

(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7009516号
(P7009516)

(45)発行日 令和4年1月25日(2022.1.25)

(24)登録日 令和4年1月14日(2022.1.14)

(51)国際特許分類 F I
G 1 6 B 30/00 (2019.01) G 1 6 B 30/00

請求項の数 42 (全71頁)

(21)出願番号	特願2019-570089(P2019-570089)	(73)特許権者	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 9 2 1 2 2 サンディエゴ イルミナ ウェイ 5 2 0 0
(86)(22)出願日	平成30年6月19日(2018.6.19)	(74)代理人	100108453 弁理士 村山 靖彦
(65)公表番号	特表2020-530261(P2020-530261 A)	(74)代理人	100110364 弁理士 実広 信哉
(43)公表日	令和2年10月22日(2020.10.22)	(74)代理人	100133400 弁理士 阿部 達彦
(86)国際出願番号	PCT/US2018/038222	(72)発明者	コンラッド・シェフラー アメリカ合衆国・カリフォルニア・9 2 1 2 2・サン・ディエゴ・イルミナ・ウ エイ・5 2 0 0
(87)国際公開番号	WO2018/236827		
(87)国際公開日	平成30年12月27日(2018.12.27)		
審査請求日	令和1年12月18日(2019.12.18)		
(31)優先権主張番号	62/522,618		
(32)優先日	平成29年6月20日(2017.6.20)		
(33)優先権主張国・地域又は機関	米国(US)		

最終頁に続く

(54)【発明の名称】 未知の遺伝子型の寄与体からのDNA混合物の正確な計算による分解のための方法

(57)【特許請求の範囲】

【請求項1】

1つ又は複数のプロセッサとシステムメモリとを含むコンピュータシステムで実現される、1つ又は複数の寄与体の核酸を含む核酸試料を定量する方法であって、

(a) コンピュータシステムによって、核酸試料から得られ、1つ又は複数の多型座位にある1つ又は複数の対立遺伝子にマッピングされた核酸配列リードを受け取る工程と、

(b) 核酸配列リードを使用して、1つ又は複数のプロセッサによって、1つ又は複数の多型座位にある1つ又は複数の対立遺伝子のそれぞれに関する対立遺伝子数を決定する工程と、

(c) 確率的混合物モデルを対立遺伝子数に適用し、確率分布を使用して1つ又は複数の多型座位にある対立遺伝子数をモデル化する、確率的混合物モデルを使用する工程であって、確率分布が核酸配列リードの誤差を説明する工程と、

(d) 確率的混合物モデルを使用して、1つ又は複数のプロセッサによって、核酸試料中の1つ又は複数の寄与体の核酸の1つ又は複数の割合を定量する工程と、

(e) 1つ又は複数の寄与体の中の特定の寄与体が特定の遺伝子型を有する事後確率を決定する工程と、

(f) 事後確率に基づいて、核酸試料が特定の寄与体由来の核酸を含むことをコールする工程と

を含み、

確率分布が、下記の通り表される第1の二項分布を含む方法。

$n_{ij} \sim \text{BN}(n_i, p_{ij})$

式中、

n_{ij} は、座位 i での対立遺伝子 j に関する対立遺伝子数であり、

n_i は、座位 i での総対立遺伝子数であり、

p_{ij} は、座位 i での対立遺伝子 j の確率を示す確率パラメータである。

【請求項 2】

1つ又は複数の寄与体が2つ以上の寄与体を含む、請求項1に記載の方法。

【請求項 3】

1つ又は複数の寄与体中の寄与体の総数を決定する工程を更に含む、請求項1に記載の方法。

【請求項 4】

1つ又は複数の寄与体の1つ又は複数の遺伝子型が未知であった、請求項1に記載の方法。

【請求項 5】

1つ又は複数の多型座位のそれぞれにある1つ又は複数の対立遺伝子構成を決定する工程を更に含み、各対立遺伝子構成が、1つ又は複数の寄与体のそれぞれに関して2つ以上の対立遺伝子の対立遺伝子状態を含む、請求項4に記載の方法。

【請求項 6】

1つ又は複数の対立遺伝子構成に関して推定された確率を決定する工程を更に含む、請求項5に記載の方法。

【請求項 7】

1つ又は複数の寄与体の中の特定の寄与体が特定の遺伝子型を有する事後確率を得る工程が、

(i) 遺伝子型構成の事前確率に、遺伝子型構成の尤度を乗じる工程と、

(ii) (i)の積を、遺伝子型空間の合計によって正規化する工程と、

(iii) 特定の遺伝子型を含有する遺伝子型構成を合計して、事後確率を得る工程とを含む、請求項1に記載の方法。

【請求項 8】

特定の遺伝子型が多座位遺伝子型を含み、方法が、

全ての寄与体上で、寄与体が全ての座位にある特定の遺伝子型を有する事後確率を合計する工程と、

合計された確率に基づいて、特定された多座位遺伝子型が任意の寄与体中に現れることを決定する工程と

を更に含む、請求項1に記載の方法。

【請求項 9】

核酸試料が法医学的試料であり、多座位遺伝子型のデータが、対象となる人間から得られ、方法が、対象となる人間が核酸試料の寄与体であることを決定する工程を更に含む、請求項8に記載の方法。

【請求項 10】

核酸試料が、DNA分子及び/又はRNA分子を含む、請求項1に記載の方法。

【請求項 11】

核酸配列リードが、一意的分子インデックスを使用してDNA分子及び/又はRNA分子を配列決定することによって得られた、請求項10に記載の方法。

【請求項 12】

確率パラメータ p_{ij} が、

(i) 核酸試料中の1つ若しくは複数の寄与体の1つの核酸の割合、若しくは β 、

(ii) 1つ若しくは複数の寄与体の遺伝子型、若しくは G 、及び/又は

(iii) 核酸配列リードの誤差、若しくは θ

の関数である、請求項1に記載の方法。

【請求項 13】

確率的混合物モデルが、ベータ分布を使用して核酸配列リードの誤差をモデル化する、請求項12に記載の方法。

10

20

30

40

50

【請求項 14】

ベータ分布が、平均パラメータ、 μ 、及び濃度パラメータ、 k によって定義される、請求項13に記載の方法。

【請求項 15】

濃度パラメータは、種々のノイズ条件を表す事前確率を有し、濃度パラメータは、座位全体にわたって様々である、請求項14に記載の方法。

【請求項 16】

(c)が、第1の二項分布及びベータ分布を組み合わせて、ベータ二項分布に従う n_{ij} の周辺分布を得る工程を含む、請求項13に記載の方法。

【請求項 17】

ベータ二項分布が、下記の形

$$BB(n_{ij} | \eta, \mu, k)$$

を有する、請求項16に記載の方法。

【請求項 18】

(c)が、核酸配列リードの尤度関数を最大限にすることによって、核酸試料中の1つ又は複数の寄与体の核酸の1つ又は複数の割合を定量する工程を含む、請求項1に記載の方法。

【請求項 19】

(c)が、

複数の潜在的割合値、及び(b)で決定された対立遺伝子数の尤度関数を使用して、複数の尤度値を計算する工程と、

最大尤度値に関連付けられた潜在的割合ベクトルを特定する工程と、

特定された潜在的割合ベクトルを使用して、核酸試料中の1つ又は複数の寄与体の核酸の1つ又は複数の割合を定量する工程と

を含む、請求項18に記載の方法。

【請求項 20】

尤度関数が、母集団対立遺伝子頻度(π)が与えられた1つ又は複数の寄与体の遺伝子型的事前確率である $P(G | \pi)$ に依存する、請求項8に記載の方法。

【請求項 21】

事前確率 $P(G | \pi)$ が、ハーディ・ワインベルグ平衡を満たす周辺分布を使用して計算される、請求項20に記載の方法。

【請求項 22】

事前確率が、機構的ドロップアウトを表す固定された事前確率を持つダミー対立遺伝子を考慮して計算される、請求項20に記載の方法。

【請求項 23】

確率的混合物モデルが、第2の二項分布を使用して対立遺伝子データのスタッターエラーをモデル化する、請求項1に記載の方法。

【請求項 24】

第2の二項分布が、下記の通り表され、

$$s_{ik} \sim BN(n_i(k+1), r_i)$$

式中、

s_{ik} は、対立遺伝子 k であるように見えるが実際には対立遺伝子 $k+1$ のスタッターエラーから生じる、スタッター対立遺伝子の座位 i でのスタッター対立遺伝子数であり、

$n_i(k+1)$ は、座位 i での対立遺伝子 $k+1$ の当初の対立遺伝子数であり、

r_i は、座位 i に関するスタッター率である、

請求項23に記載の方法。

【請求項 25】

スタッター率 r が、座位全体にわたって様々であり、種々のノイズ条件を表す事前確率を有し、事前確率が座位全体にわたって共用される、請求項24に記載の方法。

【請求項 26】

(c)が、非スタッター対立遺伝子数の尤度とスタッター対立遺伝子数の尤度との積を含む尤

10

20

30

40

50

度関数を使用して、核酸試料中の1つ又は複数の寄与体の核酸の割合を定量する工程を含む、請求項24に記載の方法。

【請求項27】

(c)が、スタッターが潜在的に生ずる可能性のある分子の数を決定するときに、分子の固定数を、対立遺伝子 $k+1$ に割り当てられた対立遺伝子数に加える工程を含む、請求項24に記載の方法。

【請求項28】

確率的混合物モデルが、ダミー試料外対立遺伝子を使用して自然のドロップアウトをモデル化する、請求項1に記載の方法。

【請求項29】

ダミー試料外対立遺伝子の事前確率が、観察されていない対立遺伝子の数に比例する、請求項28に記載の方法。

【請求項30】

観察されていない対立遺伝子の数が、最短及び最長の観察された整数値対立遺伝子間の全ての整数を内挿し、任意の観察された非整数値対立遺伝子を付加し、得られた値及び閾値の最大値を戻すことによって推定される、請求項29に記載の方法。

【請求項31】

(c)が、核酸試料中の1つ又は複数の寄与体の核酸の割合を定量するのに使用されるデータから遺伝子型構成を刈り込む工程を含む、請求項1に記載の方法。

【請求項32】

遺伝子型構成を刈り込む工程が、必要とされる対立遺伝子のリストを構築することによって妥当とされる遺伝子型構成を限定する工程と、必要とされる全ての対立遺伝子を説明するのに十分ではない寄与体を持つ座位を除外する工程とを含む、請求項31に記載の方法。

【請求項33】

必要とされる対立遺伝子のリストが、閾値よりも高く且つスタッタードロップインに起因して妥当とされるには高すぎる対立遺伝子数を有する対立遺伝子から本質的になる、請求項32に記載の方法。

【請求項34】

閾値が、(i)最大非スタッター対立遺伝子数、及び(ii)潜在的スタッタードナー対立遺伝子の数を乗じた値の合計である、請求項33に記載の方法。

【請求項35】

遺伝子型構成を刈り込む工程が、対立遺伝子データと予想される対立遺伝子数との間の不十分なマッチを有する遺伝子型構成を除去する工程を含む、請求項31に記載の方法。

【請求項36】

不十分なマッチを有する遺伝子型構成が、1つ又は複数の閾値よりも大きい二乗平均平方根誤差(RMSE)値を有する、請求項35に記載の方法。

【請求項37】

1つ又は複数の多型座位にある対立遺伝子が、単一ヌクレオチド多型(SNP)対立遺伝子及び/又は短鎖タンデム反復(STR)対立遺伝子を含む、請求項1に記載の方法。

【請求項38】

システムメモリと、

(a)核酸試料から得られ、1つ又は複数の多型座位にある1つ又は複数の対立遺伝子にマッピングされた核酸配列リードを受け取り、

(b)核酸配列リードを使用して、1つ又は複数の多型座位にある1つ又は複数の対立遺伝子のそれぞれに関する対立遺伝子数を決定し、

(c)確率的混合物モデルを対立遺伝子数に適用し、核酸配列リードの誤差を説明する確率分布を使用して1つ又は複数の多型座位にある対立遺伝子数をモデル化する、確率的混合物モデルを使用し、

10

20

30

40

50

(d)確率的混合物モデルを使用して、核酸試料中の1つ又は複数の寄与体の核酸の1つ又は複数の割合を定量し、

(e)1つ又は複数の寄与体の中の特定の寄与体が特定の遺伝子型を有する事後確率を決定し、

(f)事後確率に基づいて、核酸試料が特定の寄与体由来の核酸を含むことをコールするように構成される1つ又は複数のプロセッサとを含み、

確率分布が、下記の通り表される第1の二項分布を含むコンピュータシステム。

$n_{ij} \sim \text{BN}(n_i, p_{ij})$

式中、

n_{ij} は、座位*i*での対立遺伝子*j*に関する対立遺伝子数であり、

n_i は、座位*i*での総対立遺伝子数であり、

p_{ij} は、座位*i*での対立遺伝子*j*の確率を示す確率パラメータである。

【請求項 39】

核酸試料から核酸を抽出するためのツールを更に含む、請求項38に記載のシステム。

【請求項 40】

1つ又は複数のプロセッサが、1つ又は複数の寄与体中の寄与体の総数を決定するように更に構成される、請求項38に記載のシステム。

【請求項 41】

1つ又は複数のプロセッサが、1つ又は複数の多型座位のそれぞれにある対立遺伝子構成を決定するように更に構成され、対立遺伝子構成が、1つ又は複数の寄与体のそれぞれに関して2つ以上の対立遺伝子の対立遺伝子状態を含む、請求項38に記載のシステム。

【請求項 42】

コンピュータシステムの1つ又は複数のプロセッサによって実行されたときに、1つ又は複数の寄与体の核酸を含む核酸試料を定量する方法をコンピュータシステムで実現させるプログラムコードを記憶する非一過性コンピュータ可読媒体であって、前記プログラムコードが、

(a)核酸試料から得られ、1つ又は複数の多型座位にある1つ又は複数の対立遺伝子にマッピングされた核酸配列リードを受け取るためのコード、

(b)核酸配列リードを使用して、1つ又は複数の多型座位にある1つ又は複数の対立遺伝子のそれぞれに関する対立遺伝子数を決定するためのコード、

(c)確率的混合物モデルを対立遺伝子数に適用し、核酸配列リードの誤差を説明する確率分布を使用して1つ又は複数の多型座位にある対立遺伝子数をモデル化する、確率的混合物モデルを使用するためのコード、

(d)確率的混合物モデルを使用して、核酸試料中の1つ又は複数の寄与体の核酸の1つ又は複数の割合を定量するためのコード、

(e)1つ又は複数の寄与体の中の特定の寄与体が特定の遺伝子型を有する確率を決定するためのコード、及び

(f)事後確率に基づいて、核酸試料が特定の寄与体由来の核酸を含むことをコールするためのコード

を含み、

確率分布が、下記の通り表される第1の二項分布を含む、非一過性コンピュータ可読媒体。

$n_{ij} \sim \text{BN}(n_i, p_{ij})$

式中、

n_{ij} は、座位*i*での対立遺伝子*j*に関する対立遺伝子数であり、

n_i は、座位*i*での総対立遺伝子数であり、

p_{ij} は、座位*i*での対立遺伝子*j*の確率を示す確率パラメータである。

【発明の詳細な説明】

【技術分野】

【0001】

関連出願の相互参照

10

20

30

40

50

本出願は、参照によりその全体が本明細書に組み込まれている、2017年6月20日に出願した米国仮特許出願第62/552,618号の優先権を主張するものである。

【背景技術】

【0002】

密接に関係するゲノムの核酸(例えば、DNA又はRNA)混合物からの配列決定データは、研究並びに臨床の場で頻繁に見出されており、混合寄与体の定量は、当初のゲノムが未知の場合に非常に難しくなっていた。例えば、微生物学及びメタゲノミクスの文脈において、研究者等及び臨床医等は、環境試料中で同じ種の密接に関係する細菌株を定量する必要があると考えられる。法医学の場で、法執行機関の職員は、多数の個体のDNAを含有する血液試料からヒト個体を定量すると共に同定する必要があると考えられる。

10

【0003】

別の適用例は、次世代配列決定(NGS)連結型液体生検である。NGS-連結型液体生検は、様々な臨床の場で潜在的な適用例がある、新たに出現した診断戦略である。臓器又は組織移植の文脈において、NGS-連結型液体生検は、レシピエントの血液中の同種異系DNAの量を定量することによって同種異系移植片の健康をモニタするための、非侵襲的手法を提供する。一部の適用例では、ドナー及びレシピエントのゲノムが未知であり又は部分的に未知である。

【先行技術文献】

【特許文献】

【0004】

【文献】米国特許出願第61/552,374号

米国特許出願第15/130,668号

米国仮特許出願第62/447,851号

【非特許文献】

【0005】

【文献】Sambrook等、「Molecular Cloning: A Laboratory Manual」、第3版(Cold Spring Harbor)、[2001]

Ausubel等、「Current Protocols in Molecular Biology」[1987]

Fan等、Proc Natl Acad Sci 105:16266-16271 [2008]

Koide等、Prenatal Diagnosis 25:604-607 [2005]

Chen等、Nature Med. 2: 1033-1035 [1996]

Lo等、Lancet 350: 485-487 [1997]

Botezatu等、Clin Chem. 46: 1078-1084, 2000

Su等、J Mol. Diagn. 6: 101-107 [2004]

【発明の概要】

【発明が解決しようとする課題】

【0006】

本明細書に提示される一部の実現例は、未知の遺伝子型の2種以上の寄与体の核酸を含む核酸混合物試料のデコンボリューションのためにコンピュータで実現される、方法及びシステムを提供する。本開示の一態様は、異なるゲノムを有する2種以上の寄与体の核酸(例えば、DNA又はRNA)を含む核酸試料中の核酸割合を定量するための方法に関する。一部の實現例では、核酸混合物試料は、以下に記述される生体組織、細胞、末梢血、唾液、尿、及びその他の生体液を含む。一部の適用例では、核酸試料は、単一の寄与体のみの核酸を含み、本明細書に記述される實現例は、単一寄与体の核酸が試料中の核酸の100%を占めることを、決定することができる。したがって、以下の記述は一部の實現例における核酸混合物試料としての核酸試料に言及するが、試料は、寄与体の割合が100%又は1である単一寄与体の核酸を含むことができると理解される。当然ながら、方法は、2種以上の寄与体の核酸を含む試料を定量するために使用することもできる。

40

【0007】

本明細書で提供される様々な方法及びシステムは、確率的混合物モデル及びベイズ推定技

50

法を使用するアルゴリズム及びプロセスを実現するので、実施形態は、核酸(例えば、DNA又はRNA)混合物試料のデコンボリューションにおいて、従来の方法に勝る技術的改善を提供する。本明細書に記述される一部の實現例はDNA試料に言及するが、實現例は、RNA試料の分析にも適用可能であることが理解される。一部の實現例は、改善された分析感度及び特性を提供し、核酸混合物試料のより正確なデコンボリューション及び定量を提供する。一部の實現例は、寄与体割合の正確な定量又は寄与体遺伝子型の決定を可能にするには低過ぎる核酸量を持つ核酸混合物試料の、正確な分析を可能にする。

【0008】

一部の実施形態では、方法は、1つ又は複数のプロセッサ及びシステムメモリであって、2種以上の寄与体の核酸を含む核酸混合物試料をデコンボリューションするように構成されたものを含む、コンピュータシステムで實現される。

10

【課題を解決するための手段】

【0009】

一部の実施形態は、寄与体の核酸と少なくとも1種のその他の寄与体とを含む核酸混合物試料中の、寄与体の核酸の割合を定量するための方法を提供する。方法は:(a)コンピュータシステムによって、核酸試料から得られ、1つ又は複数の多型座位にある1つ又は複数の対立遺伝子にマッピングされた核酸配列リードを受け取る工程;(b)核酸配列リードを使用して、1つ又は複数のプロセッサによって、1つ又は複数の多型座位にある1つ又は複数の対立遺伝子のそれぞれに関する対立遺伝子数を決定する工程;(c)確率的混合物モデルを対立遺伝子数に適用し、確率分布を使用して1つ又は複数の多型座位で対立遺伝子数をモデル化する、確率的混合物モデルを使用する工程であって、確率分布が核酸配列リードの誤差を説明する工程;(d)確率的混合物モデルを使用して、1つ又は複数のプロセッサによって、核酸試料中の1つ又は複数の寄与体の核酸の1つ又は複数の割合を定量する工程;(e)1つ又は複数の寄与体の中の特定の寄与体が、特定の遺伝子型を有する確率を決定する工程;及び(f)事後確率に基づいて、核酸試料が特定の寄与体由来の核酸を含むことをコールする工程を含む。

20

【0010】

一部の實現例では、1つ又は複数の寄与体が、2つ以上の寄与体を含む。

【0011】

一部の實現例では、方法が、1つ又は複数の寄与体中の寄与体の総数を決定する工程を更に含む。

30

【0012】

一部の實現例では、1つ又は複数の寄与体の1つ又は複数の遺伝子型が未知であった。一部の實現例では、方法は更に、1つ又は複数の多型座位のそれぞれで、1つ又は複数の対立遺伝子構成を決定する工程であって、各対立遺伝子構成が、1つ又は複数の寄与体のそれぞれに関して2つ以上の対立遺伝子の対立遺伝子状態を含む工程を含む。一部の實現例では、方法は更に、1つ又は複数の対立遺伝子構成に関して推定確率を決定する工程を含む。

【0013】

一部の實現例では、1つ又は複数の寄与体の中の特定の寄与体が特定の遺伝子型を有する事後確率を得る工程は:(i)遺伝子型構成の事前確率に、遺伝子型構成の尤度を乗じる工程;(ii)(i)の積を遺伝子型空間の合計によって正規化する工程;及び(iii)特定の遺伝子型を含有する遺伝子型構成上で合計して、事後確率を得る工程を含む。

40

【0014】

一部の實現例では、特定の遺伝子型が多座位遺伝子型を含み、方法は:全ての寄与体にわたり、寄与体が全ての座位で特定の遺伝子型を有する事後確率を合計する工程と;合計した確率に基づいて、特定された多座位遺伝子型が任意の寄与体に現れることを決定する工程とを含む。一部の實現例では、核酸試料が法医学的試料であり、多座位遺伝子型のデータが、対象となる人間から得られ、方法は、対象となる人間が核酸試料の寄与体であることを決定する工程を更に含む。

【0015】

50

一部の実現例では、核酸試料は、DNA分子及び/又はRNA分子を含む。一部の実現例では、核酸配列リードは、一意的分子インデックスを使用したDNA分子及び/又はRNA分子の配列決定によって得られた。

【0016】

一部の実現例では、確率分布は、第1の二項分布を含む。一部の実現例では、第1の二項分布は下記の通り表される：

$$n_{ij} \sim \text{BN}(n_i, p_{ij})$$

n_{ij} は、座位*i*での対立遺伝子*j*の対立遺伝子数であり； n_i は、座位*i*での総対立遺伝子数であり； p_{ij} は、座位*i*での対立遺伝子*j*の確率を示す確率パラメータである。

【0017】

一部の実現例では、確率パラメータ p_{ij} は：(i)核酸試料中の1つ若しくは複数の寄与体の1つの核酸の割合、若しくは β ；(ii)1つ若しくは複数の寄与体の遺伝子型、若しくは*G*；及び/又は(iii)核酸配列リードの誤差、若しくは θ の関数である。

【0018】

一部の実施形態では、確率的混合物モデルは、ベータ分布を使用して核酸配列リードの誤差をモデル化する。一部の実現例では、ベータ分布は、平均パラメータ、 μ 、及び濃度パラメータ、 k によって定義される。一部の実現例では、濃度パラメータは、種々のノイズ条件を表す事前確率を有し、濃度パラメータは、座位全体にわたって様々である。

【0019】

一部の実現例では、(c)は、第1の二項分布とベータ分布とを組み合わせ、ベータ二項分布に従う n_{ij} の周辺分布を得る工程を含む。一部の実現例では、ベータ二項分布は下記の形を有する：

$$\text{BB}(n_{ij} \mid \mu, k)。$$

【0020】

一部の実現例では、(c)は、核酸配列リードの尤度関数を最大限にすることによって、核酸試料中の1つ又は複数の寄与体の核酸の1つ又は複数の割合を定量する工程を含む。一部の実現例では、(c)は：複数の潜在的割合値と、最大尤度値に関連付けられた潜在的割合ベクトルを特定する(b)で決定された対立遺伝子数の尤度関数とを使用して、複数の尤度値を計算する工程と；特定された潜在的割合ベクトルを使用して、核酸試料中の1つ又は複数の寄与体の核酸の1つ又は複数の割合を定量する工程とを含む。

【0021】

一部の実現例では、尤度関数は、母集団対立遺伝子頻度(π)が与えられた1つ又は複数の寄与体の遺伝子型の事前確率である $P(G \mid \pi)$ に依存する。一部の実現例では、事前確率 $P(G \mid \pi)$ は、ハーディ・ワインベルグ平衡を満たす周辺分布を使用して計算される。一部の実現例では、事前確率は、機構的ドロップアウトを表す固定された事前確率を持つダミー対立遺伝子を考慮して、計算される。一部の実現例では、確率的混合物モデルは、第2の二項分布を使用して対立遺伝子データでのスタッターエラーをモデル化する。一部の実現例では、第2の二項分布は下記の通り表される：

$$s_{ik} \sim \text{BN}(n_i(k+1), r_i)$$

s_{ik} は、対立遺伝子*k*であるように見えるが実際には対立遺伝子*k+1*のスタッターエラーから生じる、スタッター対立遺伝子の座位*i*でのスタッター対立遺伝子数であり； $n_i(k+1)$ は、座位*i*での対立遺伝子*k+1*の当初の対立遺伝子数であり； r_i は、座位*i*に関するスタッター率である。

【0022】

一部の実現例では、スタッター率*r*は、座位全体にわたって様々であり、種々のノイズ条件を表す事前確率を有し、この事前確率は、座位全体にわたって共用される。一部の実現例では、(d)は、非スタッター対立遺伝子数の尤度とスタッター対立遺伝子数の尤度との積を含む尤度関数を使用して、核酸試料中の1つ又は複数の寄与体の核酸の割合を定量する工程を含む。一部の実現例では、(c)は、スタッターが潜在的に生ずる可能性のある分子の数を決定するときに、分子の固定数を、対立遺伝子*k+1*に割り当てられた対立遺伝子数に加

10

20

30

40

50

える工程を含む。

【0023】

一部の実現例では、確率的混合物モデルは、ダミー試料外対立遺伝子を使用して自然ドロップアウトをモデル化する。一部の實現例では、ダミー試料外対立遺伝子の事前確率は、観察されていない対立遺伝子の数に比例する。一部の實現例では、観察されていない対立遺伝子の数は:最短及び最長の観察された整数値対立遺伝子の間の全ての整数を内挿し;任意の観察された非整数値対立遺伝子を付加し、得られた値の最大値及び閾値を戻すことによって推定される。

【0024】

一部の實現例では、(c)は、核酸試料中の1つ又は複数の寄与体の核酸の割合を定量するのに使用されるデータからの遺伝子型構成を、刈り込む工程を含む。一部の實現例では、遺伝子型構成を刈り込む工程は:必要とされる対立遺伝子のリストを構築することによって、妥当とされる遺伝子型構成を限定する工程と、必要とされる全ての対立遺伝子を説明するのに十分な寄与体を持たない座位を除外する工程とを含む。一部の實現例では、必要とされる対立遺伝子のリストは、本質的に、閾値よりも高く且つスタッタードロップインに起因して妥当とされるには高過ぎる対立遺伝子数を有する対立遺伝子からなる。一部の實現例では、閾値は、(i)最大非スタッター対立遺伝子数、及び(ii)潜在的スタッタードナー対立遺伝子の数で乗じた値の合計である。一部の實現例では、遺伝子型構成を刈り込む工程は:対立遺伝子データと予想される対立遺伝子数との間のマッチが不十分である遺伝子型構成を除去する工程を含む。一部の實現例では、マッチが不十分な遺伝子型構成は、1つ又は複数の閾値よりも大きい二乗平均平方根誤差(RMSE)を有する。

【0025】

一部の實現例では、1つ又は複数の多型座位の対立遺伝子は、単一ヌクレオチド多型(SNP)対立遺伝子及び/又は短鎖タンデム反復(STR)対立遺伝子を含む。

【0026】

開示される実施形態は、列挙された操作及び本明細書に記述されるその他の演算操作を行うためのプログラム命令が提供される非一過性コンピュータ可読媒体を含む、コンピュータプログラム製品も提供する。

【0027】

一部の実施形態は、寄与体及び少なくとも1つのその他の寄与体の核酸を含む核酸混合物試料中の寄与体の核酸の割合を定量するためのシステムを提供する。システムは、試料から核酸配列情報を提供する試験試料から核酸を受容するためのシーケンサ、プロセッサ;及び本明細書に列挙された方法を使用してプロセッサ上で実行するための命令をそこに記憶させる1つ又は複数のコンピュータ可読記憶媒体を含む。

【0028】

本開示の一態様は、システムメモリと、1つ又は複数のプロセッサとを含むコンピュータシステムを提供する。プロセッサは:(a)核酸試料から得られ、1つ又は複数の多型座位にある1つ又は複数の対立遺伝子にマッピングされた核酸配列リードを受け取り;(b)核酸配列リードを使用して、1つ又は複数の多型座位にある1つ又は複数の対立遺伝子のそれぞれに関する対立遺伝子数を決定し;(c)確率的混合物モデルを対立遺伝子数に適用し、核酸配列リードの誤差を説明する確率分布を使用して1つ又は複数の多型座位にある対立遺伝子数をモデル化する、確率的混合物モデルを使用し;(d)確率的混合物モデルを使用して、核酸試料中の1つ又は複数の寄与体の核酸の1つ又は複数の割合を定量し;(e)1つ又は複数の寄与体の中の特定の寄与体が特定の遺伝子型を有する事後確率を決定し;(f)事後確率に基づいて、核酸試料が、特定の寄与体由来の核酸を含むことをコールするように構成される。

【0029】

一部の實現例では、システムは更に、核酸試料から核酸を抽出するためのツールを含む。

【0030】

一部の實現例では、1つ又は複数のプロセッサは更に、1つ又は複数の寄与体中の寄与体の総数を決定するように構成される。

【0031】

一部の実現例では、1つ又は複数のプロセッサは更に、1つ又は複数の多型座位のそれぞれでの対立遺伝子構成を決定するように構成され、この対立遺伝子構成は、1つ又は複数の寄与体のそれぞれに関して2つ以上の対立遺伝子の対立遺伝子状態を含む。

【0032】

本開示の別の態様は、コンピュータシステムの1つ又は複数のプロセッサによって実行されたときに、1つ又は複数の寄与体の核酸を含む核酸試料を定量する方法をコンピュータシステムで実現させるプログラムコードを記憶する非一過性コンピュータ可読媒体を提供し、前記プログラムコードは:(a)核酸試料から得られ、1つ又は複数の多型座位にある1つ又は複数の対立遺伝子にマッピングされた核酸配列リードを受け取るためのコード;(b)核酸配列リードを使用して、1つ又は複数の多型座位にある1つ又は複数の対立遺伝子のそれぞれに関する対立遺伝子数を決定するためのコード;(c)確率的混合物モデルを対立遺伝子数に適用し、核酸配列リードの誤差を説明する確率分布を使用して1つ又は複数の多型座位にある対立遺伝子数をモデル化する、確率的混合物モデルを使用するためのコード;(d)確率的混合物モデルを使用して、核酸試料中の1つ又は複数の寄与体の核酸の1つ又は複数の割合を定量するためのコード;(e)1つ又は複数の寄与体の中の特定の寄与体が特定の遺伝子型を有する事後確率を決定するためのコード;及び(f)事後確率に基づいて、核酸試料が特定の寄与体由来の核酸を含むことをコールするコードを含む。

10

【0033】

本明細書の実施例はヒトに関し、言語は主にヒトの問題を対象とするが、本明細書に記述される概念は、任意の植物又は動物からのゲノムに適用可能である。本開示のこれら及びその他の目的及び特徴は、以下の記述及び添付される特許請求の範囲からより完全に明らかになり、又は以下に述べる本開示の実施によって学ぶことができる。

20

【0034】

参照による組込み

本明細書で言及される、全ての特許、特許出願、及びこれらの参考文献中に開示された全ての配列を含むその他の刊行物は、個々の刊行物、特許、又は特許出願のそれぞれが参照により組み込まれることをあたかも特別に且つ個々に示すのと同じ程度まで、参照により本明細書に明らかに組み込まれる。引用される全ての文献は、関連する部分が、本明細書のそれらの引用の文脈によって示される目的のために、それらの全体が参照により本明細書に組み込まれる。しかし、いかなる文献の引用も、本開示に関する先行技術であることを認めるものと解釈されるべきではない。

30

【図面の簡単な説明】

【0035】

【図1】 バイオインフォマティクスアルゴリズムの概要、及び寄与体のDNA定量のために設計された統計モデルを示す。

【図2A】 核酸試料中の1つ又は複数の寄与体の1つ又は複数の核酸画分(例えば、DNA又はRNA)を定量するプロセスを示すブロック図を示す。

【図2B】 確率的混合物モデルの様々な構成要素を示すブロック図を示す。

【図2C】 1つの対立遺伝子を別の対立遺伝子に変換し、真の対立遺伝子を予期しない対立遺伝子に変換する配列決定エラーを概略的に示す。

40

【図3】 1つ又は複数の寄与体の核酸を含む核酸試料を評価するプロセスを示すブロック図を示す。

【図4】 特定の実施形態による計算装置として機能し得る典型的なコンピュータシステムのブロック図を示す。

【図5】 試験試料からコール又は診断を生成するための分散システムの一実現例を示している。

【図6】 異なる位置で一部の実現例の様々な動作を実行するための選択肢を示している。

【図7A】 実際のDNA混合物試料から得られたデータを使用して、一部の実現例がDNA混合物試料を効果的に定量及びデコンボリューションし得るという実施例の結果を示す。

50

【図 7 B】実際の DNA 混合物試料から得られたデータを使用して、一部の實現例が DNA 混合物試料を効果的に定量及びデコンボリューションし得るという実施例の結果を示す。

【図 7 C】実際の DNA 混合物試料から得られたデータを使用して、一部の實現例が DNA 混合物試料を効果的に定量及びデコンボリューションし得るという実施例の結果を示す。

【図 7 D】実際の DNA 混合物試料から得られたデータを使用して、一部の實現例が DNA 混合物試料を効果的に定量及びデコンボリューションし得るという実施例の結果を示す。

【図 7 E】実際の DNA 混合物試料から得られたデータを使用して、一部の實現例が DNA 混合物試料を効果的に定量及びデコンボリューションし得るという実施例の結果を示す。

【図 7 F】実際の DNA 混合物試料から得られたデータを使用して、一部の實現例が DNA 混合物試料を効果的に定量及びデコンボリューションし得るという実施例の結果を示す。

10

【図 8 A】シミュレーションデータを使用して、一部の實現例が DNA 混合物試料を効果的に定量及びデコンボリューションし得るという実施例の結果を示している。

【図 8 B】シミュレーションデータを使用して、一部の實現例が DNA 混合物試料を効果的に定量及びデコンボリューションし得るという実施例の結果を示している。

【図 8 C】シミュレーションデータを使用して、一部の實現例が DNA 混合物試料を効果的に定量及びデコンボリューションし得るという実施例の結果を示している。

【図 8 D】シミュレーションデータを使用して、一部の實現例が DNA 混合物試料を効果的に定量及びデコンボリューションし得るという実施例の結果を示している。

【発明を実施するための形態】

【0036】

20

定義

他に指示しない限り、本明細書に開示される方法及びシステムの実施は、当業者の範囲内にある分子生物学、微生物学、タンパク質精製、タンパク質工学、タンパク質及び DNA 配列決定、及び組換え DNA の分野で一般に使用される従来の技法及び装置を含む。そのような技法及び装置は当業者に公知であり、数多くのテキスト及び参考資料に記載されている(例えば、Sambrook 等、「Molecular Cloning: A Laboratory Manual」、第3版(Cold Spring Harbor)、[2001]);及び Ausubel 等、「Current Protocols in Molecular Biology」[1987]参照)。

【0037】

数値範囲は、その範囲を画定する数値を含む。本明細書の全体を通して与えられる全ての最大数値限定は、全ての下限数値を、そのような下限数値が本明細書に明確に書かれたかのように含むことが意図される。本明細書の全体を通して与えられる全ての最小数値限定は、全ての上限数値を、そのような上限数値が本明細書に明確に書かれたかのように含むことになる。本明細書の全体を通して与えられる全ての数値範囲は、そのようなより広い数値範囲内に含まれる全てのより狭い数値範囲を、そのようなより狭い数値範囲が本明細書に全て明確に書かれたかのように含むことになる。

30

【0038】

本明細書に提示される見出しは、本開示を限定するものではない。

【0039】

他に本明細書で定義されない限り、本明細書で使用される全ての技術的及び科学的用語は、当業者により一般に理解されるのと同じ意味を有する。本明細書に含まれる用語を含む様々な科学辞書が周知であり、当業者が利用可能である。本明細書に記述されるものに類似の又は同等の任意の方法及び材料は、本明細書に開示される実施形態の実施又は試験での使用を見出す、いくつかの方法及び材料について記述する。

40

【0040】

すぐ下に定義される用語は、全体として本明細書を参照することによってより完全に記述される。本開示は、記述される特定の方法、プロトコル、及び試薬に限定されないことが理解されるが、それはこれらが当業者に使用される文脈に応じて様々であってもよいからである。本明細書で使用される単数形を示す用語「a」、「an」、及び「the」は、文脈が他に明示しない限り複数の指示対象を含む。

50

【0041】

他に指示しない限り、核酸は、左から右に5'から3'の向きに書かれ、アミノ酸配列は、左から右にアミノからカルボキシの向きに、それぞれ書かれる。

【0042】

ドナーDNA(dDNA)という用語は、移植片のドナーの細胞に由来したDNA分子を指す。様々な実現例において、dDNAは、ドナーから移植された組織/臓器を受け取るドナーから得られる試料に見出される。一部の實現例では、dDNAは、含む

【0043】

循環細胞のないDNA又は単に無細胞DNA(cfDNA)は、細胞内に閉じ込められていないDNA断片であり、血流又はその他の体液中で自由に循環している。cfDNAは種々の出所を有することが公知であり、ある場合には、腫瘍細胞又は腫瘍罹患細胞に由来し、その他の場合には、妊娠中の母親によって保持された胎児の胎児細胞に由来し、母体血中を循環する。一般に、cfDNAは断片化され、ゲノムの僅かな部分しか含まず、そのゲノムは、cfDNAが得られた生物のゲノムとは異なっているけれどもよいものである。

【0044】

非循環ゲノムDNA(gDNA)又は細胞DNAという用語は、細胞に閉じ込められたDNA分子を指すのに使用され、しばしば全ゲノムを含む。

【0045】

ベータ分布は、確率変数の指数として現れ且つ分布の形状を制御する、例えば α 及び β によって示される、2個のポジ形状パラメータによってパラメータ化された、間隔[0,1]で画定された連続確率分布のファミリーである。ベータ分布は、広く様々な分野において有限長の間隔に制限された確率変数の挙動をモデル化するのに適用されてきた。ベイズ推定において、ベータ分布は、ベルヌーイ、二項、負の二項、及びゲノム分布に関する共役事前確率分布である。例えばベータ分布は、ベイズ解析で使用して、成功の確率に関する初期知識について記述することができる。確率変数 X がベータ分布に従う場合、確率変数 X は、 $X \sim \text{ベータ}(\alpha, \beta)$ と書かれる。

【0046】

二項分布は、一連の n 個の独立した実験で成功した数の個別の確率分布であり、それぞれが二項選択を問い、それ自体のブール値の結果を持つ情報のシングルビットを含有する確率変数:正(確率 p を持つ)又は負(確率 $q=1-p$ を持つ)。単一試験、即ち $n=1$ では、二項分布がベルヌーイ分布である。二項分布は、サイズ N の集団からの置換えにより描かれたサイズ n の試料における成功の数をモデル化するのに頻繁に使用される。確率変数 X が、パラメータ $n \in \mathbb{N}$ 及び $p \in [0,1]$ を持つ二項分布に従う場合、確率変数 X は、 $X \sim B(n, p)$ と書かれる。

【0047】

本明細書で $\text{Pois}()$ と示されるポアソン分布は、時間及び/又は空間の固定された間隔で生ずる所与の数の事象が、公知の平均速度で且つ最後の事象からの時間とは無関係に生ずる場合に、それらの事象の確率を表す個別の確率分布である。ポアソン分布は、距離、面積、又は体積等、その他の特定された間隔における事象の数に関して使用することもできる。ポアソン分布に従う間隔で観察している k 事象の確率は、方程式:

【数1】

$$P(\text{間隔での } k \text{ 事象}) = \frac{\lambda^k e^{-\lambda}}{k!}$$

によって与えられ、式中、 λ は、間隔における事象の平均数又は事象率であり、速度パラメータとも呼ばれ、 e は2.71828、オイラー数であり、又は自然対数の底であり、 k は値0

、1、2、…をとり、 $k!$ は k の階乗である。

【0048】

ガンマ分布は、連続確率分布の2パラメータファミリーである。一般に使用される3つの異なるパラメータ化があり:形状パラメータ k 及びスケールパラメータ θ があり;形状パラメータ $\alpha=k$ 及び逆スケールパラメータ $\beta=1/\theta$ があり、速度パラメータと呼ばれ;又は形状パラメータ k 及び平均パラメータ $\mu=k/\beta$ がある。これら3つの形のそれぞれにおいて、両方のパラメータは正の実数である。ガンマ分布は、確率変数 X の最大エントロピー確率分布であり、 $E[X]=k\theta=\alpha/\beta$ は固定され且つゼロよりも大きく、 $E[\ln(X)]=\psi(k)+\ln(\theta)=\psi(\alpha)-\ln(\beta)$ は固定されている(ψ は、ディガンマ関数である)。

【0049】

多型及び遺伝子多型は、それぞれが認め得る頻度を持つ、1つのゲノム座位で2つ以上の対立遺伝子の同じ集団での出現を指すのに、本明細書では同義で使用される。

【0050】

多型部位及び多型性部位は、2つ以上の対立遺伝子が存在するゲノム上の座位を指すのに、本明細書では同義で使用される。

【0051】

対立遺伝子頻度又は遺伝子頻度は、遺伝子のその他の対立遺伝子に対する、遺伝子(又は遺伝子の変種)の対立遺伝子の頻度であり、割合又はパーセンテージとして表すことができる。対立遺伝子頻度は、遺伝子が1つ又は複数の座位にしばしば位置付けられるので、特定のゲノム座位にしばしば関連付けられる。しかし、本明細書で使用される対立遺伝子頻度は、DNA断片のサイズベースのビンに関連付けることもできる。この意味で、対立遺伝子を含むcfDNA等のDNA断片が、種々のサイズベースのビンに割り当てられる。その他の対立遺伝子の頻度に対する、サイズベースのビンにおける対立遺伝子の頻度は、対立遺伝子頻度である。

【0052】

本明細書の「パラメータ」という用語は、その値又はその他の特徴が試料又はDNA分子等の関連ある状態に影響を及ぼす物理的特徴等、系の性質を特徴付ける数値を指す。ある場合には、パラメータという用語は、数学的関係又はモデルの出力に影響を及ぼす変数に関して使用され、その変数は、独立変数(即ち、モデルへの入力)でも1つ又は複数の独立変数をベースにした中間変数であってもよい。モデルの範囲に応じて、1つのモデルの出力が別のモデルの入力になってもよく、それによってその他のモデルに対するパラメータになる。

【0053】

「複数の」という用語は、1つよりも多い要素を指す。

【0054】

「ペアエンドリード」という用語は、核酸断片の各端部から1つのリードを得るペアエンド配列決定のリードを指す。ペアエンド配列決定は、ポリヌクレオチドのストランドを、インサートと呼ばれる短い配列に断片化する工程を含んでもよい。断片化は、無細胞DNA分子等の比較的短いポリヌクレオチドに関しては任意選択であり又は不必要である。

【0055】

「ポリヌクレオチド」、「核酸」、及び「核酸分子」という用語は同義で使用され、ヌクレオチドの共有結合配列を指し(即ち、RNAに関するリボヌクレオチド及びDNAに関するデオキシリボヌクレオチド)、1つのヌクレオチドのペントースの3'位が、次のペントースの5'位にホスホジエステル基によって接合されている。ヌクレオチドは、RNA及びDNA分子、例えばcfDNA又は細胞DNA分子を含むがこれらに限定することのない、核酸の任意の形の配列を含む。「ポリヌクレオチド」という用語は、限定するものではないが一本鎖及び二本鎖ポリヌクレオチドを含む。

【0056】

本明細書の「試験試料」という用語は、核酸又は核酸の混合物を含む、生体液、細胞、組織、臓器、又は生物に典型的には由来する試料を指す。そのような試料は、痰/口内流体、

10

20

30

40

50

羊水、血液、血液画分、又は細針生検試料(例えば、外科生検、細針生検等)、尿、腹水、胸膜液、及び同様のものを含むがこれらに限定するものではない。試料はヒト対象(例えば、患者)からしばしば採取されるが、アッセイは、イヌ、ネコ、ウマ、ヤギ、ヒツジ、ウシ、ブタ等を含むがこれらに限定されない任意の哺乳動物からの試料で使うことができる。試料は、生物学的供給源から得られたままで直接、又は試料の特性が修正されるよう前処理した後で、使用されてもよい。例えばそのような前処理は、血液からの血漿を調製し、粘性流体を希釈し、及び同様の工程を含んでいてもよい。前処理の方法は、濾過、沈殿、希釈、蒸留、混合、遠心分離、凍結、凍結乾燥、濃縮、増幅、核酸断片化、妨害成分の不活性化、試薬の添加、溶解等を含んでいてもよいがこれらに限定するものではない。前処理のそのような方法を、試料に関連して用いる場合、そのような前処理方法は、典型的には、対象となる核酸が試験試料中に残されたままになるようにされ、時には、未処理の試験試料(例えば、即ち任意のそのような前処理方法に供されていない試料)における濃度に比例する濃度になるようにされる。そのような「処理された」又は「プロセッシングされた」試料は、依然として本明細書に記述される方法に関して生物学的「試験」試料と見なされる。

10

【0057】

本明細書の「次世代配列決定(NGS)」という用語は、クローン増幅分子の及び単一核酸分子の超並列配列決定を可能にする配列決定方法を指す。NGSの非限定的例は、可逆的ダイ・ターミネーターを使用した合成時配列決定(sequencing-by-synthesis)、及びライゲーション配列決定(sequencing-by-ligation)を含む。

20

【0058】

「リード」という用語は、核酸試料の一部から得られた配列を指す。典型的には、必ずしも必要ではないが、リードは、試料中で近接する塩基対の短い配列を表す。リードは、試料部分の塩基対配列によって(A、T、C、又はGで)記号により表されてもよい。リードは、メモリデバイスに記憶され且つ必要に応じて処理されて、参照配列にマッチするか否か又はその他の基準を満たすか否かを決定してもよい。リードは、配列決定装置から直接、又は試料に関して記憶された配列情報から間接的に得られてもよい。ある場合には、リードは、より大きい配列又は領域を特定するのに使用することができる、例えば染色体又はゲノム領域又は遺伝子にアラインメントし特異的に割り当てることができる、十分な長さ(例えば、少なくとも約25bp)のDNA配列である。

30

【0059】

「ゲノムリード」という用語は、個体の全ゲノムにおける任意のセグメントのリードを参照するのに使用される。

【0060】

本明細書で使われる「アラインメントされた」、「アラインメント」、又は「アラインメントする」という用語は、リード又はタグを参照配列と比較し、それによって参照配列がリード配列を含有するか否かを決定するプロセスを指す。参照配列がリードを含有する場合、リードは参照配列にマッピングされてもよく、又はある特定の形態では、参照配列の特定の場所にマッピングされてもよい。ある場合には、アラインメントは単に、リードが特定の参照配列のメンバーか否か(即ち、リードが参照配列中に存在するか又は不在であるか)を知らせる。例えば、ヒト13番染色体に関する参照配列に対するリードのアラインメントは、リードが13番染色体に関する参照配列中に存在するか否かを知らせることになる。この情報を提供するツールを、セットメンバーシップテスターと呼んでもよい。ある場合には、アラインメントは追加として、リード又はタグがマッピングされる、参照配列中の場所を示す。例えば、参照配列が全ヒトゲノム配列である場合、アラインメントは、リードが13番染色体上に存在することを示してもよく、リードが13番染色体の特定のストランド及び/又は部位上にあることを更に示してもよい。

40

【0061】

アラインメントされたリード又はタグは、参照ゲノムからの公知の配列に対する、それら核酸分子の順序のマッチとして特定される、1つ又は複数の配列である。アラインメント

50

は手作業で行うことができるが、本明細書に開示される方法を実現するのに妥当な期間でリードをアラインメントするのは不可能と考えられるので、典型的にはコンピュータアルゴリズムによって実現される。配列をアラインメントする工程からのアルゴリズムの一例は、Illumina Genomics Analysis pipelineの一部として配布されるEfficient Local Alignment of Nucleotide Data(ELAND)コンピュータプログラムである。或いは、ブルームフィルタ又は類似のセットメンバーシップテスターを用いて、リードを参照ゲノムにアラインメントしてもよい。参照によりその全体が本明細書に組み込まれる2011年10月27日出願の米国特許出願第61/552,374号を参照されたい。アラインメントにおける配列リードのマッチは、100%配列マッチ又は100%未満(不完全なマッチ)とすることができる。

10

【0062】

本明細書で使用される「マッピング」という用語は、配列リードをより大きい配列、例えば参照ゲノムに特異的に割り当ててを指し、より大きい配列の部分配列は、アラインメント又はメンバーシップアサイメントを使用する。

【0063】

本明細書で使用される「参照ゲノム」又は「参照配列」という用語は、任意の生物又はウイルスの部分的な又は完全な任意の特定の公知のゲノム配列を指し、これを用いて、対象からの特定された配列を参照してもよい。例えば、ヒト対象並びに多くのその他の生物に関して使用される参照ゲノムは、ncbi.nlm.nih.govのNational Center for Biotechnology Informationに見出される。「ゲノム」は、核酸配列中に発現する、生物又はウイルスの完全遺伝子情報を指す。

20

【0064】

様々な実施形態では、参照配列は、それにアラインメントされたリードよりも著しく大きい。例えば参照配列は、少なくとも約100倍大きくてもよく、又は少なくとも約1000倍大きくてもよく、又は少なくとも約10,000倍大きくてもよく、又は少なくとも約 10^5 倍大きくてもよく、又は少なくとも約 10^6 倍大きくてもよく、又は少なくとも約 10^7 倍大きくてもよい。

【0065】

一例において、参照配列は、完全長ヒトゲノムの配列である。そのような配列は、ゲノム参照配列と呼んでもよい。別の例では、参照配列は、13番染色体等の特定のヒト染色体に限定される。一部の実施形態では、参照Y染色体は、ヒトゲノム版hg19からのY染色体配列である。そのような配列は、染色体参照配列と呼んでもよい。参照配列のその他の例には、その他の種のゲノム、並びに任意の種の染色体、サブ染色体領域(ストランド等)等が含まれる。

30

【0066】

様々な実施形態において、参照配列は、多数の個体に由来するコンセンサス配列又はその他の組合せである。しかし、ある特定の適用例では、参照配列は特定の個体から採取されてもよい。

【0067】

「由来する」という用語は、本明細書において核酸又は核酸の混合物の文脈で使用されるとき、核酸がその出所源となる源から得られる手段を指す。例えば、一実施形態において、2つの異なるゲノムに由来する核酸の混合物は、その核酸、例えばcfDNAが、ネクロシス又はアポトーシス等の天然に生ずるプロセスを通して細胞によって自然に放出されたことを意味する。別の実施形態では、2つの異なるゲノムに由来する核酸の混合物は、その核酸が、対象からの2つの異なるタイプの細胞から抽出されたことを意味する。例えば、核酸の混合物は、ドナー細胞、及び臓器移植対象から得られたドナー細胞から生ずる核酸を含む。一部の實現例では、核酸の混合物は、2つ以上の寄与体個体の生体材料を含む。例えば、2つ以上の個体の生体材料を含む法医学的試料は、2つ以上の個体のDNAを含む。

40

【0068】

50

特定の定量値を得る文脈で使用される「基づく(based on)」という用語は、本明細書では、出力として特定の定量値を計算するための入力として別の量を使用することを指す。

【0069】

本明細書の「生体液」という用語は、生物学的供給源から採取された液体を指し、例えば、血液、血清、血漿、痰、洗浄液、脳脊髄液、尿、精液、汗、涙、唾液、及び同様のものを含む。本明細書で使用される「血液」、「血漿」、及び「血清」という用語は、それらの画分又はプロセッシング部分を明らかに包含する。同様に、試料が生検、スワブ、スメア等から採取される場合、「試料」は、生検、スワブ、スメア等に由来する、プロセッシングされた画分又は部分を明らかに包含する。

【0070】

本明細書で使用される「対応する(corresponding to)」という用語は、核酸配列、例えば遺伝子又は染色体であって、種々の対象のゲノムに存在する、及び全てのゲノムに必ずしも同じ配列を有するものではない、しかし対象となる配列の遺伝子情報以外の詳細を提供する働きをする、核酸配列、例えば遺伝子又は染色体を、時には指す。

【0071】

本明細書の「寄与体」という用語は、ヒト寄与体、並びに非ヒト寄与体、例えば哺乳動物、無脊椎動物、脊椎動物、真菌、酵母、細菌、及びウイルスを指す。本明細書の例はヒトに関し且つ言語は主にヒトの問題を対象とするが、本明細書に開示される概念は、任意の植物又は動物からのゲノムに適用可能であり、獣医学、動物科学、研究実験室等の分野で有用である。

【0072】

本明細書で使用される「感受性」という用語は、対象となる状態が存在するときに試験結果が陽性になる可能性を指す。真陽性の数を、真陽性及び偽陰性の合計で割った値として計算されてもよい。

【0073】

本明細書で使用される「特異性」という用語は、対象となる状態が存在しないときに試験結果が陰性になる可能性を指す。真陰性の数を、真陰性及び偽陽性の合計で割った値として計算されてもよい。

【0074】

本明細書で使用される「プライマー」という用語は、伸長生成物の合成を誘導する条件下に置かれた場合に、合成の開始点として作用することが可能な単離オリゴヌクレオチドを指す(例えば、その条件は、ヌクレオチド、DNAポリメラーゼ等の誘導剤、並びに適切な温度及びpHを含む)。プライマーは、増幅の最大効率のために好ましくは一本鎖であるが、代替として二本鎖であってもよい。二本鎖の場合、プライマーは、伸長生成物を調製するのに使用される前に、そのストランドを分離するように最初に処理される。好ましくはプライマーは、オリゴデオキシリボヌクレオチドである。プライマーは、誘導剤の存在下で伸長生成物の合成を刺激するのに十分長くなければならない。プライマーの正確な長さは、温度、プライマーの供給源、方法の使用、及びプライマー設計に使用されるパラメータを含む多くの因子に依存することになる。

【0075】

導入

本開示は、様々な利点及び技術的改善をもたらす、未知の遺伝子型の2つ以上の寄与体の核酸を含む核酸混合物試料のデコンボリューションのための方法及びシステムを提供する。例えば、一部の実現例は、確率的混合物モデル化、ベイズ推定技法、及び数値最適化アルゴリズムを適用して、寄与体の遺伝子型を知ることなく混合物中の寄与体DNAを定量する。

【0076】

密接に関係するゲノムの核酸(例えば、DNA又はRNA)混合物からの配列決定データは、研究並びに臨床の場で頻繁に見出され、混合寄与体の定量は、当初のゲノムが未知である場合に非常に難しくなっていた。当技術分野では、限られた成功率で、DNA混合物をデコン

10

20

30

40

50

ポリューションする試みがなされてきた。そのような試みは、キャピラリー電気泳動(CE)対立遺伝子データを使用してなされ、それらのデータは、臨床の場で有用となり得る対立遺伝子の配列情報を提供しないものである。更に、キャピラリー電気泳動をベースにした分析は、データベース内で公知の比較的少数の対立遺伝子にしばしば制限され、それらの対立遺伝子以外の情報を獲得することができない。DNA混合物試料を分析するのに、次世代配列決定技術を使用することが望ましい。しかし、DNA試料をデコンポリューションするための従来の方法は、NGSデータを解析するのに実現されていない。またNGSデータ解析のための従来の方法を修正しようとする場合であっても、その修正は自明ではない可能性があり、そのような修正の成功は疑わしい。例えば、対立遺伝子に関するCEデータは連続的であり、一方、配列決定データをベースにした対立遺伝子数は離散的である。当業者なら、連続データに関するモデルが離散データに関して全く作用できず、又は最適な状態に至らずに行われる可能性があることが、理解されよう。したがって、DNA混合物試料に関する配列決定データ(例えば、NGSデータ)をデコンポリューションするための新しい方法を開発することが望ましい。

【0077】

一部の実現例は、血液cfDNA又はgDNA試料の多重マーカー標的再配列決定データからの寄与体DNAを定量するための、方法及びシステムを提供する。一部の実現例は、新規な確率的モデル及び数値最適化アルゴリズムを使用して、血液cfDNA又はgDNA試料の多重マーカー標的再配列決定データから寄与体DNAを定量するための方法及びシステムを提供する。一部の実現例は、遺伝的関係性をコードする事前分布のベイズのモデル化を使用して、未知の遺伝子型の遺伝的に関係あるドナー及びレシピエントに関して寄与体DNAを定量するための、方法及びシステムを提供する。ベイズのフレームワークにおける事前情報を提供するのに遺伝的関係性の情報を使用することによって、DNA混合物の定量を、遺伝的関係性の情報を使用しない方法に比べて改善することができる。

【0078】

一部の実現例は、対数尤度関数の推定されたヘッシアン行列に結合されたクラメール・ラオを使用することにより、DNA定量の信頼区間を推定するための方法及びシステムを提供する。

【0079】

短い配列決定リードのマッピングにおける対立遺伝子バイアスは、DNA定量化を混乱させる。一部の実現例では、混乱した作用は、バリエーション部位に跨るリードの非バイアスマッピングを使用することによって、低減される。

【0080】

本明細書に記述される実現例は、寄与ゲノムに関する遺伝子型が全体として未知であるが、寄与体DNAの割合を正確に推定することができる。PCR増幅後のマーカー部位の対立遺伝子の割合は、ベータ分布により、信頼性を持ってモデル化することができる。

【0081】

非バイアス参照DNA配列データベースを使用して、参照対立遺伝子に対するバイアスを除去することができ、バリエーション部位での対立遺伝子数及び配列決定誤差を、信頼性を持って推定することができる。

【0082】

本明細書に記述される実現例は、混合物DNA試料の単一の配列決定の実行で、予測された寄与体DNAの割合の信頼区間を推定することができる。

【0083】

実験パイプライン

図1A～図1Cは、寄与体DNA定量のために設計された、バイオインフォマティクスアルゴリズム及び統計的モデルの概略を示す。図1Aは、配列決定をベースにした同種異系DNA検出のための実験パイプラインを示す。図1Bは、対立遺伝子計数の非バイアスリードマッピングワークフローを示す。図1Cは、マーカー座位当たりの対立遺伝子数に関する階層的、確率的混合物モデルを示す。

【0084】

一部の実現例は、図1Aに示される実験パイプラインを適用する。この一般的実験パイプラインには、以下の工程がある。

- 1)血液試料は、2つの個体から、DNAを含有するものを得る。
- 2)NAの適切なタイプ、例えば細胞DNA又は無細胞DNA(cfDNA)を、適用例に応じて抽出する。
- 3)ゲノムの特定のバリエーション部位又は多型部位を標的とし、PCR増幅及びハイブリダイゼーション等の手法によって濃縮する。バリエーション部位は、ヒト又は細菌の多様な集団の中で変化し得るように事前に選択される。或いは、非標的的全ゲノム配列決定を行うことができ、全てのバリエーション部位がカバーされることになる。
- 4)濃縮されたDNAを、標的ゲノム領域に関して濃縮される配列決定リードを得るために、以下に記述されるものの一部等のNGS技法によって配列決定する。

10

【0085】

正式な問題の記述

正式には、寄与体DNA定量(CDQ)の問題は、下記の通り述べられる: 2つの寄与体から構成されるDNA試料の配列決定データにより、試料中の各寄与体の割合が決定される。寄与体ゲノムの遺伝子型が未知である場合、CDQ問題を、盲検寄与体DNA定量(盲検-CDQ)と呼び;その反対を、非盲検-CDQと呼ぶ。一部の実現例に関する一部の記述は、ドナー及びレシピエントとしての2つの寄与体に言及するが、それらは方法の適用例を、臓器提供の場

20

に限定するものではない。一部の実現例に関する以下の一部の記述において、寄与体又はその寄与体はドナーと均等であり、その他の寄与体は、ドニーと均等である。

【0086】

盲検-CDQは、非盲検CDQに比べて、より困難な問題であるが、混合物試料の単一の配列決定実験のみが実現可能な場合には、全てのシナリオに対してより広い適用例があるものであり、一方、非盲検-CDQは、寄与体(例えば、臓器ドナー及びレシピエント)の遺伝子型を決定するために、事前配列決定実験を必要とする。

【0087】

本明細書に記述される計算方法は盲検-CDQ問題に対処し、方法の構成要素は、非盲検-CDQ問題で使用するために容易に単純化し又は適合させることができる。

【0088】

計算方法の概略

盲検-CDQに関する計算方法には、2つの主な工程がある:

- 1)対立遺伝子計数:各標的マーカー部位に関する各対立遺伝子からの配列決定リードの非バイアス計数に関するバイオインフォマティクスワークフロー(図1B)、及び
- 2)寄与体DNA定量:寄与体DNAの割合を定量するために、階層的確率的モデル及び関連した数値最適化アルゴリズムを使用する(図1C)。

30

【0089】

一部の実現例は、本明細書では「相対的定量」に対処するだけであるが、これは実現例が、絶対量ではなくて寄与体供給源から生じたDNA試料のパーセンテージ又は割合を推定することを意味する(質量又はコピー数に関して)。追加の工程は、入力されたDNAの総量がわかっている場合、相対的存在量を絶対的存在量に変換するのに行うことができる。

40

【0090】

核酸試料中の寄与体の割合を定量するための、又は核酸混合物試料をデコンボリューションするためのプロセスの概略

図2Aは、核酸試料中の1つ又は複数の寄与体の核酸(例えば、DNA又はRNA)の1つ又は複数の割合を定量するためのプロセス200を例示するブロック図を示す。核酸試料は、寄与体及び少なくとも1つのその他の寄与体の核酸(例えば、DNA又はRNA)を含む。方法は、1つ又は複数のプロセッサ及びシステムメモリ、例えば以下に記述されるシステムを含むコンピュータシステムで実現される。本明細書の記述は、一部の実現例及び適用例でDNAに言及するが、当業者なら、本明細書に記述される実現例を使用して、RNA及びその他の形

50

の核酸も分析できることが理解される。本明細書に記述される様々な実現例は、1つ又は複数の寄与体由来の核酸の核酸試料を分析するのに使用することができる。一部の実現例では、方法及びシステムは、1つ又は複数の寄与体の核酸の1つ又は複数の割合を定量するために提供される。本明細書の一部の記述では、試料が2つ以上の寄与体由来の核酸を含むことができるので、核酸試料を混合物試料と呼ぶ。しかし、「混合物」という用語の使用は、試料が2つ以上の寄与体の核酸を含む可能性を示し、試料が単一の寄与体のみからの核酸を含む可能性を排除しないことが理解される。後者の場合、1又は100%のパーセンテージの割合(又は誤差の範囲内の値)が1つの寄与体に関して決定され得る。

【0091】

一部の実現例では、核酸試料の1つ又は複数の寄与体は、移植片のドナーにおける移植片のドナーを含む。一部の実現例では、移植片は、同種異系又は異種移植片を含む。一部の実現例では、核酸試料は、ドナーから得られた生体試料である。一部の実現例では、核酸試料は無細胞核酸を含む。一部の実現例では、試料は細胞DNAを含む。

10

【0092】

プロセス200では、本明細書に記述されるもののような技法を使用して、核酸試料から核酸分子を抽出する。ブロック202を参照されたい。

【0093】

プロセス200では更に、抽出された核酸分子を増幅する。ブロック204を参照されたい。本明細書に記述されるもののような様々な増幅技法が使用され得る。一部の実現例では、PCRを使用して、抽出された核酸分子を増幅する。一部の実現例では、増幅は、特定の多型を標的とする。その他の実現例では、全ゲノム増幅を行ってもよく、特定の多型部位に関する対立遺伝子データは配列決定によって得られてもよい。

20

【0094】

プロセス200では、核酸配列リードを生成するために核酸シーケンサを使用して、増幅された核酸分子の配列決定もする。ブロック206を参照されたい。様々な配列決定技法及びデバイスについて、以下に更に記述するが、これは操作206に適用され得るものである。

【0095】

プロセス200では更に、核酸配列リードを、参照配列上の1つ又は複数の多型座位にマッピングする。一部の実現例では、アラインメント技法を使用して、核酸配列リードを1つ又は複数の多型座位にマッピングしてもよい。他の実現例では、非バイアスマッピング技法を使用して、核酸配列リードを多型座位にマッチさせてもよい。ブロック208を参照されたい。一部の実現例では、核酸配列リードを、多型座位の特定の対立遺伝子にマッピングする。非バイアスマッピング技法について更に、以下に記述する。一部の実現例では、1つ又は複数の多型座位(又は多型性座位)が二対立遺伝子座位を含む。一部の実現例では、1つ又は複数の多型座位にある対立遺伝子は、単一ヌクレオチド多型(SNP)対立遺伝子を含む。

30

【0096】

一部の実現例では、一意的分子インデックス(UMI)を、抽出された核酸分子に付着させ、次いで増幅し、配列決定し、多型座位又は対立遺伝子にマッピングする。一意的分子インデックスは、試料のプロセッシング及び分析工程中に生ずる可能性のある誤差を低減させるメカニズムを提供する。例えば、同じ一意的分子インデックス(UMI)を共用する種々のリードを、組み合わせ又は崩壊させて、リードが誘導された配列を決定することができ、試料のプロセッシング及び配列決定中に生じていた誤差が効果的に除去される。

40

【0097】

プロセス200では更に、核酸配列リードの方法を使用して、1つ又は複数の多型座位にある対立遺伝子に関する核酸配列リードの対立遺伝子数を決定する。ブロック210を参照されたい。

【0098】

またプロセス200では、確率的混合物モデルを、核酸配列リードの対立遺伝子数に適用する。確率的混合物モデルは、1つ又は複数の多型座位にある核酸配列リードの対立遺伝子

50

数をモデル化するために、確率分布を使用する。確率分布は、核酸配列リードの誤差を説明する。確率的混合物モデルは、確率分布からのランダム試料として、核酸配列リードの各対立遺伝子数を処理する。

【0099】

一部の実現例では、確率分布は、第1の二項分布を含む。一部の実現例では、第1の二項分布は、座位にある全対立遺伝子数を示す量パラメータと、座位にある第1の対立遺伝子の確率を示す確率パラメータとを含む。一部の実施形態では、第1の二項分布は下記の通り表される：

$n_{ij} \sim \text{BN}(n_i, p_{ij})$

式中、 n_{ij} は、座位*i*での対立遺伝子*j*に関する核酸配列リードの対立遺伝子数であり； n_i は、座位*i*での総リード数であり； p_{ij} は、座位*i*での対立遺伝子*j*の確率を示す確率パラメータである。座位にある全ての可能性のある対立遺伝子に関する対立遺伝子の確率は、合計して1になる。

【0100】

一部の実現例では、確率パラメータは、寄与体の核酸の割合の関数、又は β である。確率パラメータは、1つ又は複数の寄与体の遺伝子型の関数でもある。確率パラメータは、206の配列決定操作から生ずる誤差の関数、又は λ でもある。一部の実現例では、確率パラメータは、以下に記述する表3の p_1 '値を使用して得られる。一部の実現例では、寄与体の1つ又は複数の遺伝子型が未知であった。一部の実現例では、確率的混合物モデルは、図2Bに示されるような様々な確率分布を含む。

【0101】

図2Aに戻ると、プロセス200は、確率的混合物モデルを使用して、核酸試料中の1つ又は複数の寄与体の核酸の1つ又は複数の割合を定量する工程を含む。ブロック214を参照されたい。一部の実現例では、定量する工程は、確率パラメータを得るために、遺伝子型の複数の可能性のある組合せにわたって周辺化する工程を含む。一部の実現例では、定量する工程は、確率的混合物モデルのパラメータに対して条件付けられた操作210で決定された核酸配列リードの対立遺伝子数の尤度関数を使用して、寄与体の核酸の割合を定量する工程を含む。

【0102】

一部の実現例では、定量は、複数の潜在的割合値及び核酸配列リードの対立遺伝子数の尤度関数を使用して、複数の尤度値を計算する工程を含む。定量では、複数の尤度値の中の最大値である尤度値に関連する、潜在的割合値の特定も行う。一部の実現例では、複数の尤度値は、複数のパラメータ及びそれらの格子内での値に関して得られる。定量では、最大尤度を有する特定された潜在的割合値での、核酸試料中の寄与体の核酸の割合の定量も行う。一部の実現例では、尤度関数は、1つ又は複数の多型座位に関する複数の周辺分布を含む。

【0103】

一部の実現例では、1つ又は複数の寄与体は、2つの寄与体としてモデル化され、尤度関数は下記に従う：

$L(n_1, \dots, n_j \mid \beta, \pi) = \prod \sum g_{1j} g_{2j} \text{BN}(n_{ij}, p_{ij}(g_{1j}, g_{2j}, \lambda, \beta)) \cdot P(g_{1j}, g_{2j} \mid \pi)$

式中、 $L(n_1, \dots, n_j \mid \beta, \pi)$ は、パラメータ β (寄与体の1つの核酸の割合)及び π (母集団対立遺伝子頻度)が与えられた、対立遺伝子1から*j*に関する対立遺伝子数 n_1 から n_j を観察する尤度であり； $p_{ij}(g_{1j}, g_{2j}, \lambda, \beta)$ は、対立遺伝子*j*(g_{1j}, g_{2j})の2つの寄与体の遺伝子型に基づく、座位*i*にある対立遺伝子*j*の確率を示す確率パラメータであり； $P(g_{1j}, g_{2j} \mid \pi)$ は、母集団対立遺伝子頻度(π)が与えられた、2つの寄与体に関する対立遺伝子*j*の遺伝子型を観察する事前接合確率である。

【0104】

一部の実現例では、確率的混合物モデルは、202で行われた核酸分子を抽出する工程から生じる誤差、並びに206での配列決定操作から生ずる誤差を説明する。

【0105】

10

20

30

40

50

一部の実現例では、確率的混合物モデルは、第2の二項分布を使用して、1つ又は複数の多型座位での対立遺伝子に関して、抽出された核酸分子の対立遺伝子数をモデル化する。一部の実現例では、第2の二項分布は下記の通り表され:

$$n_{ij} \sim \text{BN}(n_i, p_{ij})$$

式中、 n_{ij} は、座位*i*での対立遺伝子*j*に関する、抽出された核酸分子の対立遺伝子数であり; n_i は、座位*i*での、合計抽出核酸分子数であり; p_{ij} は、座位*i*での対立遺伝子*j*の確率を示す確率パラメータである。

【0106】

一部の実現例では、第1の二項分布は、対立遺伝子の割合 n_{ij}/n_i に対して条件付けられる。一部の実現例では、第1の二項分布は、下記の通り再度パラメータ化され:

$$n_{ij} | n_i, n_i \sim \text{BN}(n_i, n_{ij}/n_i)$$

式中、 n_{ij} は、座位*i*での対立遺伝子*j*に関する核酸配列リードの対立遺伝子数であり; n_i は、座位*i*での抽出された核酸分子の総数であって、全ゲノムコピー数 n に等しいものであり; n_i は、座位*i*での総リード数であり; n_{ij} は、座位*i*での対立遺伝子*j*に関する抽出された核酸分子の数である。

【0107】

一部の実現例では、確率的混合物モデルは、 n_{ij}/n の分布を近似するのに、第1のベータ分布を使用する。一部の実現例では、第1のベータ分布は、第2の二項分布の平均及び分散に一致した平均及び分散を有する。

【0108】

一部の実現例では、座位*i*は二対立遺伝子としてモデル化され、第1のベータ分布は下記の通り表され:

$$n_{i1}/n \sim \text{Beta}((n-1)p_{i1}, (n-1)p_{i2})$$

式中、 p_{i1} は、座位*i*での第1の対立遺伝子の確率を示す確率パラメータであり; p_{i2} は、座位*i*での第2の対立遺伝子の確率を示す確率パラメータである。

【0109】

一部の実現例では、プロセスは、第1のベータ二項分布に従う n_{i1} の周辺分布を得るために、第1の二項分布と第1のベータ分布とを組み合わせる工程を含む。

【0110】

一部の実現例では、1つ又は複数の寄与体は2つの寄与体としてモデル化され、第1のベータ二項分布は下記の形を有し:

$$\text{BB}(n_{i1}, n_{i2} | \pi, (n-1) \cdot p_1(g_{11}, g_{21}, \lambda, \beta), (n-1) \cdot p_2(g_{11}, g_{21}, \lambda, \beta))$$

式中、 n_{i2} は、座位*i*での第2の対立遺伝子に関する核酸配列リードの対立遺伝子数であり; $p_1(g_{11}, g_{21}, \lambda, \beta)$ は、第1の対立遺伝子に関する第1の寄与体の遺伝子型(g_{11})及び第1の対立遺伝子に関する第2の寄与体の遺伝子型(g_{21})、並びに配列決定誤差 λ 及び寄与体割合 β に基づく第1の対立遺伝子の確率を示す、確率パラメータであり; $p_2(g_{11}, g_{21}, \lambda, \beta)$ は、第1の対立遺伝子に関する第1の寄与体の遺伝子型(g_{11})及び第1の対立遺伝子に関する第2の寄与体の遺伝子型(g_{21})、並びに配列決定誤差 λ 及び寄与体割合 β に基づく第2の対立遺伝子の確率を示す、確率パラメータである。

【0111】

一部の実現例では、操作214は、第1のベータ二項分布を使用して得られた尤度関数を使用して、核酸試料中の1つ又は複数の寄与体の核酸の1つ又は複数の割合を定量する工程を含み、この尤度関数は、下記の通りであり:

$$L(n_1, n_2 | \beta, n, \lambda, \pi) =$$

$$\prod_i \sum_{g_{11} g_{21}} \text{BB}(n_{i1}, n_{i2} | \pi, (n-1) \cdot p_1(g_{11}, g_{21}, \lambda, \beta), (n-1) \cdot p_2(g_{11}, g_{21}, \lambda, \beta)) \cdot P(g_{11}, g_{21} | \pi)$$

式中、 $L(n_1, n_2 | \beta, n, \lambda, \pi)$ は、パラメータ β 、 n 、 λ 、及び π が与えられた、第1の対立遺伝子に関する対立遺伝子数(n_1)及び第2の対立遺伝子に関する対立遺伝子数(n_2)を観察する尤度であり; $P(g_{11}, g_{21} | \pi)$ は、母集団対立遺伝子頻度(π)が与えられた第1の対立遺伝子に関する第1の寄与体の遺伝子型(g_{11})及び第1の対立遺伝子に関する第2の寄与体

10

20

30

40

50

の遺伝子型(g_{21})を観察する、事前接合確率である。

【0112】

一部の実現例では、操作214は、抽出された核酸分子の質量から、全抽出ゲノムコピー数 n を推定する工程を含む。

【0113】

一部の実現例では、確率的混合物モデルは、204で行われた核酸分子の増幅から生じる誤差、並びに206での配列決定操作から生ずる誤差を説明する。一部の実現例では、増幅プロセスは、下記のようにモデル化され：

$$x_{t+1} = x_t + y_{t+1}$$

式中、 x_{t+1} は、増幅の $t+1$ サイクル後の所与の対立遺伝子の核酸コピーであり； x_t は、増幅の t サイクル後の所与の対立遺伝子の核酸コピーであり； y_{t+1} は、 $t+1$ サイクルで発生した新しいコピーであって、二項分布 $y_{t+1} \sim \text{BN}(x_t, r_{t+1})$ に従うものであり； r_{t+1} は、 $t+1$ サイクルに関する増幅率である。

10

【0114】

一部の実現例では、確率的混合物モデルは、第2のベータ分布を使用して、1つ又は複数の多型座位での対立遺伝子に関して増幅された核酸分子の対立遺伝子割合をモデル化する。一部の実現例では、座位 i は二対立遺伝子としてモデル化され、第2のベータ分布は下記の通り表され：

$$n_{i1}' / (n_{i1}' + n_{i2}') \sim \text{Beta}(n \cdot p_i \cdot p_{i1}, n \cdot p_i \cdot p_{i2})$$

式中、 n_{i1}' は、座位 i での第1の対立遺伝子に関する増幅核酸分子の対立遺伝子数であり； n_{i2}' は、座位 i での第2の対立遺伝子に関する増幅核酸分子の対立遺伝子数であり； n は、任意の座位での全抽出核酸分子数であり； p_i は、平均増幅率 r に関する定数であり； p_{i1} は、座位 i での第1の対立遺伝子の確率であり； p_{i2} は、座位 i での第2の対立遺伝子の確率である。一部の実現例では、 p_i が $(1+r)/(1-r)/[1-(1+r)^{-t}]$ である。一部の実現例では、 p_i は、 $(1+r)/(1-r)$ として近似される。

20

【0115】

一部の実現例では、操作214は、第2のベータ二項分布に従う n_{i1} の周辺分布を得るために、第1の二項分布と第2のベータ分布とを組み合わせる工程を含む。一部の実現例では、第2のベータ二項分布は、下記の形を有し：

$$\text{BB}(n_{i1}, n_{i2} \mid n, n \cdot p_i \cdot p_{i1}, n \cdot p_i \cdot p_{i2})$$

式中、 n_{i2} は、座位 i での第2の対立遺伝子に関する核酸配列リードの対立遺伝子数であり； p_{i1} は、座位 i での第1の対立遺伝子の確率を示す確率パラメータであり； p_{i2} は、座位 i での第2の対立遺伝子の確率を示す確率パラメータである。

30

【0116】

一部の実現例では、操作214は、1つ又は複数の多型座位が同じ増幅率を有すると仮定することによって、第2のベータ二項分布を再度パラメータ化する工程を含み：

$$\text{BB}(n_{i1}, n_{i2} \mid n, (1+r)/(1-r) \cdot p_{i1}(g_{11}, g_{21}, \lambda, \beta), (1+r)/(1-r) \cdot p_{i2}(g_{11}, g_{21}, \lambda, \beta))$$

式中 r は増幅率であり； p_{i2} は、座位 i での第2の対立遺伝子の確率を示す確率パラメータである。

【0117】

一部の実現例では、操作214は、第2のベータ二項分布を使用して得られた尤度関数を使用して、核酸試料中の1つ又は複数の寄与体の核酸の1つ又は複数の割合を定量する工程を含み、この尤度関数は下記の通りになり：

$$L(n_1, n_2 \mid \beta, r, \lambda, \pi) =$$

$$\prod_i \sum_{g_{11} g_{21}} [\text{BB}(n_{i1}, n_{i2} \mid n, (1+r)/(1-r) \cdot p_1(g_{11}, g_{21}, \lambda, \beta), (1+r)/(1-r) \cdot p_2(g_{11}, g_{21}, \lambda, \beta)) \cdot P(g_{11}, g_{21} \mid \pi)]$$

式中、 $L(n_1, n_2 \mid \beta, r, \lambda, \pi)$ は、パラメータ β 、 r 、 λ 、及び π が与えられた、第1の対立遺伝子に関する対立遺伝子数(n_1)及び第2の対立遺伝子に関する対立遺伝子数(n_2)を観察する尤度である。

40

【0118】

50

一部の実現例では、操作214は、各多型座位の相対増幅率が座位当たりの総リードに比例すると定義することによって、第2のベータ二項分布を再度パラメータ化する工程を含み：
 $BB(n_{i1}, n_{i2} \mid \eta, c' \cdot n_i \cdot p_{i1}(g_{11}, g_{21}, \lambda, \beta), c' \cdot n_i \cdot p_{i2}(g_{11}, g_{21}, \lambda, \beta))$
 式中、 c' は、最適化されるパラメータであり； p_{i2} は、座位*i*での第2の対立遺伝子の確率を示す、確率パラメータである。

【0119】

一部の実現例では、操作214は、第2のベータ二項分布を使用して得られた尤度関数を使用して、核酸試料中の1つ又は複数の寄与体の核酸の1つ又は複数の割合を定量する工程を含み、この尤度関数は下記の通りである：

$L(n_1, n_2 \mid \beta, c', \lambda, \pi) =$

$\prod_i \sum g_{11} g_{21} [BB(n_{i1}, n_{i2} \mid \eta, c' \cdot n_i \cdot p_1(g_{11}, g_{21}, \lambda, \beta), c' \cdot n_i \cdot p_2(g_{11}, g_{21}, \lambda, \beta)) \cdot P(g_{11}, g_{21} \mid \pi)]。$

【0120】

一部の実現例では、確率的混合物モデルは、202で行われた核酸分子を抽出する工程及び204で行われた核酸分子を増幅する工程から生ずる誤差、並びに206での配列決定操作から生ずる誤差を説明する。一部の実現例では、確率的混合物モデルは、第3のベータ分布を使用して1つ又は複数の多型座位にある対立遺伝子に関して増幅された核酸分子の対立遺伝子割合をモデル化し、これは202で行われた核酸分子の抽出及び204で行われた核酸分子の増幅から生ずるサンプリングエラー、並びに206での配列決定操作から生ずる誤差を説明する。

【0121】

一部の実現例では、座位*i*が二対立遺伝子としてモデル化され、第3のベータ分布が下記の形を有し：

$n_{i1}' / (n_{i1}' + n_{i2}') \sim \text{Beta}([n'' \cdot (1 + r_i) / 2 - 1] p_{i1}, [n'' \cdot (1 + r_i) / 2 - 1] p_{i2})$

式中、 n_{i1}' は、座位*i*での第1の対立遺伝子に関して増幅された核酸分子の対立遺伝子数であり； n_{i2}' は、座位*i*での第2の対立遺伝子に関して増幅された核酸分子の対立遺伝子数であり； n'' は、全抽出核酸分子数であり； r_i は、座位*i*での増幅率であり； p_{i1} は、座位*i*での第1の対立遺伝子の確率であり； p_{i2} は、座位*i*での第2の対立遺伝子の確率である。

【0122】

一部の実現例では、操作214は、第3のベータ二項分布に従う n_{i1} の周辺分布を得るために、第1の二項分布と第3のベータ分布とを組み合わせる工程を含む。一部の実現例では、第3のベータ二項分布は下記の形を有する：

$BB(n_{i1}, n_{i2} \mid \eta, (n'' \cdot (1 + r_i) / 2 - 1) \cdot p_1(g_{11}, g_{21}, \lambda, \beta), (n'' \cdot (1 + r_i) / 2 - 1) \cdot p_2(g_{11}, g_{21}, \lambda, \beta))。$

【0123】

一部の実現例では、操作214は、第3のベータ二項分布を使用して得られた尤度関数を使用して、核酸試料中の1つ又は複数の寄与体の核酸の1つ又は複数の割合を定量する工程を含み、尤度関数は：

$L(n_1, n_2 \mid \beta, n'', \lambda, \pi) = \prod_i \sum g_{11} g_{21} BB(n_{i1}, n_{i2} \mid \eta, (n'' \cdot (1 + r_i) / 2 - 1) \cdot p_1(g_{11}, g_{21}, \lambda, \beta), (n'' \cdot (1 + r_i) / 2 - 1) \cdot p_2(g_{11}, g_{21}, \lambda, \beta)) \cdot P(g_{11}, g_{21} \mid \pi)$

を含み、式中、 $L(n_1, n_2 \mid \beta, n'', \lambda, \pi)$ は、パラメータ β 、 n'' 、 λ 、及び π が与えられた、第1の対立遺伝子に関する対立遺伝子数 n_1 及び第2の対立遺伝子に関する対立遺伝子数 n_2 を観察する尤度である。

【0124】

一部の実現例では、プロセス200は更に、クラメール・ラオの不等式を使用して、1つ又は複数の寄与体の核酸の1つ又は複数の割合の1つ又は複数の信頼区間を推定する工程を含む。

【0125】

一部の実現例では、208のマッピング操作は、複数の非バイアス標的配列の任意の配列にマッチする、核酸配列リードの中のリードを特定する工程を含み、これら複数の非バイア

10

20

30

40

50

ス標的配列は、参照配列の部分配列、及び単一のヌクレオチドだけ部分配列が異なる配列を含むものである。

【0126】

一部の実現例では、複数の非バイアス標的配列は、配列の5つのカテゴリを含む: (i) 参照配列の部分配列である、参照標的配列であって、各参照標的配列が、多型座位を包含し且つ参照配列上に見出される参照対立遺伝子を有する、参照標的配列; (ii) 参照標的配列に対応する代替標的配列であって、各代替標的配列が、対応する参照標的配列上に見出される参照対立遺伝子とは異なる代替対立遺伝子を有し、対応する配列が、参照配列上に同じ長さ及び同じ場所を有する、代替標的配列; (iii) それぞれが、参照対立遺伝子と代替対立遺伝子との間での相違を定めるヌクレオチド以外のただ1つのヌクレオチドだけ、参照標的配列とは異なっている、全ての可能性のある配列を含む、変異参照標的配列; (iv) それぞれが、参照対立遺伝子と代替対立遺伝子との間の相違を定めるヌクレオチド以外のただ1つのヌクレオチドだけ、代替標的配列とは異なっている、全ての可能性のある配列を含む、変異代替標的配列; 及び (v) 参照標的配列に対応する、予想外の対立遺伝子標的配列であって、各予想外の対立遺伝子標的配列が、対応する参照標的配列上に見出される参照対立遺伝子及び対応する代替標的配列上に見出される代替対立遺伝子とは異なる予想外の対立遺伝子を有し、対応する配列が、参照配列上で同じ長さ及び同じ場所を有する、予想外の対立遺伝子標的配列。

10

【0127】

一部の実現例では、操作208は、1つ又は複数の多型座位での対立遺伝子に関する核酸配列リードの対立遺伝子数を決定するために、特定されたリード及びそれらのマッチング非バイアス標的配列を使用する工程を含む。一部の実現例では、複数の非バイアス標的配列は、核酸配列リードと同じ長さを有するように、切断された配列を含む。一部の実現例では、複数の非バイアス標的配列は、1つ又は複数のハッシュテーブルに記憶された配列を含み、リードはハッシュテーブルを使用して特定される。

20

【0128】

一部の実現例では、プロセス200は更に、1つ又は複数の多型座位のそれぞれで、対立遺伝子構成を決定する工程を含み、このそれぞれの対立遺伝子構成は、1つ又は複数の寄与体のそれぞれに関して2つ以上の対立遺伝子の対立遺伝子状態を含む。

【0129】

図3は、1つ又は複数の寄与体の核酸を含む核酸試料を評価するための、プロセス300を例示するブロック図を示す。プロセス300は、核酸試料から得られた1つ又は複数の多型座位にある1つ又は複数の対立遺伝子の核酸配列リードを受け取ることによって、開始する。ブロック302を参照されたい。一部の実現例では、核酸配列リードは、本明細書に記述される様々な技法を使用して、核酸試料中の核酸を配列決定することによって得られた。

30

【0130】

一部の実現例では、一意的分子インデックス(UMI)を、抽出された核酸分子に取着し、次いで増幅し、配列決定し、多型座位又は対立遺伝子にマッピングする。一意的分子インデックスは、試料のプロセッシング及び分析工程中に生ずる可能性のある誤差を低減させるメカニズムを提供する。例えば、同じ一意的分子インデックス(UMI)を共用する種々のリードを組み合わせ又は崩壊させて、そこからリードが誘導される配列を決定することができ、試料のプロセッシング及び配列決定中に生じていた誤差が効果的に除去される。参照によりその全体が全ての目的で組み込まれる、2016年4月16日出願の米国特許出願第15/130,668号及び2017年1月18日出願の米国仮特許出願第62/447,851号は、一意的分子インデックスを使用して核酸を配列決定するための様々な方法及びシステムについて記述する。

40

【0131】

プロセス300では更に、核酸配列リードを使用して、1つ又は複数の多型座位にある1つ又は複数の対立遺伝子に関する対立遺伝子数を決定する。

【0132】

またプロセス300では、確率的混合物モデルを対立遺伝子数に適用する。確率的モデルは

50

、確率分布を使用して、1つ又は複数の多型座位にある対立遺伝子の対立遺伝子数をモデル化する。確率分布は、対立遺伝子データにおける誤差に関して意味がある。誤差は、核酸抽出、試料のプロセッシング、及び配列決定の操作から生ずる誤差を含む。

【0133】

一部の実現例では、確率分布は第1の二項分布を含む。一部の実現例では、第1の二項分布は、座位での全対立遺伝子数を示すパラメータと、座位での第1の対立遺伝子の確率を示す確率パラメータとを含む。一部の実現例では、確率パラメータは、核酸試料中の1つ又は複数の寄与体の核酸の割合の関数である。確率パラメータは、1つ又は複数の寄与体の遺伝子型の関数、又はGでもあり、核酸配列リードデータの誤差の関数、又は θ でもある。一部の実現例では、リードデータの誤差は、核酸抽出、試料のプロセッシング、及び配列決定操作から生ずる誤差を含む。

10

【0134】

プロセス300では、モデルパラメータが与えられた対立遺伝子データを観察する尤度値、及び潜在的核酸割合値も得る。ブロック308を参照されたい。

【0135】

一部の実現例では、プロセス300では、尤度値を使用して、核酸試料中の1つ又は複数の寄与体の核酸の割合を定量する。ブロック310を参照されたい。

【0136】

一部の実現例では、プロセス300では更に、尤度値を使用して、寄与体の少なくとも1つに関する少なくとも1つの遺伝子型を決定する。ブロック312を参照されたい。

20

【0137】

一部の実現例では、寄与体の遺伝子型は、プロセス300の前は未知であった。

【0138】

一部の実現例では、確率的混合物モデルは、ベータ分布を使用して対立遺伝子データの誤差をモデル化する。一部の実現例では、ベータ分布は、平均パラメータ及び濃度パラメータによって定義される。一部の実現例では、濃度パラメータが、種々のノイズ条件を表す離散事前確率を有する。濃度パラメータは、座位全体にわたって様々である。

【0139】

一部の実現例では、操作310の定量は、ベータ二項分布に従う周辺分布を得るために、第1の二項分布とベータ分布とを組み合わせる工程を含む。

30

【0140】

一部の実現例では、310の定量は、対立遺伝子データの尤度関数を使用して、核酸試料中の1つ又は複数の寄与体の核酸の割合を定量する工程を含む。一部の実現例では、定量では、複数の潜在的割合値、及び対立遺伝子数の尤度関数を使用して、複数の尤度値を計算する。定量では、最大尤度値に関連した潜在的割合ベクトルも特定し、特定された潜在的割合ベクトルを使用して、核酸試料中の1つ又は複数の寄与体の核酸の割合も定量する。

【0141】

一部の実現例では、尤度関数は $P(G \mid \pi)$ に依存するが、これは母集団対立遺伝子頻度(π)が与えられた1つ又は複数の寄与体の遺伝子型の事前確率である。一部の実現例では、事前確率は、機構的ドロップアウトを表す固定された事前確率を持つ、ダミー対立遺伝子を考慮して計算される。

40

【0142】

一部の実現例では、1つ又は複数の寄与体は、2つ以上の寄与体を含む。一部の実現例では、プロセス300は、1つ又は複数の寄与体中の寄与体の総数を決定する操作を含む。一部の実現例では、1つ又は複数の寄与体の1つ又は複数の遺伝子型は未知であり、プロセス300は、1つ又は複数の多型座位のそれぞれにある対立遺伝子構成を決定する操作を含み、この対立遺伝子構成は、寄与体のそれぞれに関して2つ以上の対立遺伝子の対立遺伝子状態を含む。一部の実現例では、プロセス300は、対立遺伝子構成に関して推定された確率を決定する操作を含む。

【0143】

50

一部の実現例では、プロセス300は更に、1つ又は複数の寄与体の中の特定の寄与体が特定の遺伝子型を有する事後確率を、得る工程を含む。一部の実現例では、プロセス300は更に、事後確率に基づいて、核酸試料が特定の寄与体由来の核酸を含むことをコールする工程を含む。一部の実現例では、1つ又は複数の寄与体の中の特定の寄与体が特定の遺伝子型を有するという事後確率を得る工程は:(i)遺伝子型構成の事前確率に、遺伝子型構成の尤度を乗じる工程;(ii)(i)と遺伝子型空間の合計との積を正規化する工程;及び(iii)特定の遺伝子型を含有する遺伝子型構成を合計して、事後確率を得る工程を含む。

【0144】

一部の実現例では、特定の遺伝子型が多座位遺伝子型を含み、方法は更に:全ての寄与体上で、寄与体が全ての座位で特定の遺伝子型を有する事後確率を合計する工程と;合計した確率に基づいて、特定された多座位遺伝子型が任意の寄与体に現れることを決定する工程とを含む。

10

【0145】

一部の実現例では、核酸試料が法医学的試料であり、多座位遺伝子型のデータは、対象となる人間から得られる。プロセスは更に、対象となる人間が核酸試料の寄与体であることを決定する工程を含む。

【0146】

一部の実現例では、確率的混合物モデルは、第2の二項分布を使用して対立遺伝子データのスタッターエラーをモデル化する。一部の実現例では、第2の二項分布は下記の通り表され:

20

$s_{ik} \sim \text{BN}(n_i(k+1), r_i)$

式中、 s_{ik} は、対立遺伝子 k であるように見えるが実際には対立遺伝子 $k+1$ のスタッターエラーから生ずる、スタッター対立遺伝子の座位 i でのスタッター対立遺伝子数であり; $n_i(k+1)$ は、座位 i での対立遺伝子 $k+1$ の当初の対立遺伝子数であり; r_i は、座位 i に関するスタッター率である。

【0147】

一部の実現例では、スタッター率 r は、座位全体にわたって様々であり、種々のノイズ条件を表す事前確率を有し、この事前確率は、座位全体にわたって共用される。

【0148】

一部の実現例では、操作310は、非スタッター対立遺伝子数の尤度とスタッター対立遺伝子数の尤度との積を含む尤度関数を使用して、核酸試料中の1つ又は複数の寄与体の核酸の割合を定量する工程を含む。

30

【0149】

一部の実現例では、確率的混合物モデルを適用する工程は、スタッターをそこから潜在的に生じさせることができる分子の数を決定するときに、対立遺伝子 $k+1$ に割り当てられた対立遺伝子数に、固定数の分子を付加する工程を含む。

【0150】

一部の実現例では、確率的混合物モデルは、ダミー試料外対立遺伝子を使用して自然ドロップアウトをモデル化する。一部の実現例では、ダミー試料外対立遺伝子の事前確率が、観察されていない対立遺伝子の数に比例する。一部の実現例では、観察されていない対立遺伝子の数は:最短と最長の間で観察された整数値対立遺伝子の全ての整数を内挿し、任意の観察された非整数値対立遺伝子を付加し、得られた値の最大値及び基準値を戻すことによって推定される。

40

【0151】

一部の実現例では、確率的混合物モデルを適用する工程は、核酸試料中の1つ又は複数の寄与体の核酸の割合を定量するのに使用されるデータから遺伝子型構成を刈り込む工程を含む。一部の実現例では、遺伝子型構成を刈り込む工程は:必要とされる対立遺伝子のリストを構築することによって妥当とされる遺伝子型構成を限定する工程と、必要とされる全ての対立遺伝子を説明するのに十分ではない寄与体を持つ座位を除外する工程とを含む。一部の実現例では、必要とされる対立遺伝子のリストは、本質的に、閾値よりも高く且つ

50

スタッタードロップインに起因して妥当とされるには高過ぎる対立遺伝子数を有する対立遺伝子からなる。一部の実現例では、閾値は、(i)最大非スタッター対立遺伝子数、及び(ii)潜在的なスタッタードナー対立遺伝子の数を乗じた値の合計である。一部の実現例では、遺伝子型構成を刈り込む工程は、対立遺伝子データと予想される対立遺伝子数との間で不十分なマッチを有する遺伝子型構成を除去する工程を含む。一部の実現例では、不十分なマッチを有する遺伝子型構成は、1つ又は複数の閾値よりも大きい二乗平均平方根誤差(RMSE)値を有する。

【0152】

一部の實現例では、1つ又は複数の多型座位での対立遺伝子は、単一ヌクレオチド多型(SNP)対立遺伝子及び/又は短鎖タンデム反復(STR)対立遺伝子を含む。

【0153】

マーカー部位に対するリードの非バイアスマッピングのための方法

核酸(例えば、DNA又はRNA)配列決定リードをゲノムにマッピングするための従来の計算方法は、使用される参照ゲノムによってバイアスさせることができる。各バリエーション部位に関するただ1つの対立遺伝子(参照対立遺伝子)が参照ゲノム中に存在するので、リードと参照との間のミスマッチは、既存のリードマッピングアルゴリズムにおける配列決定誤差として処理される。問題は、非参照対立遺伝子を含むリードが、配列決定誤差を含むとして処理されたとき、アラインメントの信頼性(スコア)が低下し、したがって、後続のフィルタリング工程で、高い信頼性でマッピングされたリードとして保持され難いことである。このマッピングバイアスは、対立遺伝子数を歪ませることになり(図1B)、その後、寄与体DNA割合の推定を損なわせる。

【0154】

マッピングバイアスの問題に対処し且つ最適なCDQを可能にするために、一部の實現例は、リードをバリエーション部位にマッピングするための新規なワークフローを提供する。新しいリードマッピング手法は、対立遺伝子の非バイアス計数と、バリエーション部位及び非バリエーション部位での配列決定誤差の推定を可能にする。

【0155】

リードマッピングワークフローは下記の通りである。ワークフローはまず、1)参照配列及び2)バリエーション部位の公知の対立遺伝子に基づいて、5つのタイプの配列(Table 1(表1)参照)を生成する。複数の単一変異が配列当たりにある場合、より多くのタイプの配列が生成されることになる。5つのタイプの配列を、それぞれref、alt、ref.mut、alt.mut、及びsnp.mutと呼ぶ。例えば、長さLの標的配列によってカバーされた各二対立遺伝子SNPマーカー部位ごとに、1つのref、1つのalt、 $[L-1] \times 3$ ref.mut、 $[L-1] \times 3$ alt.mut、及び2snp.mut配列がある。次いで5つの全てのタイプの配列は、「非バイアス標的配列」のデータベース(図1B)に含まれる。次いで、シーケンサからのリードの長さに応じて、非バイアス標的配列は2つのバージョンに切断される。rをリード長とする。切断された標的配列のバージョン1は、全ての非バイアス標的配列のr 5'塩基を含み、一方、切断された標的配列のバージョン2は、全ての非バイアス標的配列のr 3'塩基の逆相補を含む。次いで切断標的配列における冗長配列を除去する。次いで2つの切断配列データベースにおける一意的な配列を、2つのハッシュテーブルに記録する。次に、配列決定リードを、ハッシュテーブルを使用して計数する。ペアエンド配列決定戦略では、R1リード及びR2リードを、それぞれ第1及び第2のハッシュテーブルを使用して計数する。非ペアエンド配列決定では、全てのリードを、第1のハッシュテーブルを使用して計数する。最後に、各マーカー部位ごとに、カウント数を、切断非バイアス標的配列がTable 1(表1)において対応するタイプに応じて、上記にて定義された5つのタイプに集計する。

【0156】

類似の戦略は、マッピングのためのハッシュテーブルを使用する代わりに配列アラインメント・ツールを使用する場合、実現することができる。各マーカー部位ごとに、ref及びaltタイプの配列を生成して、非バイアス配列データベースを形成する。次いで各配列決定リードをこのデータベースにアラインメントして、所定の数の配列決定誤差になるようにす

10

20

30

40

50

る。次いでマッピングされたリードを、Table 1(表1)に基づいて分類する。SNPマーカーの場合、二対立遺伝子SNPのみがここでは考慮される。

【0157】

【表1】

Table 1. バリアント部位付近で参照配列から生成される5つのタイプの配列の定義。

タイプ	定義
ref	参照対立遺伝子をとる SNP 部位
alt	代替対立遺伝子をとる SNP 部位
ref.mut	SNP 部位が ref であるときの非 SNP 部位上の単一変異
alt.mut	SNP 部位が alt であるときの非 SNP 部位上の単一変異
snp.mut	参照も代替の対立遺伝子もとらない SNP 部位

10

【0158】

提示されたリードマッピングワークフローは、実データを使用して試験をしたときのリードマッピングバイアスの問題に対処する。ワークフローによれば、参照の代替誤差に対する、及び代替の参照誤差に対する、観察された誤差率は、同一である。参照DNAコピー上の非バリアント部位での、及び代替のDNAコピー上の非バリアント部位での、配列決定誤差率も、同一である。

20

【0159】

寄与体DNA割合と、対立遺伝子割合との連結

配列決定誤差がないと仮定する

本発明者等は、DNAを試料に供給する、 n_d ドナー細胞及び n_r レシピエント細胞があると仮定する。これらの細胞に基づいて、実現例は、少量の寄与体割合を、 $\beta = n_d / (n_d + n_r)$ と定義する。特定の座位のそれぞれにあるドナー及びレシピエントの遺伝子型に応じて、2つの対立遺伝子は異なる割合を有し(詳細は、Table 2(表2)参照)、それらを計算するための一般式は、 $p_1 = [g_{11}(1 - \beta) + c \cdot g_{21} \cdot \beta] / 2$ 及び $p_2 = [g_{12}(1 - \beta) + g_{22} \cdot \beta] / 2$ である。 g_{11} 及び g_{12} はレシピエントの遺伝子型であり、即ちレシピエントゲノムの対立遺伝子1及び2のコピーであり; g_{21} 及び g_{22} はドナーの遺伝子型であり、即ちドナーゲノムの対立遺伝子1及び2のコピーであることに留意されたい。

30

【0160】

40

50

【表 2】

Table 2: 所与のバリエーション部位に関するドナー及びレシピエント対の間の 9 つの可能性のある遺伝子型組合せに関する二項モデルパラメータ p_1 及び p_2

g_{11}	g_{21}	p_1	p_2
0	0	0	1
0	1	$\beta/2$	$1-\beta/2$
0	2	β	$1-\beta$
1	0	$(1-\beta)/2$	$(1+\beta)/2$
1	1	$1/2$	$1/2$
1	2	$(1+\beta)/2$	$(1-\beta)/2$
2	0	$1-\beta$	β
2	1	$1-\beta/2$	$\beta/2$
2	2	1	0

【0161】

配列決定誤差のモデル化

2つの公知の対立遺伝子がバリエーション部位にある場合、配列決定誤差は、2つの公知の対立遺伝子を2つの残りのヌクレオチドにこの座位で変換することに加え、1つの対立遺伝子を別の対立遺伝子に変換することになる。その結果、配列決定されたリード内の対立遺伝子割合は、NGS入力DNA試料中の対立遺伝子割合から逸脱することになる。

【0162】

図2Cは、1つの対立遺伝子を別の対立遺伝子に、及び真の対立遺伝子を予測外の対立遺伝子に変換する、配列決定誤差を概略的に示す。パネル(A)はヌクレオチド依存性配列決定誤差を示し、パネル(B)は均一な配列決定誤差を示す。

【0163】

N_1 、 N_2 を対立遺伝子1及び対立遺伝子2のヌクレオチドとする。 p_1' 、 p_2' を、それぞれ、対立遺伝子1及び対立遺伝子2のリードを観察する確率とし、これらが実際のものか又は配列決定誤差に起因するものかを問わず、且つ $p_0' = 1 - p_1' - p_2'$ は、配列決定誤差に起因した、2つの予測外の対立遺伝子を観察する確率である。 $\lambda_{N_1N_2}$ を、 N_1 から N_2 までの変異率(確率)とし、この場合、 N_1 及び N_2 は各SNP部位に一意的であり、且つ $\lambda_{N_1\#}$: N_1 から、3ヌクレオチド非 N_1 ヌクレオチドのいずれかまでの変異確率である。

【0164】

SNP部位の4つのヌクレオチドでの遷移図を、図2Cに示す。これに基づいて、実現例は、真の対立遺伝子割合 p_1 、 p_2 から観察された対立遺伝子割合 p_1' 、 p_2' 、及び p_0' に変換するための、下記の方程式を得る:

$$p_1' = p_1 - p_1 \cdot \lambda_{N_1\#} + p_2 \cdot \lambda_{N_2N_1}$$

$$p_2' = p_2 - p_2 \cdot \lambda_{N_2\#} + p_1 \cdot \lambda_{N_1N_2}$$

$$p_0' = p_1 \cdot (\lambda_{N_1\#} - \lambda_{N_1N_2}) + p_2 \cdot (\lambda_{N_2\#} - \lambda_{N_2N_1})$$

【0165】

実現例が、ヌクレオチドの詳細とは無関係に均一な配列決定誤差率を仮定する場合、実現例は下記を有する。

$$p_1' = p_1 \cdot (1 - 3 \cdot \lambda) + p_2 \cdot \lambda$$

$$p_2' = p_2 \cdot (1 - 3 \cdot \lambda) + p_1 \cdot \lambda$$

$$p_0' = 2 \cdot \lambda$$

【0166】

実現例が、予測外の対立遺伝子を見逃す場合、下記の通りになる。

$$p_1' = (p_1 \cdot (1 - 3 \cdot \lambda) + p_2 \cdot \lambda) / (1 - 2 \cdot \lambda)$$

$$p_2' = (p_2 \cdot (1 - 3 \cdot \lambda) + p_1 \cdot \lambda) / (1 - 2 \cdot \lambda),$$

ここで $o(\lambda^2)$ は近似誤差であり、これは下記のように書き直される。

$$p_1' = p_1 \cdot (1 - \lambda) + p_2 \cdot \lambda$$

$$p_2' = p_2 \cdot (1 - \lambda) + p_1 \cdot \lambda$$

次いで、寄与体の遺伝子型に応じて、寄与体割合 β を、観察された対立遺伝子割合 p_1' と連結する式を、Table 3(表3)に列挙する。

【0167】

【表3】

10

Table 3: 各ドナー/レシピエント遺伝子型組合せに対して条件付けられた、配列決定誤差を考慮した後に対立遺伝子1及び2を観察する予想確率 $p_{i1}(g_{11}, g_{21}, \lambda, \beta)$ 及び $p_{i2}(g_{11}, g_{21}, \lambda, \beta)$ 。ここでは、全ての N_1 及び N_2 に関して $\lambda_{N1N2} = \lambda$ と仮定する。変異率 λ は小さいので、一次近似を使用する。

g_{11}	g_{21}	p_1'	p_2'
0	0	λ	$1 - \lambda$
0	1	$\beta/2 + \lambda - \beta\lambda$	$1 - \beta/2 - \lambda + \beta\lambda$
0	2	$\beta + \lambda - 2\beta\lambda$	$1 - \beta - \lambda + 2\beta\lambda$
1	0	$(1 - \beta)/2 + \beta\lambda$	$(1 + \beta)/2 - \beta\lambda$
1	1	$1/2$	$1/2$
1	2	$(1 + \beta)/2 - \beta\lambda$	$(1 - \beta)/2 + \beta\lambda$
2	0	$1 - \beta - \lambda + 2\beta\lambda$	$\beta + \lambda - 2\beta\lambda$
2	1	$1 - \beta/2 - \lambda + \beta\lambda$	$\beta/2 + \lambda - \beta\lambda$
2	2	$1 - \lambda$	λ

20

30

【0168】

DNA抽出、PCR(増幅)、及び配列決定モデルの概観

3つの確率的モデル(図1C)は、一般的実験パイプライン(図1A)での3つの主要な構成要素をモデル化するために提示する: 1) DNA抽出; 2) 標的DNAを濃縮するための手法としてのDNA増幅(例えば、PCR); 3) 配列決定(例えば、NGS配列決定)。

【0169】

下記の注釈は、Table 4(表4)で詳述する数学的モデルで使用される。

p_1 、 p_{1i} : 座位 i に関する対立遺伝子1の確率。下付き文字 i は、実現例が単一座位に焦点を当てるときには省略されることに留意されたい。

p_1 、 p_{2i} : 座位 i に関する対立遺伝子2の確率

40

n_1 、 n_{1i} 、 n_2 、 n_{2i} : 座位 i に関する対立遺伝子1及び対立遺伝子2のリード数

n 、 $n_i = n_{1i} + n_{2i}$: 座位 i に関する2つの公知の対立遺伝子の総リード数

g_{11} 、 g_{12} 、 g_{11i} 、 g_{12i} : レシピエントの遺伝子型、即ちレシピエントのゲノムにおける対立遺伝子1及び2のコピー

g_{21} 、 g_{22} 、 g_{21i} 、 g_{22i} : ドナーの遺伝子型、即ちドナーのゲノムにおける対立遺伝子1及び2のコピー

$B(\cdot)$: ベータ関数

$Beta(\cdot)$ 、 $BN(\cdot)$ 、 $Pois(\cdot)$ 、 $Gamma(\cdot)$: ベータ分布、二項分布、及びポアソン分布、及びガンマ分布

N : 試料中でDNAを供給した細胞の数

50

β : 試料中のドナー由来のものであるDNAのパーセンテージと定義された、ドナーDNA割合。

$n_r = N \cdot (1 - \beta)$ 、 $n_d = N \cdot \beta$: 試料中のDNAを供給したレシピエント及びドナー細胞の数
【0170】

【表4】

Table 4: 一般的実験パイプラインでの 3 つの主な構成要素に関する統計的モデル。各構成要素に関するモデルは、事前の構成要素に対して条件付けられる。モデルは、各座位当たりのものである。

血液試料から抽出された gDNA 又は cfDNA	対立遺伝子 1 のコピー: $n_1'' \sim \text{Pois}(c \cdot p_1)$	対立遺伝子 2 のコピー: $n_2'' \sim \text{Pois}(c \cdot p_2)$
	座位の総コピーが与えられた対立遺伝子 1 のコピー: $n_1'' n'' \sim \text{BN}(n'', p_1)$ 、ここで $n'' = n_1'' + n_2''$ 。	
PCR 増幅 DNA	対立遺伝子 1 のコピー: $n_1' \sim \text{Gamma}(n_1'' \cdot p, \theta)$	対立遺伝子 2 のコピー: $n_2' \sim \text{Gamma}(n_2'' \cdot p, \theta)$
	PCR 産物中の座位の対立遺伝子 1 の割合であり、抽出された DNA 中の対立遺伝子 1 及び対立遺伝子 2 のコピーに対して条件付けている: $n_1' / n' n_1'', n_2'' \sim \text{Beta}(n_1'' \cdot p, n_2'' \cdot p)$ 、ここで $n' = n_1' + n_2'$ 。 DNA サンプリングの変動を無視する(したがって、 $n_1'' = n'' \cdot p_1$, $n_2'' = n'' \cdot p_2$): $n_1' / n' \sim \text{Beta}(n'' \cdot p \cdot p_1, n'' \cdot p \cdot p_2)$	
座位にマッピングされた配列決定リードのコピー(配列決定誤差なし)	座位の総コピーが与えられた対立遺伝子 1 のコピーであり、PCR 産物中の座位の対立遺伝子 1 の割合に対して条件付けている: $n_1 n, n_1' / n' \sim \text{BN}(n, p = n_1' / n')$ 、ここで $n = n_1 + n_2$ 。	

【0171】

DNA抽出モデル

cfDNA又は細胞DNAが血液試料から抽出されたとき、得られたDNAは、DNAの大きいプールからの小さい試料であり、したがって実現例は、2つのポアソン分布として、各座位で、2つの対立遺伝子の数をモデル化する。したがって、総数 n'' に対して条件付けられる座位での対立遺伝子1に関するDNAコピー(n_1'')は、二項分布: $n_1'' \sim \text{BN}(n'', p_1)$ に従い、平均 $\mu_0 = n'' \cdot p_1$ 及び分散 $\delta_0^2 = n'' \cdot p_1 \cdot p_2$ である。ドナー割合 $\beta < 0.2$ 、 $\delta_0^2 \ll \mu_0$ 。

【0172】

gDNAが試料から抽出されたとき、各座位ごとに得られたgDNAの量は、抽出損失に起因して再び変数とすることができる。 p_1 を、入力試料中の対立遺伝子1の割合として見れば、抽出されたDNA中の対立遺伝子1の量は、二項分布: $n_1'' \sim \text{BN}(n'', p_1)$ によって再びモデル化することができる。

【0173】

PCR増幅モデル

本発明者等は、PCR産物中の対立遺伝子1の数の確率分布を得るために、確率過程としてPCR増幅プロセスをモデル化する。 x_t を、PCR増幅の t サイクル後の所与の対立遺伝子のDN

Aコピーとし、 r_t を、 t サイクルでの増幅率とし、 y_t を、 t サイクルで発生した新しいコピーとする。DNAの各小片が、増幅されDNAプールに加えられた確率 r_t を有すると仮定することにより、実現例は、増幅に関する下記のモデルを有する:

$x_{t+1}=x_t+y_{t+1}$ 、ここで $y_{t+1} \sim \text{BN}(x_t, r_{t+1})$ は、パラメータとして x_t 及び r_{t+1} を持つ二項分布に従う。

【0174】

このモデルに基づけば、実現例は、PCR産物中の座位に関するDNAコピー数が、ガンマ分布に近似的に従うと主張される。以下に、その正当性を提示する。

【0175】

工程1:ユール過程(連続時間確率過程)を使用して、PCRを近似する(離散時間確率過程)。PCRプロセス $x_{t+1}=x_t+y_{t+1}$ (式中、 $y_{t+1} \sim \text{BN}(x_t, r_{t+1})$)は、離散時間純出生過程であり: 時間 t の所与のサイクルで、DNAの各コピーは、一部の率 r_t で独立して「出生する」。純出生過程の連続時間バージョンは、ユール・ファリー過程として周知である。連続時間出生過程では、所与の時間 t での座位に関する最終コピー数は、負の二項分布に従うことが公知である。実現例は、PCRサイクル数が1に近くないとき、離散時間出生過程を近似するのに同じ分布を使用することができる。

【0176】

工程2:ガンマ分布(連続分布)を使用して、負の二項分布(離散分布)を近似する。負の二項確率変数(r.v.)は、i.i.d.幾何r.v.sの合計として書くことができる。指数分布は、幾何分布の連続バージョンであることが公知である。したがって、ガンマ分布に従うi.i.d.指数r.v.sの合計は、負の二項である二項r.v.sの合計の連続バージョンである。

【0177】

以下に、実現例は、PCR産物における対立遺伝子数のガンマ分布のパラメータを推定する。

【0178】

総分散の法則 $\text{var}(x_{t+1})=\text{var}(E(x_{t+1} | x)) + E(\text{var}(x_{t+1} | x))$ に基づいて、実現例は、下記の通り x_t の平均及び分散を誘導することができる:

$$\begin{aligned} \mu_{t+1} &= \mu_t \cdot (1+r_{t+1}) \\ \delta_{t+1}^2 &= \mu_t \cdot r_{t+1} \cdot (1-r_{t+1}) + \delta_t^2 \cdot (1+r_{t+1})^2, \\ \text{ここで } \mu_t &= E(x_t), \delta_t^2 = \text{var}(x_t). \end{aligned}$$

【0179】

PCRサイクル当たりの平均増幅率を $r_{t+1}=r$ と仮定すると、実現例は、下記を有する。

$$\begin{aligned} \mu_t &= \mu_0 \cdot (1+r)^t \\ \delta_t^2 &= \mu_0 \cdot (1+r)^t \cdot [(1+r)^t - 1] \cdot (1-r) / (1+r) + \delta_0^2 \cdot (1+r)^{2t} \end{aligned}$$

【0180】

μ_0 及び δ_0^2 は、PCR増幅入力におけるDNA対立遺伝子数の平均及び分散であり、それらは上述のDNA抽出モデルに基づいて計算できることに留意されたい。或いは、実現例がcf DNA/細胞DNA対立遺伝子数を確率変数として処理しない場合、実現例は、 $\mu_0=n_1$ "又は n_2 "、及び $\delta_0^2=0$ を有する。

【0181】

この平均及び分散に一致する、対応するガンマ分布 $G(x_t | k, \theta) = k \cdot 1e^{-x/\theta} / [\theta^k \cdot \Gamma(k)]$ は、下記のパラメータを有する:

$$\begin{aligned} \theta &= [(1+r)^t - 1] \cdot (1-r) / (1+r) + \delta_0^2 / \mu_0 \cdot (1+r)^t \\ k &= \mu_0 \cdot (1+r)^t / [(1+r)^t - 1] \cdot (1-r) / (1+r) + \delta_0^2 / \mu_0 \cdot (1+r)^t. \end{aligned}$$

【0182】

2つの対立遺伝子及び2つの初期コピー(n_1 ", n_2 ")を持つ所与の座位では、各座位ごとに2つの対立遺伝子に関して同一の増幅率 $r_1=r_2=r$ を仮定すると、2つの対応するガンマ分布 $G(n_1' | k_1, \theta_1)$ 及び $G(n_2' | k_2, \theta_2)$ が下記のパラメータを有する:

$$\begin{aligned} \theta_1 &= [(1+r)^t - 1] \cdot (1-r) / (1+r) + p_2 \cdot (1+r)^t \\ \theta_2 &= [(1+r)^t - 1] \cdot (1-r) / (1+r) + p_1 \cdot (1+r)^t \end{aligned}$$

$$k_1 = n \cdot p_1 / [(1 - (1+r)^{-t}) \cdot (1-r)/(1+r) + p_2]$$

$$k_2 = n \cdot p_2 / [(1 - (1+r)^{-t}) \cdot (1-r)/(1+r) + p_1]。$$

【0183】

実現例が、PCRモデルをDNA抽出モデルに対して条件付ける場合、s.t. $\mu_0 = n_1''$ 又は n_2'' であり、 $\delta_0^2 = 0$ であり、次いで実現例は下記を有する。

$$\theta_1 = [(1+r)^t - 1] \cdot (1-r)/(1+r)$$

$$\theta_2 = [(1+r)^t - 1] \cdot (1-r)/(1+r)$$

$$k_1 = n_1'' \cdot (1+r)/(1-r) / [1 - (1+r)^{-t}]$$

$$k_2 = n_2'' \cdot (1+r)/(1-r) / [1 - (1+r)^{-t}]。$$

【0184】

したがって、PCR産物中の対立遺伝子コピー n_1' 及び n_2' は、同一のスケールパラメータ θ_1 及び θ_2 を持つ2つのガンマ分布に従い、これらはPCRプロセス(サイクル数及び増幅率)のみに依存する。したがって、

$$n_1' / (n_1' + n_2') \sim \text{Beta}(n_1'' \cdot \rho, n_2'' \cdot \rho)$$

であり、

式中、 $\rho = (1+r)/(1-r) / [1 - (1+r)^{-t}]$ 、又は近似的に $\rho = (1+r)/(1-r)$ は、サイクル数 t が大きいとき、増幅率 r に関する定数であり、これはPCRプロセスのみに依存する。特定の座位では、座位特異的PCR増幅率を捕獲するために $n_{i1}' / (n_{i1}' + n_{i2}') \sim \text{Beta}(n_{i1}'' \cdot \rho_i, n_{i2}'' \cdot \rho_i)$ と書かれる。

【0185】

実現例がDNAサンプリングを無視し、全ての座位が同じ総DNAコピー数 $n_i'' = n$ を有すると仮定する場合、 $n_{i1}'' = n \cdot \rho_i \cdot p_{i1}$ であり、 $n_{i2}'' = n \cdot \rho_i \cdot p_{i2}$ である。PCR産物における座位での対立遺伝子割合は、下記の通りである：

$$n_{i1}' / (n_{i1}' + n_{i2}') \sim \text{Beta}(n \cdot \rho_i \cdot p_{i1}, n \cdot \rho_i \cdot p_{i2})。$$

【0186】

ガンマ分布近似なしでは、実現例は、 $n_1' \sim \text{NB}(r_1, p)$ 及び $n_2' \sim \text{NB}(r_2, p)$ を有し、比 $n_1' / (n_1' + n_2')$ は、閉鎖形分布を持たないことに留意されたい。ガンマ分布近似によれば、 $n_1' \sim \text{Gamma}(n_1'' \cdot \rho, \theta)$ 及び $n_2' \sim \text{Gamma}(n_2'' \cdot \rho, \theta)$ 、及び $n_1' / (n_1' + n_2')$ はベータ分布に従う。

【0187】

リード数に関する配列決定モデル

NGS配列決定は、シーケンサに供給されたDNA分子のプールからサンプリングし、これらの分子の配列を読み出すプロセスである。PCR産物中の座位 i に関する対立遺伝子1の割合は、 $n_{i1}' / (n_{i1}' + n_{i2}')$ である。この割合は、対立遺伝子1リードが配列決定結果で生ずる確率を決定する。座位当たりのリードの総数 n_i に対して条件付けると、 n_{i1} の分布、座位の対立遺伝子1リード数は、二項分布 $n_{i1} \sim \text{BN}(n_i, n_1' / (n_1' + n_2'))$ とモデル化される。

【0188】

事前分布としての寄与体間の遺伝的関連性のモデル化

寄与体(ドナー/レシピエント)遺伝子型が完全に公知である場合、それらは、上述の構成要素モデルのパラメータとして直接組み込むことができる(Table 2(表2)又はTable 3(表3)を使用して)。しかし、遺伝子型が未知である場合、実現例は、臨床の適用例でしばしば入手可能なドナーとレシピエントとの間の遺伝的関係性情報を利用することができる。

【0189】

本発明者等は、種々のタイプのドナー-レシピエント関係性を、ドナー及びレシピエントの可能性のある遺伝子型組合せの空間上での全く別の事前分布として公式化する。ハーディ・ワインベルグ平衡を仮定すると、単一個体に関する遺伝子型分布は $P(g_{\text{Mother}} = [0, 1, 2]) = [(1-\pi)^2, 2\pi(1-\pi), \pi^2]$ であり、このとき対立遺伝子2の母集団頻度を π と仮定する。全ての遺伝的関連性は、親子関係の結果であることに留意されたい。与えられた二対立遺伝子マーカー部位に関する親と子との間の遺伝的関係性に基つき(Table 5(表5))、実現例は、任意の遺伝的関係性に関する接合分布を計算することができる。

10

20

30

40

50

【0190】

【表5】

Table 5: 所与の座位に関する、親の遺伝子型が与えられた子供の遺伝子型の確率分布、並びに類縁ではないと仮定したときの父と母との間の接合分布。

gFather	gMother	遺伝子型に関する子供の確率 それぞれ[0, 1, 2]	P(gFather, gMother)
0	0	[1, 0, 0]	$(1-\pi)^4$
0	1	[1/2, 1/2, 0]	$2\pi(1-\pi)^3$
0	2	[0, 1, 0]	$\pi^2(1-\pi)^2$
1	0	[1/2, 1/2, 0]	$2\pi(1-\pi)^3$
1	1	[1/4, 1/2, 1/4]	$4\pi^2(1-\pi)^2$
1	2	[0, 1/2, 1/2]	$2\pi^3(1-\pi)$
2	0	[0, 1, 0]	$\pi^2(1-\pi)^2$
2	1	[0, 1/2, 1/2]	$2\pi^3(1-\pi)$
2	2	[0, 0, 1]	π^4

【0191】

以下は、下記のタイプの遺伝的関係性:親-子、子-親、兄弟、叔父/叔母-甥、甥-叔父/叔母、及び無関係に関する事前分布である。

【0192】

父と子との間の遺伝子型の接合分布

例として、父-子ドナー-レシピエント遺伝子型(GT)接合分布を、下式を使用して計算する:
 $P(\text{レシピエント}=\text{Me GT}, \text{ドナー}=\text{Father GT}) = \sum \text{mother GT} [P(\text{Me GT} \mid \text{Father GT}, \text{Mother GT}) \cdot P(\text{Father GT}, \text{Mother GT})]$ 、

式中、 $P(\text{Me GT} \mid \text{Father GT}, \text{Mother GT})$ 及び $P(\text{Father GT}, \text{Mother GT})$ の値は、それぞれ、Table 5(表5)の第3及び4欄から得られる。

【0193】

兄弟間の遺伝子型の接合分布

例として、Me-兄弟ドナー-レシピエント遺伝子型接合分布を、親ゲノムが与えられた2つの兄弟遺伝子型条件付き独立性に基づいて、下式を使用して計算する:

$P(\text{レシピエント}=\text{Me GT}, \text{ドナー}=\text{Sibling GT}) = \sum \text{mother GT} \sum \text{father GT} [P(\text{Me GT} \mid \text{Father GT}, \text{Mother GT}) \cdot P(\text{Sibling GT} \mid \text{Father GT}, \text{Mother GT}) \cdot P(\text{Father GT}, \text{Mother GT})]$ 、

式中、 $P(\text{Me GT} \mid \text{Father GT}, \text{Mother GT})$ 、 $P(\text{Sibling GT} \mid \text{Father GT}, \text{Mother GT})$ 、及び $P(\text{Father GT}, \text{Mother GT})$ の値は、Table 5(表5)、第3欄及び第4欄からそれぞれ得られる。

【0194】

叔父-甥の間の遺伝子型の接合分布

例として、叔父/叔母-甥/姪ドナー-レシピエント遺伝子型接合分布を、下式を使用して計算する:

$P(\text{レシピエント}=\text{Me GT}, \text{ドナー}=\text{Uncle GT})$
 $= \sum \text{grand mother GT} \sum \text{grand father GT} \sum \text{mother GT} \sum \text{father GT} [P(\text{Me GT} \mid \text{Father GT}, \text{Mother GT}) \cdot P(\text{Mother GT}) \cdot P(\text{Father GT} \mid \text{GrandFather GT}, \text{GrandMother GT}) \cdot P(\text{Uncle GT} \mid \text{GrandFather}, \text{GrandMother GT}) \cdot P(\text{GrandFather GT}, \text{GrandMother GT})]$

$= \sum \text{mother GT} \sum \text{father GT} P(\text{Me GT} \mid \text{Father GT}, \text{Mother GT}) \cdot P(\text{Mother GT}) \cdot P(\text{Father GT}, \text{Uncle GT})$ 、

式中、 $P(\text{Me GT} \mid \text{Father GT}, \text{Mother GT})$ の値はtable 5(表5)の第3欄から得られ、 $P(\text{Father GT}, \text{Uncle GT})$ は、 $P(\text{レシピエント}=\text{Me GT}、\text{ドナー}=\text{Sibling GT})$ と同じである。

【0195】

上記誘導過程の結果をTable 6(表6)にまとめ、母集団SNP対立遺伝子頻度値 $\pi=0.5$ が与えられた特定の場合をTable 7(表7)に提示する。祖父母-孫、孫-祖父母、片親が異なる兄弟、及びいとこ等の追加の関係性は、同じ基礎をなす原理に基づいて誘導することができる。

【0196】

【表6】

Table 6: 関係する又は無関係のゲノムの事前分布 $P(g_{11}, g_{21})$ 。1)全ての SNP は常染色体由来であり、2)全ての夫婦は遺伝的に無関係であると仮定する。 g_{11} はレシピエントゲノムであり、 g_{21} はドナーゲノムである。

g_{11}	g_{21}	レシピエントに対するドナーの関係性			
		親又は子	兄弟	叔父/叔母又は甥/姪	無関係
0	0	$(1-\pi)^3$	$(1-\pi)^2(1-\pi/2)^2$	$(1-\pi)^3(1-\pi/2)$	$(1-\pi)^4$
0	1	$\pi(1-\pi)^2$	$\pi(1-\pi)^2(1-\pi/2)$	$\pi(1-\pi)^2(3/2-\pi)$	$2\pi(1-\pi)^3$
0	2	0	$\pi^2(1-\pi)^2/4$	$\pi^2(1-\pi)^2/2$	$\pi^2(1-\pi)^2$
1	0	$\pi(1-\pi)^2$	$\pi(1-\pi)^2(1-\pi/2)$	$\pi(1-\pi)^2(3/2-\pi)$	$2\pi(1-\pi)^3$
1	1	$\pi(1-\pi)$	$\pi(1-\pi)[1+\pi-\pi^2]$	$\pi(1-\pi)[1/2+2\pi-2\pi^2]$	$4\pi^2(1-\pi)^2$
1	2	$\pi^2(1-\pi)$	$\pi^2(1-\pi)(1/2+\pi/2)$	$\pi^2(1-\pi)(1/2+\pi)$	$2\pi^3(1-\pi)$
2	0	0	$\pi^2(1-\pi)^2/4$	$\pi^2(1-\pi)^2/2$	$\pi^2(1-\pi)^2$
2	1	$\pi^2(1-\pi)$	$\pi^2(1-\pi)(1/2+\pi/2)$	$\pi^2(1-\pi)(1/2+\pi)$	$2\pi^3(1-\pi)$
2	2	π^3	$\pi^2(1/2+\pi/2)^2$	$\pi^3(1/2+\pi/2)$	π^4

【0197】

10

20

30

40

50

【表 7】

Table 7: SNP 母集団事前 $\pi=0.5$ の場合の関係する又は無関係のゲノムの分布 $P(g_{11}, g_{21})$ 。1)全ての SNP は常染色体由来であり、2)全ての対は遺伝的に無関係であると仮定する。

g_{11}	g_{21}	レシピエントに対するドナーの関係性			
		親又は子	兄弟	叔父/叔母又は甥/姪	無関係
0	0	8/64	9/64	6/64	4/64
0	1	8/64	6/64	8/64	8/64
0	2	0	1/64	2/64	4/64
1	0	8/64	6/64	8/64	8/64
1	1	16/64	20/64	16/64	16/64
1	2	8/64	6/64	8/64	8/64
2	0	0	1/64	2/64	4/64
2	1	8/64	6/64	8/64	8/64
2	2	8/64	9/64	6/64	4/64

【0198】

親/子、兄弟に関する分布は、無関係の場合と全く異なり、一方、叔父/叔母/甥/姪は、無関係に近いことに留意されたい。ドナー遺伝子型が未知の場合、実現例は、上記遺伝的関係性のそれぞれの当て嵌められたモデルの尤度関数を評価することによって、遺伝的関係性を推測することができる。或いは実現例は、遺伝的事前分布における多数の自由パラメータを可能にし(周辺分布がハーディ・ワインベルグ平衡に従うべきであるという追加の拘束がある)、これらのパラメータを、ドナー割合の推定により一緒に推定することができる。

【0199】

モデル化構成要素の一体化

確率的混合物モデルの構成要素を一体化して、寄与体DNA定量(CDQ)の問題の解決策を得る。各SNP部位ごとの母集団対立遺伝子頻度 π は、dbSNP等の公開データベースから得ることができる。最も情報量の多いSNPを選択する場合、即ち $\pi=0.5$ のSNPを選択する場合、全ての座位に関して $\pi=0.5$ を設定することができ、 $P(g_{11}, g_{21})$ は、先のセクションで記述したように遺伝的関連性事前分布とすることができる。

【0200】

概略的レベルで、図2Bは、確率的混合物モデル200の様々な構成要素を例示するブロック図を示す。一部の構成要素は、一部の實現例において任意選択である。確率的混合物モデル200は、配列決定リードの対立遺伝子数をモデル化するための二項分布208を含む。一部の實現例では、確率的混合物モデルは、遺伝的関連性事前分布202を使用して、ドナー・ドニー(又はレシピエント)の関係性をモデル化するための構成要素も含む。一部の實現例では、確率的混合物モデルは、DNA抽出対立遺伝子数をモデル化するための二項分布204も含む。一部の實現例では、確率的混合物モデル200は、PCR産物又は増幅産物対立遺伝子割合をモデル化するためのベータ分布206も含む。ブロック206を参照されたい。

【0201】

一部の實現例では、混合物モデルは、二項分布208と二項分布204とを組み合わせ、DNA抽出誤差と配列決定誤差の両方をモデル化する。そのような實現例では、混合物モデルは、二項分布210を使用して配列決定リードの対立遺伝子数をモデル化し、この配列決定リードの対立遺伝子数は、DNA抽出の対立遺伝子数に依存する。

【0202】

一部の実現例では、確率的混合物モデル200は、ベータ分布206と二項分布208とを組み合わせ、ベータ二項分布212を使用して、PCR又は増幅プロセスでの誤差と配列決定プロセスの誤差との両方をモデル化する。

【0203】

一部の実現例では、確率的混合物モデル200は、二項分布204、ベータ分布206、及び二項分布208を組み合わせ、DNA抽出、増幅プロセス、及び配列決定プロセスからそれぞれ生ずる分散を説明する。そのような実現例では、確率的混合物モデル200はまず、ベータ分布214を使用して、二項分布204及びベータ分布206の作用を近似する。次いで確率的混合物モデル200は、ベータ二項分布216を使用して、ベータ分布214と二項分布208とを組み合わせる。

10

【0204】

配列モデル

完全モデルの基本バージョンは、DNA抽出モデル及びPCRモデルを無視し、配列決定モデルを考慮するだけである。各座位ごとに、参照対立遺伝子に関する配列決定リード数を二項分布(図1C)、 $n_{i1} \sim \text{BN}(n_{i1}, p_{i1})$ によってモデル化し、ここでパラメータ $p_{i1}(g_{11}, g_{21}, \lambda, \beta)$ の値は、座位に関するドナー-レシピエント遺伝子型組合せの関数である(Table 2(表2)及びTable 3(表3))。遺伝子型が未知であるとすれば、実現例は、事前分布として $P(g_{11}, g_{21} | \pi)$ を持つ各座位ごとに9つの可能性のある遺伝子型組合せを周辺化する(Table 6(表6)及びTable 7(表7))。全座位にわたる完全尤度関数は、全ての座位に関する周辺分布の積である:

20

$L(n_1, n_2 | \beta, \pi) = \prod \sum g_{11} g_{21} \text{BN}(n_{i1}, p_{i1}(g_{11}, g_{21}, \lambda, \beta)) \cdot P(g_{11}, g_{21} | \pi)$ 、ここで π は公知のパラメータであり、 β はドナーDNA割合である。

【0205】

抽出-配列複合モデル

より進化したモデルは、DNA抽出モデル並びに配列決定モデルを組み合わせる。実現例は、PCR工程を無視し(即ち、各座位ごとに、PCR産物中の対立遺伝子割合は、DNA試料中の対立遺伝子割合と同じであると仮定する)、DNAサンプリング及び配列決定工程をモデル化するだけである。各座位ごとに、入力DNA試料中の対立遺伝子数に関する二項分布がある。これはNGS配列決定に提供された入力DNA中の対立遺伝子割合の座位ごとの変動を獲得する。

30

【0206】

DNA抽出モデルに関し、実現例は、DNA抽出モデルに対して条件付けながら $n_{i1} \sim \text{BN}(n, p_{i1})$ を有し、配列決定モデルは $n_{i1} | n_{i1}''$ 、 $n_{i1}'' \sim \text{BN}(n_{i1}, n_{i1}''/n_{i1}'')$ であり、ここで $n_{i1}''=n$ は、入力DNAが対応する一倍体ゲノムのコピーである。残念ながら、 n_{i1} の周辺分布は閉鎖形の式を持たず、実現例は、ベータ分布Beta(a, b)で n_{i1}''/n の分布を近似するように選択し、最良のベータ分布は、 n_{i1}''/n の平均及び分散を、二項モデル $n_{i1} \sim \text{BN}(n, p_{i1})$ から誘導されたものと一致させることによって選択される:

$$p_{i1} = a / (a + b)$$

$$p_{i1} \cdot (1 - p_{i1}) / n = ab / (a + b)^2 / (a + b + 1).$$

40

【0207】

方程式を解くことにより、ベータ分布Beta((n-1)p_{i1}, (n-1)p_{i2})が最良の近似として得られる。次いでDNA抽出モデルに対するこの近似により、 n_{i1} の周辺分布は下記の形のベータ二項分布に従う:

$$\text{BB}(n_{i1}, n_{i2} | n, (n-1) \cdot p_1(g_{11}, g_{21}, \lambda, \beta), (n-1) \cdot p_2(g_{11}, g_{21}, \lambda, \beta)).$$

【0208】

次いで遺伝的関連性の事前確率を考慮する、対応する完全尤度関数は、下記の通りである:

$$L(n_1, n_2 | \beta, n, \lambda, \pi) = \prod \sum g_{11} g_{21} \text{BB}(n_{i1}, n_{i2} | n, (n-1) \cdot p_1(g_{11}, g_{21}, \lambda, \beta), (n-1) \cdot p_2(g_{11}, g_{21}, \lambda, \beta)) \cdot P(g_{11}, g_{21} | \pi).$$

【0209】

50

n'' 及び $\pi=0.5$ は共に公知のパラメータであり、最終完全尤度関数は単一の未知のパラメータ β 、ドナーDNA割合のみ有することに留意されたい。

【0210】

入力DNA(一倍体)コピー数 n'' は、入力DNA質量から誘導することができる。入力DNA量が8ngである場合、 $n''=8\text{ng}/[3.59 \times 10^{-3}\text{ng/コピー}]=2228.412$ である。

【0211】

PCR-配列複合モデル

DNA抽出モデルを無視し、所与の座位に関する公知の遺伝子型組合せを仮定すると、PCRモデル: $n_{i1}'/(n_{i1}'+n_{i2}') \sim \text{Beta}(n'' \cdot p_i \cdot p_{i1}, n'' \cdot p_i \cdot p_{i2})$ 及び配列決定モデル $n_{i1} \sim \text{BN}(n_i, n_{i1}'/(n_{i1}'+n_{i2}'))$ は、ベータ二項分布: $\text{BB}(n_{i1}, n_{i2} \mid \eta, n'' \cdot p_i \cdot p_{i1}, n'' \cdot p_i \cdot p_{i2})$ に組み合わせることができる。両方の基礎をなす座位特異的PCR増幅率 p_i は、未知であることに留意されたい。実現例が、全ての座位が同じ固有の増幅率を有すると仮定する場合、実現例は、 $\text{BB}(n_{i1}, n_{i2} \mid \eta, c \cdot p_{i1}(g_{11}, g_{21}, \beta), c \cdot p_{i2}(g_{11}, g_{21}, \beta))$ を有する。

10

【0212】

次いで全座位にわたる完全尤度モデルは下記の通りであり: $L(n_1, n_2 \mid \beta, c, \pi) = \prod \sum g_{11} g_{21} [\text{BB}(n_{i1}, n_{i2} \mid \eta, c \cdot p_1(g_{11}, g_{21}, \lambda, \beta), c \cdot p_2(g_{11}, g_{21}, \lambda, \beta)) \cdot P(g_{11}, g_{21} \mid \pi)]$ 、ここで c 及び β は、推定される2つのパラメータである。

【0213】

或いは、実現例は、各座位の相対増幅率が座位当たりの総リードに比例すると定義することができ、 $\text{BB}(n_{i1}, n_{i2} \mid \eta, c' \cdot n_i \cdot p_{i1}(g_{11}, g_{21}, \beta), c' \cdot n_i \cdot p_{i2}(g_{11}, g_{21}, \beta))$ としてベータ二項を再度パラメータ化することができる。

20

【0214】

次いで全ての座位にわたる完全尤度モデルは下記の通りであり: $L(n_1, n_2 \mid \beta, c', \pi) = \prod \sum g_{11} g_{21} [\text{BB}(n_{i1}, n_{i2} \mid \eta, c' \cdot n_i \cdot p_1(g_{11}, g_{21}, \lambda, \beta), c' \cdot n_i \cdot p_2(g_{11}, g_{21}, \lambda, \beta)) \cdot P(g_{11}, g_{21} \mid \pi)]$ 、ここで c 及び β は、推定される2つのパラメータである。

【0215】

抽出-PCR-配列複合モデル

抽出-PCR-配列決定の一般的実験パイプラインにおける3つ全ての構成要素は、実現例がDNA抽出とPCRモデルとを1つのモデルに組み合わせ且つそれを単一ベータ分布によって近似する場合、ベータ二項によって一緒にモデル化することができる。直感的に、PCR産物中の対立遺伝子1割合の予想値(n_{i1}'/n' 、Table 4(表4)参照)は p_1 のままであるが、 n_{i1}'/n' の不確かさ(分散)はDNA抽出及びPCR工程の両方に由来する。DNA抽出及びPCRと一緒にモデル化しようベータ分布 $\text{beta}(a, b)$ を得るために、実現例は、下記の法則: $E(n_{i1}'/n') = E(E(n_{i1}'/n' \mid \eta_{i1}'/n'))$ 、及び $\text{var}(n_{i1}'/n') = \text{var}(E(n_{i1}'/n' \mid \eta_{i1}'/n')) + E(\text{var}(n_{i1}'/n' \mid \eta_{i1}'/n'))$ に基づいて、 n_{i1}'/n' の無条件平均及び分散を計算する。これは: $E(n_{i1}'/n') = p_{i1}$ 、及び $\text{var}(n_{i1}'/n') = p_{i1}p_{i2}/n'' + p_{i1}p_{i2}/(n'' \cdot p_{i1} + 1) - p_{i1}p_{i2}/[n'' \cdot (n'' \cdot p_{i1} + 1)]$ を与え、式中、 $p_i = (1+r_i)/(1-r_i) > 1$ は、増幅率 r_i に関連する定数である。 n'' は大きいので、実現例は下記の近似 $\text{var}(n_{i1}'/n') = p_{i1}p_{i2}/[n'' \cdot (1+r_i)/2]$ を有する。次いでDNA抽出及びPCRをモデル化する最良のベータ分布は、 $\text{Beta}([n'' \cdot (1+r_i)/2 - 1]p_{i1}, [n'' \cdot (1+r_i)/2 - 1]p_{i2})$ である。これは、cfDNA/gDNA抽出に関するベータ分布 $\text{Beta}((n''-1)p_{i1}, (n''-1)p_{i2})$ に近く、それでも分散はここでより大きいことに留意されたい。 $r_i=0.8$ から0.95である典型的なPCR反応では、実現例は $n'' \cdot (1+r_i)/2 = 0.9 \cdot n''$ から $0.975 \cdot n''$ を有する。

30

40

【0216】

cfDNA-PCR-Seqモデルに関する完全尤度関数は、下記の通りである:

$L(n_1, n_2 \mid \beta, n'', \pi) = \prod \sum g_{11} g_{21} \text{BB}(n_{i1}, n_{i2} \mid \eta, (n'' \cdot (1+r_i)/2 - 1) \cdot p_1(g_{11}, g_{21}, \lambda, \beta), (n'' \cdot (1+r_i)/2 - 1) \cdot p_2(g_{11}, g_{21}, \lambda, \beta)) \cdot P(g_{11}, g_{21} \mid \pi)$ 。

【0217】

寄与体核酸割合及びそれらの信頼区間の推定に関するアルゴリズム

50

寄与体DNA割合を推定するための数値の最適化

寄与体DNA割合 β は、全尤度関数 $L(n_1, n_2 | \beta)$ を最大限にする値と推定される。上述のように、DNAがここで及びその他の実施例で言及されるが、RNA及びその他の核酸分子を同様にプロセッシングし分析してもよい。また、実施例は核酸混合物試料に言及するが、試料は、単一寄与体の核酸のみを含んでいてもよく、その場合、寄与体割合は1と推定され又は1からの誤差の範囲内と推定することが可能である。 $L(n_1, n_2 | \beta)$ の計算中、多数の小さい確率値が乗じられる。小さい確率を乗じたときに数値のアンダーフローを回避するために、実現例は、全ての合算及び乗算を対数スケールで行う。対数スケール上での小さい確率の合計は、下記の通り行われる。1) 対数確率の最大値を $x_{\max}$ として得る; 2) 最大値により全ての対数確率値から差し引く; 3) 得られた値を累乗し次いで合計する; 4) 得られた合計を対数変換する; 5) 対数確率の最大値を加減する。 $\log(\exp(x_1 - x_{\max}) + \exp(x_2 - x_{\max}) + \dots + \exp(x_n - x_{\max})) + x_{\max}$ 。

【0218】

負の値を回避するために、変換式 $\beta = 1/(1 + e^{-\eta})$ を使用し、極小値を回避するために、完全尤度関数を $\beta_0 = 1/(1 + e^{-\eta_0})$ で初期化し、ここで η_0 は、-10、-9.9、-9.8、...、-0.1、0 の中の値であり、これは $L(n_1, n_2 | 1/(1 + e^{\eta_0}))$ を最大限にするものである。 η の、他の数値最適化は、BFGS-準ニュートン法を使用して行われ、これは $-\log_2(L)$ を最小限に抑えるのに使用する。

【0219】

信頼区間の推定

推定値の信頼区間の下限は、クラメール・ラオの不等式: $\text{var}(\theta_{ML}) \geq 1/I(\theta_{ML})$ に基づいて決定され、式中、 θ_{ML} は、パラメータ θ の最大尤度推定値であり、 $I(\theta_{ML})$ は、 θ_{ML} でのフィッシャー情報である。これに基づいて、上述の尤度関数における β 及び c の分散を推定することができる。標準誤差は、クラメール・ラオの限界に従って $\sqrt{1/H}$ と推定され、ここで H は、近似することができ且つ BFGS-準ニュートン法で推定されるヘッシアン行列である。

【0220】

本発明者等は、 β 及び c を推定するために、数値最適化の最中に以下の再パラメータ化を使用する。

$$\beta = 1/(1 + e^{-\eta}),$$

$$c = e^{\kappa}.$$

【0221】

$I(\eta)$ 及び $I(\kappa)$ を、パラメータ化 η 及び κ の下でのフィッシャー情報とし、次いで当初のパラメータのフィッシャー情報は下記の通りである。

$$I(\beta) = I(\eta)(1/(\beta(1-\beta)))^2$$

$$I(c) = I(\kappa)(1/c)^2.$$

【0222】

したがって実現例は、推定される標準偏差に関して下記の変換式を有する。

$$\text{std}(\beta) = \text{std}(\eta) \cdot \beta \cdot (1 - \beta)$$

$$\text{std}(\beta) = \text{std}(\eta) \cdot c.$$

【0223】

核酸試料をデコンボリューションする法医学的適用

混合物デコンボリューション: 観察されたカウント数を D とし、寄与体頻度 f 及び座位当たりの遺伝子型構成 G を推測する

下記の実現例は、法医学的適用例に適切である。本明細書に記述されるプロセスは、まず、それらの事前確率上でシータ及び G を周辺化する間に(以下に記述する)最大尤度推定値 f を得、次いでそれらの推定値に対して条件付けて全ての遺伝子型構成に関する事後確率を計算する。プロセスは、妥当とされる遺伝子型構成のトップ- N リスト及び関連する確率と共に、 f の MLE を報告する。遺伝子型構成は、座位当たりで(全ての寄与体にわたって)及び座位当たりの寄与体当たりで報告される。このセクションの下での寄与体頻度 f は、上述の

10

20

30

40

50

寄与体割合 β に対応する。Gは、寄与体の遺伝子型を示し、上述の g_{11} 、 g_{21} に対応する。データDは、上記対立遺伝子数nに対応する。

【0224】

試料包含クエリ: 観察されたカウント数D、推測された点推定値f、及びクエリ遺伝子型を得て、クエリ遺伝子型が試料中に存在するか否かを推測する。

手法: プロセスでは、一般母集団から引き出された(公知の母集団対立遺伝子頻度を使用して)N-寄与体試料中にあるクエリ遺伝子型の事前確率 $P(G_Q)$ と、観察された試料中にある(母集団からランダムに引き出されたその他の寄与体と共に)クエリ遺伝子型の事後確率 $P(G_Q | D)$ とを計算し、次いでこれら2つの確率間の対数比を、証拠の尺度として報告する。 $P(G_Q | D) = P(D | G_Q) P(G_Q) / P(D)$ であることに留意され、したがって報告された証拠は(均等に)尤度比 $P(D | G_Q) / P(D | G_{\text{random}})$ として記述することもできるようになるが、それはデータ $P(D)$ の周辺確率が、母集団からランダムに引き出された寄与体に対して条件付けられるデータの確率 $P(D | G_{\text{random}})$ と同じものだからである。

【0225】

推測手法

中心となる計算は、入力としてf及びシータの値を得る関数であり、全ての座位に関して、f及びシータに対して条件付けられるデータの周辺対数確率(即ち、G上で周辺化された対数尤度)を、遺伝子型特異的確率上で合計(遺伝子型事前確率によって重みが付けられた)を行うことによって計算する。次いでこれらを、シータに関する(離散型の同じ重みの)分布上で合計して、G及びシータ上で周辺化された対数尤度を得る。実現例は、Gの個々の値に関して合計(シータ上)も保持し、したがって(正規化後に)実現例は、全ての座位にある全ての遺伝子型に関して事後確率を有するようになるが、これはfを条件とし、しかしシータ上で周辺化されたものである。座位当たりの周辺対数尤度は累積され、全データ集合に関する単一对数尤度として戻され、それでもfに対して条件付けられるものである。実現例は、この計算を、頻度ベクトルfに関して可能性のある値の格子の全ての点で行い、尤度が最大である格子点をピックアップすることによってfに関するMLEを得る。頻度格子を、均等に間を空けた2.5%の間隔で設定し、それに加えて全ての頻度での余分な点を均等に設定し(既に表れていない場合)、このとき頻度は非昇順で列挙され且つ合計で1になるという拘束がある。これは2つの寄与体に関して21点の格子を、又は3つの寄与体に関して155点の格子をもたらし(4つの寄与体の場合はまだ実現されておらず、より粗い格子が必要となる可能性があり; 3つの寄与体の場合は、最初により粗い格子を使用し、次いで格子の興味ある部分でズーム・インした後に第2の工程として精製することによって、高速化することもできる)。

【0226】

fのMLEに対して条件付けられる、遺伝子型構成の座位当たりの事後確率は、混合物デコンボリューションクエリのために及び試料クエリのために使用される。

【0227】

周辺尤度計算

全対数尤度は、座位特異的対数尤度の合計であり: $\log P(D | \theta, f) = \sum \log P(D_i | \theta, f)$ である。

【0228】

座位特異的周辺尤度は、妥当とされる遺伝子型構成の大きい集合を合計することによって計算される: $P(D_i | \theta, f) = \sum_{G_i} P(D_i | \theta, f, G_i) P(G_i)$ 。計算を容易にするために、下記の操作を用いる。

【0229】

閾値外対立遺伝子数 ≤ 1 。プロトタイプ1では、歴史的理由のため、実現例は、遺伝子型構成の刈り込んでいない及び刈り込んだリストの両方を構築する。刈り込んでいないリストのみが、閾値未満の対立遺伝子を持つ構成を含有し; 一部の実現例は、以下に記述する刈り込んだリスト及び更に短い(より積極的に刈り込んだ)バージョンのみを使用する。

【0230】

「必要とされる」対立遺伝子のリストを構築し、それらの、少なくとも1回だけ全ての必要とされる対立遺伝子を含む遺伝子型構成のみを列挙することによって、妥当と見なされる遺伝子型構成を限定する。対立遺伝子は、一般的な「ドロップイン」N-1スタッターを介してそのUMI数が妥当とされるには高過ぎると判断された場合、必要とされる対立遺伝子のリスト上に配置される。下記のハードな閾値が使用される。

【0231】

絶対閾値(カウント_閾値; 10に設定)。これは非スタッターUMIの最大数であり、実現例は、任意の寄与体に存在しない対立遺伝子を説明しようとするものである。

【0232】

相対閾値(スタッター_閾値; 0.1に設定)。潜在的なスタッタードナーの数を乗じたこの値(下記の「スタッターの取扱い」参照)はUMIの最大数であり、実現例は、N-1スタッターとして説明しようとするものである。

【0233】

観察されたカウント数が、上記2つの閾値の合計よりも高い場合、実数と見なされ、全ての遺伝子型構成に存在しなければならない。そうでない場合には省略されてもよい(より大きい数の遺伝子型構成を考慮したままにする)。

【0234】

より計算上で求められている部分で使用される(即ち、最終デコンボリューション工程ではなく、寄与体頻度を推測する場合)、積極的に刈り込んだリストを構築する:

【0235】

各構成ごとに:

- ・ 各寄与体頻度ベクトルごとに、実現例は、予想される対立遺伝子当たりのリード数と観察された数とを比較し、対立遺伝子に関してRMSEを計算する
- ・ この計算は、スタッターを考慮に入るが、完全尤度計算よりも非常に安価である
- ・ 頻度ベクトルの最小RMSEを使用する(構成が妥当とされる頻度ベクトルがある?)。「最良のRMSE」(下記参照)では、実現例は、頻度ベクトルの平均を使用する(全体としてデータ集合が与えられた、完全に非現実的なものとなり得る最良の当て嵌め頻度ベクトルを使用する場合はブレイクする)。
- ・ 観察が、以下の基準の両方により予想に十分近い場合は、構成を保持する:
- ・ 絶対閾値(0.2): RMSEが最大リード数のこの割合よりも大きい構成を刈り込む
- ・ 相対閾値(5): RMSEが「最良のRMSE」(上記参照)からのこの因子よりも大きい構成を刈り込む

【0236】

妥当とされる遺伝子型構成の集合を、それらの事前確率 $P(G_i)$ (下記参照)と共に、予備プロセシング中に構築し、尤度計算がコールされる度に再使用する。

【0237】

遺伝子型特異的尤度

生成モデルは、「当初の」試料中の座位当たりの潜在的に検出可能な分子の固定数を規定する(犯罪現場から又はプロセシングの後半段階で収集された物理的試料に対応し得る)。実現例は、それら対立遺伝子が遺伝子型構成に割り当てられた寄与体の寄与体頻度に比例して、これらの分子が対立遺伝子当たりで分割されると仮定する。次いでこれらの分子のそれぞれが、検出され又は検出されず、したがって所与の対立遺伝子に関して検出された分子の数は二項プロセスによって支配されるようになる。検出確率(即ち、二項パラメータ)は、対立遺伝子ごと(及び座位ごと)に様々になってもよく、全ての座位で、実現例は、2つのパラメータによって支配されるベータ事前確率を割り当てる:

平均(平均検出確率)

包含範囲(検出されたUMIの総数)は座位ごとに様々である。原則として、このことは、ベータ分布の平均が座位ごとに変わるべきであることを意味する。一部の實現例は、座位全体にわたって固定された平均を保持し、代わりに試料中の分子の数を座位ごとに変化させる。このことは、同じにものになるべきである(2つのパラメータは高度に相関することが

10

20

30

40

50

予想されるので、それらを個別のパラメータとして処理することはそれらをほとんど確認されないものとすると考えられる)。平均パラメータは、値0.1にハードコードされ;分子の総数は、この平均パラメータ並びにスタッター率(下記参照)を考慮して、観察された包含範囲から外挿することによって設定される。

【0238】

濃度パラメータ(検出確率が対立遺伝子ごとにどの程度変化するか:これは、実験室で測定された対立遺伝子バランスに密接に関係するが同一ではない)

【0239】

このパラメータは座位ごとに変化し、3成分離散事前確率(座位全体にわたり共用される)が割り当てられ、低、中、及び高ノイズ条件を表す。

【0240】

その他のノイズパラメータは下記の通りである:

【0241】

スタッター_確率:上記外挿された分子の総数のそれぞれに関し(対立遺伝子N)、生成モデルは、確率スタッター_確率でUMI(対立遺伝子N-1)を生成することになると規定する。

【0242】

このパラメータは、同じ座位にある対立遺伝子間で共用されるが、座位ごとに変化する。3成分離散事前確率(座位全体で共用される)が割り当てられ、低、中、及び高ノイズ条件を表す。

【0243】

予想_ドロップイン:これは、対立遺伝子で観察された見せ掛けUMIの、予想された数である(その対立遺伝子の又はスタッタードナーの分子によって生成されない)。パラメータは、ハードコードされ(推測されない)、対立遺伝子及び座位全体にわたり共用される。

【0244】

次いで上記ノイズパラメータ(シータ₁を含む)が与えられると、遺伝子型特異的尤度 $P(D_i | t, \theta_i, f, G_i)$ を、対立遺伝子特異的尤度の積として計算する。

【0245】

座位に関する尤度計算中に、一部の実現例は多くの遺伝子型構成を試み、その一部は、少しの対立遺伝子で互いに異なるだけである。その結果、対立遺伝子特異的尤度はしばしば、以前使用されていたものと全く同じカウント数に正確に必要なとされる。実現例は、ルックアップテーブルに全ての対立遺伝子特異的尤度計算の結果を記憶し、それらが既にテーブルにない場合には尤度を計算するだけである。

【0246】

本発明者等は、2つの対立遺伝子特異的尤度計算を裏付ける:単純(より速い)スタッターフリー計算、及びスタッターを考慮する完全計算。

【0247】

対立遺伝子特異的尤度:スタッターフリーバージョン

対立遺伝子kに関する尤度計算のスタッターフリーバージョンは、「実際の」(非スタッター)検出の予想量が非ゼロであるとき(少なくとも1つの寄与体の遺伝子型でのkに起因して)、及びスタッター検出の予想量(スタッター率、及び「スタッター隣接」対立遺伝子k+1に割り当てられた分子の数から計算された)が閾値よりも下であるときに、スタッター率がゼロであり且つ計算上のショートカットとしても使用される場合に、適用可能である。

【0248】

最終的に、対立遺伝子kに割り当てられた分子のそれぞれは、二項としてモデル化されたプロセスを介して、UMIとして検出され又は検出されず、即ち、個々の分子の検出は独立して行われる。ベータ分布頻度パラメータ及び潜在的に検出可能な分子の対立遺伝子特異的数を持つ二項プロセスは、全ての対立遺伝子で観察されたUMI数に関するベータ二項分布を示唆する。したがって対立遺伝子特異的尤度は、ベータ二項分布に関する式を使用して計算される。

【0249】

10

20

30

40

50

本発明者等は、観察された対立遺伝子のためだけに確率を計算し;観察されている非ゼロ確率を有する観察されていない対立遺伝子から生ずるべきペナルティは、無視される。

【0250】

スタッター及びその他のドロップインの取扱い

第2の尤度計算は、スタッターが関連あるものとして決定される少数の場合に使用される(上記参照)。UMI数がMである対立遺伝子kで、一部の実現例は、0からMまでの全ての値mを、問題となっている対立遺伝子由来のUMIの数に関して可能性のある値と見なし、残りの(M-m)UMIは対立遺伝子k+1由来である。これらの場合の1つに関する尤度は、真のカウント数に関する尤度と、スタッター数に関する尤度との積である(Nが、対立遺伝子k+1での当初のUMIの数に等しく、二項頻度がスタッター率に等しい、二項モデルの下で)。全尤度は、これらの場合の全ての(線形-ドメイン)合計である。

10

【0251】

実際に、一部の實現例は、上記合計の項の全てを計算することを必要としない。これはスタッター率が小さいからであり、したがってスタッター観察の数の分布は、ゼロに素早く到達する(多数のスタッター観察は本質的に不可能であり、合計における対応する項は、事実上ゼロになる)。一部の實現例は、スタッター観察の数の累積分布の追跡を保持し、残りの確率の重みが閾値よりも低くなったときに合計を終了する。

【0252】

一般的ドロップインは、スタッターが潜在的に生ずる可能性のある分子数を決定するときに、分子の固定数を、対立遺伝子k+1に割り当てられたUMIの数に加えることによって取り扱う。数は、予想_ドロップイン/スタッター_確率になるように設定され、したがってドロップインUMIの予想数は、予想_ドロップインに等しい。

20

【0253】

ドロップアウトの取扱い:

本発明者等は、2つのタイプのドロップアウトを区別する:

自然ドロップアウト:これは、二項プロセスが、寄与体の遺伝子型に存在する対立遺伝子に関してゼロのカウント数をもたらすときである。全ての潜在的対立遺伝子を明らかに表すのではなく(潜在的対立遺伝子の網羅的リストがないので不可能、又は費用がかかる可能性もある)、一部の實現例は、特別な「ダミー」試料外対立遺伝子を使用する。この対立遺伝子は、任意の遺伝子型構成で存在してもよく、ゼロのUMI数を有し、通常の対立遺伝子のように処理される。自然ドロップアウトは、低頻度寄与体の対立遺伝子に生ずる可能性があるが、高頻度寄与体の対立遺伝子には非常に生じ難い。

30

【0254】

試料外対立遺伝子に感覚的事前確率を割り当てるために、一部の實現例は、潜在的対立遺伝子の総数を推量し、これらの潜在的な対立遺伝子に関して均一な事前確率を設定する。したがって試料外事前確率は、観察されていない対立遺伝子の数に比例する。現在、潜在的な対立遺伝子の数に関する推量は、最短及び最長の観察された整数値対立遺伝子の間の全ての整数を内挿し、任意の観察された非整数値対立遺伝子を付加し、得られた値及び5の最大値を戻すことによって得られる。

【0255】

機構的ドロップアウト:一部の實現例は、特別なメカニズムをモデルに組み込み、それによって、対立遺伝子はシーケンサから「見えなくなる」可能性があり(例えば、プライマー領域での変異に起因して)、その場合、一部の實現例は、その総分子数とは無関係であるため、UMIを観察しない(即ち、機構的ドロップアウトは、低頻度寄与体の場合のように、高頻度寄与体で見られるようである)。全ての見えない対立遺伝子の集合は、第2のダミー対立遺伝子によって表される。

40

【0256】

この対立遺伝子は、その尤度がデータとは無関係に1であることを観察することが不可能であるので;接合確率へのその単なる直接寄与は、その事前確率を介する(したがって、低く設定しなければならない)。計算の便宜上、一部の實現例は、見えない対立遺伝子の事前

50

確率として、ハードコードされたパラメータドロップアウト_確率を使用し、見える対立遺伝子の事前確率を、1-ドロップアウト_確率への合計にスケーリングする。このため、予備プロセッシング中に遺伝子型事前確率の計算が可能になり(下記参照)、ドロップアウト確率は、余分な母集団対立遺伝子頻度値として作用する。

【0257】

単一供給源の試料の場合、ホモ接合型対立遺伝子の推測される事後確率は、ドロップアウト_確率の値に強く依存するが、それはモデルが考慮しなければならない主要な代替仮説の確率(1つの対立遺伝子が見えないヘテロ接合体)を決定するからである。一部の実現例は、単一供給源のホモ接合体対立遺伝子に関して得られる事後確率(0.999程度)が妥当であるという直感に基づいて、ドロップアウト_確率を $1e-4$ に校正した。

【0258】

この対立遺伝子は、任意の遺伝子型構成で存在し得る。構成の事前確率を計算するときに特別な処理が与えられ、尤度計算中は無視される。

【0259】

遺伝子型構成の積極的な刈込み中(上記参照;リード数に基づく)、この対立遺伝子の出現には、それらの尤度ペナルティ(ドロップアウト_確率に依存する)から、通常の対立遺伝子に関して推定されたミスマッチ数と同等であることを意味する「カウント」値への、その場に限る変換を使用して、ペナルティが科される。変換は、 $-\log(\text{ドロップアウト_確率})$ の尤度ペナルティを受けるのにピークからどの程度離れていなければならないかを計算することによる、二項に対する通常の近似に基づく。変換は、 $1.2876 * \sqrt{\max(\text{リード_数}) / \text{avg_det_prob}}$ であり、定数は $\sqrt{-2p(1-p)\ln(D)}$ であり、ここで p は、 $\text{avg_det_prob} = 0.1$ であり、 D はドロップアウト-確率 $= 1e-4$ である。包含範囲への依存に留意されたい;一部の實現例は、最も豊富な対立遺伝子の包含範囲を使用する。

【0260】

予備プロセッシング中に生成される遺伝子型構成のリストは、1つ又は複数の対立遺伝子がドロップアウト対立遺伝子である遺伝子型を含む。ドロップアウト対立遺伝子に関して観察されたカウント数は、常に0であり、その尤度は常に1である。高い尤度を有するにも関わらず、モデルは、データを説明するのにドロップアウトメカニズムを頻繁に使用しないが、それは低ドロップアウト確率が低い事前確率を誘発させるからである。

【0261】

事前及び事後計算

個々の寄与体の遺伝子型の事前確率は、ハーディ・ワインベルグのモデルの下で、母集団対立遺伝子頻度から計算する:対立遺伝子頻度が p であるホモ接合遺伝子型に関して $P(G) = p^2$ 、及び対立遺伝子頻度が p 及び q であるヘテロ接合遺伝子型に関して $P(G) = 2pq$ 。多数の寄与体遺伝子型構成の事前確率は、寄与体当たりの事前確率の積である。これは、遺伝子型構成が構築される時に予備プロセッシング中に計算され、全ての周辺尤度計算で再使用される。ドロップアウトを説明するために、一部の實現例は、固定された事前確率を持つダミードロップアウト対立遺伝子を付加して、実験測定に基づいて設定されるようにし;母集団対立遺伝子頻度は、それに応じてディスカウントされる。

【0262】

遺伝子型構成の事後確率は、事前確率に尤度を乗じ、遺伝子型空間全体の合計によって正規化したものである(明示的に計算されるが、それは一部の實現例が全ての遺伝子型構成に関する項を如何様にも計算するからである)。特定の寄与体(例えば、主な寄与体)が特定の遺伝子型を有する事後確率は、その遺伝子型を含有する遺伝子型構成を合計することによって得られる。

【0263】

試料包含クエリの場合、医師は、全ての座位で同じ寄与体であることを前提として、特定された多座位遺伝子型が任意の寄与体で現れるという事後確率に興味を持つ。これは、全ての寄与体 i 上で、寄与体 i が全ての座位で特定された遺伝子型を有する確率(即ち、ループが入れ子状態になった事項である順序)を合計することによって得られる。

【0264】

本明細書で使用される試料は、「無細胞」(例えば、cfDNA)又は細胞結合(例えば、細胞DNA)である核酸を含有する。無細胞DNAを含む無細胞核酸は、血漿、血清、及び尿を含むがこれらに限定されない生体試料から、当技術分野で公知の様々な方法によって得ることができる(例えば、Fan等、Proc Natl Acad Sci 105:16266-16271 [2008]; Koide等、Prenatal Diagnosis 25:604-607 [2005]; Chen等、Nature Med. 2: 1033-1035 [1996]; Lo等、Lancet 350: 485-487 [1997]; Botezatu等、Clin Chem. 46: 1078-1084, 2000;及びSu等、J Mol. Diagn. 6: 101-107 [2004]参照)。無細胞DNAを試料中の細胞から分離するため、分別、遠心分離(例えば、密度勾配遠心分離)、DNA特異的沈殿、若しくは高スループット細胞分別、及び/又はその他の分離方法を含むがこれらに限定されない様々な方法を、使用することができる。cfDNAの手動及び自動化分離に関する市販のキットが入手可能である(Roche Diagnostics社、Indianapolis、IN、Qiagen社、Valencia、CA、Macherey-Nagel社、Duren、DE)。cfDNAを含む生体試料は、染色体異数性及び/又は様々な多型を検出することができる配列決定アッセイによって、染色体異常、例えばトリソミー21の存在又は不在を決定するアッセイで使用されてきた。

【0265】

試料

本明細書で使用される試料は、「無細胞」(例えば、cfDNA)又は細胞結合(例えば、細胞DNA)された核酸を含有する。無細胞DNAを含む無細胞核酸は、血漿、血清、及び尿を含むがこれらに限定されない生体試料から、当技術分野で公知の様々な方法によって得ることができる(例えば、Fan等、Proc Natl Acad Sci 105:16266-16271 [2008]; Koide等、Prenatal Diagnosis 25:604-607 [2005]; Chen等、Nature Med. 2: 1033-1035 [1996]; Lo等、Lancet 350: 485-487 [1997]; Botezatu等、Clin Chem. 46: 1078-1084, 2000;及びSu等、J Mol. Diagn. 6: 101-107 [2004]参照)。無細胞DNAを試料中の細胞から分離するため、分別、遠心分離(例えば、密度勾配遠心分離)、DNA特異的沈殿、若しくは高スループット細胞分別、及び/又はその他の分離方法を含むがこれらに限定されない様々な方法を、使用することができる。cfDNAの手動及び自動化分離に関する市販のキットが入手可能である(Roche Diagnostics社、Indianapolis、IN、Qiagen社、Valencia、CA、Macherey-Nagel社、Duren、DE)。cfDNAを含む生体試料は、染色体異数性及び/又は様々な多型を検出することができる配列決定アッセイによって、染色体異常、例えばトリソミー21の存在又は不在を決定するアッセイで使用されてきた。

【0266】

様々な実施形態では、試料中に存在するDNAは、使用前に特異的に又は非特異的に濃縮することができる(例えば、配列決定ライブラリを調製する前に)。試料DNAの非特異的濃縮は、DNA配列決定ライブラリを調製する前に試料DNAのレベルを上昇させるのに使用することができる試料のゲノムDNA断片の全ゲノム増幅を指す。非特異的濃縮は、複数のゲノムを含む試料中に存在する2つのゲノムのうちの1つの選択的濃縮とすることができる。例えば、非特異的濃縮は、血漿試料中のがんゲノムで選択的とすることができ、これは、試料中の正常なDNAに対するがんの相対的割合を増大させる公知の方法によって得ることができるものである。或いは、非特異的濃縮は、試料中に存在する両方のゲノムの非選択的増幅とすることができる。例えば非特異的増幅は、がん及び正常なゲノムからのDNAの混合物を含む試料中の、がん及び正常なDNAのものとして行うことができる。全ゲノム増幅のための方法は、当技術分野で公知である。縮重オリゴヌクレオチド-プライムPCR(DOP)、プライマー伸長PCR技法(PEP)、及び多置換増幅(MDA)は、全ゲノム増幅方法の例である。一部の実施形態では、種々のゲノムからのcfDNAの混合物を含む試料は、混合物中に存在するゲノムのcfDNAに関して濃縮しない。その他の実施形態では、種々のゲノムからのcfDNAの混合物を含む試料は、試料中に存在するゲノムのいずれか1つに関して非特異的に濃縮する。

【0267】

本明細書に記述される方法が適用される、核酸を含む試料は、例えば上述のように、典型

的には生体試料(「試験試料」)を含む。

【0268】

したがって、ある特定の実施形態では、試料は、精製された又は単離されたポリヌクレオチドを含み若しくはそのようなポリヌクレオチドからなり、又は組織試料、生体液試料、細胞試料、及び同様のもの等の試料を含むことができる。適切な生体液試料には、血液、血漿、血清、汗、涙、痰、尿、痰、耳の流動液、リンパ、唾液、脳脊髄液、洗浄液(ravag e)、骨髓懸濁液、膿液、大腿骨頸部洗浄液、脳液、腹水、乳、呼吸器、腸及び泌尿生殖器の分泌液、羊水、乳、及び白血球フェレーシス試料が含まれるがこれらに限定するものではない。一部の実施形態では、試料は、非侵襲的手順によって容易に得られる試料であり、例えば血液、血漿、血清、汗、涙、痰、尿、痰、耳の流動液、唾液、又は便である。ある特定の実施形態では、試料が末梢血試料、或いは末梢血試料の血漿及び/又は血清画分である。他の実施形態では、生体試料は、スワブ又はスメア、生検試験片、又は細胞培養物である。別の実施形態では、試料は、2つ以上の生体試料の混合物であり、例えば生体試料は、生体液試料、組織試料、及び細胞培養物試料の2つ以上を含むことができる。本明細書で使用される「血液」、「血漿」、及び「血清」という用語は、画分又はそのプロセシング部分を明らかに包含する。同様に、試料が生検、スワブ、スメア等から採取される場合、「試料」は、生検、スワブ、スメア等に由来する、プロセシングされた画分又は部分を明らかに包含する。

10

【0269】

ある特定の実施形態では、試料は、種々の個体からの試料、同じ又は異なる個体の種々の発達段階からの試料、種々の罹患個体(例えば、がんを持ち又は遺伝的障害を有することが疑われる個体)からの試料、正常な個体、個体における疾患の種々の段階で得られた試料、疾患に関する種々の治療を受けた個体から得られた試料、種々の環境因子に供された個体からの試料、病理に対する素因を持つ個体からの試料、感染性疾患物質(例えば、HIV)に曝された個体の試料、及び同様のものを含むがこれらに限定されない供給源から得ることができる。

20

【0270】

例示的な、しかし非限定的な一実施形態では、試料は、妊娠したメス、例えば妊娠した女性から得られる母系試料である。この場合、試料は、胎児における潜在的染色体異常の出生前診断を提供するために、本明細書に記述される方法を使用して分析することができる。母系試料は、組織試料、生体液試料、又は細胞試料とすることができる。生体液には、非限定的な例として、血液、血漿、血清、汗、涙、痰、尿、痰、耳の流動液、リンパ、唾液、脳脊髄液、洗浄液、骨髓懸濁液、膿液、大腿骨頸部洗浄液、脳液、腹水、乳、呼吸器、腸及び泌尿生殖器の分泌液、及び白血球フェレーシス試料が含まれる。

30

【0271】

別の例示的な、しかし非限定的な実施形態では、母系試料は、2つ以上の生体試料混合物であり、例えば生体試料は、生体液試料、組織試料、及び細胞培養物試料の2つ以上を含むことができる。一部の実施形態では、試料は、非侵襲的手順によって容易に得ることが可能な試料、例えば血液、血漿、血清、汗、涙、痰、尿、乳、痰、耳の流動液、唾液、及び便である。一部の実施形態では、生体試料は、末梢血試料、及び/又はその血漿及び血清画分である。他の実施形態では、生体試料は、スワブ又はスメア、生検試験片、又は細胞培養物の試料である。上記で開示されたように、「血液」、「血漿」、及び「血清」という用語は、それらの画分又はプロセシング部分を明らかに包含する。同様に、試料が生検、スワブ、スメア等から採取される場合、「試料」は、生検、スワブ、スメア等に由来する、プロセシングされた画分又は部分を明らかに包含する。

40

【0272】

ある特定の実施形態では、試料は、in vitro培養組織、細胞、又はその他のポリヌクレオチド含有供給源から得ることもできる。培養試料は、種々の媒体及び条件(例えば、pH、圧力、又は温度)で維持された培養物(例えば、組織又は細胞)、種々の期間にわたり維持された培養物(例えば、組織又は細胞)、種々の因子又は試薬(例えば、薬物候補又はモジュレ

50

ーター)で処理された培養物(例えば、組織又は細胞)、又は種々のタイプの組織及び/又は細胞の培養物を含むがこれらに限定されない供給源から採取することができる。

【0273】

生物学的供給源から核酸を単離する方法は、周知であり、供給源の性質に応じて異なることになる。当業者なら、本明細書に記載される方法のために必要に応じて供給源から核酸を容易に単離することができる。ある場合には、核酸試料中の核酸分子を断片化することが有利とすることができる。断片化は、ランダムにすることができ、又は例えば制限エンドヌクレアーゼ消化を使用して実現されるように特異的にすることができる。ランダムな断片化のための方法は当技術分野で周知であり、例えば、限定されたDNase消化、アルカリ処理、及び物理的剪断を含む。一実施形態では、試料である核酸は、断片化に供されないcfDNAから得られる。

10

【0274】

配列決定ライブラリ調製

一実施形態では、本明細書に記載の方法は、次世代配列決定技術(NGS)を利用し得、複数の試料を、ゲノム分子として(すなわち、シングルプレックス配列決定)、又は単一の配列決定実行において、インデックス付きゲノム分子を含むプールされた試料として(例えば、マルチプレックス配列決定)、個別に配列決定することを可能にする。これらの方法は、DNA配列の最大数億のリードを生成し得る。様々な実施形態において、ゲノム核酸の配列、及び/又はインデックス付きゲノム核酸の配列は、例えば、本明細書に記載の次世代配列決定技術(NGS)を使用して決定され得る。様々な実施形態において、NGSを使用して取得された大量の配列データの分析は、本明細書に記載される1つ又は複数のプロセッサを使用して実行され得る。

20

【0275】

様々な実施形態において、そのような配列決定技術の使用は、配列決定ライブラリの調製を必要としない。

【0276】

しかし、特定の実施形態では、本明細書で企図される配列決定方法は、配列決定ライブラリの調製を伴う。一つの例示的なアプローチでは、配列決定ライブラリ調製は、配列決定される準備ができているアダプター修飾DNA断片(例えば、ポリヌクレオチド)のランダムな収集の生成を含む。ポリヌクレオチドの配列決定ライブラリは、DNA又はRNA、例としては、DNA又はcDNAのいずれかの等価物、類似体、例えば、RNAテンプレートから生成された、相補性であるかコピーDNAであるDNA又はcDNAから、逆転写酵素の作用により、調製し得る。ポリヌクレオチドは、二本鎖形態(例えば、ゲノムDNA断片、cDNA、PCR増幅産物等のdsDNA)に由来してもよいし、又は特定の実施形態では、ポリヌクレオチドは一本鎖形態(例えば、ssDNA、RNA等)に由来してもよく、dsDNA形式に変換されている。例として、特定の実施形態では、一本鎖mRNA分子は、配列決定ライブラリの調製における使用に適した二本鎖cDNAにコピーされ得る。一次ポリヌクレオチド分子の正確な配列は、一般にライブラリ調製の方法にとって重要ではなく、既知でも未知でもよい。一実施形態では、ポリヌクレオチド分子は、DNA分子である。より具体的には、特定の実施形態では、ポリヌクレオチド分子は、生物の遺伝的補体全体又は実質的に生物の遺伝的補体全体を表し、ゲノムDNA分子(例えば、細胞DNA、無細胞DNA(cfDNA)等)であり、通常、イントロン配列とエキソン配列(コーディング配列)の両方、及びプロモーターやエンハンサー配列等の非コーディング調節配列を含む。特定の実施形態では、一次ポリヌクレオチド分子は、ヒトゲノムDNA分子、例えば、妊娠対象の末梢血に存在するcfDNA分子を含む。

30

40

【0277】

特定の範囲の断片サイズを含むポリヌクレオチドを使用することにより、いくつかのNGS配列決定プラットフォーム用の配列決定ライブラリの調製が容易になる。そのようなライブラリの調製は、典型的には、所望のサイズ範囲のポリヌクレオチドを得るための大きなポリヌクレオチド(例えば、細胞ゲノムDNA)の断片化を伴う。

50

【0278】

断片化は、当業者に公知の任意の多数の方法によって達成され得る。例えば、断片化は、噴霧、超音波処理、及びハイドロシャー(hydroshear)を含むがこれらに限定されない機械的手段によって達成され得る。しかし、機械的断片化は通常、C-O、P-O、及びC-C結合でDNA骨格を切断し、結果として、切断されたC-O、P-O、及びC-C結合を有する平滑末端並びに3'及び5'オーバーハング末端の不均一な混合が生じ(例えば、Alnemri及びLiwack, J Biol.Chem 265:17323~17333頁[1990];Richards and Boyer, J Mol Biol 11:327~240頁[1965]を参照)、その後の酵素反応、例えば、配列決定のためのDNAの調製に必要な配列決定アダプターの連結に必要な5'-リン酸が不足している可能性があるため、修復が必要な場合がある。

10

【0279】

対照的に、cfDNAは、典型的には約300塩基対未満の断片として存在し、その結果、断片化は通常、cfDNA試料を使用して配列決定ライブラリを生成するために必要ではない。

【0280】

典型的には、ポリヌクレオチドが強制的に断片化される(例えば、インビトロで断片化される)か、又は断片として天然に存在するかにかかわらず、それらは5'-リン酸及び3'-ヒドロキシルを有する平滑末端DNAに変換される。標準プロトコル、例えば、本明細書の他のいずれかに記載されているIlluminaプラットフォームを使用した配列決定のプロトコルは、ユーザに対して、試料DNAの最終修復、dAテーリングの前の最終修復生成物の精製、及びライブラリ調製のアダプター結合ステップの前のdAテーリング生成物の精製を指示する。

20

【0281】

本明細書に記載の配列ライブラリ調製方法の様々な実施形態は、標準プロトコルによって典型的に義務付けられる1つ又は複数のステップを実行して、NGSによって配列決定され得る修飾DNA生成物を得る必要性を省く。略式法(ABB法)、1ステップ法、及び2ステップ法は、その全体が参照により組み込まれる、2012年7月20日に出願された特許出願13/555,037に見出され得る、配列決定ライブラリの調製方法の実施例である。

【0282】

配列決定方法

一部の實現例では、調製された試料(例えば、配列決定ライブラリ)は、核酸の混合物をデコンボリューションするための手順の一部として配列決定される。任意の多数の配列決定技術が利用され得る。

30

【0283】

Affymetrix Inc.社(Sunnyvale, CA)のハイブリダイゼーションによる配列決定プラットフォーム、及び454 Life Sciences社(Bradford, CT)、Illumina/Solexa社(Hayward, CA)及びHelicos Biosciences社(Cambridge, MA)による合成時配列決定プラットフォーム等、並びにApplied Biosystems社(Foster City, CA)のライゲーションによる配列決定プラットフォーム等、後述のいくつかの配列決定技術が市販されている。Helicos Biosciences社の合成時配列決定を使用して実行される単一分子配列決定に加えて、他の単一分子配列決定技術としては、限定するものではないが、Pacific Biosciences社のSMRT(商標)技術、ION TORRENT(商標)技術、及び例えばOxford Nanopore Technologies社が開発したナノポア配列決定が挙げられる。

40

【0284】

自動サンガー法は「第一世代」技術とみなされるが、自動サンガー配列決定を含むサンガー配列決定も、本明細書に記載の方法で使用され得る。追加の適切な配列決定方法としては、限定するものではないが、核酸画像化技術、例えば、原子間力顕微鏡法(AFM)又は透過型電子顕微鏡法(TEM)が挙げられる。例示的な配列決定技術については、以下で詳しく説明する。

【0285】

例示的であるが非限定的な一実施形態において、本明細書に記載の方法は、試験試料中の核酸、例えば母体試料中のcfDNA、癌に関してスクリーニングされている対象のcfDNA又

50

は細胞DNA等の配列情報を、Illuminaの合成時配列決定及びリバーシブルターミネーターベースの配列決定化学を使用して取得することを包含する(例えば、Bentleyら、Nature 6:53~59頁[2009]に記載されているように)。テンプレートDNAは、ゲノムDNA、例えば細胞DNA又はcfDNAであってもよい。一部の実施形態では、単離された細胞由来のゲノムDNAがテンプレートとして使用され、数百塩基対の長さに断片化される。他の実施形態では、cfDNAがテンプレートとして使用され、cfDNAが短い断片として存在するため、断片化は必要とされない。例えば、胎児のcfDNAは、血流中を約170塩基対(bp)の長さの断片として循環し(Fanら、Clin Chem 56:1279~1286頁[2010])、配列決定の前にDNAの断片化は必要としない。循環腫瘍DNAはまた短い断片で存在し、サイズ分布は約150~170bpでピークに達する。Illuminaの配列決定技術は、断片化されたゲノムDNAを、オリゴヌクレオチドアンカーが結合する光学的に透明な平面に付着させることに依存している。テンプレートDNAの末端を修復して5'リン酸化平滑末端を生成し、Klenow断片のポリメラーゼ活性を使用して、単一のA塩基を平滑リン酸化DNA断片の3'末端に付加する。この付加により、ライゲーション効率を高めるために3'末端に単一のT塩基のオーバーハングを有するオリゴヌクレオチドアダプターへのライゲーション用のDNA断片が調製される。アダプターオリゴヌクレオチドは、フローセルアンカーオリゴに相補的である(繰り返し拡大の分析でアンカー/アンカーリードと混同しないよう)。限界希釈条件下では、アダプター修飾された一本鎖テンプレートDNAがフローセルに追加され、アンカーオリゴへのハイブリダイゼーションによって固定される。付着したDNA断片を伸長し、ブリッジ増幅して、それぞれが同じテンプレートの約1,000コピーを含む数億個のクラスターを有する超高密度配列決定フローセルを作成する。一実施形態では、ランダムに断片化されたゲノムDNAは、クラスター増幅に供される前にPCRを使用して増幅される。或いは、増幅フリー(例えば、PCRフリー)のゲノムライブラリ調製物を使用し、ランダムに断片化されたゲノムDNAを、クラスター増幅のみを使用して富化する(Kozarewaら、Nature Methods 6:291~295頁[2009])。テンプレートは、取り外し可能な蛍光色素を備えたりリバーシブルターミネーターを使用する、堅調な4色DNA合成配列決定技術を使用して配列決定される。高感度の蛍光検出は、レーザー励起及び全内部反射光学を使用して実現される。約数十から数百塩基対の短い配列リードが参照ゲノムに対してアラインメントされ、特別に開発されたデータ解析パイプラインソフトウェアを使用して、参照ゲノムへの短い配列リードの一意的なマッピングが特定される。最初のリードの完了後、テンプレートをインサイチュで再生成して、断片の反対側の端からの2番目のリードを可能にし得る。したがって、DNA断片のシングルエンド配列決定を使用しても、又はペアエンド配列決定を使用してもよい。

【0286】

本開示の様々な実施形態は、ペアエンド配列決定を可能にする合成時配列決定を使用してもよい。一部の実施形態では、Illuminaによる合成プラットフォームによる配列決定は、断片のクラスタリングを伴う。クラスタリングとは、各断片分子が等温で増幅されるプロセスである。一部の実施形態では、本明細書で説明する実施例のように、断片は、その断片の2つの末端に付着した2つの異なるアダプターを有し、アダプターによって、その断片がフローセルレーンの表面上の2つの異なるオリゴとハイブリダイズすることが可能になる。断片は、断片の両端にある2つのインデックス配列を更に含むか、又はそこに接続されており、インデックス配列は、マルチプレックス配列決定で異なる試料を特定するための標識を提供する。一部の配列決定プラットフォームでは、配列決定される断片は挿入とも呼ばれる。

【0287】

一部の実現例では、Illuminaプラットフォーム内でのクラスタリングのためのフローセルは、レーンを備えたガラススライドである。各レーンは、2種類のオリゴのローンでコーディングされたガラスチャネルである。ハイブリダイゼーションは、表面上の2種類のオリゴの最初のものによって可能になる。このオリゴは、断片の一端の最初のアダプターに相補的である。ポリメラーゼは、ハイブリダイズした断片の相補鎖を作成する。二本鎖分

子は変性され、元のテンプレート鎖は洗い流される。残りの鎖は、他の多くの残りの鎖と並行して、ブリッジアプリケーションを介してクローン的に増幅される。

【0288】

ブリッジ増幅では、鎖が折り畳まれ、鎖の第2の端の第2のアダプター領域が、フローセル表面の第2のタイプのオリゴとハイブリダイズする。ポリメラーゼは相補鎖を生成し、二本鎖架橋分子を形成する。この二本鎖分子が変性され、2つの異なるオリゴを介してフローセルにつながれた2つの一本鎖分子をもたらす。その後、このプロセスが何度も繰り返され、数百万のクラスターに対して同時に発生し、全ての断片のクローン増幅が生じる。ブリッジ増幅後、逆方向鎖が切断され、洗い流されて、順方向鎖のみが残る。不要なプライミングを防ぐために、3'末端はブロックされている。

10

【0289】

クラスタリング後、配列決定は、最初の配列決定プライマーを伸長して最初のリードを生成することから始まる。各サイクルで、蛍光的にタグ付けされたヌクレオチドは、成長している鎖への付加について競う。テンプレートの配列に基づいて組み込まれるのは1つだけである。各ヌクレオチドの添加後、クラスターは光源によって励起され、特徴的な蛍光シグナルが放出される。サイクル数によってリードの長さが決まる。放射波長及びシグナル強度がベースコールを決定する。所定のクラスターでは、全ての同一の鎖が同時に読み取られる。数億個のクラスターが、超並列方式で配列決定される。最初のリードが完了すると、読み取られた生成物は洗い流される。

【0290】

20

2つのインデックスプライマーを含むプロトコルの次のステップでは、インデックス1プライマーを導入し、テンプレート上のインデックス1領域にハイブリダイズする。インデックス領域は、断片の特定を提供し、これは、マルチプレックス配列決定プロセスで試料を脱多重化(de-multiplexing)するのに有用である。インデックス1のリードは、最初のリードと同様に生成される。インデックス1のリードが完了すると、リード生成物は洗い流され、鎖の3'末端が脱保護される。次に、テンプレート鎖が折り畳まれ、フローセル上の2番目のオリゴに結合する。インデックス2の配列は、インデックス1と同じ方法で読み取られる。次に、ステップの完了時に、インデックス2のリード生成物が洗い流される。

【0291】

30

2つのインデックスを読み取った後、リード2は、ポリメラーゼを使用して2番目のフローセルオリゴを伸長し、二本鎖架橋を形成することによって開始する。この二本鎖DNAは変性されており、3'末端はブロックされている。元の順方向鎖は切断され、洗い流され、逆方向鎖が残る。リード2は、リード2の配列決定プライマーの導入から始まる。リード1と同様に、所望の長さが達成されるまで配列決定ステップが繰り返される。リード2生成物は、洗い流される。このプロセス全体で、全ての断片に相当する数百万のリードが生成される。プールされた試料ライブラリ由来の配列は、試料の調製中に導入された一意的なインデックスに基づいて分離される。各試料について、ベースコールの同様のストレッチのリードは局所的にクラスター化される。順方向と逆方向のリードがペアになって、連続した配列が作成される。これらの連続した配列は、バリエーションの同定のために参照ゲノムにアラインメントされる。

40

【0292】

上述の合成時配列決定の例は、開示された方法の実施形態の多くで使用される対になったペアエンドリードを含む。ペアエンド配列決定には、断片の両端からの2つのリードが含まれる。1対のリードが参照配列にマッピングされると、2つのリード間の塩基対距離が決定され得、次いで、この距離を使用して、リードが取得された断片の長さを決定し得る。場合によっては、2つのピンにまたがる断片のペアエンドのリードの一方が1つのピンに、もう一方が隣接するピンにアラインメントされる。これは、ピンが長くなるか、リードが短くなるにつれてまれになる。様々な方法を使用して、これらの断片のピンメンバーシップを説明し得る。例えば、ピンの断片サイズの頻度を決定する際にそれらを省略し得る。隣接する両方のピンについて計数可能であり;これらは、2つのピンのより多くの塩基対を

50

含むピンに割り当てられてもよい;又は、各ピンの塩基対の部分に関連する重みを使用して、両方のピンに割り当てられてもよい。

【0293】

ペアエンドリードは、異なる長さ(すなわち、配列決定される異なる断片サイズ)の挿入を使用してもよい。本開示におけるデフォルトの意味として、ペアエンドリードを使用して、様々な挿入長から取得されたリードを指す。場合によっては、ショートインサートペアエンドリードとロングインサートペアエンドリードを区別するために、後者はメイトペアリード(mate pair read)とも呼ばれる。メイトペアリードを含む一部の実施形態では、最初に2つのビオチン接合アダプターが、比較的長いインサート(例えば、数kb)の2つの端部に取り付けられる。次に、ビオチン接合アダプターが、インサートの両端をリンクして、環状化された分子を形成する。次に、環状分子を更に断片化することにより、ビオチン接合アダプターを含むサブ断片を取得し得る。次に、逆の順序で元の断片の2つの端を含むサブ断片を、上記のショートインサートペアエンド配列決定と同じ手順で配列決定し得る。Illuminaプラットフォームを使用したメイト・ペア配列決定の更に詳細は、次のURLのオンライン出版物に示されており、その全体が参照により組み込まれている: res.illumina.com/documents/products/technotes/technote_nextera_matepair_data_processing。ペアエンド配列決定に関する追加情報は、米国特許第7601499号及び米国特許公開第2012/0,053,063号に見出すことができ、これらは、ペアエンド配列決定の方法及び装置に関する資料に関して参照により組み込まれる。

【0294】

DNA断片の配列決定後、所定の長さ、例えば100bpの配列リードが、公知の参照ゲノムにマッピング又はアラインメントされる。マッピングされたリード又はアラインメントされたリード、及び参照配列上のそれらの対応する位置はまた、タグとも呼ばれる。一実施形態において、参照ゲノム配列は、NCBI36/hg18配列であり、これは、genome.ucsc.edu/cgi-bin/hgGateway?org=Human&db=hg18&hgside=166260105で、ワールドワイドウェブ上で入手可能である。或いは、参照ゲノム配列は、GRCh37/hg19であり、これはgenome.ucsc.edu/cgi-bin/hgGatewayのワールドワイドウェブで入手し得る。パブリック配列情報の他のソースには、GenBank、dbEST、dbSTS、EMBL(欧州分子生物學研究所)、及びDDBJ(日本のDNAデータバンク)が挙げられる。限定するものではないが、BLAST(Altschulら、1990)、BLITZ(MPsrch)(Sturrock&Collins、1993)、FASTA(Person&Lipman、1988)、BOWTIE(Langmeadら、Genome Biology 10:R25.1-R25.10[2009])、又はELAND(Illumina, Inc.社、San Diego, CA, USA)を含む、配列をアラインメントするための多数のコンピュータアルゴリズムが利用可能である。一実施形態において、血漿cfDNA分子のクローン的に増殖されたコピーの一端は、ヌクレオチドデータベースの効率的な大規模アラインメント(Efficient Large-Scale Alignment of Nucleotide Databases:ELAND)ソフトウェアを使用する、イルミナゲノムアナライザー(Illumina Genome Analyzer)のバイオインフォマティクスアラインメント分析により配列決定され、プロセッシングされる。

【0295】

例示的であるが、非限定的な一実施形態では、本明細書に記載の方法は、試験試料中の核酸、例えば母体試料中のcfDNA、癌に関してスクリーニング中の対象中のcfDNA又は細胞DNA等の配列情報を、Helicos True Single Molecule Sequencing(tSMS)技術の単一分子配列決定技術(例えば、Harris T.D.ら、Science 320:106~109頁[2008]に記載されるような)を使用して、取得することを包含する。tSMS技術では、DNA試料は、約100~200ヌクレオチドの鎖に切断され、各DNA鎖の3'末端にポリA配列が付加される。各鎖は、蛍光標識されたアデノシンヌクレオチドの添加により標識される。次に、DNA鎖をフローセルにハイブリダイズさせ、フローセルには、フローセル表面に固定された数百万のオリゴT捕捉部位が含まれる。特定の実施形態において、テンプレートは、約1億テンプレート/cm²の密度であり得る。次に、フローセルを機器、例えばHeliScope(商標)シーケンサにロードし、レーザーでフローセルの表面を照らし、各テンプレートの位置を明らかに

する。CCDカメラは、フローセル表面上のテンプレートの位置をマッピングし得る。次いで、テンプレート蛍光標識を切断し、洗い流す。配列決定反応は、DNAポリメラーゼと蛍光標識ヌクレオチドを導入することから始まる。オリゴT核酸は、プライマーとして機能する。ポリメラーゼは、標識されたヌクレオチドをテンプレート指向の方法でプライマーに組み込む。ポリメラーゼ及び組み込まれていないヌクレオチドは除去される。蛍光標識ヌクレオチドの組み込みを指示したテンプレートは、フローセル表面を画像化することにより識別される。画像化後、切断ステップで蛍光標識を除去し、望ましいリード長が得られるまで、他の蛍光標識ヌクレオチドを使用してプロセスを繰り返す。配列情報は、ヌクレオチドの追加ステップごとに収集される。単一分子配列決定技術による全ゲノム配列決定では、配列決定ライブラリの調製におけるPCRベースの増幅が除外又は通常は不要になり、この方法により、その試料のコピーの測定ではなく、試料の直接測定が可能になる。

【0296】

別の例示的であるが非限定的な実施形態では、本明細書に記載の方法は、試験試料中の核酸、例えば、母体試験試料中のcfDNA、癌についてスクリーニングされる対象のcfDNA又は細胞DNA等の配列情報を、454配列(Roche)を使用して取得することを包含する(例えば、Margulies, M.ら、Nature 437:376~380頁[2005]に記載)。454配列決定には通常2つのステップが含まれる。最初のステップで、DNAは、約300~800塩基対の断片に切断され、断片は平滑末端になる。次に、オリゴヌクレオチドアダプターを断片の末端に連結する。アダプターは、断片の増幅及び配列決定のプライマーとして機能する。この断片は、5'-ビオチンタグを含む、アダプターB等を使用して、ストレプトアビジンコーティングビーズ等のDNAキャプチャビーズに取り付けてもよい。ビーズに付着した断片は、油水エマルジョンの液滴内でPCR増幅される。その結果、各ビーズ上にクローン増幅されたDNA断片の複数のコピーが生じる。2番目のステップでは、ビーズをウェル(例えば、ピコリットルサイズのウェル)に捕捉する。パイロシーケンシングは、各DNA断片に対して並行して実行される。1つ又は複数のヌクレオチドを追加すると、配列決定機器のCCDカメラで記録される光信号が生成される。シグナル強度は、組み込まれたヌクレオチドの数に比例する。パイロシーケンシングでは、ヌクレオチドの添加時に放出されるピロリン酸(Pi)を利用する。Piは、アデノシン5'ホスホ硫酸の存在下でATPスルフィラーゼによってATPに変換される。ルシフェラーゼはATPを使用してルシフェリンをオキシルシフェリンに変換し、この反応により光が生成され、測定及び分析される。

【0297】

別の例示的であるが非限定的な実施形態では、本明細書に記載の方法は、試験試料中の核酸、例えば、母体試験試料中のcfDNA、癌に関してスクリーニングされている対象のcfDNA又は細胞DNA等の配列情報を、SOLiD(商標)技術(Applied Biosystems社)を使用して、取得することを包含する。SOLiD(商標)のライゲーションによる配列決定では、ゲノムDNAを断片に切断し、アダプターを断片の5'及び3'末端に付着させて断片ライブラリを生成する。或いは、断片の5'及び3'末端にアダプターを連結し、断片を環状化し、環状化した断片を消化して内部アダプターを生成し、得られた断片の5'及び3'末端にアダプターを取り付けることにより、内部アダプターを導入して、断片を作成して、メイト・ペアード・ライブラリを生成してもよい。次に、ビーズ、プライマー、テンプレート、及びPCR成分を含むマイクロリアクターで、クローンビーズ集団を調製する。PCRに続いて、テンプレートを変性し、ビーズを濃縮して、拡張テンプレートでビーズを分離する。選択したビーズのテンプレートは、ガラススライドへの結合を可能にする3'の変更を受ける。配列は、特定のフルオロフォアによって特定される中央決定塩基(又は塩基のペア)と部分的にランダムなオリゴヌクレオチドの順次ハイブリダイゼーション及びライゲーションによって決定し得る。色が記録された後、連結されたオリゴヌクレオチドが切断されて除去され、次いで、プロセスが繰り返される。

【0298】

別の例示的であるが非限定的な実施形態では、本明細書に記載の方法は、試験試料中の核酸、例えば、母体試験試料中のcfDNA、癌に関してスクリーニングされている対象のcfD

NA又は細胞DNA等の配列情報を、Pacific Biosciences社の単一分子リアルタイム(SMRT(商標))配列決定技術を使用して、取得することを包含する。SMRT配列決定では、色素標識ヌクレオチドの連続的な組み込みが、DNA合成中に画像化される。単一のDNAポリメラーゼ分子は、個々のゼロモード波長検出器(ZMW検出器)の底面に取り付けられ、ホスホ結合されたヌクレオチドが成長中のプライマー鎖に組み込まれている間に配列情報を取得する。ZMW検出器は、ZMWの外に急速に拡散する蛍光ヌクレオチド(例えば、マイクロ秒)を背景に、DNAポリメラーゼによる単一ヌクレオチドの組み込みを観察し得る閉じ込め構造を備えている。通常、成長中の鎖にヌクレオチドを組み込むには数ミリ秒かかる。この間、蛍光標識は、励起されて蛍光シグナルを生成し、この蛍光タグは切断される。色素の対応する蛍光の測定によって、どの塩基が組み込まれたかが示される。このプロセスを繰り返して、配列を得る。

10

【0299】

別の例示的であるが非限定的な実施形態では、本明細書に記載の方法は、試験試料中の核酸、例えば、母体試験試料中のcfDNA、癌についてスクリーニングされている対象のcfDNA又は細胞DNA等の配列情報を、ナノポア配列決定を使用して(例えば、Soni GV及びMellner A.Clin Chem 53:1996~2001頁[2007]に記載されているように)取得することを包含する。ナノポア配列決定DNA分析技術は、例えば、Oxford Nanopore technologies社(Oxford, United Kingdom)、Sequenom社、NABsys社等を含む多くの企業によって開発されている。ナノポア配列決定は、単一分子のDNAが、ナノポアを通過する際に直接配列決定される単一分子配列決定技術である。ナノポアは、一般に直径1ナノメートルのオーダーの小さな穴である。導電性流体へのナノポアの浸漬、及びそこへの電位(電圧)の印加により、ナノポアを通じたイオンの伝導に起因したわずかな電流が発生する。流れる電流の量は、ナノポアのサイズと形状に敏感である。DNA分子がナノポアを通過すると、DNA分子の各ヌクレオチドがナノポアを様々な程度に妨害し、ナノポアを通過して流れる電流の大きさを様々な程度で変化させる。したがって、DNA分子がナノポアを通過する際の電流のこの変化によって、DNA配列のリードが提供される。

20

【0300】

別の例示的であるが非限定的な実施形態では、本明細書に記載の方法は、試験試料中の核酸、例えば、母体試験試料中のcfDNA、癌についてスクリーニングされている対象のcfDNA又は細胞DNA等についての配列情報を、化学物質に感受性の電界効果トランジスタ(chemFET)アレイを使用して取得することを包含する(例えば、米国特許出願公開第2009/0026082号に記載されているように)。この技術の一例では、DNA分子を、反応チャンバーに入れて、テンプレート分子をポリメラーゼに結合した配列決定プライマーにハイブリダイズさせ得る。配列決定プライマーの3'末端にある新しい核酸鎖への1つ又は複数の三リン酸の組み込みは、chemFETによって電流の変化として認識し得る。アレイには複数のchemFETセンサーを備えてもよい。別の実施例では、単一の核酸を、ビーズに付着させてもよく、核酸をビーズ上で増幅してもよく、個々のビーズを、chemFETアレイ上の個々の反応チャンバーに移してもよく、各チャンバーにchemFETセンサーを搭載し、核酸を配列決定し得る。

30

【0301】

別の実施形態において、本方法は、透過型電子顕微鏡法(TEM)を使用して、試験試料中の核酸、例えば母体試験試料中のcfDNAの配列情報を取得することを包含する。個別分子配置高速ナノ転送(Individual Molecule Placement Rapid Nano Transfer:IMPRNT)と呼ばれるこの方法は、重原子マーカで選択的に標識された高分子量(150kb以上)DNAの単一原子分解能透過型電子顕微鏡画像化を利用すること、及びこれらの分子を、超薄膜上に、一貫したベース間間隔の超高密度(3nm鎖間)並列アレイで配置することを包含する。電子顕微鏡を使用してフィルム上の分子を画像化し、重原子マーカの位置を決定し、DNAから塩基配列情報を抽出する。この方法は、PCT特許公開WO2009/046445に更に記載されている。この方法では、10分未満で完全なヒトゲノムの配列決定が可能である。

40

【0302】

50

別の実施形態では、DNA配列決定技術は、半導体技術と単純な配列決定化学を組み合わせ、化学的にコード化された情報(A、C、G、T)を半導体チップ上のデジタル情報(0、1)に直接変換するIon Torrent単一分子配列決定である。自然界では、ヌクレオチドがポリメラーゼによってDNAの鎖に組み込まれると、水素イオンが副産物として放出される。Ion Torrentは、微細加工されたウェルの高密度アレイを使用して、この生化学プロセスを大規模に並行して実行する。各ウェルは異なるDNA分子を保持している。ウェルの下にはイオン感受性層があり、その下にはイオンセンサーがある。ヌクレオチド、例えばCが、DNAテンプレートに追加され、次いで、DNAの鎖に組み込まれると、水素イオンが放出される。そのイオンからの電荷により溶液のpHが変化し、これはIon Torrentのイオンセンサーによって検出され得る。シーケンサ(基本的には世界最小の固体pHメーター)は、化学情報からデジタル情報に直接移行するベースをコールする)。次いで、Ionパーソナルゲノムマシン(PGM(商標))シーケンサは、チップを次々とヌクレオチドと順番にフラッディングする。チップをフラッディングする次のヌクレオチドが一致しない場合、電圧変化は記録されず、ベースはコールされない。DNA鎖に2つの同一の塩基がある場合、電圧は2倍になり、チップはコールされた2つの同一の塩基を記録する。直接検出により、ヌクレオチドの組み込みを数秒で記録可能になる。

【0303】

別の実施形態において、本方法は、ハイブリダイゼーションによる配列決定を使用して、試験試料中の核酸、例えば母体試験試料中のcfDNAの配列情報を取得することを包含する。ハイブリダイゼーションによる配列決定は、複数のポリヌクレオチド配列を複数のポリヌクレオチドプローブと接触させることを包含し、複数のポリヌクレオチドプローブのそれぞれは、必要に応じて基質につながれてもよい。基質は、既知のヌクレオチド配列のアレイを含む平坦な表面であり得る。アレイへのハイブリダイゼーションのパターンを使用して、試料に存在するポリヌクレオチド配列を決定してもよい。他の実施形態では、各プローブは、ビーズ、例えば磁気ビーズ等につながれている。ビーズへのハイブリダイゼーションを決定して、試料内の複数のポリヌクレオチド配列を特定するために使用してもよい。

【0304】

本明細書に記載の方法の一部の実施形態では、マッピングされた配列タグは、約20bp、約25bp、約30bp、約35bp、約40bp、約45bp、約50bp、約55bp、約60bp、約65bp、約70bp、約75bp、約80bp、約85bp、約90bp、約95bp、約100bp、約110bp、約120bp、約130、約140bp、約150bp、約200bp、約250bp、約300bp、約350bp、約400bp、約450bp、又は約500bpという配列のリードを含む。技術的進歩により、500bpを超えるシングルエンドリードが可能になり、ペアエンドリードが生成されたときに約1000bpを超えるリードが可能になると予想される。一実施形態では、マッピングされた配列タグは、36bpの配列リードを含む。配列タグのマッピングは、タグの配列を参照の配列と比較して、配列決定された核酸(例えば、cfDNA)分子の染色体起源を決定することで実現され、特定の遺伝子配列情報は必要がない。参照ゲノムと混合物試料中のゲノム間に存在し得るマイナーな多型を説明するために、わずかな不一致(配列タグごとに0~2個の不一致)が許可される場合もある。

【0305】

通常、試料ごとに複数の配列タグが取得される。一部の実施形態では、少なくとも約 3×10^6 個の配列タグ、少なくとも約 5×10^6 個の配列タグ、少なくとも約 8×10^6 個の配列タグ、少なくとも約 10×10^6 個の配列タグ、少なくとも約 15×10^6 個の配列タグ、少なくとも約 20×10^6 個の配列タグ、少なくとも約 30×10^6 個の配列タグ、少なくとも約 40×10^6 個の配列タグ、又は少なくとも約 50×10^6 配列タグであって、20~40bpのリード(例えば36bp)を含むタグが、1試料ごとの参照ゲノムに対するリードのマッピングから得られる。一実施形態では、全ての配列リードは、参照ゲノムの全ての領域にマッピングされる。一実施形態では、参照ゲノムの全ての領域、例えば全ての染色体にマッピングされたタグが分析される。

【0306】

複数の供給源からの核酸の混合物をデコンボリューションするための装置及びシステム配列決定データの分析及びそこから得られる診断は、典型的には、様々なコンピュータ実行アルゴリズム及びプログラムを使用して実行される。したがって、特定の実施形態は、1つ又は複数のコンピュータシステム又は他のプロセッシングシステムに記憶又は転送されるデータを含むプロセスを使用する。本明細書で開示される実施形態は、これらの動作を実行するための装置にも関する。この装置は、必要な目的のために特別に構築されてもよいし、又はコンピュータに記憶されたコンピュータプログラム及び/又はデータ構造によって選択的に起動又は再構成される汎用コンピュータ(又はコンピュータのグループ)であってもよい。一部の実施形態では、プロセッサのグループは、列挙された分析操作の一部又は全てを共同で(例えば、ネットワーク又はクラウドコンピューティングを介して)及び/又は並行して実行する。本明細書に記載の方法を実行するためのプロセッサ又はプロセッサのグループは、プログラマブルデバイス(例えば、CPLD及びFPGA)等のマイクロコントローラ及びマイクロプロセッサ、並びにゲートアレイASIC又は汎用マイクロプロセッサ等の非プログラマブルデバイスを含む様々なタイプのものであり得る。

10

【0307】

更に、特定の実施形態は、様々なコンピュータ実現の操作を実行するためのプログラム命令及び/又はデータ(データ構造を含む)を含む、有形及び/又は非一時的なコンピュータ可読媒体又はコンピュータプログラム製品に関する。コンピュータ可読媒体の例としては、限定するものではないが、半導体メモリデバイス、磁気媒体、例えば、ディスクドライブ、磁気テープ、光媒体、例えば、CD、光磁気媒体、並びにプログラム命令を記憶及び実行するように特別に構成されたハードウェアデバイス、例えば、読み出し専用メモリデバイス(ROM)及びランダムアクセスメモリ(RAM)が挙げられる。コンピュータ可読媒体は、エンドユーザによって直接制御されてもよいし、媒体はエンドユーザによって間接的に制御されてもよい。直接制御される媒体の例としては、ユーザ施設にある媒体、及び/又は他のエンティティと共有されていない媒体が挙げられる。間接的に制御される媒体の例としては、外部ネットワーク及び/又は「クラウド」等の共有リソースを提供するサービスを介してユーザが間接的にアクセスし得る媒体が挙げられる。プログラム命令の例としては、コンパイラによって生成されるようなマシンコード、及びインタープリターを使用してコンピュータによって実行され得る、より高いレベルのコードを含むファイルが挙げられる。

20

30

【0308】

様々な実施形態において、開示された方法及び装置で使用されるデータ又は情報は、電子形式で提供される。そのようなデータ又は情報には、核酸試料に由来するリード及びタグ、参照配列の特定の領域にアラインメントする(例えば、染色体又は染色体セグメントにアラインメントする)そのようなタグのカウント又は密度、参照配列(単独又は主に多型を提供する、参照配列を含む)、染色体及びセグメントの量、SNV又は異数性のコール等のコール、正規化された染色体及びセグメント値、染色体又はセグメントの対、及び対応する正規化染色体又はセグメント、カウンセリングの推奨事項、診断等が挙げられる。本明細書で使用される場合、電子形式で提供されるデータ又はその他の情報は、マシンへの記憶及びマシン間の送信に使用し得る。従来、電子形式のデータはデジタルで提供され、様々なデータ構造、リスト、データベース等にビット及び/又はバイトとして記憶され得る。データは電子的、光学的等で具現化され得る。

40

【0309】

一実施形態は、試験試料において、SNV又は癌に関連する異数性の有無を示す出力を生成するためのコンピュータプログラム製品を提供する。コンピュータ・プロダクトは、染色体異常を判定するための上記の方法のいずれか1つ又は複数を実行するための指示書を含んでもよい。説明したとおり、コンピュータ・プロダクトは、プロセッサが核酸の混合物をデコンボリューションし得るように、コンピュータ実行可能又はコンパイル可能なロジック(例えば、命令)をそこに記録した非一時的及び/又は有形のコンピュータ可読媒体を備え得る。一例では、コンピュータ・プロダクトは、プロセッサが核酸の混合物をデコンボ

50

リユーシオンすることを可能にするためにコンピュータ実行可能又はコンパイル可能なロジック(例えば、命令)が記録されたコンピュータ可読媒体を備える。

【0310】

検討中の試料由来の配列情報を染色体参照配列にマッピングして、対象となる任意の1つ又は複数の染色体のそれぞれの多数の配列タグを特定し、このような対象となる任意の1つ又は複数の染色体のそれぞれの正規化セグメント配列の多数の配列タグを特定し得る。様々な実施形態において、参照配列は、例えば、リレーショナルデータベース又はオブジェクトデータベース等のデータベースに記憶される。

【0311】

助けを受けない人間が本明細書に開示された方法の計算操作を実行することは実際的ではないか、ほとんどの場合可能でさえないことを理解されたい。例えば、試料由来の単一の30bpのリードを人間の染色体のいずれかにマッピングするには、計算装置の支援なしで何年もの努力が必要になる場合がある。

10

【0312】

本明細書に開示される方法は、1つ又は複数の寄与体の核酸を含む核酸試料を定量するためのシステムを使用して実施し得る。このシステムは:(a)試料から核酸配列情報を提供する試験試料から核酸を受け取るためのシーケンサと;(b)プロセッサと;(c)核酸の混合物をデコンボリユーシオンする方法を実行するために、前記プロセッサ上で実行するための命令が記憶された1つ又は複数のコンピュータ可読記憶媒体とを備える。

【0313】

一部の実施形態では、この方法は、核酸の混合物をデコンボリユーシオンするための方法を実行するためのコンピュータ可読命令を記憶したコンピュータ可読媒体によって指示される。したがって、一実施形態は、コンピュータシステムの1つ又は複数のプロセッサによって実行されると、コンピュータシステムに、1つ又は複数の寄与体の核酸を含む核酸試料を定量するための方法を実施させる、コンピュータ実行可能命令が記憶された1つ又は複数のコンピュータ可読非一時的記憶媒体を備えるコンピュータプログラム製品を提供する。この方法は、(a)コンピュータシステムによって、核酸試料から得られ、1つ又は複数の多型座位にある1つ又は複数の対立遺伝子にマッピングされた核酸配列リードを受け取る工程と、(b)核酸配列リードを使用して、1つ又は複数のプロセッサによって、1つ又は複数の多型座位にある1つ又は複数の対立遺伝子のそれぞれに関する対立遺伝子数を決定する工程と、(c)確率的混合物モデルを使用して、1つ又は複数のプロセッサによって、核酸試料中の1つ又は複数の寄与体の核酸の1つ又は複数の割合を定量する工程とを含み、確率的混合物モデルを使用する工程は、確率的混合物モデルを対立遺伝子数に適用する工程を含み、確率的混合物モデルは、確率分布を使用して1つ又は複数の多型座位にある対立遺伝子数をモデル化し、確率分布は核酸配列リードの誤差を説明する。

20

30

【0314】

一部の実施形態では、指示は、母体試験試料を提供するヒト対象の患者医療記録における方法に関連する情報を自動的に記録することを更に包含し得る。この患者医療記録は、例えば、研究室、医療従事者のオフィス、病院、健康維持組織、保険会社、又は個人の医療記録ウェブサイトによって維持され得る。更に、プロセッサで実施される分析の結果に基づいて、この方法は、母体試験試料が採取されたヒト対象の処置を処方、開始、及び/又は変更することを更に包含し得る。これには、対象から採取した追加の試料に対して1つ又は複数の追加の試験又は分析を実行することを包含し得る。

40

【0315】

開示された方法はまた、1つ又は複数の寄与体の核酸を含む核酸試料を、定量する方法を実行するように適合又は構成されているコンピュータプロセッシングシステムを使用して実行されてもよい。一実施形態は、本明細書に記載の方法を実行するように適合又は構成されたコンピュータプロセッシングシステムを提供する。一実施形態では、この装置は、試料中の核酸分子の少なくとも一部を配列決定して、本明細書の他の場所に記載の配列情報のタイプを取得するように、適合又は構成された配列決定デバイスを備える。この装置はま

50

た、試料をプロセッシングするための構成要素を備えてもよい。そのような構成要素は、本明細書の他の場所で説明されている。

【0316】

配列又は他のデータは、コンピュータに入力するか、コンピュータ可読媒体に直接又は間接的に記憶され得る。一実施形態において、コンピュータシステムは、試料から核酸の配列を読み取るか、及び/又は分析する配列決定デバイスに直接連結される。このようなツールからの配列又はその他の情報は、コンピュータシステムのインターフェースを介して提供される。或いは、システムによってプロセッシングされる配列は、データベース又は他のリポジトリ等の配列記憶装置ソースから提供される。プロセッシング装置で一旦利用可能になれば、メモリデバイス又は大容量記憶装置デバイスは、核酸の配列を少なくとも一時的にバッファリング又は記憶する。加えて、メモリデバイスは、様々な染色体又はゲノム等のタグカウントを記憶し得る。メモリはまた、配列又はマッピングされたデータの提示を分析するための様々なルーチン及び/又はプログラムも記憶してもよい。そのようなプログラム/ルーチンには、統計分析等を実行するためのプログラムが含まれ得る。

10

【0317】

一例では、ユーザは試料を配列決定装置に提供する。データは、コンピュータに接続された配列決定装置によって収集及び/又は分析される。コンピュータ上のソフトウェアによって、データの収集及び/又は分析が可能になる。データを記憶しても、表示(モニター又は他の同様のデバイス経由)してもよいし、及び/又は別の場所に送信してもよい。コンピュータは、遠隔地のユーザ(例えば、医師、科学者、又はアナリスト)が利用する携帯用デバイスにデータを送信するために使用されるインターネットに接続されてもよい。データは、送信前に記憶及び/又は分析し得ることが理解される。一部の実施形態では、生データが収集され、データを分析及び/又は保存する遠隔地のユーザ又は装置に送信される。送信は、インターネット経由で発生する可能性があるが、衛星又は他の接続経由でも発生してもよい。或いは、データをコンピュータ可読媒体に記憶し、その媒体をエンドユーザに(例えば、メールで)搬送してもよい。遠隔地のユーザは、建物、都市、州、国又は大陸を含むがこれらに限定されない同じ又は異なる地理的場所にいてもよい。

20

【0318】

一部の実施形態では、この方法はまた、複数のポリヌクレオチド配列(例えば、リード、タグ及び/又は参照染色体配列)に関するデータを収集すること、及びコンピュータ又は他の計算システムにデータを送信することも包含する。例えば、コンピュータは、実験装置、例えば、試料収集装置、ヌクレオチド増幅装置、ヌクレオチド配列決定装置、又はハイブリダイゼーション装置に接続されてもよい。その後、コンピュータは、実験デバイスによって収集された適切なデータを収集し得る。データは、任意のステップでコンピュータに記憶されてもよく、例えば、リアルタイムで収集されている間、送信前、送信中若しくは送信と組み合わせて、又は送信後に記憶されてもよい。データは、コンピュータから抽出可能であるコンピュータ可読媒体に記憶し得る。収集又は記憶されたデータは、例えばローカルネットワーク又はインターネット等のワイドエリアネットワークを介して、コンピュータから遠隔地に送信され得る。以下に説明するように、遠隔地では、送信されたデータに対して様々な操作が実行され得る。

30

【0319】

本明細書で開示されるシステム、装置、及び方法で記憶、送信、分析、及び/又は操作され得る電子的にフォーマットされたデータのタイプとしては、以下である:

試験試料中の核酸を配列決定することにより得られたリード

リードを参照ゲノム又は他の参照配列(単数又は複数)にアラインメントすることにより得られたタグ

参照ゲノム又は配列

配列タグ密度-参照ゲノム又は他の参照配列の2つ以上の領域(通常は染色体又は染色体セグメント)のそれぞれのタグのカウント又は数

特定の染色体又は対象となる染色体セグメントに関して染色体又は染色体セグメントを正

40

50

規化する同一性

対象となる染色体又はセグメントから取得した染色体又は染色体セグメント(又は他の領域)の量と、対応する正規化染色体又はセグメント

影響を受けるか、影響を受けないか、又はコールなしとして染色体量をコールするための閾値

染色体量の実際のコール

診断(コールに関連する臨床状態)

コール及び/又は診断から派生したさらなる試験の推奨事項

コール及び/又は診断から派生した処置及び/又はモニタリング計画

【0320】

これらの様々なタイプのデータは、別個の装置を使用して1つ又は複数の場所で取得、記憶、送信、分析、及び/又は操作され得る。プロセッシングの選択肢は広範囲におよぶ。スペクトルの一方の端で、この情報の全て又は大部分が記憶され、試験試料がプロセッシングされる場所、例えば医師のオフィス又はその他の臨床の場で使用される。他の極端な場合、試料は1つの場所で取得され、これはプロセッシングされ、必要に応じて別の場所で配列決定され、リードがアラインメントされ、1つ又は複数の異なる場所でコールが行われ、そして診断、推奨事項、及び/又は計画がまだ別の場所(試料が取得された場所の場合もある)で準備されている。

【0321】

様々な実施形態において、リードは、配列決定装置で生成され、次いで遠隔サイトに送信され、そこでコールを生成するためにプロセッシングされる。この遠隔地では、一例として、リードは参照配列にアラインメントされてタグを生成し、このタグがカウントされて、対象となる染色体又はセグメントに割り当てられる。また、この遠隔地では、関連する正規化染色体又はセグメントを使用して、カウントが量に変換される。更に、遠隔地では、その量を使用してコールを生成する。

【0322】

別個の場所で使用され得るプロセッシング操作の中には、以下のものがある:

試料収集

配列決定に先立つ試料プロセッシング

配列決定

配列データを分析し、1つ又は複数の寄与体の核酸を含む核酸試料を定量すること

診断

診断及び/又はコールを、患者又は医療提供者に報告すること

さらなる処置、検査、及び/又はモニタリングのための計画を開発すること

計画の実行

カウンセリング

【0323】

これらの操作のいずれか1つ又は複数のを、本明細書の他の場所で説明するように自動化してもよい。通常、配列データの配列決定及び分析、並びにDNA混合物試料のデコンボリューションは、計算によって実行される。他の操作は、手動で実行されても、又は自動で実行されてもよい。

【0324】

試料収集が実行され得る場所の例としては、医療従事者のオフィス、診療所、患者の家(試料収集ツール又はキットが提供される)、及び移動式ヘルスケア車両が挙げられる。配列決定の前に試料プロセッシングが実行され得る場所の例としては、医療従事者のオフィス、診療所、患者の家(試料プロセッシング装置又はキットが提供される)、移動式ヘルスケア車両、及びDNA分析提供者の施設が挙げられる。配列決定が実行され得る場所の例としては、医療従事者のオフィス、診療所、医療従事者のオフィス、診療所、患者の家(試料配列決定装置及び/又はキットが提供される)、移動式ヘルスケア車両、及びDNA分析提供者の施設が挙げられる。配列決定が行われる場所には、電子形式で配列データ(通常はリード)を送

10

20

30

40

50

信するための専用ネットワーク接続が提供され得る。そのような接続は、有線でも無線でもよく、プロセッシングサイトへの送信前にデータがプロセッシング及び/又は集約され得るサイトにデータを送信するように構成されてもよい。データ集積者は、Health Maintenance Organizations(HMO)等の保健組織によって維持され得る。

【0325】

分析及び/又は導出操作は、前述の位置のいずれかで、或いは核酸配列データの計算及び/又は分析サービス専用のさらなる遠隔サイトで実行されてもよい。このようなサイトとしては、例えば、汎用サーバファーム等のクラスター、DNA分析サービスビジネスの施設等が挙げられる。一部の実施形態では、この分析を実行するために使用される計算装置はリース又はレンタルされる。計算リソースは、口語でクラウドと呼ばれるプロセッシングリソース等、インターネットでアクセス可能なプロセッサの収集の一部である場合がある。場合によっては、計算は、相互に関連するか又は関連しないプロセッサの並列又は大規模並列グループによって実行される。プロセッシングは、クラスターコンピューティング、グリッドコンピューティング等の分散プロセッシングを使用して実行され得る。そのような実施形態では、計算リソースのクラスター又はグリッドは、本明細書で説明される分析及び/又は導出を実行するために、一緒に作用する複数のプロセッサ又はコンピュータで構成されるスーパーバーチャルコンピュータを集合的に形成する。本明細書で説明するように、これらの技術及び更に従来型のスーパーコンピュータを使用して、配列データをプロセッシングし得る。それぞれは、プロセッサ又はコンピュータに依存する並列コンピューティングの形式である。グリッドコンピューティングの場合、これらのプロセッサ(多くの場合、コンピュータ全体)は、イーサネット(登録商標)等の従来のネットワークプロトコルによってネットワーク(プライベート、パブリック、又はインターネット)で接続されている。対照的に、スーパーコンピュータには、ローカルの高速コンピュータバスで接続された多くのプロセッサがある。

【0326】

特定の実施形態では、診断は、分析動作と同じ場所で生成される。他の実施形態では、異なる場所で実行される。いくつかの例では、診断の報告は、試料が採取された場所で実行されるが、そうである必要はない。診断を生成又は報告し得る場所、及び/又は計画の開発が行われる場所の例としては、医療従事者のオフィス、診療所、コンピュータでアクセス可能なインターネットサイト、及び携帯用デバイス、例えば、ネットワークへの有線又は無線接続を有する携帯電話、タブレット、スマートフォン等が挙げられる。カウンセリングが行われる場所の例としては、医療従事者のオフィス、診療所、コンピュータによってアクセス可能なインターネットサイト、携帯用デバイス等が挙げられる。

【0327】

一部の実施形態では、試料収集、試料プロセッシング、及び配列決定操作は、第1の場所で実行され、分析及び導出操作は、第2の場所で実行される。ただし、場合によっては、試料収集は1つの場所(例えば、医療従事者のオフィス又は診療所)で収集され、試料のプロセッシングと配列決定は、別の場所で実行され、これは必要に応じて分析と導出が行われる同じ場所である。

【0328】

様々な実施形態において、上に列挙された動作の配列は、試料収集、試料プロセッシング、及び/又は配列決定を開始するユーザ又はエンティティによって誘発され得る。1つ又は複数のこれらの操作が実行を開始した後、他の操作が自然に続く場合がある。例えば、配列決定操作により、リードが自動的に収集され、プロセッシング装置に送られ、その後、多くの場合自動的に、可能性としてはさらなるユーザの介入なしに、配列分析を実行する。一部の実現例では、このプロセッシング操作の結果は、次いで、おそらく診断として再フォーマットされて、医療専門家及び/又は患者への情報の報告をプロセッシングするシステムコンポーネント又はエンティティに自動的に送達される。説明したように、このような情報は、おそらくカウンセリング情報とともに、処置、検査、及び/又はモニタリング計画を作成するために自動的にプロセッシングされてもよい。したがって、早期段階の操作を開始する

と、医療専門家、患者、又はその他の関係者に、身体状況に基づいて行動するのに役立つ診断、計画、カウンセリング、及び/又はその他の情報が提供される、エンド・ツー・エンド配列が誘発される。これは、システム全体の一部が物理的に分離されており、例えば試料及び配列装置の場所から離れている可能性がある部分でも達成される。

【0329】

図4は、適切に構成又は設計されたとき、特定の実施形態による計算装置として機能し得る典型的なコンピュータシステムを単純なブロック形式で示している。コンピュータシステム2000は、一次記憶装置2006(典型的にはランダムアクセスメモリ、又はRAM)、一次記憶装置2004(典型的には読み出し専用メモリ、すなわちROM)を含む記憶装置に連結される任意の数のプロセッサ2002(中央処理装置、又はCPUとも呼ばれる)を備える。CPU2002は、マイクロコントローラ及びマイクロプロセッサ、例えば、プログラマブルデバイス(例えば、CPLD及びFPGA)及び非プログラマブルデバイス、例えば、ゲートアレイASIC又は汎用マイクロプロセッサを含む様々なタイプのものであり得る。図示された実施形態では、一次記憶装置2004は、データ及び命令をCPUに一方向に転送するように機能し、一次記憶装置2006は、通常、データ及び命令を双方向に転送するために使用される。これらの一次記憶デバイスの両方とも、上記のもの等の任意の適切なコンピュータ可読媒体を備えてもよい。大容量記憶装置デバイス2008も一次記憶装置2006に双方向に連結され、追加のデータ記憶容量を提供し、上記の任意のコンピュータ可読媒体を備えてもよい。大容量記憶装置デバイス2008を使用して、プログラム、データ等を記憶してもよく、通常はハードディスク等の二次記憶媒体である。多くの場合、このようなプログラム、データ等は、CPU 2002で実行するために一時的に一次メモリ2006にコピーされる。大容量記憶装置デバイス2008内に保持される情報は、適切な場合、標準的な方法で一次記憶装置2004の一部として組み込まれてもよい。CD-ROM 2014等の特定の大容量記憶装置デバイスはまた、CPU又は一次記憶装置に一方向にデータを渡す場合もある。

【0330】

CPU2002はまた、インターフェース2010に連結されており、これが1つ又は複数の入力/出力デバイス、例えば、核酸シーケンサ(2020)、ビデオモニター、トラックボール、マウス、キーボード、マイク、タッチセンサー式ディスプレイ、トランスデューサカードリーダー、磁気又は紙テープリーダー、タブレット、スタイラス、音声又は手書き認識周辺機器、USBポート、又はもちろん他のコンピュータ等のその他の周知の入力デバイスに接続する。最後に、CPU 2002は、必要に応じて2012に一般的に示されている外部接続を使用して、データベース又はコンピュータ又は通信ネットワーク等の外部デバイスに連結されてもよい。このような接続では、CPUがネットワークから情報を受信する可能性があるか、又は、本明細書に記載される方法のステップを実行する過程で、ネットワークに情報を出力する場合があると考えられる。一部の實現例では、核酸シーケンサ(2020)は、インターフェース2010の代わりに、又はそれに加えて、ネットワーク接続2012を介してCPU 2002に通信可能にリンクされてもよい。

【0331】

一実施形態では、コンピュータシステム2000等のシステムは、本明細書で説明するタスクの一部又は全てを実行し得るデータインポート、データ相関、及びクエリシステムとして使用される。データファイルを含む情報及びプログラムは、研究者によるアクセス又はダウンロードのためにネットワーク接続2012を介して提供されてもよい。或いは、そのような情報、プログラム、及びファイルは、記憶装置デバイスで研究者に提供され得る。

【0332】

特定の実施形態では、コンピュータシステム2000は、データ取得システム、例えば、試料からデータを捕捉する、マイクロアレイ、ハイスループットスクリーニングシステム、又は核酸シーケンサ(2020)に直接連結される。そのようなシステムからのデータは、システム2000による分析のためにインターフェース2010を介して提供される。或いは、システム2000によってプロセッシングされるデータは、データベース又は関連データの他のリポジトリ等のデータ記憶装置ソースから提供される。一旦、装置2000に入れば、一次記

憶装置2006又は大容量記憶装置2008等のメモリデバイスは、少なくとも一時的に関連データをバッファリング又は記憶する。メモリにはまた、データをインポート、分析、及び表示するための様々なルーチン及び/又はプログラムも記憶し得、これには、配列リード、UMI、配列リードを決定するコード、配列リードの折りたたみ、及びリードエラーの修正等が挙げられる。

【0333】

特定の実施形態では、本明細書で使用されるコンピュータとしては、ユーザ端末が挙げられ、これは任意のタイプのコンピュータ(例えば、デスクトップ、ラップトップ、タブレット等)、メディアコンピューティングプラットフォーム(例えば、ケーブル、衛星セットトップボックス、デジタルビデオレコーダー等)、携帯用コンピューティングデバイス(例えば、PDA、電子メールクライアント等)、携帯電話、又は任意の他のタイプのコンピューティングプラットフォーム又は通信プラットフォームであってもよい。

10

【0334】

特定の実施形態では、本明細書で使用されるコンピュータは、ユーザ端末と通信するサーバシステムも備えてもよく、そのサーバシステムは、サーバデバイス又は分散型サーバデバイスを備えてもよく、これには、メインフレームコンピュータ、ミニコンピュータ、スーパーコンピュータ、パーソナルコンピュータ、又はそれらの組み合わせが挙げられる。本発明の範囲から逸脱することなく、複数のサーバシステムを使用してもよい。ユーザ端末及びサーバシステムは、ネットワークを介して相互に通信し得る。ネットワークは、本発明の範囲を限定することなく、例えば、LAN(ローカルエリアネットワーク)、WAN(ワイドエリアネットワーク)、MAN(メトロポリタンエリアネットワーク)、ISDN(統合サービスデジタルネットワーク)等の有線ネットワーク、並びに無線LAN、CDMA、ブルートゥース(登録商標)、及び衛星通信ネットワーク等の無線ネットワークを含み得る。

20

【0335】

図5は、試験試料からコール又は診断を生成するための分散システムの一実現例を示している。試料収集場所01は、試験試料の取得に使用される。次いで、その試料はプロセッシング及び配列決定の場所03に提供され、そこで上記のように試験試料がプロセッシング及び配列決定される。場所03には、試料をプロセッシングする装置、及びプロセッシングされた試料を配列決定する装置が含まれる。本明細書の他の箇所でも説明するように、配列決定の結果は、通常は電子形式で提供され、図5の参照番号05で示されるインターネット等のネットワークに提供される、リードの収集である。

30

【0336】

配列データは、分析及びコール生成が実行される、遠隔地07に提供される。この場所には、コンピュータ又はプロセッサ等の1つ又は複数の強力な計算デバイスが含まれ得る。場所07の計算リソースが分析を完了し、受信した配列情報からコールを生成した後、そのコールはネットワーク05に中継される。一部の実現例では、場所07でコールが生成されるだけでなく、関連する診断も行われる。次に、コール及び/又は診断がネットワーク経由で送信され、図5に示されるように、試料収集場所01に戻される。説明したとおり、これは、コール又は診断の生成に関連する様々な操作が、様々な場所の間で分割され得る方法の多くのバリエーションの1つにすぎない。1つの一般的な変化型は、試料の収集及びプロセッシング、並びに単一の場所での配列決定が含まれる。別のバリエーションには、分析及びコール生成と同じ場所でのプロセッシングと配列決定の提供が含まれる。

40

【0337】

図6は、異なる場所で様々な操作を実行するための選択肢を詳しく説明している。図6に示す最もきめ細かい意味では、次の各操作は別々の場所で実行される:試料収集、試料プロセッシング、配列決定、リードアラインメント、コーリング(calling)、診断、並びに報告及び/又は計画開発。

【0338】

これらの操作のいくつかを集約する一実施形態では、試料プロセッシング及び配列決定は、1つの場所で実行され、リードアラインメント、コーリング、及び診断は別個の場所で実

50

行される。参照文字Aで特定される図6の部分を参照のこと。図6で、文字Bで特定される別の実現例では、試料収集、試料プロセッシング、及び配列決定は全て同じ場所で行われる。この実現例では、リードアライメント及びコーリングは2番目の場所で行われる。最終的に、診断及び報告、並びに/又は計画開発が3番目の場所で行われる。図6に文字Cで示した実現例では、試料収集は、最初の場所で行われ、試料プロセッシング、配列決定、リードアライメント、コーリング、及び診断は全て2番目の場所と一緒に実行され、報告及び/又は計画開発は3番目の場所で行われる。最後に、図6でDと標識された実現例では、試料収集は、最初の場所で行われ、試料プロセッシング、配列決定、リードアライメント、及びコーリングは全て2番目の場所で行われ、診断及び報告並びに/又は計画管理は3番目の場所で行われる。

10

【0339】

一実施形態は、腫瘍に関連する単純なヌクレオチドバリエーションの無細胞DNA(cfDNA)を分析するためのシステムを提供し、このシステムは、核酸試料を受け取り、その核酸試料から核酸配列情報を提供するためのシーケンサと;プロセッサと;このプロセッサでの実行のための指示を含むマシン可読記憶媒体とを備え、この指示は、(a)核酸試料から得られ、1つ又は複数の多型座位にある1つ又は複数の対立遺伝子にマッピングされた核酸配列リードを受け取るためのコード、(b)核酸配列リードを使用して、1つ又は複数の多型座位にある1つ又は複数の対立遺伝子のそれぞれに関する対立遺伝子数を決定するためのコード、及び(c)確率的混合物モデルを使用して、核酸試料中の1つ又は複数の寄与体の核酸の1つ又は複数の割合を定量するためのコードを含む。一部の實現例では、確率的混合物モデルを使用する工程は、確率的混合物モデルを対立遺伝子数に適用する工程を含む。確率的混合物モデルは、確率分布を使用して1つ又は複数の多型座位にある対立遺伝子数をモデル化し、確率分布は核酸配列リードの誤差を説明する。

20

【0340】

本明細書で提供される任意のシステムの一部の実施形態では、シーケンサは、次世代配列決定(NGS)を実行するように構成される。一部の實施形態では、シーケンサは、可逆的色素ターミネーターを用いた合成時配列決定を使用して大規模並列配列決定を実行するように構成される。他の実施形態では、このシーケンサは、ライゲーションによる配列決定を実行するように構成される。更に他の実施形態では、このシーケンサは、単一分子配列決定を実行するように構成される。

30

【実施例】

【0341】

(実施例1)

この実施例は、実際のDNA混合物試料から得られたデータを使用して、一部の實現例が、本明細書に開示された確率的アプローチを使用しない従来の技術よりも、DNA混合物試料の定量において、より高い精度及び信頼性、並びにより低い経験的バイアスを提供し得ることを示す。

【0342】

DNA混合物試料には、ゲノム(寄与体)由来の2つのDNAが含まれ、マイナー画分は、異なる試料で0.1%、0.2%、0.4%、及び2%である。一部の試料には3ngの入力DNAが含まれ、他の試料には10ngが含まれていた。試料は、2つのデザインで標的座位の数が異なる、2つのプライマー設計を示すために、Nack又はNack2として表示された2つの実験手順でプロセッシングされた。いくつかの試料は、MiSeq配列決定プラットフォームを使用してプロセッシングされ、一部はMiniSeqプラットフォームを使用してプロセッシングされた。

40

【0343】

試料データは、3つの異なる方法を使用して分析された。Table 2(表2)は、複数の混合画分にわたる変動係数の平均(CV、 $\text{standard_deviation_of_predictions/true_fraction}$ として定義)値、及び3つの異なる方法の複数の混合画分にわたって変動係数の平均+バイアス(CVB、一般にCV(RMSD)として示され、 $\text{RMSD/true_fraction}$ として定義されている)値を、様々な試料と実験手順を使用して、示す。最初の方法は、配列決定エラーをモデ

50

ル化するための二項分布を含む確率モデルを適用する。最初の方法は、上記でSeqモデルと呼ばれる一部の実現例に対応している。最初の方法(Seq)のデータを、Table 8(表8)の3行目に示す。2番目の方法は、DNA抽出エラー、PCR増幅エラー、配列決定エラーを説明する確率分布を含む確率的混合物モデルを適用する。2番目の方法は、上記のExtraction-PCR-Seq Modelとして説明されている一部の實現例に対応している。2番目の方法(EPS)のデータは、Table 8(表8)の4行目に示されている。

【0344】

第3の方法は、決定論的線形回帰モデルを適用して、対立遺伝子アカウントデータを記述する。次のように、データの合計二乗誤差を推定する。

$$E=[r_i-p_i(\beta)]T \cdot [r_i-p_i(\beta)]$$

ここで、 r_i は観測された対立遺伝子の画分、 $p_i=G \cdot \beta$ は、座位 i の予想される対立遺伝子の画分であり、これは β の線形関数であり、 G は n 座位及び d ドナーの遺伝子型の行列であり、 β は未知の寄与体画分の長さ d ベクトルである。3番目の方法(NaiveLM)のデータは、Table 8(表8)の5行目に示す。

Seq又はEPS方法で寄与体の画分を定量するために、寄与体の遺伝子型情報を使用しなかったが、NaiveLM方法で使用されたことは注目に値する。Seq方法とEPS方法は、寄与体の遺伝子型情報を使用する必要がなかったという事実にもかかわらず、NaiveLM方法よりも小さい変動係数値で示されるように、より信頼され得る結果を生み出した。更に、Seq方法及びEPS方法は、NaiveLM方法よりも小さいCVB値で示されるように、バイアスが低くなった。Table 8(表8)では、3つの方法の中で最良の結果を太字で示している。つまり、確率的混合物モデルを使用する2つの方法は、線形回帰法よりも信頼性が高く、正確で、偏りの少ない結果を生じた。

【0345】

【表8】

Table 8.

アルゴリズム	遺伝子型	Nack (3ng, 10ng)		Nack2 (3ng, 10ng)		Nack2 MiniSeq (3ng, 10ng)		Nack2 検証 (3ng, 10ng)	
		CV	CVB	CV	CVB	CV	CVB	CV	CVB
Seq	使用せず	0.151	0.796	0.214	0.688	0.175	0.414	0.21	0.488
EPS	使用せず	0.139	0.207	0.126	0.193	0.125	0.216	0.165	0.253
NaiveLM	使用	0.771	1.117	0.889	1.388	3.818	1.03	8.83	1.407

【0346】

(実施例2)

図7A～図7Fは、実際のDNA混合物試料から得られたデータを使用して、一部の實現例がDNA混合物試料を効果的に定量及びデコンボリューションし得ることを示す実施例の結果を示す。この実施例は、一部の實現例がDNA混合デコンボリューションのシグナルレベルを改善し得ることを示している。この実施例では、データは狭い事前確率を使用して分析した。

【0347】

試料は、2つの寄与体由来のDNAを含み、様々な試料が2つの寄与体について60%～40%、75%～25%、90%～10%、及び95%～5%の画分を有する。その試料には、対象NA12878及びNA18507のそれぞれ3つの複製が含まれていた。

【0348】

図7Aは、一部の実現例によって定量されたメジャーな寄与体画分(又は図では「メジャー頻度」と呼ばれる)を示している。横軸は、メジャーな寄与体の実際の寄与体頻度を示している。縦軸は、確率的混合物モデルによって推定されたメジャーな寄与体画分(2.5%に最も近い)を示している。このデータは、確率的混合物モデルが、アイデンティティラインの近くに配置されたデータポイントによって示されるように、真の画分に非常に近い予測を提供することを示している。

【0349】

図7Bは、4つのサブプロットにおける4つの異なる対立遺伝子の確率的混合物モデルにより予測される、メジャーな寄与体及びマイナーな寄与体の遺伝子型を示す。左の2つのサブプロットは、75~25の寄与体画分の試料から得られた結果を示している。右の2つのサブプロットは、60~40の画分の混合物試料から得られた結果を示している。横軸は、ある座位の異なる対立遺伝子の表示を示す。縦軸は、座位の対立遺伝子数を示している。モデルによって予測された全ての遺伝子型は、左上のサブプロットに示された座位D4S2408のマイナーな寄与体の1つの対立遺伝子を除いて正しいものであった。その座位では、真のマイナーな寄与体の遺伝子型は(10、10)であるが、モデルは(8、10)と予測した。興味深いことに、マイナーな寄与体に対するこの座位の予測の信頼レベルは、68.6%という比較的低いレベルである。この実施例では、コール基準を70%超に設定することにより、誤った予測を削除し得る。

【0350】

図7Cは、寄与体の遺伝子型の正しいコールと誤ったコールの数を示している。横軸は、メジャーな寄与体の実際の寄与体の画分を示す(図では「メジャー頻度」と標識されている)。縦軸は、正しいコールと誤ったコールの数を示している。「x」記号はメジャーな寄与体のデータを示し、円記号はマイナーな寄与体のデータを示す。黒い記号は、正しいコールを示すが、灰色の記号は誤ったコールのデータを示す。28の水平線は、理論上の最大の正しいコールを示している。図7Cは、信頼度90%の閾値で行われたコールのデータを示している。図7Cのデータは、異なる寄与体の画分にまたがり、正しいコール数が比較的多いのに対し、誤ったコールは比較的少なく、一貫して5未満であることを示している。また、図7Cは、寄与体の画分が60%から95%に増加するにつれて、メジャーな寄与体の正しいコールが増大し、理論上の最大レベルに近づいたことを示している。

【0351】

図7Dは、同じデータを使用するが、より高いコール基準を99%とする正しいコール及び誤ったコールを示している。繰り返すが、正しいコールの数は、異なるメジャーな寄与体の画分にまたがり一貫して高く、誤ったコールの数は比較的少なく、一貫して5未満である。図7Dのコール基準は、図7Cの基準よりも高いため、正しいコールと誤ったコールの両方の数は低くなる。しかし、90%及び95%の画分でのメジャーな寄与体の正しいコールは依然として高く、理論上の最大値に近かった。

【0352】

図7Eは、コール基準が99.9%に増加したことを除いて、図7C及び図7Dと同様のデータを示している。コール閾値が更に高いため、正しいコールと誤ったコールの数は、図7Dの結果よりもわずかに少なくなる。重要なことは、既知の理由でモデルの仮定から逸脱することが知られている3つの座位を除いて、この信頼レベルでは誤ったコールはない。これらの座位は、分析では回避され得る。これらの場合とは別に、このモデルは誤ったコールに高い信頼度を与えることはない。したがって、遺伝子型コールの不確実性を適切に定量している。

【0353】

図7C~図7Eの結果、確率的混合物モデルが寄与体の遺伝子型を正確に決定し得ることが示されている。様々なアプリケーションの様々なニーズに基づいて、様々なコール基準値を採用して、望ましい感度及び選択性を達成してもよい。

【0354】

図7Fは、既知の寄与体のDNAがDNA混合物試料に含まれているか否かに関する正しいコ

10

20

30

40

50

ールと誤ったコールの数を示している。横軸は、メジャーな寄与体の実際の寄与体の画分を示す(「真のメジャー頻度」と標識されている)。縦軸は、試料に遺伝子型が含まれている座位ごとの証拠値を示している。「x」記号は、メジャーな寄与体のデータを示しており、円記号はマイナーな寄与体のデータを示す。図7Fのデータは、試料に2つの寄与体が含まれているという比較的高いレベルの証拠があったことを示している。驚くことではないが、メジャーな寄与体の画分が90%と95%の場合、マイナーな寄与体の証拠レベルは比較的低かった。

【0355】

(実施例3)

図8A~図8Dは、シミュレーションデータを使用して、一部の実現例がDNA混合物試料を効果的に定量及びデコンボリューションし得ることを示す実施例の結果を示している。この実施例は、一部の實現例がDNA混合物デコンボリューションのシグナルレベルを改善し得ることを示している。

10

【0356】

シミュレーションには、4つの異なる設計がある。Easy2-寄与体(80-20);Difficult2-寄与体(55-45);Easy3-寄与体(60-30-10);及びDifficult3-寄与体(50-30-20)。簡単なデザインには、難しいデザインよりも更に離れた寄与体画分がある。

【0357】

シミュレーションは、50個の座位及び6個の対立遺伝子のデータを含む。対立遺伝子のバランスは、以下に依存する:元の試料の分子数(固定:6000)、平均分子検出率(10%に固定;すなわち600分子が平均して座位を検出)、分子検出率の対立遺伝子間での変動(ある範囲で変化する)、及びサンプリングノイズ。スタッター率は1%又は2%としてシミュレートされ、ドロップアウト率は1%である。結果は、幅広い事前確率を想定して取得した。

20

【0358】

easy3-寄与体(60-30-10)混合物試料について、全ての寄与体頻度は、真の値の2.5%以内で推測された。difficult3-寄与体(50-30-20)の混合物試料では、全ての寄与体頻度が真の値の7.5%以内で推測された。

【0359】

図8A~図8Dは、easy3-寄与体(60-30-10)混合物試料のデータを示す。図8Aは、寄与体の遺伝子型の正しいコールと誤ったコールの数を示している。横軸は、メジャーな寄与体の対立遺伝子のバランスを示している。縦軸は、正しいコールと誤ったコールの数を示している。黒い記号は1%のスタッター率のデータを示し、灰色の記号は2%のスタッター率のデータを示す。実線は正しいコールを示し、一方、破線は誤ったコールのデータを示す。図8Aは、信頼度90%という閾値で行われたコールのデータを示している。図8Aのデータは、異なる対立遺伝子バランス値全体で正しいコール数が比較的多いのに対し、誤ったコールは一貫してゼロに近いことを示している。

30

【0360】

図8Bは、図8Aと同じデータを使用するが、より高いコール基準を99%とする正しいコールと誤ったコールを示している。正しいコールの数は、図8Aの数よりも有意に少ないが、誤ったコールは底打ちされており、このアプリケーションでは99%の閾値が厳しすぎる場合があることが示されている。図8Cは、コール基準が99.9%に増大したことを除いて、図8A及び図8Bと同様のデータを示している。コール閾値が更に高いために、正しいコールの数は更に減少する。図3A~図3Cの結果、確率的混合物モデルが寄与体の遺伝子型を正確に決定し得、この実施例の適切な閾値が90%近く又は99%未満に設定され得ることが示されている。

40

【0361】

図8Dは、3つの寄与体のDNAのうちの1つが、DNA混合物試料に含まれているか否かに関する正しいコール及び誤ったコールの数を示している。横軸は、メジャーな寄与体の対立遺伝子のバランスを示している。縦軸は、試料に遺伝子型が含まれている、座位ごとの証拠値を示している。実線は1%のスタッターエラーのデータを示し、破線は2%のスタッタ

50

【図 3】

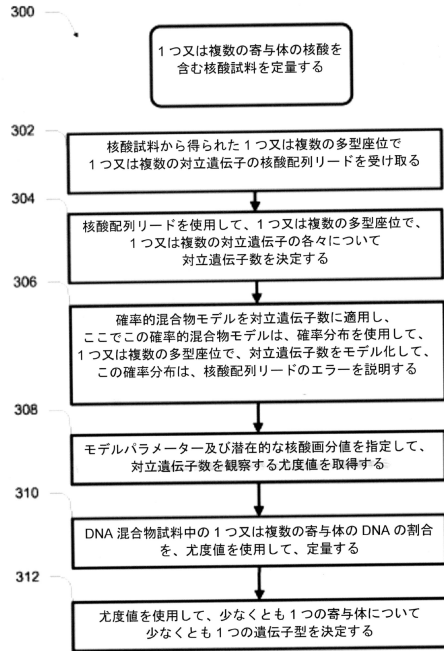


Figure 3

【図 4】

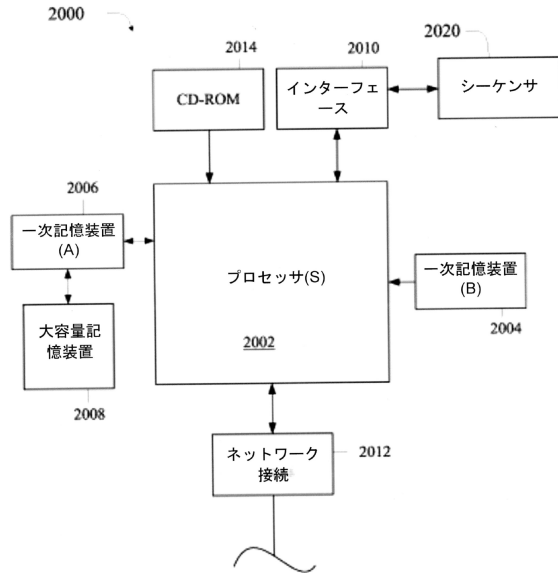


Figure 4

【図 5】

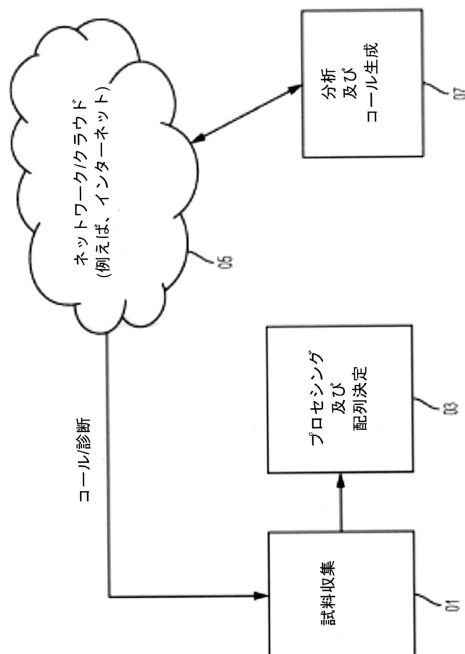


Figure 5

【図 6】

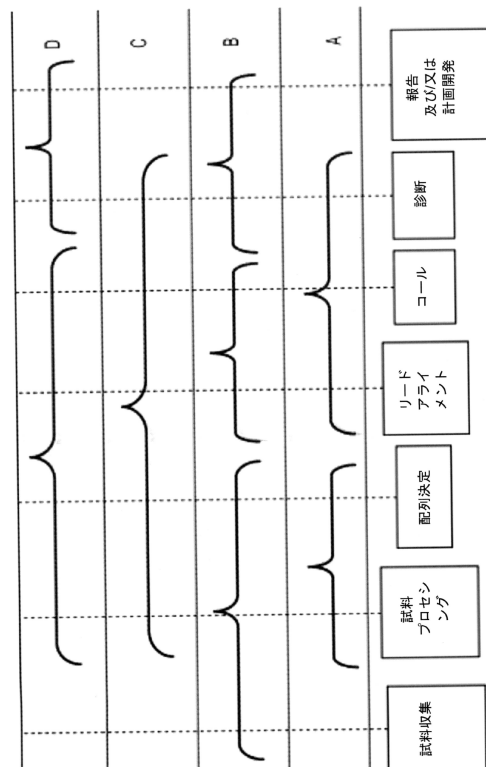


Figure 6

【図 7 A】

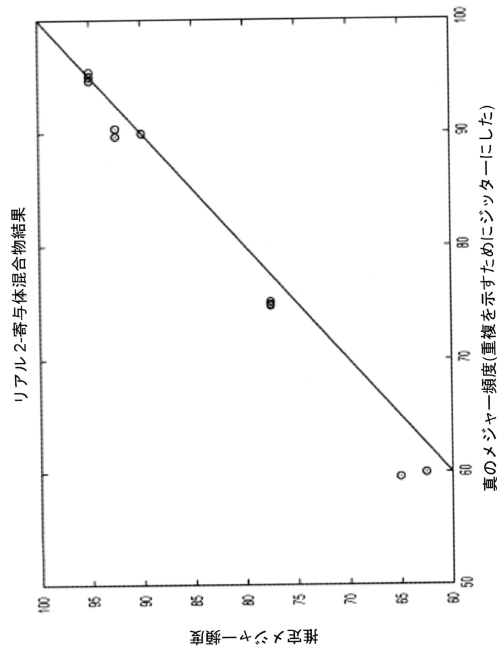


Figure 7A

【図 7 B】

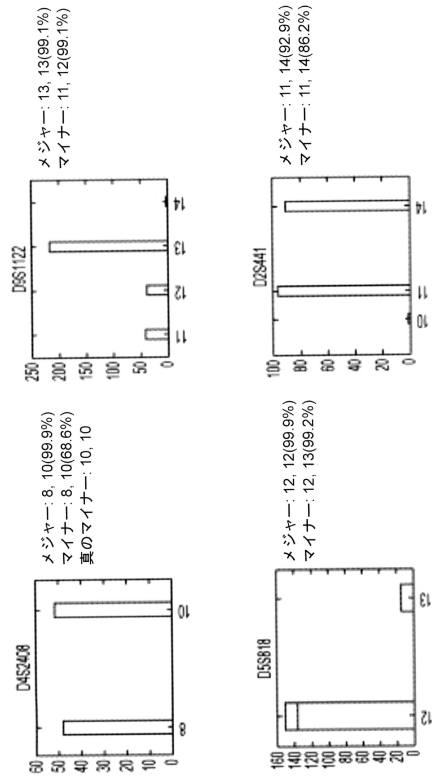


Figure 7B

【図 7 C】

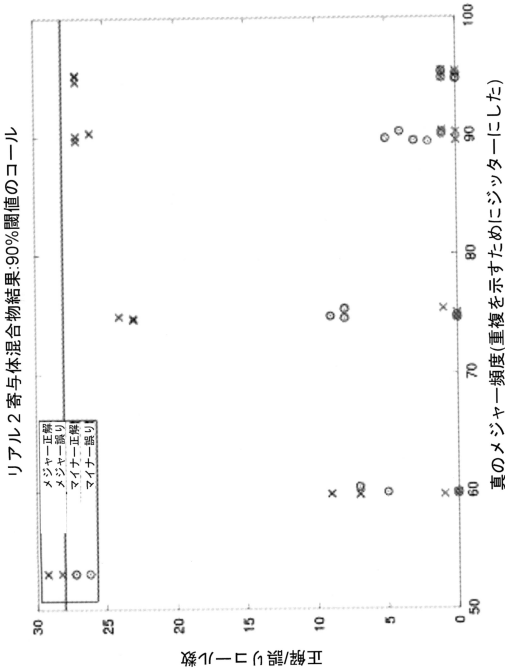


Figure 7C

【図 7 D】

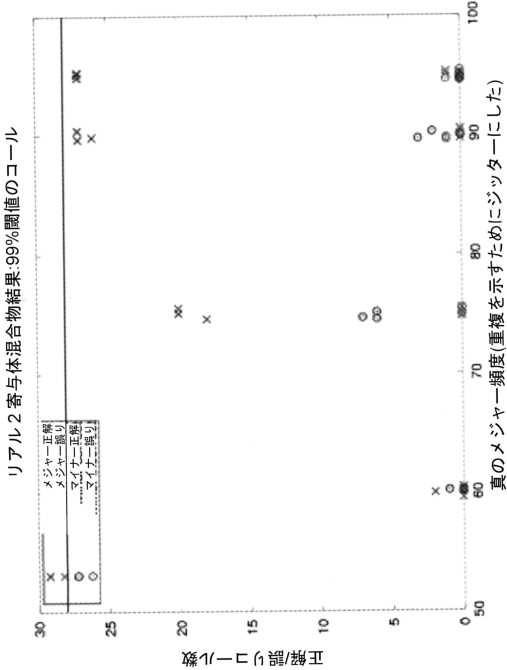


Figure 7D

10

20

30

40

50

【図 7 E】

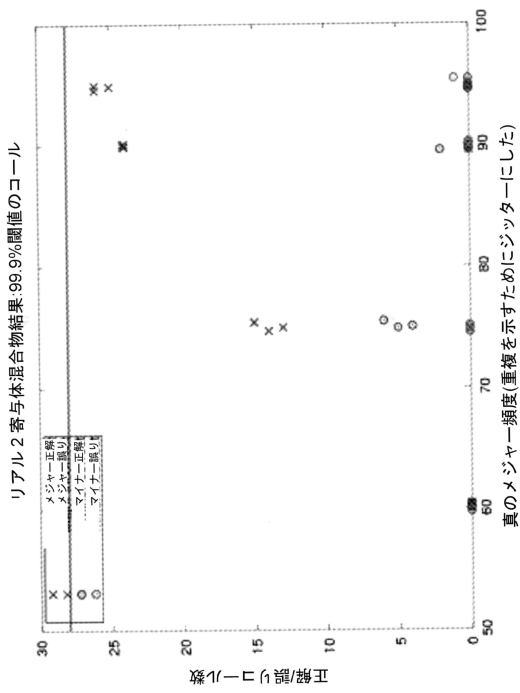


Figure 7E

【図 7 F】

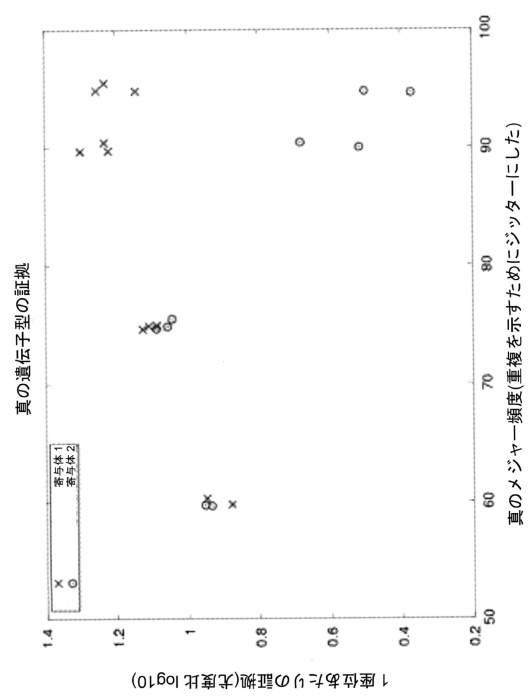


Figure 7F

【図 8 A】

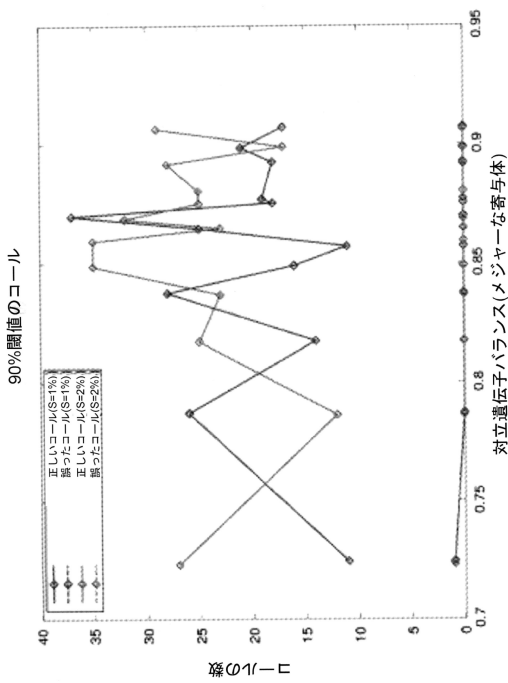


Figure 8A

【図 8 B】

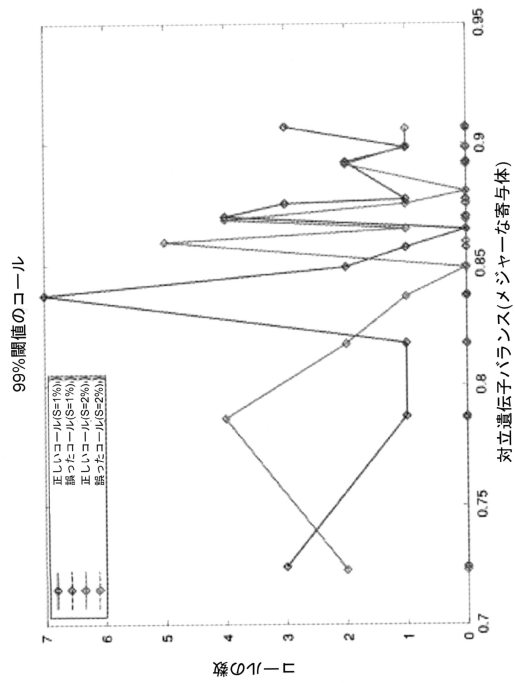


Figure 8B

10

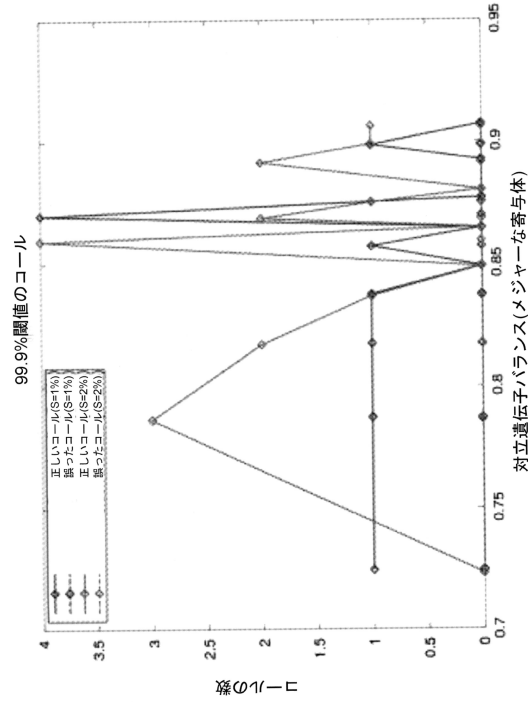
20

30

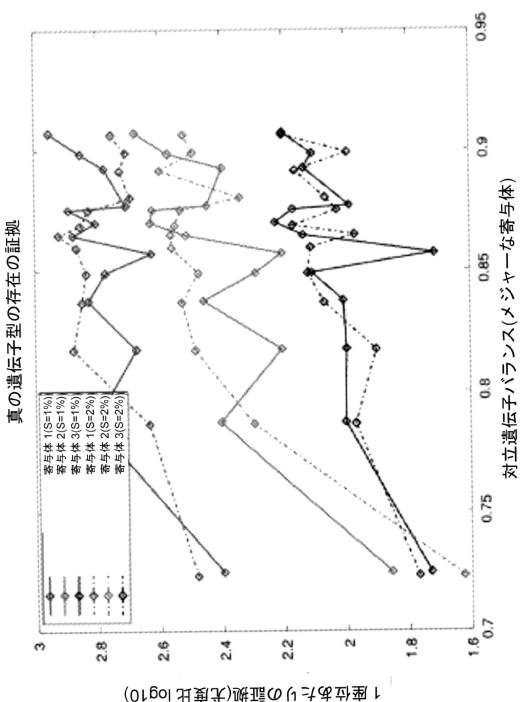
40

50

【図 8 C】



【図 8 D】



10

20

30

40

50

フロントページの続き

- (72)発明者 ヨハン・フィリックス・シュレジンジャー
 アメリカ合衆国・カリフォルニア・９２１２２・サン・ディエゴ・イルミナ・ウェイ・５２００
- (72)発明者 ライアン・ケリー
 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
- 審査官 山内 裕史
- (56)参考文献 米国特許出願公開第２０１２／０２６４１２１（ＵＳ，Ａ１）
 特開２０１６－１８４４２９（ＪＰ，Ａ）
 特表２００６－５１９９７７（ＪＰ，Ａ）
- (58)調査した分野 (Int.Cl.，ＤＢ名)
 Ｇ１６Ｂ ５／００ － ９９／００