各群ｒに対するカウントをプールする。即ち、上記と同様のアプローチを使用して、Ｎｇｍ，ｒはここで、突然変異Ｘｇｍを有しており、突然変異Ｘｇｍ及び表現型がある群ｒ中の親類を有する、データベース中の人数を表し；Ｎｇｍ，ｒ，ａｆｆｅｃｔｅｄはここで、影響される者の数を表す。例えばｒ＝１／２は、対象の遺伝学的情報の半分がある群－母、父、兄弟、姉妹、息子、娘；ｒ＝１／４は遺伝学的情報の４分の１がある群－祖父、祖母、異父又は異母兄弟、異父又は異母姉妹、叔母、叔父、姪、甥、男の孫、女の孫など；ｒ＝１／８は遺伝学的情報の８分の１がある群を表す。このアプローチにおいて、Ｘｇｍ及び表現型を有し、同じ親類群ｒにある親類を有する何れか２人の対象は、同じ
【数１７】

$$\hat{p}_{gm,r}$$

を有する。この同じアプローチは、親類が対象と同量の遺伝学的情報を共有し、群の他のメンバーと同じ性別であるか否かに従い、親類を群分けするために適用され得る。この場合、例えば対象の遺伝学的情報の１／４を有する群は、男性群：祖父、異父又は異母兄弟、叔父、甥、男の孫など及び女性群：祖母、異父又は異母姉妹、叔母、姪、女の孫など、に分けられる。ｒにより表される場合、親類の多くの異なる組み合わせ又はセットが使用され得、Ｘｇを有するセット中の親類の多くの異なるサブセットは、カウントＮｇｍ，ｒ中に対象を含むために、単に１つ以上というのではなく、表現型を有することを必要とされ得る。

【００５６】

実施例４：遺伝子レベル突然変異

別のアプローチは、切り離して各変異体を処理するのではなく、遺伝子レベルで突然変異の存在に対処することである。即ちＸｇが突然変異した遺伝子ｇに相当するものとし、これは例えば機能喪失など、機能遺伝子ｇにおいて同じ効果を有することが知られる突然変異Ｘｇｍ，ｍ＝１．．．Ｍ全てを組み込む。この場合、機能遺伝子ｇにおいて機能喪失突然変異を有する人及び遺伝子ｇにおいて機能喪失突然変異など、そのタイプの突然変異をまた有する群ｒ中の親類の数である、Ｎｇ，ｒをカウントし得る。次に遺伝子レベルでの確率を計算し得る：

【数１８】

$$\hat{p}_{g,r} = \frac{N_{g,r,affected}}{N_{g,r}},\quad \hat{\sigma}_{g,r}{}^{2} = \frac{\hat{p}_{g,r}(1-\hat{p}_{g,r})}{N_{g,r}}$$

【００５７】

実施例５：年齢の組み込み

別のアプローチは、データベース中の人々の年齢に対処し、Ｎｇｍ，ｒの算出において死亡している人を考慮することのみ必要であることを取り除く。突然変異ではなく遺伝子レベルで作業するので、Ｎｇｍ，ｒの代わりにＮｇ，ｒを計算し得る。

【００５８】

【数１９】

$$\hat{p}_{g,r}(A)$$

は、年齢Ａ、突然変異Ｘｇ、及び突然変異Ｘｇを有する親類ｒの対象が、現在のところ表現型を有していない場合に表現型を発現する確率の推定値とする。データの入手可能性に依存して、突然変異Ｘｇを有する親類が表現型を発現しているか又は発現するであろう要件を組み込んでもよいし又は組み込まなくてもよい。Ｎｇ，ｒ，Ａは、年齢Ａよりも長く生き、年齢Ａで表現型を有しなかった、突然変異Ｘｇ及びＸｇがある親類ｒを有する全対象とする。Ｎｇ，ｒ，Ａ，ａｆｆｅｃｔｅｄは年齢Ａ以降に表現型を発現したＮｇ，ｒ，Ａ対象の人数とする。

【数２０】

$$\hat{p}_{g,r}(A) = \frac{N_{g,r,A,affected}}{N_{g,r,A}},\quad \hat{\sigma}_{g,r}(A)^{2} = \frac{\hat{p}_{g,r}(A)\left(1-\hat{p}_{g,r}(A)\right)}{N_{g,r}}$$

【００５９】

基本的な概念を変更することなく、まだ表現型を発現していない対象に対してｐｇ，ｒ（Ａ）を近似するための多くの他の方法があることに注意されたい。例えば、限定データに対して、ｐｒ（Ａ）又はｐｇ（Ａ）を算出することによってｐｇ，ｒ（Ａ）を近似し得、即ち彼らが突然変異Ｘｇを有するか又は突然変異Ｘｇがある親類ｒを有することを必要とすることに基づいてデータベース中で対象をフィルタリングしない。

【００６０】

限定データによる別のアプローチは、それらが突然変異Ｘｇ又は親類ｒを有するか否か

とは独立に、表現型を発現したデータベース中の全員を考慮し、表現型が発現されたときのヒストグラムを算出することである。このような模擬例のヒストグラムを、平均発症年齢６０歳の表現型について図１のバーにおいて示す。年齢の関数としての表現型を発現する個人の累積確率が算出され得、赤で示され、これは、表現型を発現する集団内頻度であるｐに漸近し、この場合ｐ＝０．２である。ｐとは異なるリスクを有する個々の対象について、表現型が発現すると思われる年齢に対する相対確率は不変であると見積もり得る。その場合、推定生涯リスク

【数２１】

$$\hat{p}_{g,r}$$

の対象について、

【数２２】

$$\frac{\hat{p}_{g,r}}{p}$$

により、累積確率を単純にスケーリングし得る。この実施例において、対象に対する累積確率は、

【数２３】

$$\hat{p}_{g,r} = 0.4$$

に漸近する灰色の線で示される。近似仮定を使用して、これは、平均６０歳の基礎的な確率分布に対する累積確率分布のままである。年齢Ａの対象の場合、対象の生涯において対象がこれから蓄積する確率がさらにどの程度大きいかを決定することによって、

【数２４】

$$\hat{p}_{g,r}(A)$$

が求められ得、図面の例では、年齢Ａ＝４０で垂直線として示され、

【数２５】

$$\hat{p}_{g,r}(40) = 0.34$$

である。対象に対する年齢により調整される集団遺伝学及び疫学由来の他の仮定及び確率分布を用いて、基本的な概念を変化させることなく、このテーマにおける多くのバリエーションが可能である。

【００６１】

実施例６：複数の親類の効果の組み合わせ

別のアプローチは、対象が、変異体及び表現型がある複数の親類を有するという状況を含む。最も単純なアプローチは、上記と同じ方法を使用することであるが、データベースにおいて１人の親類のみを有する場合をカウントするのではなく、複数の親類の同じセットを有する全例をカウントし、この場合、親類は、対象と共通の遺伝学的データの同量を共有している、及び特定の性別であるなど、上記の群分けｒに関して分類される。例えば、性別により、並びに共通する遺伝学的情報の量により群分けする場合、全てが変異体及び疾患を有する１人の父、１人の叔父及び１人の祖父を有する対象は、例えば変異体及び疾患を有する２人の息子及び１人の叔父を有する対象と一緒にカウントされ得る。別の例として、共通する遺伝学的情報の量によってのみ群分けする場合、全てが変異体及び疾患を有する１人の父、１人の叔母及び１人の祖母を有する対象は、例えば変異体及び疾患を有する２人の息子及び１人の叔父を有する対象と一緒にカウントされ得る。

【００６２】
限定データの場合、リスクを近似し得、この結果、一般的に、変異体及び疾患を有する対象の親類の一部を無視することにより下界が得られ、より多くのデータがプールされ得るようになる。この場合、一般的にはより多くの遺伝学的情報を対象と共有する親類を優先する。例えば、全てが変異体及び疾患を有する１人の父、１人の叔父及び１人の祖父を有する対象は、変異体及び疾患を有する唯一の親類、父、を有する対象として処理され得る。
【００６３】
別のアプローチは、親類のいくつかのカテゴリーにわたりデータを組み合わせることである。この概念に対して多くの経験的又はヒューリスティックなアプローチがある。例えば、ある代表的なアプローチは、Ｘｇの浸透率に影響する遺伝子の数が非常に多く、これらの各遺伝子の個々の効果量が非常に小さい場合に適切である。
【数２６】
$\Delta\hat{p}_{g,r}$

は、親類から関連がある突然変異遺伝子の全てを受け継ぐ場合の、確立された確率ｐｇからの差を表すものとする。ここで、確率の変化が、受け継がれる関連突然変異遺伝子の数に対して比例して増減するという、非常に単純化した及び不正確な（ｎｏｎ－ａｃｃｕｒａｔｅ）仮定を作り得る。
【数２７】
$\hat{p}_{g,r} - p_g = r\Delta\hat{p}_{g,r}$

（式中、各親類群に対して上で記載するように、ｒ＝１／２、１／４、１／８．．．である）。
【００６４】
次に、各親類群に対する一連の式を使用して、
【化１】

$\Delta\hat{p}_{g,r}$

を解き得、これは各群の個々の分散：
【数２８】
$$\Delta\hat{p}_{g,r} = \frac{\sum_{r=\frac{1}{2},\frac{1}{4},\frac{1}{8}\cdots} \frac{r}{\hat{\sigma}_{g,r}^{\,2}}\left(\hat{p}_{g,r} - p_g\right)}{\sum_{r=\frac{1}{2},\frac{1}{4},\frac{1}{8}\cdots} \frac{r^2}{\hat{\sigma}_{g,r}^{\,2}}}$$

により重み付けされ得る。次に、
【数２９】
$\Delta\hat{p}_{g,r}$

及び既知のｐｇを使用して、
【数３０】
$\hat{p}_{g,r}$

を推定し得る。

【００６５】

実施例７：多遺伝子リスクスコアに対する本方法の適用

表現型を発現する確率を述べる多遺伝子リスクスコア又は回帰モデルという観点から、又は表現型の確率を決定するための他の機械学習モデルにおいて、上記の技術が使用され得る。例えば、突然変異又は遺伝子レベルで以下の多遺伝子又は多変量、回帰モデルに基づいて表現型をモデル化し得る：

【数３１】

$$P = b_0 + \sum_{g=1\ldots G} \sum_{m=1\ldots M_g} b_{gm} X_{gm}$$
$$= b_0 + \sum_{g=1\ldots G} b_g X_g$$

【００６６】

既に記載のように、遺伝子レベルで指標変数X_gを仮定し、機能喪失又は特定のタイプの機能獲得などの類似タイプの全ての突然変異X_{gm}を合わせる。遺伝子が突然変異を有する場合は$X_g = 1$であり、ない場合は$X_g = 0$である。この同じ概念は、機能喪失又は機能獲得突然変異の異なるクラスなどの突然変異の異なる分類に拡大され得る。

【００６７】

以下の例は普遍性を失うことなく突然変異レベルで有効である。上記などの回帰モデルは、本明細書中で概説する方法を使用して、特定の個体に対して誘導される確率に基づいて調整され得る。Ｐが、本来は確率ではないが、対象の遺伝学的リスクスコアがどのパーセンタイルに入るかを決定するためなど、他のスコアとの関連において意義を有する多遺伝子リスクスコア（ＰＲＳ）である場合を考える。この場合、各遺伝子又は変異体の効果量に対してバイアスパラメータ$b_0 = 0$などを設定し得る。この効果量b_{gm}は、突然変異X_{gm}あり及びなしで、疾患表現型Ｄを発現する確率の比率の対数をとることによって推定され得る。

【数３２】

$$b_{gm} = \log\left(\frac{P(D|X_{gm})}{P(D|\overline{X_{gm}})}\right)$$

Ｐ（Ｄ｜X_{gm}）は、突然変異が与えられる場合の疾患の確率であり、上で計算した確率

【数３３】

$$P(D|X_{gm}) = \hat{p}_{gm}$$

により近似される。

【数３４】

$$P(D|\overline{X_{gm}})$$

を計算するために、展開：

【数３５】

$$P(D) = P(D|X_{gm})P(X_{gm}) + P(D|\overline{X_{gm}})P(\overline{X_{gm}})$$

を使用する。

【数３６】

$$P(\overline{X_{gm}}) = 1 - P(X_{gm})$$

を置き換え、上のものに

【数３７】

$$P(D|\overline{X_{gm}})$$

に代入し、

【数３８】

$$b_{gm} = log\left(\frac{P(D|X_{gm})\left(1 - P(X_{gm})\right)}{P(D) - P(D|X_{gm})P(X_{gm})}\right)$$

$$b_{gm} = \log\left(\frac{\hat{p}_{gm}\left(1 - P(X_{gm})\right)}{P(D) - \hat{p}_{gm}P(X_{gm})}\right)$$

を得る（式中、Ｐ（Ｘｇｍ）は、集団における突然変異の度数であり、Ｐ（Ｄ）は、ｐとして以前に定義された、集団における表現型の度数である。明確にするために、ここでＰ（Ｄ）を使用する。１つのアプローチは、オッズ比の対数に対してモデルパラメータを設定することである。集団において突然変異が稀である場合、即ちＰ（Ｘｇｍ）が小さい場合、これは、

【数３９】

$$b_{gm} \approx log\left(\frac{\hat{p}_{gm}}{P(D)}\right) = log\left(\frac{\hat{p}_{gm}}{p}\right)$$

に単純化し、これが実際によく使用されるものである。

【数４０】

$$\hat{p}_{gm}$$

がｐに近い場合、一般的な場合のように、特定の変異体Ｘｇｍ効果量が小さく、

【数４１】

$$b_{gm} \approx \frac{\hat{p}_{gm}}{p} - 1$$

を使用し得る。

【００６８】

関心のある個体が、影響される親類ｒを有することが分かっているとき、影響される親類ｒが与えられる場合、表現型を発現する確率であるｐｒに対して効果量を使用して、これを考慮に入れるためにパラメータを変化させ得る。

【数４２】

$$b_{gm,r} \approx log\left(\frac{\hat{p}_{gm,r}}{p_r}\right) - 1$$

（式中、

【数４３】

$$\hat{p}_{gm,r}$$

は上記のとおりである）。発明者らは、これらのパラメータがなぜｐではなくｐｒに対して定められるか及びこのアプローチの長所が何であるかを以下に記載する。しかし、この概念の多くのバリエーションがあることにまず注意することである。例えば、発明者らは、それらの分散の逆数によりパラメータに重み付けし得る：

【数４４】

$$var(b_{gm,r}) \approx \frac{1}{{p_r}^2} var(\hat{p}_{gm,r}) = \frac{{\hat{\sigma}_{gm,r}}^2}{{p_r}^2}$$

従って、

【数４５】

$$b_{gm,r,weighted} = \frac{{p_r}^2}{{\hat{\sigma}_{gm,r}}^2}\left(\frac{\hat{p}_{gm,r}}{p_r} - 1\right)$$

である。

【００６９】

なぜパラメータがｐではなくｐｒに対して定められるかを理解するために、多遺伝子モデルが、複数の遺伝学的変数から得られる表現型の確率をモデル化しようとしていることを考える。ここで次のように３つの遺伝学的変数Ｘ１、Ｘ２、Ｘ３があると仮定する。

【数４６】

$$P(D|X_1X_2X_3) = \frac{P(DX_1X_2X_3)}{P(X_1X_2X_3)} = \frac{P(X_1|DX_2X_3)P(DX_2X_3)}{P(X_1X_2X_3)} = \frac{P(X_1|DX_2X_3)P(DX_2X_3)}{P(X_1X_2X_3)}$$

しかし、Ｘ１、Ｘ２及びＸ３　がほぼ独立であると仮定する場合、

【数４７】

$$P(X_1|DX_2X_3) \approx P(X_1|D)\text{及び}P(X_1X_2X_3) \approx P(X_1)P(X_2)P(X_3)$$

故に

【数４８】

$$P(D|X_1X_2X_3) \approx \frac{P(X_1|D)P(DX_2X_3)}{P(X_1)P(X_2)P(X_3)}$$

となり、ここでＰ（ＤＸ２Ｘ３）は、独立仮定ゆえに分解され得る。

【数４９】

$$P(DX_2X_3) \approx P(X_2|DX_3)P(DX_3) \approx \frac{P(X_2|D)P(DX_3)}{P(X_1)P(X_2)P(X_3)} = \frac{P(X_2|D)P(X_3|D)P(D)}{P(X_1)P(X_2)P(X_3)}$$

【数５０】

$$P(D|X_1X_2X_3) = \frac{P(X_1|D)P(X_2|D)P(X_3|D)P(D)}{P(X_1)P(X_2)P(X_3)}$$

において置き換える。ここでベイズの規則Ｐ（Ｘ１｜Ｄ）／Ｐ（Ｘ１）＝Ｐ（Ｄ｜Ｘ１）／Ｐ（Ｄ）を適用する：

【数５１】

$$P(D|X_1X_2X_3) \approx P(D)\frac{P(D|X_1)P(D|X_2)P(D|X_3)}{P(D)P(D)P(D)}$$

この引数（ａｒｇｕｍｅｎｔ）は、変数Ｘ$_1$．．．Ｘ$_G$の何れの数にも適用され得る。これらの独立変数は、遺伝学的表現型である必要がないだけでなく、ライフスタイル又は他の表現型でもあり得ることにも注意すべきである。

【数５２】

$$P(D|X_1 \ldots X_G) \approx P(D)\frac{P(D|X_1)P(D|X_2) \ldots P(D|X_G)}{P(D)P(D) \ldots P(D)}$$

$$logP(D|X_1 \ldots X_G) \approx logP(D) + log\frac{P(D|X_1)}{P(D)} + \cdots log\frac{P(D|X_G)}{P(D)}$$

【００７０】

ｌｏｇＰ（Ｄ｜Ｘ$_1$．．．Ｘ$_G$）を算出するための上の説明は、ｌｏｇＰ（Ｄ｜Ｘ$_1$．．．Ｘ$_G$）を推定するために、各ＳＮＰに対する対数オッズ比又はそれに対する近似を合計する、多遺伝子予測モデルの背後の導出及び概念を概説する。

【数５３】

$$\frac{P(D|X_g)}{P(X_g)}$$

の形の因数のそれぞれは、多遺伝子リスクモデルにおける遺伝子座ｇに適用されるオッズ比の使用に対して理論的背景を与える。Ｘ$_g$＝１である場合、ベースライン集団確率Ｐ（Ｄ）は

【数５４】

$$\frac{P(D|X_g = 1)}{P(D)}$$

によりスケーリングされるが、Ｘｇ＝０の場合、Ｐ（Ｄ）は、

【数５５】

$$\frac{P(D|X_g = 0)}{P(D)}$$

によりスケーリングされる。これは、上記のように多くのＰＲＳモデルで行われることと同様であり、効果量ｂ$_g$：

【数５６】

$$b_g = log\left(\frac{P(D|X_g = 1)}{P(D|X_g = 0)}\right)$$

を算出し、次いで、個々の遺伝学的データに従い効果量を合計することによってＰＲＳスコアを算出する：

【数５７】

$$PRS = \sum\nolimits_{g=1 \ldots G} b_g X_g$$

【００７１】
X_g＝１である場合、上記のように
【数５８】

$$\frac{P(D|X_g = 1)}{P(D)}$$

によりスケーリングされるのではなく、ｌｏｇＰ（Ｄ｜X_g＝１）加算及びｌｏｇＰ（Ｄ｜X_g＝０）減算の両方を行っている。一般的には疾患の確率を直接推論するためにＰＲＳを使用しないので、これらの２つのシナリオ間の相違は一般に、実際には有意ではない。むしろ、対象は一般的には、それらのＰＲＳに基づきビンへとバケット化され、各ビンは、実際に疾患を有するビン中の個体の割合をカウントすることに基づいて個別に特定のリスクで特徴付けられる。言い換えると、マッピング－通常は線形マッピング－は一般的に、ＰＲＳと疾患を有する個体の実際のリスクとの間で作成される。結果的に、ＰＲＳの算出に適用される、スケーリングに関する問題の何れも又は効果量の増加は、重要ではない。

【００７２】
ＰＲＳ又はＰ（Ｄ｜X_1．．．X_g）の推定の目的は、対象に対する疾患又は表現型の確率を可能な限り厳密に再現すること及び疾患の異なる確率を有する対象間で可能な限り徹底的な区別を行うことである。親類情報の使用の価値を示すために、下の説明及び以下で論じるＭＡＴＬＡＢシミュレーションコードにおいて、より理論的な確率の式を使用し得る。即ち、以下の説明は、一般的に行われるように、変数X_rにおいて得られる親類情報を組み込む疾患確率を推定する有効性と、親類情報を使用することなくＰ（Ｄ｜X_1．．．X_g）を推定する有効性を比較する。

【００７３】
上記Ｐ（Ｄ｜X_1．．．X_g）を推定するための導出において、変数X_1．．．X_gの独立性に関する強い仮定に基づいていくつかの近似を行った。ここで、X_r変数は、親類又は親類のセットが関心のある疾患又は表現型を有するか否かを表すものとする。この変数は一般的にX_1．．．X_Gと独立ではない。例えばこれらが遺伝学的変数である場合、影響される親類の存在は、対象が遺伝子を有する確率又はX_1＝１，．．．，X_G＝１である確率にかなりの影響を及ぼす。しかし、集団平均に対するリスク、Ｐ（Ｄ）を計算する代わりに、関心のある疾患又は表現型を有する確率に対するリスクを代わりに計算する場合、疾患又は表現型を有する親類のセットＰ（Ｄ｜X_r）が与えられると、変数X_1．．．X_Gを超える状況において独立仮説を拡張することなく、家族歴に含有される情報を利用して、より強力な多遺伝子予測モデルを作成し得る。X_rが与えられるとき、X_1、X_2及びX_3の間で同様の独立仮説を使用して、X_r及びX_1X_2．．．X_3の間の依存性を無視する必要なく、Ｐ（Ｄ｜$X_1X_2X_3$）に対して上記のように同じ導出引数（ｄｅｒｉｖａｔｉｏｎ　ａｒｇｕｍｅｎｔ）を使用して、リスクを計算し得る。
【数５９】

$$P\big((D|X_r)|X_1X_2X_3\big) = P(D|X_rX_1X_2X_3) \approx P(D|X_r)\frac{P(D|X_rX_1)}{P(D|X_r)}\frac{P(D|X_rX_2)}{P(D|X_r)}\frac{P(D|X_rX_3)}{P(D|X_r)}$$

【００７４】
同様に、遺伝学的、ライフスタイル、環境又は表現型変数X_1．．．X_Gのあらゆる数に対してこの方法を拡張し得る。これらの変数間の独立性を仮定し得る場合：
【数６０】

$$P\big((D|X_r)|X_1X_2 \ldots X_G\big) = P(D|X_rX_1X_2 \ldots X_G) \approx P(D|X_r)\frac{P(D|X_rX_1)}{P(D|X_r)}\frac{P(D|X_rX_2)}{P(D|X_r)}\ldots\frac{P(D|X_rX_G)}{P(D|X_r)}$$

である。

【００７５】

上で記載したことと同様に、あるアプローチは、次のように、ＰＲＳを作成して効果量ｂｇ，ｒを算出することである：

【数６１】

$$b_{g,r} = log\left(\frac{P(D|X_r X_g = 1)}{P(D|X_r X_g = 0)}\right)$$

式中、Ｐ（Ｄ｜ＸｒＸｇ＝１）及びＰ（Ｄ｜ＸｒＸｇ＝０）は、経験的データから算出される。次に、合計することによって、関連する影響される親類又は一連の影響される親類を有する人に対してＰＲＳスコアを算出する:

【数６２】

$$PRS_{X_r} = \sum_{g=1\ldots G} b_{g,r} X_g$$

【００７６】

続く説明は、独立であるものとして近似される３つの遺伝学的変数の場合に焦点を当てる。Ｐ（Ｄ｜Ｘ１Ｘ２Ｘ３）ではなくモデルＰ（Ｄ｜ＸｒＸ１Ｘ２Ｘ３）に対して親類Ｘｒからの利用可能なデータを使用する価値を説明するために、ＭＡＴＬＡＢシミュレーションを記載するが、これは、各個体に対する疾患の確率をモデル化するその能力において正確度がより低く、一般的にその結果、より多くの間違った結果、医療費増大、転帰の悪化などが起こる。続く説明は、ＰＲＳの代わりにＰＲＳＸｒを算出するために上の式を同様に利用し得るが、これは、より一層理論的にＰ（Ｄ｜Ｘ１Ｘ２Ｘ３Ｘｒ）の推定に基づいて使用する。

【００７７】

発明者らが、集団における個々の罹患率が１／２０及び１／５０である２つの遺伝子Ｘ１及びＸ２を有し、Ｘ２がＸ１に対するスイッチとして働き、対象が、Ｘ１＝１及びＸ２＝１の両方である場合、表現型を有するようになる例を考える。この実施例をより例示的にするために、これらだけが疾患を引き起こし得る因子ではなく、存在する場合に１００％浸透率で疾患を引き起こす別の遺伝子Ｘ３があるとさらに仮定する。さらに発明者らは、この概念の一般性を失うことなく、各対象に対して考慮される親類のセットがその両親だけである、即ち、何れかの親が疾患を有する場合はＸｒ＝１、両親のどちらも疾患がない場合はＸｒ＝０と仮定する。付録ＡにおけるＭＡＴＬＡＢコードは、このシナリオに適用される発明概念を実行する。モデルを作成し、モデルを試験するために、シミュレーションが同じデータを使用することに注意すること。これは、シミュレーションされる対象の数と比較して推定されているパラメータが非常に少ないからであり、そのため、新しい試験データを生成させる概ね同じ結果を得る。即ちこのＭＡＴＬＡＢで実施するための削減は、限定されるデータの効果に焦点を当てるのではなく、モデル化アプローチのそれぞれの多用途性又は上で記載される及びデータにおいて捕捉される疾患確率を正確に推定するためのモデルの能力に焦点を当てる。

【００７８】

図３Ａ及び３Ｂは、一般的な集団において遺伝子Ｘ３が１／１００の度数を有する場合の対象のそれぞれに対する、ｙ軸対数スケール上での、予測のヒストグラムを示し、関連遺伝子のサブセットのみがモデルにおいて利用可能である。即ち、図３Ａは、遺伝学的変数Ｘ１及びＸ２のみを使用するモデルを記載し、図３Ｂは遺伝学的変数Ｘ１及びＸ３のみを使用するモデルを記載する。このようなシナリオは、例えば多遺伝子モデルが遺伝子のサブセットにおいてある一定の関連するＳＮＰのみをカバーし、一方で他の関連遺伝子がそのモデル中に含まれない場合によくある。これが生じるのは、例えば、遺伝学的変数の効

果及び独立性の線形性を仮定するモデルにおいて、排除される遺伝学的変数が統計学的有意性に到達しないから、又は排除される遺伝子が、有意な効果を一緒に有するがＳＮＰ又は「単ヌクレオチドポリモルフィズム」として認識されるのに十分高い度数を有する何れか１つの共通変異体と関連しない、多くの稀な変異体により影響を受けるからである。両図面において、対象のそれぞれに対する真実、即ち、それぞれ１又は０として捉えられる、各対象が疾患を実際に発現したか否か、が含まれる。図３Ａは、Ｐ（Ｄ｜X_1X_2）及びＰ（Ｄ｜$X_rX_1X_2$）を推定することによって、そのデータのモデル化を例示する。図３Ｂは、Ｐ（Ｄ｜X_1X_3）及びＰ（Ｄ｜$X_rX_1X_3$）を推定することによって、そのデータのモデル化を例示する。大抵の場合、親類情報を含めることによって、モデルが正しい根本的な統計学的モデルをより厳密に捉え、真実をより正確に模倣することが可能になることが分かり得る。図３Ｃは、全ての遺伝学的変数、即ちX_1X_2及びX_3が含まれる場合の正確度を例示し、その結果、推定値Ｐ（Ｄ｜$X_1X_2X_3$）及びＰ（Ｄ｜$X_rX_1X_2X_3$）が得られる。図３ＣもＰ（X_3）＝１／１００と仮定する。

【００７９】

表１は、この実施例において、両親である親類X_rに関する情報あり及びなしで多遺伝子リスクモデルにおいて遺伝子の異なる組み合わせが使用される場合の、遺伝学的変数の異なる組み合わせを用いた、シミュレーションからのいくつかのモデルの二乗平均平方根誤差（ＲＭＳＥ）を記載する。

【表１】

表 1: RMSE 推定値

	推定値の二乗平均平方根誤差		
	P(X3)=1/100	P(X3)=1/500	P(X3)=1/2000
P(D|XrX1)	0.0769	0.0429	0.0330
P(D|X1)	0.1041	0.0536	0.0383
P(D|XrX1X2)	0.0769	0.0427	0.0317
P(D|X1X2)	0.1030	0.0486	0.0251
P(D|XrX1X3)	0.0313	0.0294	0.0288
P(D|X1X3)	0.0509	0.0686	0.0800
P(D|XrX1X2X3)	0.0312	0.0290	0.0279
P(D|X1X2X3)	0.0846	0.0853	0.0540

【００８０】

図３Ｃにより表される後者のケースにおいて、両親の疾患歴、即ちX_rの組み込みは、ＲＭＳＥを０．０８４６から０．０３１２へ変化させるか又は６３％低下させる。

【００８１】

図４Ａ～Ｃは、Ｐ（X_3）＝１／５００であることを除き、図３Ａ～３Ｃと同様の状況を表す。図５Ａ～Ｃは、Ｐ（X_3）＝１／２０００であることを除き、図３Ａ～３Ｃと同様の状況を表す。図３、４及び５で記載されるこれらのシナリオの全てに対するＲＭＳＥは、他のシナリオとともに表１で捉えられる。一般には、親類情報X_rの組み込みによって一般的に真実のデータとの一致において性能が向上するということに注意すること。

【００８２】

実施例８：表現型確率のモデル化に対する他のアプローチ

（リスクスコアそれ自身ではなく）表現型の確率をモデル化するとき、本明細書中に記載のアプローチを使用して、例えばロジスティック回帰分析に基づくアプローチを使用して、個体に対してパラメータを修正することもできる。遺伝子レベルでは、ロジスティック回帰分析モデルは：

【数６３】

$$P(D|X_r X_1 \ldots X_G) = \frac{1}{1 + exp\left(-b_0 - a_0 \sum_{g=1\ldots G} b_{g,r} X_g\right)}$$

であり得る。

【００８３】

パラメータａ$_0$及びｂ$_0$をデータにフィットさせ得る場合、ｂ$_g$を選択するために、上で概説される概念を使用している。

【００８４】

遺伝子又は変異体の非線形的組み合わせを使用してＰ（Ｄ｜Ｘ$_r$Ｘ$_1$．．．Ｘ$_G$）の推定に対して、同じ概念が適用され得る。ここで、再び一般性を失わずに、発明者らは変異体レベルではなく遺伝子において作業する。遺伝子間の相互作用を捉えたいと仮定し、２つの遺伝子相互作用に注目しているだけであると仮定する（データについて課題がある可能性があるものの、３つ以上の遺伝子相互作用に同じ概念が適用され得る）。２つの遺伝子Ｘ$_1$及びＸ$_2$の何らかの論理結合：

【数６４】

X_1X_2 $(X_1 AND X_2)$、$\overline{X_1}\,\overline{X_2}$、$X_1\overline{X_2}$及び$\overline{X_1}X_2$

から回帰モデルに対して独立変数を作成し得る。回帰モデルについては、独立変数のセットにおけるＸ$_1$及びＸ$_2$の存在は、Ｘ$_1$Ｘ$_2$及び

【数６５】

$\overline{X_1}\,\overline{X_2}$

などの独立変数のような２つの付加的な論理結合の使用のみを必要とすることを念頭に置くべきであるが、それは、

【数６６】

$X_1\overline{X_2}$又は$\overline{X_1}X_2$

などの他の結合の独立変数は、既に含まれる変数に対して線形に依存するからである。限定されるデータを用いて、例えば標準的方法を使用して最初に線形回帰モデルを確立し、次に、有意であることが見出される全ての遺伝子ｇ＝１．．．Ｇをまとめ、これらの遺伝子の非線形相互作用を記載することによって、遺伝子相互作用を探すモデルを作成し得る。他の機械学習法、例えば、主成分、サポートベクター機器、ニューラルネットワーク、ディープラーニングニューラルネットワーク及び遺伝学的変数をモデルＰ（Ｄ｜Ｘ$_r$Ｘ$_1$．．．Ｘ$_G$）と組み合わせるための他の関数も使用し得る。

【００８５】

付録A: MATLAB式

```
% rel_sim
% simulates training polygenic prediction using relative relationships
% simulation parameters
n = 1000000; % 1000000; % number of families
p_x1 = 1/20; %1/20; % P(X1) the probability of X1 variant in the general pop
```

```
ulation
p_x2 = 1/50; %1/50; % P(X2) the probability of X2 variant in the general pop
ulation
p_x3 = 1/2000; %1/100; %1/500; %1/2000; % P(X3) the probability of X3 var
iant in the general population
% setting up variables
% assume no denovo variants
% assume no homozygotes of variant in parents
% ph_x1 = min(roots([1 -2 p_x1])); % probability per homolog; comment out if
assume no homozygotes of variant in parents
% ph_x2 = min(roots([1 -2 p_x2])); % probability per homolog; comment out if
assume no homozygotes of variant in parents
% create parents
par1_vec_x1 = (rand(n,1) < p_x1); % 1 if have variant 0 if don't
par1_vec_x2 = (rand(n,1) < p_x2); % 1 if have variant 0 if don't
par1_vec_x3 = (rand(n,1) < p_x3); % 1 if have variant 0 if don't
par2_vec_x1 = (rand(n,1) < p_x1); % 1 if have variant 0 if don't
par2_vec_x2 = (rand(n,1) < p_x2); % 1 if have variant 0 if don't
par2_vec_x3 = (rand(n,1) < p_x3); % 1 if have variant 0 if don't
par1_vec_dis = (par1_vec_x1 & par1_vec_x2) | par1_vec_x3;
par2_vec_dis = (par2_vec_x1 & par2_vec_x2) | par2_vec_x3;
par_vec_dis = par1_vec_dis | par2_vec_dis;
% create children
p_inh_x1 = 0.5*par1_vec_x1 + 0.5*par2_vec_x1 - 0.25*par1_vec_x1.*par2_vec
_x1;
chi_vec_x1 = (rand(n,1) < p_inh_x1);
p_inh_x2 = 0.5*par1_vec_x2 + 0.5*par2_vec_x2 - 0.25*par1_vec_x2.*par2_vec
_x2;
chi_vec_x2 = (rand(n,1) < p_inh_x2);
p_inh_x3 = 0.5*par1_vec_x3 + 0.5*par2_vec_x3 - 0.25*par1_vec_x3.*par2_vec
_x3;
chi_vec_x3 = (rand(n,1) < p_inh_x3);
chi_vec_dis = (chi_vec_x1 & chi_vec_x2) | chi_vec_x3; % child gets sick if eit
her (x1 and x2) or x3
%%%% train model for phenotype using standard method: P(D/X1X2) = P(D)*P
(D/X1)/P(D)*P(D/X2)/P(D)*P(D/X3)/P(D)
% just using child data for now; can do this also for parents
p_dis_h = length(find(chi_vec_dis==1))/n
chi_vec_x1e1_ind = find(chi_vec_x1==1);
p_dis_x1e1_h=length( find(chi_vec_dis(chi_vec_x1e1_ind)==1) )/length(chi_v
ec_x1e1_ind);
chi_vec_x1e0_ind = find(chi_vec_x1==0);
p_dis_x1e0_h=length( find(chi_vec_dis(chi_vec_x1e0_ind)==1))/length(chi_ve
c_x1e0_ind);
chi_vec_x2e1_ind = find(chi_vec_x2==1);
p_dis_x2e1_h=length( find(chi_vec_dis(chi_vec_x2e1_ind)==1) )/length(chi_v
ec_x2e1_ind);
chi_vec_x2e0_ind = find(chi_vec_x2==0);
p_dis_x2e0_h=length( find(chi_vec_dis(chi_vec_x2e0_ind)==1) )/length(chi_v
ec_x2e0_ind);
```

```
chi_vec_x3e1_ind = find(chi_vec_x3==1);
p_dis_x3e1_h=length( find(chi_vec_dis(chi_vec_x3e1_ind)==1) )/length(chi_v
ec_x3e1_ind);
chi_vec_x3e0_ind = find(chi_vec_x3==0);
p_dis_x3e0_h=length( find(chi_vec_dis(chi_vec_x3e0_ind)==1) )/length(chi_v
ec_x3e0_ind);
% prediction on the training data
% can also implement this on test data
p_dis_x1_h = zeros(n,1);
p_dis_x1_h(chi_vec_x1e1_ind)=p_dis_x1e1_h;
p_dis_x1_h(chi_vec_x1e0_ind)=p_dis_x1e0_h;
p_dis_x2_h = zeros(n,1);
p_dis_x2_h(chi_vec_x2e1_ind)=p_dis_x2e1_h;
p_dis_x2_h(chi_vec_x2e0_ind)=p_dis_x2e0_h;
p_dis_x3_h = zeros(n,1);
p_dis_x3_h(chi_vec_x3e1_ind)=p_dis_x3e1_h;
p_dis_x3_h(chi_vec_x3e0_ind)=p_dis_x3e0_h;
% prediction using x1 and x2
p_dis_x1x2_h = p_dis_h*(p_dis_x1_h/p_dis_h).*(p_dis_x2_h/p_dis_h);
% prediction using x1 and x3
p_dis_x1x3_h = p_dis_h*(p_dis_x1_h/p_dis_h).*(p_dis_x3_h/p_dis_h);
% prediction using x1,x2 and x3
p_dis_x1x2x3_h=p_dis_h*(p_dis_x1_h/p_dis_h).*(p_dis_x2_h/p_dis_h).*(p_dis
_x3_h/p_dis_h);
%%%% train model for phenotype using relative method: P(D/Xr/X1X2) = P(D/
Xr) * P(D/XrX1)/P(D/Xr) * P(D/XrX2)/P(D/Xr)
% just using child data for now to train; can train and test also for parents
par_vec_dis_ind = find(par_vec_dis==1);
p_dis_xr_h=length( find(chi_vec_dis(par_vec_dis_ind)==1) )/length(par_vec_d
is_ind);
% computing P(D/XrX1) for all states
chi_vec_xre1_x1e1_ind = find(par_vec_dis==1 & chi_vec_x1==1);
p_dis_xre1_x1e1_h=length( find(chi_vec_dis(chi_vec_xre1_x1e1_ind)==1) )/le
ngth(chi_vec_xre1_x1e1_ind);
chi_vec_xre0_x1e1_ind = find(par_vec_dis==0 & chi_vec_x1==1);
p_dis_xre0_x1e1_h=length( find(chi_vec_dis(chi_vec_xre0_x1e1_ind)==1) )/le
ngth(chi_vec_xre0_x1e1_ind);
chi_vec_xre0_x1e0_ind = find(par_vec_dis==0 & chi_vec_x1==0);
p_dis_xre0_x1e0_h=length( find(chi_vec_dis(chi_vec_xre0_x1e0_ind)==1) )/le
ngth(chi_vec_xre0_x1e0_ind);
chi_vec_xre1_x1e0_ind = find(par_vec_dis==1 & chi_vec_x1==0);
p_dis_xre1_x1e0_h=length( find(chi_vec_dis(chi_vec_xre1_x1e0_ind)==1) )/le
ngth(chi_vec_xre1_x1e0_ind);
% computing P(D/XrX2) for all states
chi_vec_xre1_x2e1_ind = find(par_vec_dis==1 & chi_vec_x2==1);
p_dis_xre1_x2e1_h=length( find(chi_vec_dis(chi_vec_xre1_x2e1_ind)==1) )/le
ngth(chi_vec_xre1_x2e1_ind);
chi_vec_xre0_x2e1_ind = find(par_vec_dis==0 & chi_vec_x2==1);
p_dis_xre0_x2e1_h=length( find(chi_vec_dis(chi_vec_xre0_x2e1_ind)==1) )/le
ngth(chi_vec_xre0_x2e1_ind);
```

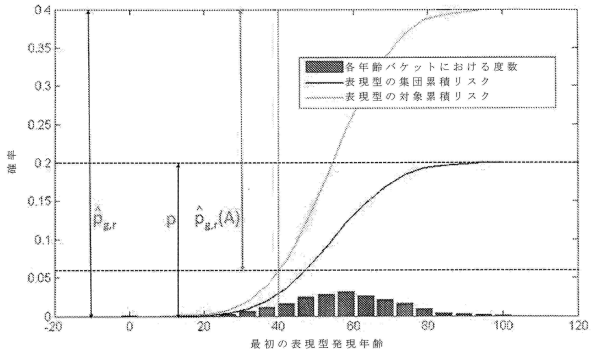
```
chi_vec_xre0_x2e0_ind = find(par_vec_dis==0 & chi_vec_x2==0);
p_dis_xre0_x2e0_h=length( find(chi_vec_dis(chi_vec_xre0_x2e0_ind)==1) )/le
ngth(chi_vec_xre0_x2e0_ind);
chi_vec_xre1_x2e0_ind = find(par_vec_dis==1 & chi_vec_x2==0);
p_dis_xre1_x2e0_h=length( find(chi_vec_dis(chi_vec_xre1_x2e0_ind)==1) )/le
ngth(chi_vec_xre1_x2e0_ind);
% computing P(D/XrX3) for all states
chi_vec_xre1_x3e1_ind = find(par_vec_dis==1 & chi_vec_x3==1);
p_dis_xre1_x3e1_h=length( find(chi_vec_dis(chi_vec_xre1_x3e1_ind)==1) )/le
ngth(chi_vec_xre1_x3e1_ind);
chi_vec_xre0_x3e1_ind = find(par_vec_dis==0 & chi_vec_x3==1);
p_dis_xre0_x3e1_h=length( find(chi_vec_dis(chi_vec_xre0_x3e1_ind)==1) )/le
ngth(chi_vec_xre0_x3e1_ind);
chi_vec_xre0_x3e0_ind = find(par_vec_dis==0 & chi_vec_x3==0);
p_dis_xre0_x3e0_h=length( find(chi_vec_dis(chi_vec_xre0_x3e0_ind)==1) )/le
ngth(chi_vec_xre0_x3e0_ind);
chi_vec_xre1_x3e0_ind = find(par_vec_dis==1 & chi_vec_x3==0);
p_dis_xre1_x3e0_h=length( find(chi_vec_dis(chi_vec_xre1_x3e0_ind)==1) )/le
ngth(chi_vec_xre1_x3e0_ind);
% prediction on the training data
% could also implement this on separate test data
% computing P(D/XrX1)
p_dis_xr_x1_h = zeros(n,1);
p_dis_xr_x1_h(chi_vec_xre1_x1e1_ind)=p_dis_xre1_x1e1_h;
p_dis_xr_x1_h(chi_vec_xre0_x1e1_ind)=p_dis_xre0_x1e1_h;
p_dis_xr_x1_h(chi_vec_xre0_x1e0_ind)=p_dis_xre0_x1e0_h;
p_dis_xr_x1_h(chi_vec_xre1_x1e0_ind)=p_dis_xre1_x1e0_h;
% computing P(D/XrX2)
p_dis_xr_x2_h = zeros(n,1);
p_dis_xr_x2_h(chi_vec_xre1_x2e1_ind)=p_dis_xre1_x2e1_h;
p_dis_xr_x2_h(chi_vec_xre0_x2e1_ind)=p_dis_xre0_x2e1_h;
p_dis_xr_x2_h(chi_vec_xre0_x2e0_ind)=p_dis_xre0_x2e0_h;
p_dis_xr_x2_h(chi_vec_xre1_x2e0_ind)=p_dis_xre1_x2e0_h;
% computing P(D/XrX3)
p_dis_xr_x3_h = zeros(n,1);
p_dis_xr_x3_h(chi_vec_xre1_x3e1_ind)=p_dis_xre1_x3e1_h;
p_dis_xr_x3_h(chi_vec_xre0_x3e1_ind)=p_dis_xre0_x3e1_h;
p_dis_xr_x3_h(chi_vec_xre0_x3e0_ind)=p_dis_xre0_x3e0_h;
p_dis_xr_x3_h(chi_vec_xre1_x3e0_ind)=p_dis_xre1_x3e0_h;
%%% computing key results
% prediction using xr, x1 and x2
p_dis_xrx1x2_h=p_dis_xr_h*(p_dis_xr_x1_h/p_dis_xr_h).*(p_dis_xr_x2_h/p_di
s_xr_h);
% prediction using xr, x1 and x3
p_dis_xrx1x3_h=p_dis_xr_h*(p_dis_xr_x1_h/p_dis_xr_h).*(p_dis_xr_x3_h/p_di
s_xr_h);
% prediction using xr, x1, x2 and x3
p_dis_xrx1x2x3_h=
p_dis_xr_h*(p_dis_xr_x1_h/p_dis_xr_h).*(p_dis_xr_x2_h/p_dis_xr_h).*(p_dis_
xr_x3_h/p_dis_xr_h);
```

```
%%% plotting key results
%%raw data
disp_vec = [1:10000];
% figure; plot(chi_vec_dis(disp_vec),'b.'); hold on; plot(chi_vec_dis(disp_vec),
'b');
%%prediction using xr, x1
% plot(p_dis_xr_x1_h(disp_vec),'gx');
% prediction using x1
% plot(p_dis_x1_h(disp_vec),'ro');
%%prediction using x1 and x2
% plot(p_dis_x1x2_h(disp_vec),'ro');
% prediction using xr, x1 and x2
% plot(p_dis_xrx1x2_h(disp_vec),'gx');
%%histograms using x1, x2 (and xr)
figure; hold on;
[t1,c1] = hist(chi_vec_dis); bar(c1, log10(t1),'b');
[t2,c2] = hist(p_dis_xrx1x2_h); bar(c2, log10(t2),'g');
[t3,c3] = hist(p_dis_x1x2_h); bar(c3, log10(t3),'r');
legend('Truth', 'Estimate of P(D ¦ XrX1X2)', 'Estimate of P(D ¦ X1X2)');
ylabel('log10(count)');
xlabel('probability estimate');
title('histogram of estimates P(D ¦ X1X2), P(D ¦ XrX1X2)');
grid;
%%prediction using x1 and x3
% plot(p_dis_x1x3_h,'ro');
% prediction using xr, x1 and x3
% plot(p_dis_xrx1x3_h,'gx');
% histograms using x1, x3 (and xr)
figure; hold on;
[tmp3,c3] = hist(p_dis_x1x3_h); bar(c3, log10(tmp3),'r');
[tmp1,c1] = hist(chi_vec_dis); bar(c1, log10(tmp1),'b');
[tmp2,c2] = hist(p_dis_xrx1x3_h); bar(c2, log10(tmp2),'g');
legend('Estimate of P(D ¦ X1X3)', 'Truth', 'Estimate of P(D ¦ XrX1X3)');
ylabel('log10(count)');
xlabel('probability estimate');
title('histogram of estimates P(D ¦ X1X3), P(D ¦ XrX1X3)');
grid;
%%prediction using x1, x2 and x3
% plot(p_dis_x1x2x3_h,'ro');
% prediction using xr, x1, x2 and x3
% plot(p_dis_xrx1x2x3_h,'gx');
% histograms using x1, x2, x3 (and xr)
figure; hold on;
[tm3,c3] = hist(p_dis_x1x2x3_h); bar(c3, log10(tm3),'r');
[tm2,c2] = hist(p_dis_xrx1x2x3_h); bar(c2, log10(tm2),'g');
[tm1,c1] = hist(chi_vec_dis); bar(c1, log10(tm1),'b');
legend('Estimate of P(D ¦ X1X2X3)','Estimate of P(D ¦ XrX1X2X3)','Truth');
ylabel('log10(count)');
xlabel('probability estimate');
title('histogram of estimates P(D ¦ X1X2X3), P(D ¦ XrX1X2X3)');
```

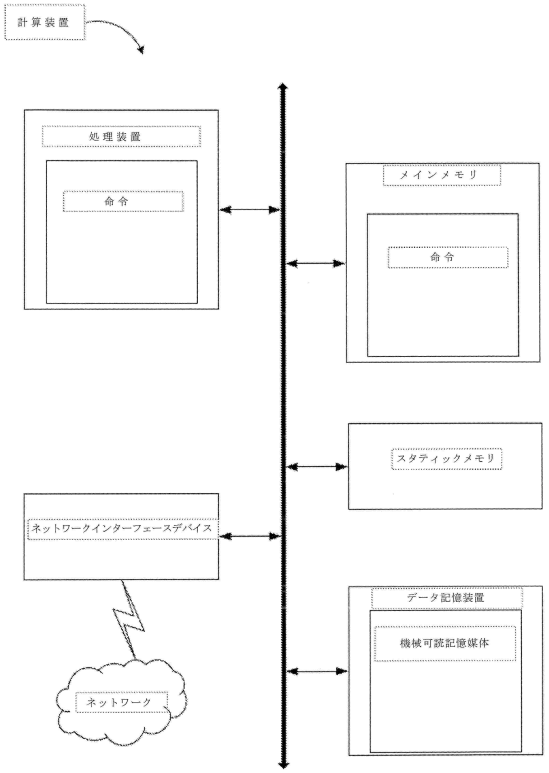
```
grid;
%%% comparing RMSE accuracy of results
% prediction using x1 (and xr)
p_dis_xr_x1_h_e = p_dis_xr_x1_h-chi_vec_dis;
p_dis_x1_h_e = p_dis_x1_h-chi_vec_dis;
p_dis_xr_x1_h_RMSE = sqrt(p_dis_xr_x1_h_e'*p_dis_xr_x1_h_e/n)
p_dis_x1_h_RMSE = sqrt(p_dis_x1_h_e'*p_dis_x1_h_e/n)
% prediction using x1 and x2 (and xr)
p_dis_xrx1x2_h_e = p_dis_xrx1x2_h-chi_vec_dis;
p_dis_x1x2_h_e = p_dis_x1x2_h-chi_vec_dis;
p_dis_xrx1x2_h_RMSE = sqrt(p_dis_xrx1x2_h_e'*p_dis_xrx1x2_h_e/n)
p_dis_x1x2_h_RMSE = sqrt(p_dis_x1x2_h_e'*p_dis_x1x2_h_e/n)
% prediction using x1, x3 (and xr)
p_dis_xrx1x3_h_e = p_dis_xrx1x3_h-chi_vec_dis;
p_dis_x1x3_h_e = p_dis_x1x3_h-chi_vec_dis;
p_dis_xrx1x3_h_RMSE = sqrt(p_dis_xrx1x3_h_e'*p_dis_xrx1x3_h_e/n)
p_dis_x1x3_h_RMSE = sqrt(p_dis_x1x3_h_e'*p_dis_x1x3_h_e/n)
% prediction using x1, x2, x3 (and xr)
p_dis_xrx1x2x3_h_e = p_dis_xrx1x2x3_h-chi_vec_dis;
p_dis_x1x2x3_h_e = p_dis_x1x2x3_h-chi_vec_dis;
p_dis_xrx1x2x3_h_RMSE = sqrt(p_dis_xrx1x2x3_h_e'*p_dis_xrx1x2x3_h_e/n)
p_dis_x1x2x3_h_RMSE = sqrt(p_dis_x1x2x3_h_e'*p_dis_x1x2x3_h_e/n)
```

【図面】

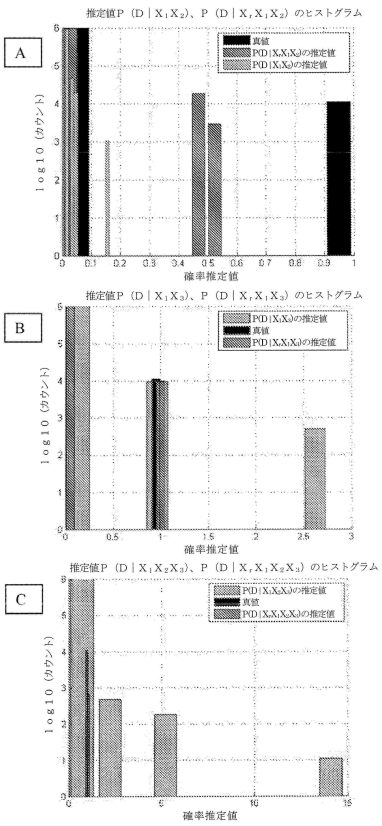
【図１】



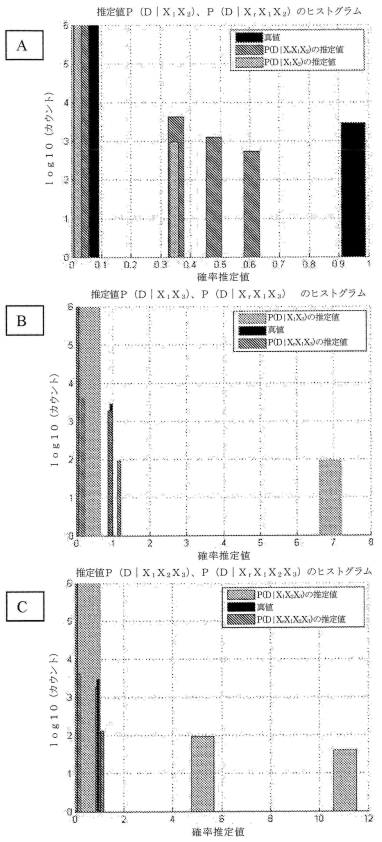
【図２】



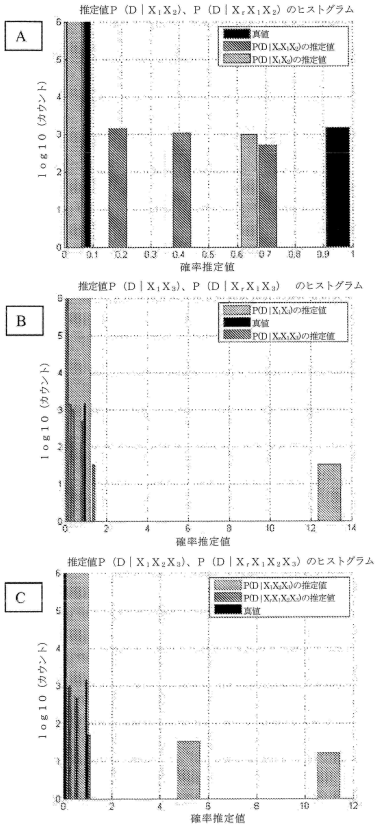
【図３】


推定値P（D｜X1X2）、P（D｜X2X1X2）のヒストグラム
A
真値
PD｜X1X2(X1X2)の推定値
PD｜X1X2(X2)の推定値
log10（カウント）
確率推定値
推定値P（D｜X1X3）、P（D｜X2X1X3）のヒストグラム
B
PD｜X1X2(X1X2)の推定値
真値
PD｜X1X2(X1X3)の推定値
log10（カウント）
確率推定値
推定値P（D｜X1X2X3）、P（D｜X2X1X2X3）のヒストグラム
C
PD｜X1X2(X1X2)の推定値
真値
PD｜X1X2(X2X3)の推定値
log10（カウント）
確率推定値

【図４】


推定値P（D｜X1X2）、P（D｜X2X1X2）のヒストグラム
A
真値
PD｜X1X2(X1X2)の推定値
PD｜X1X2(X2)の推定値
log10（カウント）
確率推定値
推定値P（D｜X1X3）、P（D｜X2X1X3）のヒストグラム
B
PD｜X1X2(X1X2)の推定値
真値
PD｜X1X2(X1X3)の推定値
log10（カウント）
確率推定値
推定値P（D｜X1X2X3）、P（D｜X2X1X2X3）のヒストグラム
C
PD｜X1X2(X1X2)の推定値
真値
PD｜X1X2(X2X3)の推定値
log10（カウント）
確率推定値

【図５】


推定値P（D｜X1X2）、P（D｜X2X1X2）のヒストグラム
A
真値
PD｜X1X2(X1X2)の推定値
PD｜X1X2(X2)の推定値
log10（カウント）
確率推定値
推定値P（D｜X1X3）、P（D｜X2X1X3）のヒストグラム
B
PD｜X1X2(X1X2)の推定値
真値
PD｜X1X2(X1X3)の推定値
log10（カウント）
確率推定値
推定値P（D｜X1X2X3）、P（D｜X2X1X2X3）のヒストグラム
C
PD｜X1X2(X1X2)の推定値
PD｜X1X2(X2X3)の推定値
真値
log10（カウント）
確率推定値

フロントページの続き

(72)発明者　ラビノウィッツ，マシュー
アメリカ合衆国，カリフォルニア州　９４０７０，サン　カルロス，スイート　４１０，２０１　インダストリアル　ロード

審査官　藤原　拓也

(56)参考文献　米国特許出願公開第２０１７／０１３７９６８（ＵＳ，Ａ１）
国際公開第２０１０／０３０９２９（WO，Ａ１）
国際公開第２００５／０３６４４３（WO，Ａ１）
米国特許第０５４６４７４２（ＵＳ，Ａ）
国際公開第２０１６／０６９７７１（WO，Ａ１）
米国特許出願公開第２０１６／０２８３４８４（ＵＳ，Ａ１）
LAUTENBACHほか，"Communicating Genetic Risk Information for Common Disorders in the Era of Genomic Medicine"，Annual Review of Genomic and Human Genetics，2013年08月31日，Vol. 14，P.491-513

(58)調査した分野　(Int.Cl.，ＤＢ名)
Ｇ１６Ｂ　５／００－９９／００