

(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号

特許第7046007号

(P7046007)

(45)発行日 令和4年4月1日(2022.4.1)

(24)登録日 令和4年3月24日(2022.3.24)

(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 28 (全141頁)

(21)出願番号	特願2018-561218(P2018-561218)	(73)特許権者	595117091
(86)(22)出願日	平成29年5月25日(2017.5.25)		ベクトン・ディキンソン・アンド・カン
(65)公表番号	特表2019-522268(P2019-522268		パニー
	A)		BECTON, DICKINSON A
(43)公表日	令和1年8月8日(2019.8.8)		ND COMPANY
(86)国際出願番号	PCT/US2017/034576		アメリカ合衆国 ニュー・ジャージー 0
(87)国際公開番号	WO2017/205691		7 4 1 7 - 1 8 8 0 フランクリン・レ
(87)国際公開日	平成29年11月30日(2017.11.30)		イクス ベクトン・ドライブ 1
審査請求日	令和2年5月25日(2020.5.25)		1 BECTON DRIVE, FRA
(31)優先権主張番号	62/381,945		NKLIN LAKES, NEW JE
(32)優先日	平成28年8月31日(2016.8.31)		RSEY 0 7 4 1 7 - 1 8 8 0, UN
(33)優先権主張国・地域又は機関			ITED STATES OF AMER
	米国(US)		ICA
(31)優先権主張番号	62/401,720	(74)代理人	100094569
(32)優先日	平成28年9月29日(2016.9.29)		弁理士 田中 伸一郎
	最終頁に続く		最終頁に続く

(54)【発明の名称】 分子標識カウントの調節方法

(57)【特許請求の範囲】

【請求項1】

サンプル中の核酸標的の数を決定する方法であって、

(a) 複数の確率バーコードを用いて、複数の核酸標的に確率バーコードを付けて、複数の確率バーコード付き核酸標的を生成する工程、ここで前記複数の確率バーコードの各々は分子標識を含む；

(b) 前記確率バーコード付き核酸標的のシーケンシングデータを取得する工程；及び

(c) 前記複数の核酸標的の1つ以上について、以下の(i)～(iv)の工程；

(i) 前記シーケンシングデータ中の前記核酸標的に関連付けられた識別可能な配列を有する分子標識の数をカウントする工程；

(ii) 方向近接性を用いて、前記核酸標的の分子標識のクラスターを同定する工程、ここで、前記同定する工程は、識別可能な配列を有する全ての分子標識について再帰的に、子供分子標識が1以上の親分子標識を含むクラスターに属するかどうかを決定することを含み、前記クラスター内の前記核酸標的の前記分子標識は、1つ以上の親分子標識と、前記1つ以上の親分子標識の子供分子標識とを含み、且つ、前記親分子標識の発生数は、所定の方向近接性発生数閾値以上である；

(iii) (ii) で同定された前記核酸標的の分子標識の前記クラスターを用いて、(b) で得られた前記シーケンシングデータを折りたたむ工程；及び

(iv) 前記核酸標的の数を推定する工程、ここで推定される前記核酸標的の数は、(ii) の前記シーケンシングデータの折りたたみ後の、(i) でカウントされた前記シー

ケンシングデータ中の前記核酸標的に関連付けられた識別可能な配列を有する分子標識の数と相関する、
を含む、方法。

【請求項 2】

クラスター内の前記核酸標的の分子標識が、互いの所定の方向近接性閾値内にある、請求項 1 に記載の方法。

【請求項 3】

前記方向近接性閾値が、1 のハミング距離である、請求項 2 に記載の方法。

【請求項 4】

前記所定の方向近接性発生数閾値が、 $2 \times (\text{子供分子標識の発生数}) - 1$ である、請求項 1 に記載の方法。

【請求項 5】

(i i) で同定された前記核酸標的の分子標識の前記クラスターを用いて、(b) で得られた前記シーケンシングデータを折りたたむ工程が、
前記子供分子標識の発生数を前記親分子標識に帰属させる工程
を含む、請求項 1 ～ 4 のいずれか一項に記載の方法。

【請求項 6】

前記核酸標的のシーケンシング深度を決定する工程をさらに含む、請求項 1 ～ 5 のいずれか一項に記載の方法。

【請求項 7】

前記核酸標的の前記シーケンシング深度が所定のシーケンシング深度閾値を超える場合、前記核酸標的の数を推定する工程が、(i) でカウントされた前記シーケンシングデータを調節する工程を含む、請求項 6 に記載の方法。

【請求項 8】

(i) でカウントされた前記シーケンシングデータを調節する工程が、
前記核酸標的の分子標識を閾値化して、(b) で得られた前記シーケンシングデータ中の前記核酸標的に関連付けられた真の分子標識および偽の分子標識を決定する工程
を含む、請求項 7 に記載の方法。

【請求項 9】

前記核酸標的の前記分子標識を閾値化する工程が、前記核酸標的の前記分子標識について統計解析を実施する工程を含む、請求項 8 に記載の方法。

【請求項 10】

前記統計解析を実施する工程が、
前記核酸標的の前記分子標識の分布およびそれらの発生数を 2 つのネガティブ二項分布に当てはめる工程；

前記 2 つのネガティブ二項分布を用いて真の分子標識の数 n を決定する工程；及び

(b) で得られた前記シーケンシングデータから前記偽の分子標識を除去する工程、ここで前記偽の分子標識は、 n 番目に豊富な分子標識の発生数よりも低い発生数を有する分子標識を含み、前記真の分子標識が、 n 番目に豊富な分子標識の発生数以上の発生数を有する分子標識を含む

を含む、請求項 9 に記載の方法。

【請求項 11】

前記ネガティブ二項分布が、前記真の分子標識に対応する第 1 のネガティブ二項分布と、前記偽の分子標識に対応する第 2 のネガティブ二項分布を含む、請求項 10 に記載の方法。

【請求項 12】

核酸標的の数を決定する方法であって、

(a) 複数の確率バーコードを用いて、複数の核酸標的に確率バーコードを付けて、複数の確率バーコード付き核酸標的を生成する工程、ここで前記複数の確率バーコードの各々は分子標識を含む；

(b) 前記確率バーコード付き核酸標的のシーケンシングデータを取得する工程；

10

20

30

40

50

(c) 前記シーケンシングデータにおける前記核酸標的のシーケンシングステータスを決定する工程；

(d) 前記複数の核酸標的の1つ以上について、以下の(i)～(iii)：

(i) 前記シーケンシングデータ中の前記核酸標的に関連付けられた識別可能な配列を有する分子標識の数をカウントする工程；

(ii) 前記シーケンシングデータ中の前記核酸標的に関連付けられた識別可能な配列を有するノイズ分子標識の数を決定する工程、ここで、当該工程は、

前記シーケンシングステータスに基づきネガティブ二項分布当てはめ条件が満たされる場合、

(1) シグナルネガティブ二項分布を、(i)でカウントされた前記シーケンシングデータ中の前記核酸標的に関連付けられた識別可能な配列を有する前記分子標識の数に当てはめること、ここで前記シグナルネガティブ二項分布は、シグナル分子標識である、(i)でカウントされた前記シーケンシングデータ中の前記核酸標的に関連付けられた識別可能な配列を有する分子標識の数に対応する；

(2) ノイズネガティブ二項分布を、(i)でカウントされた前記シーケンシングデータ中の前記核酸標的に関連付けられた識別可能な配列を有する前記分子標識の数に当てはめること、ここで前記ノイズネガティブ二項分布は、ノイズ分子標識である、(i)でカウントされた前記シーケンシングデータ中の前記核酸標的に関連付けられた識別可能な配列を有する分子標識の数に対応する；及び

(3) 前記(1)で当てはめた前記シグナルネガティブ二項分布および(2)で当てはめた前記ノイズネガティブ二項分布を用いて、前記ノイズ分子標識の数を決定すること、を含む；及び

(iii) 前記核酸標的の数を推定する工程、ここで推定される前記核酸標的の数は、(ii)で決定された前記ノイズ分子標識の数に応じて調節された、(i)でカウントされた前記シーケンシングデータ中の前記核酸標的に関連付けられた前記識別可能な配列を有する分子標識の数と相関する、を含む、方法。

【請求項13】

前記シーケンシングデータ中の前記核酸標的の前記シーケンシングステータスが、飽和シーケンシング、過少シーケンシング、または過剰シーケンシングである、請求項12に記載の方法。

【請求項14】

前記飽和シーケンシングステータスが、所定の飽和閾値よりも大きい、識別可能な配列を含む分子標識の数を有する前記核酸標的によって決定される、請求項13に記載の方法。

【請求項15】

前記シーケンシングデータ中の前記核酸標的の前記シーケンシングステータスが、前記飽和シーケンシングステータスである場合、(ii)で決定される前記ノイズ分子標識の数が、ゼロである、請求項13または14に記載の方法。

【請求項16】

前記過少シーケンシングステータスが、所定の過少シーケンシング閾値より小さい深度を有する前記核酸標的によって決定され、前記核酸標的の前記深度が、前記シーケンシングデータ中の前記核酸標的に関連付けられた識別可能な配列を有する前記分子標識の、平均、最小、または最大深度を含む、請求項13～15のいずれか一項に記載の方法。

【請求項17】

前記過少シーケンシング閾値は、識別可能な配列を有する前記分子標識の数とは無関係である、請求項16に記載の方法。

【請求項18】

前記シーケンシングデータ中の前記核酸標的の前記シーケンシングステータスが、前記過少シーケンシングステータスである場合、(ii)で決定される前記ノイズ分子標識の数が、ゼロである、請求項13～17のいずれか一項に記載の方法。

10

20

30

40

50

【請求項 19】

前記過剰シーケンシングステータスが、所定の過剰シーケンシング閾値より大きい深度を有する前記核酸標的によって決定され、前記核酸標的の前記深度が、前記シーケンシングデータ中の前記核酸標的に関連付けられた識別可能な配列を有する前記分子標識の、平均、最小、または最大深度を含む、請求項 13～18 のいずれか一項に記載の方法。

【請求項 20】

前記シーケンシングデータ中の前記核酸標的の前記シーケンシングステータスが、前記過剰シーケンシングステータスである場合、

前記シーケンシングデータ中の前記核酸標的に関連付けられた識別可能な配列を有する前記分子標識の数を、前記所定の過剰シーケンシング閾値付近にサブサンプリングする工程をさらに含む、請求項 19 に記載の方法。

10

【請求項 21】

前記ネガティブ二項分布当てはめ条件が、前記シーケンシングデータ中の前記核酸標的の前記シーケンシングステータスが、前記過剰シーケンシングステータスまたは前記過剰シーケンシングステータスではないことを含む、請求項 12 に記載の方法。

【請求項 22】

(v) で当てはめた前記シグナルネガティブ二項分布および (v i) で当てはめた前記ノイズネガティブ二項分布を用いて、前記ノイズ分子標識の数を決定する工程が、

前記シーケンシングデータ中の前記核酸標的に関連付けられた前記識別可能な配列の各々について、

20

前記識別可能な配列のシグナル確率が、前記シグナルネガティブ二項分布内であることを決定する工程と；

前記識別可能な配列のノイズ確率が、前記ノイズネガティブ二項分布内であることを決定する工程と；

前記シグナル確率が前記ノイズ確率より小さければ、前記識別可能な配列がノイズ分子標識であることを決定する工程と、

を含む、請求項 12～21 のいずれか一項に記載の方法。

【請求項 23】

前記シーケンシングデータ中の前記核酸標的に関連付けられた識別可能な配列を有する前記ノイズ分子標識の数を決定する工程が、

30

前記シーケンシングデータ中の前記核酸標的の前記シーケンシングステータスが、前記過剰シーケンシングステータスまたは前記過剰シーケンシングステータスではなく、かつ、

(i) でカウントされた前記シーケンシングデータ中の前記核酸標的に関連付けられた識別可能な配列を有する前記分子標識の数が、疑似点閾値より少ない場合、(i i) で前記シーケンシングデータ中の前記核酸標的に関連付けられた識別可能な配列を有する前記ノイズ分子標識の数を決定する前に、前記シーケンシングデータ中の前記核酸標的に関連付けられた識別可能な配列を有する前記分子標識の数の疑似点を加える工程を含む、

請求項 13～22 のいずれか一項に記載の方法。

【請求項 24】

前記シーケンシングデータ中の前記核酸標的に関連付けられた識別可能な配列を有する前記ノイズ分子標識の数を決定する工程が、

40

前記シーケンシングデータ中の前記核酸標的の前記シーケンシングステータスが、前記過剰シーケンシングステータスまたは前記過剰シーケンシングステータスではなく、かつ、

(i) でカウントされた前記シーケンシングデータ中の前記核酸標的に関連付けられた識別可能な配列を有する前記分子標識の数が、疑似点閾値以上である場合、(i i) で前記シーケンシングデータ中の前記核酸標的に関連付けられた識別可能な配列を有する前記ノイズ分子標識の数を決定する際に、非ユニーク分子標識を除去する工程を含む、

請求項 13～23 のいずれか一項に記載の方法。

【請求項 25】

前記非ユニーク分子標識を除去する工程が、前記シーケンシングデータ中の前記核酸標的

50

に関連付けられた識別可能な配列を有する前記分子標識の数が、所定の再使用分子標識閾値より大きい場合、(i i) で前記シーケンシングデータ中の前記核酸標的に関連付けられた識別可能な配列を有する前記ノイズ分子標識の数を決定する際に、前記非ユニーク分子標識を除去する工程を含む、請求項 2 4 に記載の方法。

【請求項 2 6】

前記非ユニーク分子標識を除去する工程が、

前記シーケンシングデータ中の前記核酸標的に関連付けられた識別可能な配列を有する前記分子標識の数について非ユニーク分子標識の理論上の数を決定する工程；及び

前記シーケンシングデータ中の前記核酸標的に関連付けられた識別可能な配列を有する分子標識のうち n 番目に豊富な前記分子標識よりも大きい発生数を有する分子標識を除去する工程、ここで n は、非ユニーク分子標識の理論数である

10

を含む、請求項 2 4 又は 2 5 に記載の方法。

【請求項 2 7】

ハードウェアプロセッサと、

前記ハードウェアプロセッサによって実行される場合、前記プロセッサに請求項 1 ～ 2 6 のいずれか一項に記載の方法を実行させる命令を記憶した非一過性メモリと、を含む、核酸標的の数を決定するためのコンピュータシステム。

【請求項 2 8】

請求項 1 ～ 2 6 のいずれか一項に記載の方法を実行するためのコードを含むソフトウェアプログラムを含む、コンピュータ読取り媒体。

20

【発明の詳細な説明】

【技術分野】

【0 0 0 1】

関連出願

本出願は、3 5 U . S . C . § 1 1 9 (e) に従い、2 0 1 6 年 5 月 2 6 日出願の米国仮特許出願第 6 2 / 3 4 2 1 3 7 号明細書；2 0 1 6 年 8 月 3 1 日出願の米国仮特許出願第 6 2 / 3 8 1 9 4 5 号明細書；および 2 0 1 6 年 9 月 2 9 日出願の米国仮特許出願第 6 2 / 4 0 1 7 2 0 号明細書に基づく優先権を主張する。これらの出願各々の内容は、本出願をもってその全体が参照により明示的に組み込まれる。

【背景技術】

30

【0 0 0 2】

本開示は、概して、核酸バーコーディング、より具体的には、分子標識を用いた P C R およびシーケンシングエラーの訂正の分野に関する。

【0 0 0 3】

関連分野の説明

確率バーコーディングなどの方法および技術は、細胞分析において、特に、たとえば、逆転写、ポリメラーゼ連鎖反応 (P C R) 増幅、および次世代シーケンシング (N G S) を用いて細胞の状態を判定するために、遺伝子発現プロファイルを解読する上で有用である。しかし、これらの方法および技術は、置換エラー (1 つ以上の塩基を含む) および非置換エラーなどのエラーを導入する恐れがあり、未訂正のままだと、過大評価された分子カウントが生じる。従って、確率バーコーディングを用いて推定される正確な分子カウントを取得するために、さまざまなエラーを訂正することができる方法および技術が求められる。

40

【発明の概要】

【課題を解決するための手段】

【0 0 0 4】

本明細書には、標的の数を決定する方法が開示される。いくつかの実施形態では、本方法は、(a) 複数の確率バーコードを用いて、複数の標的に確率バーコードを付けて、複数の確率バーコード付き標的を生成する工程であって、複数の確率バーコードの各々が分子標識を含む工程と；(b) 確率バーコード付き標的のシーケンシングデータを取得する工

50

程と；（c）複数の標的の1つ以上について：（i）シーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数をカウントする工程と；（ii）（b）で得られたシーケンシングデータ中の標的のクオリティステータスを決定する工程と；（iii）（b）で得られたシーケンシングデータ中の1つ以上のシーケンシングデータエラーを決定する工程であって、シーケンシングデータ中の1つ以上のシーケンシングデータエラーを決定する工程が、以下：シーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数、シーケンシングデータ中の標的のクオリティステータス、および複数の確率バーコード中の識別可能な配列を有する分子標識の数のうち1つ以上を決定することを含む工程と；（iv）標的の数を推定する工程であって、推定された標的の数が、（iii）で決定された1つ以上のシーケンシングデータエラーに応じて調節された、（i）でカウントされたシーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数と相関する工程と、を含む。工程（i）、（ii）、（iii）、および（iv）は、複数の標的の各々について実施することができる。本方法は、多重化することができる。

【0005】

いくつかの実施形態では、本方法は、1つ以上のシーケンシングデータエラーを決定する前に、（b）で得られたシーケンシングデータを折りたたむ工程をさらに含む。（b）で得られたシーケンシングデータを折りたたむ工程は、類似した分子標識を有し、かつ、所定の折りたたみ発生数閾値よりも少ない発生数を有する標的のコピーを、複数の標的について同じ分子標識を有するものとして帰属させる工程を含み、ここで、標的の2つのコピーは、標的の2つのコピーの分子標識の配列が少なくとも1塩基相違する場合、類似の分子標識を有する。

【0006】

いくつかの実施形態では、所定の折りたたみ発生数閾値は、確率バーコードが、識別可能な配列を有する約6561の分子標識を含む場合、7となりうる。確率バーコードが、識別可能な配列を有する約65536の分子標識を含む場合、所定の折りたたみ発生数閾値は17となりうる。標的の2つのコピーは、標的の2つのコピーの分子標識の配列が少なくとも1塩基相違する場合、類似の分子標識を有する。いくつかの実施形態では、分子標識は、5～20個のヌクレオチドを含む。さまざまな確率バーコードの分子標識は、互いに異なっていてよい。複数の確率バーコードは、識別可能な配列を有する約6561の分子標識を含む。複数の確率バーコードは、識別可能な配列を有する約65536の分子標識を含む。

【0007】

いくつかの実施形態では、50ヌクレオチド以上のリード長を有する複数の標的の配列を含む。シーケンシングデータは、75ヌクレオチド以上のリード長を有する複数の標的の配列を含む。シーケンシングデータは、100ヌクレオチド以上のリード長を有する複数の標的の配列を含む。（b）で得られたシーケンシングデータは、複数の確率バーコード付き標的に対してポリメラーゼ連鎖反応（PCR）増幅を実施することによって生成することができる。

【0008】

いくつかの実施形態では、1つ以上のシーケンシングデータエラーは、PCR導入エラー、シーケンシング導入エラー、バーコード混入に起因するエラー、ライブラリー作製エラー、またはそれらの任意の組合せでありうる。PCR導入エラーは、PCR増幅エラー、PCR増幅バイアス、不十分なPCR増幅、またはそれらの任意の組合せの結果でありうる。シーケンシング導入エラーは、不正確なベースコーリング、不十分なシーケンシング、またはそれらの任意の組合せの結果でありうる。

【0009】

いくつかの実施形態では、シーケンシングデータ中の標的のクオリティステータスは、完全シーケンシング、不完全シーケンシング、または飽和シーケンシングでありうる。シーケンシングデータ中の標的のクオリティステータスは、複数の確率バーコード中に識別可

10

20

30

40

50

能な配列を有する分子標識の数と、カウントされたシーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数によって決定することができる。シーケンシングデータ中の標的のクオリティステータスは、(b)で得られたシーケンシングデータ中の標的のクオリティステータスが、完全シーケンシングではなく、かつ、飽和シーケンシングではない場合に、不完全シーケンシングとして分類することができる。

【0010】

いくつかの実施形態では、完全シーケンシングクオリティステータスは、所定の完全シーケンシング散布閾値以上のポアソン分布と比較した散布指数によって決定され、ここで、所定の完全シーケンシング散布閾値は、0.9、1、または4でありうる。完全シーケンシングクオリティステータスは、さらに、(b)で得られたシーケンシングデータ中の所定の完全シーケンシング発生数閾値以上の発生数を有する分子標識によって決定することもでき、ここで、所定の完全シーケンシング発生数閾値は、10または18でありうる。

10

【0011】

いくつかの実施形態では、飽和シーケンシングクオリティステータスは、所定の飽和閾値よりも大きい、識別可能な配列を含む分子標識の数を有する標的によって、決定することができる。飽和シーケンシングクオリティステータスは、さらに、所定の飽和閾値よりも大きい、識別可能な配列を含む分子標識の数を有する複数の標的のうちの1つの他の標的によって、決定することもできる。所定の飽和閾値は、確率バーコードが、識別可能な配列を有する約6561の分子標識を含む場合、6557でありうる。所定の飽和閾値は、確率バーコードが、識別可能な配列を有する約65536の分子標識を含む場合、65532でありうる。

20

【0012】

いくつかの実施形態では、(i)でカウントされたシーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数は、(iv)において、標的が完全シーケンシングクオリティステータスを有していれば、1つ以上の親分子標識についてすべての子供分子標識を決定する工程と；少なくとも1つの子供分子標識および親分子標識について第1の統計解析を実施する工程と；第1の統計解析の帰無仮説が容認されれば、子供分子標識の発生数を親分子標識に帰属させる工程と、によって調節される。

【0013】

いくつかの実施形態では、1つ以上の親分子標識は、所定の完全シーケンシング親閾値以上の発生数を有する分子標識を含み、ここで、所定の完全シーケンシング親閾値は、所定の完全シーケンシング発生数閾値と等しい。子供分子標識は、親分子標識と1塩基相違し、かつ、所定の完全シーケンシング子供閾値以下の発生数を有する分子標識を含み、ここで、所定の完全シーケンシング子供閾値は、3または5でありうる。第1の統計解析の帰無仮説は、真であるという帰無仮説の確率が偽発見率を下回れば、容認することができ、ここで、偽発見率は、5%または10%である。第1の統計解析は、多重二項検定であってよい。

30

【0014】

いくつかの実施形態では、(i)でカウントされたシーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数は、(iv)において、標的が完全シーケンシングクオリティステータスを有していれば、標的の分子標識を閾値化して、(b)で得られたシーケンシングデータ中の標的に関連付けられた真の分子標識および偽の分子標識を決定する工程によって調節される。標的の分子標識を閾値化する工程は、標的の分子標識について第2の統計解析を実施する工程を含む。

40

【0015】

いくつかの実施形態では、第2の統計解析を実施する工程は、以下：標的の分子標識の分布およびそれらの発生数を2つのポアソン分布に当てはめる工程と；2つのポアソン分布を用いて真の分子標識の数nを決定する工程と；(b)で得られたシーケンシングデータから偽の分子標識を除去する工程と、を含み、ここで、偽の分子標識は、n番目に豊富な分子標識の発生数よりも低い発生数を有する分子標識を含み、また、真の分子標識は、n

50

番目に豊富な分子標識の発生数以上の発生数を有する分子標識を含む。2つのポアソン分布は、真の分子標識に対応する第1のポアソン分布と、偽の分子標識に対応する第2のポアソン分布を含む。

【0016】

いくつかの実施形態では、(i)でカウントされたシーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数は、(iv)において、(b)で得られたシーケンシングデータ中の標的のクオリティステータスが、不完全シーケンシングクオリティステータスである場合、標的が、(b)で得られたシーケンシングデータにおいてノイズであるか否かを決定する工程と；(b)で得られたシーケンシングデータからノイズ標的を除去する工程とによって調節されうる。ノイズ標的の分子標識の発生数が、不完全シーケンシングクノイズ標的閾値以下であれば、標的はノイズである可能性があり、ここで、不完全シーケンシングノイズ遺伝子閾値は、5である。不完全シーケンシングノイズ標的閾値は、完全シーケンシングのクオリティステータスを有する複数の標的の分子標識の中央または平均発生数と等しくてもよい。

10

【0017】

いくつかの実施形態では、(i)でカウントされたシーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数は、(iv)において、(b)で得られたシーケンシングデータ中の標的のクオリティステータスが不完全シーケンシングクオリティステータスであれば、標的の分子標識を閾値化して、(b)で得られたシーケンシングデータ中の真の分子標識および偽の分子標識を決定する工程によって調節することができる。

20

【0018】

いくつかの実施形態では、標的の分子標識を閾値化する工程は、分子標識について第3の統計解析を実施する工程を含む。分子標識について第3の統計解析を実施する工程は、ゼロ切断ポアソンモデルを用いて、真の分子標識の数nを決定する工程と；(b)で得られたシーケンシングデータから偽の分子標識を除去する工程と、を含み、ここで、偽の分子標識は、n番目に豊富な分子標識の発生数よりも低い発生数を有する分子標識を含み、また、真の分子標識は、n番目に豊富な分子標識の発生数以上の発生数を有する分子標識を含む。

【0019】

いくつかの実施形態では、(i)でカウントされたシーケンシングデータが、(iii)で決定された1つ以上のシーケンシングデータエラーに応じて調節された後、(b)で得られたシーケンシングデータ中の分子標識の少なくとも50%または80%を保持することができる。

30

【0020】

いくつかの実施形態では、複数の標的に確率バーコードを付ける工程は、複数の確率バーコードを複数の標的とハイブリダイズさせて、確率バーコード付き標的を生成する工程を含む。複数の標的に確率バーコードを付ける工程は、確率バーコード付き標的のインデックス付きライブラリーを作製する工程を含む。確率バーコード付き標的のインデックス付きライブラリーを作製する工程は、複数の確率バーコードを含む固体担体を用いて実施することができる。固体担体は、複数の確率バーコードと関連した複数の合成粒子を含む。固体担体は、2次元または3次元の複数の確率バーコードを含む。固体担体は、ポリマー、マトリックス、ヒドロゲル、ニードルアレイデバイス、抗体、またはそれらの任意の組合せを含む。

40

【0021】

いくつかの実施形態では、複数の確率バーコードの各々は、サンプル標識、ユニバーサル標識および細胞標識の1つ以上を含み、ここで、サンプル標識は、固体担体上の複数の確率バーコードに対するものと同じであってよく、ユニバーサル標識は、固体担体上の複数の確率バーコードに対するものと同じであってよく、細胞標識は、固体担体上の複数の確率バーコードに対するものと同じであってよい。サンプル標識は、5〜20ヌクレオチド

50

を含む。ユニバーサル標識は、5～20ヌクレオチドを含む。細胞標識は、5～20ヌクレオチドを含む。

【0022】

いくつかの実施形態では、合成粒子はビーズであってよい。ビーズは、シリカゲルビーズ、多孔性ガラスビーズ、磁気ビーズ、ダイナビーズ、セファデックス／セファロースビーズ、セルロースビーズ、ポリスチレンビーズ、またはそれらの任意の組合せであってよい。

【0023】

いくつかの実施形態では、複数の標的をサンプル中に含有させることができる。サンプルは、1つ以上の細胞を含む。サンプルは、単一細胞であってよい。1つ以上の細胞は、1つ以上の細胞型を含む。1つ以上の細胞型の少なくとも1つは、脳細胞、心細胞、癌細胞、循環腫瘍細胞、臓器細胞、上皮細胞、転移細胞、良性細胞、一次細胞、循環細胞、またはそれらの任意の組合せである。

10

【0024】

いくつかの実施形態では、複数の標的は、リボ核酸(RNA)、メッセンジャーRNA(mRNA)、microRNA、低分子干渉RNA(siRNA)、RNA分解産物、ポリ(A)テールを各々含むRNA、またはそれらの任意の組合せを含む。

【0025】

いくつかの実施形態では、本方法は、さらに、1つ以上の細胞を溶解する工程を含みうる。1つ以上の細胞を溶解する工程は、サンプルを加熱する工程、サンプルを洗剤と接触させる工程、サンプルのpHを変える工程、またはそれらの任意の組合せを含む。

20

【0026】

本明細書には、標的の数を決定する方法が開示される。いくつかの実施形態では、本方法は、(a) 複数の確率バーコードを用いて、複数の標的に確率バーコードを付けて、複数の確率バーコード付き標的を生成する工程であって、複数の確率バーコードの各々が分子標識を含む工程と；(b) 確率バーコード付き標的のシーケンシングデータを取得する工程と；(c) 複数の標的の1つ以上について：(i) シーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数をカウントする工程と；(ii) 方向近接性を用いて、標的の分子標識のクラスターを同定する工程と；(iii) (ii) で同定された標的の分子標識のクラスターを用いて、(b) で得られたシーケンシングデータを折りたたむ工程と；(iv) 標的の数を推定する工程であって、推定された標的の数が、(ii) のシーケンシングデータの折りたたみ後に、(i) でカウントされたシーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数と関連する工程と、を含む。複数の標的は、細胞の全トランスクリプトームの標的を含む。

30

【0027】

いくつかの実施形態では、クラスター内の標的の分子標識は、互いの所定の方向近接性閾値内にある。方向近接性閾値は、1のハミング距離である。クラスター内の標的の分子標識は、1つ以上の親分子標識と、1つ以上の親分子標識の子供分子標識を含み、ここで、親分子標識の発生数は、所定の方向近接性発生数閾値以上である。所定の方向近接性発生数閾値は、 $2 \times (\text{子供分子標識の発生数}) - 1$ であってよい。

【0028】

いくつかの実施形態では、(ii) で同定された標的の分子標識のクラスターを用いて、(b) で得られたシーケンシングデータを折りたたむ工程は、子供分子標識の発生数を親分子標識に帰属させる工程を含む。

40

【0029】

いくつかの実施形態では、本方法は、さらに、標的のシーケンシング深度を決定する工程も含みうる。標的の数を推定する工程は、標的のシーケンシング深度が所定のシーケンシング深度閾値を超える場合、(i) でカウントされたシーケンシングデータを調節する工程を含む。所定のシーケンシング深度閾値は、15～20であってよい。(i) でカウントされたシーケンシングデータを調節する工程は、標的の分子標識を閾値化して、(b) で得られたシーケンシングデータ中の標的に関連付けられた真の分子標識および偽の分子

50

標識を決定する工程を含む。標的の分子標識を閾値化する工程は、標的の分子標識について統計解析を実施する工程を含む。統計解析を実施する工程は、以下：標的の分子標識の分布およびそれらの発生数を2つのポアソン分布に当てはめる工程と；2つのポアソン分布を用いて真の分子標識の数 n を決定する工程と；(b) で得られたシーケンシングデータから偽の分子標識を除去する工程と、を含み、ここで、偽の分子標識は、 n 番目に豊富な分子標識の発生数よりも低い発生数を有する分子標識を含み、また、真の分子標識は、 n 番目に豊富な分子標識の発生数以上の発生数を有する分子標識を含む。

【0030】

本明細書には、標的の数を決定するためのコンピュータシステムが開示される。いくつかの実施形態では、コンピュータシステムは、実行可能命令を保存するコンピュータが可読メモリと；コンピュータ可読メモリと連絡する1つ以上のコンピュータプロセッサを含み、ここで、1つ以上のコンピュータプロセッサは、実行可能命令によりプログラムされて、(a) 複数の確率バーコードを用いて、複数の標的に確率バーコードを付けて、複数の確率バーコード付き標的を生成する工程であって、複数の確率バーコードの各々が分子標識を含む工程と；(b) 確率バーコード付き標的のシーケンシングデータを取得する工程と；(c) 複数の標的の1つ以上について：(i) シーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数をカウントする工程と；(ii) (b) で得られたシーケンシングデータ中の標的のクオリティステータスを決定する工程と；(iii) (b) で得られたシーケンシングデータ中の1つ以上のシーケンシングデータエラーを決定する工程であって、シーケンシングデータ中の1つ以上のシーケンシングエラーを決定する工程が、以下：シーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数、シーケンシングデータ中の標的のクオリティステータス、および複数の確率バーコード中の識別可能な配列を有する分子標識の数のうち1つ以上を決定することを含む工程と；(iv) 標的の数を推定する工程であって、推定された標的の数が、(iii) で決定された1つ以上のシーケンシングデータエラーに応じて調節された、(i) でカウントされたシーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数と相関する工程と、を含む。工程(i)、(ii)、(iii)、および(iv)は、複数の標的の各々について実施することができる。工程(a)、(b)、(c)、(i)、(ii)、(iii)、および(iv)は、多重化することができる。

【0031】

いくつかの実施形態では、実行可能命令は、さらに、1つ以上のシーケンシングデータエラーを決定する前に、(b) で得られたシーケンシングデータを折りたたむ工程を実施するように、1つ以上のコンピュータプロセッサをプログラムすることもできる。(b) で得られたシーケンシングデータを折りたたむ工程は、類似分子標識を有し、かつ、所定の折りたたみ発生数閾値よりも少ない発生数を有する標的のコピーを、複数の標的について同じ分子標識を有するものとして帰属させる工程を含み、ここで、標的の2つのコピーは、標的の2つのコピーの分子標識の配列が、少なくとも1塩基相違する場合、類似の分子標識を有する。

【0032】

いくつかの実施形態では、所定の折りたたみ発生数閾値は、確率バーコードが、識別可能な配列を有する約6561の分子標識を含む場合、7となりうる。確率バーコードが、識別可能な配列を有する約65536の分子標識を含む場合、所定の折りたたみ発生数閾値は17となりうる。標的の2つのコピーは、標的の2つのコピーの分子標識の配列が、少なくとも1塩基相違する場合、類似の分子標識を有する。いくつかの実施形態では、分子標識は、5～20ヌクレオチドを含む。さまざまな確率バーコードの分子標識は、互いに異なっていてよい。複数の確率バーコードは、識別可能な配列を有する約6561の分子標識を含む。複数の確率バーコードは、識別可能な配列を有する約65536の分子標識を含む。

【0033】

いくつかの実施形態では、50ヌクレオチド以上のリード長を有する複数の標的の配列を含む。シーケンシングデータは、75ヌクレオチド以上のリード長を有する複数の標的の配列を含む。シーケンシングデータは、100ヌクレオチド以上のリード長を有する複数の標的の配列を含む。(b)で得られたシーケンシングデータは、複数の確率バーコード付き標的に対してポリメラーゼ連鎖反応(PCR)増幅を実施することによって生成することができる。

【0034】

いくつかの実施形態では、1つ以上のシーケンシングデータエラーは、PCR導入エラー、シーケンシング導入エラー、バーコード混入に起因するエラー、ライブラリー作製エラー、またはそれらの任意の組合せでありうる。PCR導入エラーは、PCR増幅エラー、PCR増幅バイアス、不十分なPCR増幅、またはそれらの任意の組合せの結果でありうる。シーケンシング導入エラーは、不正確なベースコーリング、不十分なシーケンシング、またはそれらの任意の組合せの結果でありうる。

10

【0035】

いくつかの実施形態では、実行可能命令は、さらに、シーケンシングデータ中の標的のクオリティステータスが、完全シーケンシング、不完全シーケンシング、または飽和シーケンシングであることを決定するように、1つ以上のコンピュータプロセッサをプログラムすることもできる。シーケンシングデータ中の標的のクオリティステータスは、複数の確率バーコード中に識別可能な配列を有する分子標識の数と、カウントされたシーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数によって決定することができる。シーケンシングデータ中の標的のクオリティステータスは、(b)で得られたシーケンシングデータ中の標的のクオリティステータスが、完全シーケンシングではなく、飽和シーケンシングでもない場合、不完全シーケンシングとして分類することができる。

20

【0036】

いくつかの実施形態では、実行可能命令は、さらに、所定の完全シーケンシング散布閾値以上のポアソン分布と比較した散布指数によって、完全シーケンシングクオリティステータスを決定するように、1つ以上のコンピュータプロセッサをプログラムすることもでき、ここで、所定の完全シーケンシング散布閾値は、0.9、1、または4でありうる。完全シーケンシングクオリティステータスは、さらに、(b)で得られたシーケンシングデータ中の所定の完全シーケンシング発生数閾値以上の発生数を有する分子標識によって決定することもでき、ここで、所定の完全シーケンシング発生数閾値は、10または18でありうる。

30

【0037】

いくつかの実施形態では、実行可能命令は、さらに、所定の飽和閾値よりも大きい、識別可能な配列を含む特定の数の分子標識を有する標的によって、飽和シーケンシングクオリティステータスを決定するように、1つ以上のコンピュータプロセッサをプログラムすることもできる。飽和シーケンシングクオリティステータスは、さらに、所定の飽和閾値よりも大きい、識別可能な配列を含む特定の数の分子標識を有する複数の標的のうちの1つの他の標的によって決定することもできる。所定の飽和閾値は、確率バーコードが、識別可能な配列を有する約6561の分子標識を含む場合、6557でありうる。所定の飽和閾値は、確率バーコードが、識別可能な配列を有する約65536の分子標識を含む場合、65532でありうる。

40

【0038】

いくつかの実施形態では、実行可能命令は、さらに、(i)でカウントされたシーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数を、(iv)において、標的が完全シーケンシングクオリティステータスを有していれば、1つ以上の親分子標識についてすべての子供分子標識を決定する工程と；少なくとも1つの子供分子標識および親分子標識について第1の統計解析を実施する工程と；第1の統計解析の帰無仮説が容認されれば、子供分子標識の発生数を親分子標識に帰属させる工程と、によって

50

調節するように、1つ以上のコンピュータプロセッサをプログラムすることもできる。

【0039】

いくつかの実施形態では、1つ以上の親分子標識は、所定の完全シーケンシング親閾値以上の発生数を有する分子標識を含み、ここで、所定の完全シーケンシング親閾値は、所定の完全シーケンシング発生数閾値と等しい。子供分子標識は、親分子標識と1塩基相違し、かつ、所定の完全シーケンシング子供閾値以下の発生数を有する分子標識を含み、ここで、所定の完全シーケンシング子供閾値は、3または5でありうる。第1の統計解析の帰無仮説は、真であるという帰無仮説の確率が偽発見率を下回れば、容認することができ、ここで、偽発見率は、5%または10%である。第1の統計解析は、多重二項検定であってよい。

10

【0040】

いくつかの実施形態では、実行可能命令は、さらに、(i)でカウントされたシーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数を、(iv)において、標的が完全シーケンシングクオリティステータスを有していれば、標的の分子標識を閾値化して、(b)で得られたシーケンシングデータ中の標的に関連付けられた真の分子標識および偽の分子標識を決定する工程によって調節するように、1つ以上のコンピュータプロセッサをプログラムすることもできる。標的の分子標識を閾値化する工程は、標的の分子標識について第2の統計解析を実施する工程を含む。

【0041】

いくつかの実施形態では、実行可能命令は、さらに、標的の分子標識の分布およびそれらの発生数を2つのポアソン分布に当てはめる工程と；2つのポアソン分布を用いて真の分子標識の数nを決定する工程と；(b)で得られたシーケンシングデータから偽の分子標識を除去する工程と、によって、第2の統計解析を実施する工程を実施するように、1つ以上のコンピュータプロセッサをプログラムすることもでき、ここで、偽の分子標識は、n番目に豊富な分子標識の発生数よりも低い発生数を有する分子標識を含み、また、真の分子標識は、n番目に豊富な分子標識の発生数以上の発生数を有する分子標識を含む。2つのポアソン分布は、真の分子標識に対応する第1ポアソン分布と、偽の分子標識に対応する第2ポアソン分布を含む。

20

【0042】

いくつかの実施形態では、実行可能命令は、さらに、(i)でカウントされたシーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数を、(iv)において、(b)で得られたシーケンシングデータ中の標的のクオリティステータスが、不完全シーケンシングクオリティステータスである場合、標的が、(b)で得られたシーケンシングデータにおいてノイズであるか否かを決定する工程と；(b)で得られたシーケンシングデータからノイズ標的を除去する工程と、によって調節するように、1つ以上のコンピュータプロセッサをプログラムすることもできる。ノイズ標的の分子標識の発生数が、不完全シーケンシングクオリティステータス閾値以下であれば、標的はノイズである可能性があり、ここで、不完全シーケンシングノイズ遺伝子閾値は5である。不完全シーケンシングノイズ標的閾値は、完全シーケンシングのクオリティステータスを有する複数の標的の分子標識の中央または平均発生数と等しくてもよい。

30

【0043】

いくつかの実施形態では、実行可能命令は、さらに、(i)でカウントされたシーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数を、(iv)において、(b)で得られたシーケンシングデータ中の標的のクオリティステータスが不完全シーケンシングクオリティステータスであれば、標的の分子標識を閾値化して、(b)で得られたシーケンシングデータ中の真の分子標識および偽の分子標識を決定する工程によって調節するように、1つ以上のコンピュータプロセッサをプログラムすることもできる。

40

【0044】

いくつかの実施形態では、実行可能命令は、さらに、分子標識について第3の統計解析を

50

実施する工程によって、標的の分子標識を閾値化するように、1つ以上のコンピュータプロセッサをプログラムすることもできる。分子標識について第3の統計解析を実施する工程は、ゼロ切断ポアソンモデルを用いて、真の分子標識の数 n を決定する工程と；(b) で得られたシーケンシングデータから偽の分子標識を除去する工程と、を含み、ここで、偽の分子標識は、 n 番目に豊富な分子標識の発生数よりも低い発生数を有する分子標識を含み、また、真の分子標識は、 n 番目に豊富な分子標識の発生数以上の発生数を有する分子標識を含む。

【0045】

いくつかの実施形態では、(i) でカウントされたシーケンシングデータが (i i i) で決定された1つ以上のシーケンシングデータエラーに応じて調節された後、(b) で得られたシーケンシングデータ中の分子標識の少なくとも50%または80%が保持される。

【0046】

本明細書には、標的の数を決定するためのコンピュータシステムが開示される。いくつかの実施形態では、コンピュータシステムは、実行可能命令を記憶するコンピュータ可読メモリと；コンピュータ可読メモリと連絡する1つ以上のコンピュータプロセッサを含み、ここで、1つ以上のコンピュータプロセッサは、以下：(a) 複数の確率バーコードを用いて、複数の標的に確率バーコードを付けて、複数の確率バーコード付き標的を生成する工程であって、複数の確率バーコードの各々が分子標識を含む工程と；(b) 確率バーコード付き標的のシーケンシングデータを取得する工程と；(c) 複数の標的の1つ以上について：(i) シーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数をカウントする工程と；(i i) 方向近接性を用いて、標的の分子標識のクラスターを同定する工程と；(i i i) (i i) で同定された標的の分子標識のクラスターを用いて、(b) で得られたシーケンシングデータを折りたたむ工程と；(i v) 標的の数を推定する工程であって、推定された標的の数が、(i i) でシーケンシングデータを折りたたんだ後、(i) でカウントされたシーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数と相関する工程と、を実施するように、実行可能命令によりプログラムされる。複数の標的は、細胞の全トランスクリプトームの標的を含む。

【0047】

いくつかの実施形態では、クラスター内の標的の分子標識は、互いの所定の方向近接性閾値内にある。方向近接性閾値は、1のハミング距離である。クラスター内の標的の分子標識は、1つ以上の親分子標識と、1つ以上の親分子標識の子供分子標識を含み、ここで、親分子標識の発生数は、所定の方向近接性発生数閾値以上である。所定の方向近接性発生数閾値は、 $2 \times (\text{子供分子標識の発生数}) - 1$ であってよい。

【0048】

いくつかの実施形態では、(i i) で同定された標的の分子標識のクラスターを用いて、(b) で得られたシーケンシングデータを折りたたむ工程は、子供分子標識の発生数を親分子標識に帰属させる工程を含む。

【0049】

いくつかの実施形態では、実行可能命令は、さらに、標的のシーケンシング深度を決定するように、1つ以上のコンピュータプロセッサをプログラムすることができる。標的の数を推定する工程は、標的のシーケンシング深度が所定のシーケンシング深度閾値を超える場合、(i) でカウントされたシーケンシングデータを調節する工程を含む。所定のシーケンシング深度閾値は、15~20であってよい。(i) でカウントされたシーケンシングデータを調節する工程は、標的の分子標識を閾値化して、(b) で得られたシーケンシングデータ中の標的に関連付けられた真の分子標識および偽の分子標識を決定する工程を含む。標的の分子標識を閾値化する工程は、標的の分子標識について統計解析を実施する工程を含む。統計解析を実施する工程は、以下：標的の分子標識の分布およびそれらの発生数を2つのポアソン分布に当てはめる工程と；2つのポアソン分布を用いて真の分子標識の数 n を決定する工程と；(b) で得られたシーケンシングデータから偽の分子標識

を除去する工程と、を含み、ここで、偽の分子標識は、n番目に豊富な分子標識の発生数よりも低い発生数を有する分子標識を含み、また、真の分子標識は、n番目に豊富な分子標識の発生数以上の発生数を有する分子標識を含む。

【0050】

本明細書には、実行可能コードを含む1つ以上の非一過性コンピュータ読取り媒体が開示され、これは、実行されると、1つ以上のコンピュータデバイスに標的の数を決定させる。いくつかの実施形態では、実行可能コードは、実行されると、1つ以上のコンピュータデバイスに、以下：(a) 複数の確率バーコードを用いて、複数の標的に確率バーコードを付けて、複数の確率バーコード付き標的を生成する工程であって、複数の確率バーコードの各々が分子標識を含む工程と；(b) 確率バーコード付き標的のシーケンシングデータを取得する工程と；(c) 複数の標的の1つ以上について：(i) シーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数をカウントする工程と；(ii) (b) で得られたシーケンシングデータ中の標的のクオリティステータスを決定する工程と；(iii) (b) で得られたシーケンシングデータ中の1つ以上のシーケンシングデータエラーを決定する工程であって、シーケンシングデータ中の1つ以上のシーケンシングエラーを決定する工程が、以下：シーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数、シーケンシングデータ中の標的のクオリティステータス、および複数の確率バーコードに識別可能な配列を有する分子標識の数のうち1つ以上を決定することを含む工程と；(iv) 標的の数を推定する工程であって、推定された標的の数が、(iii) で決定された1つ以上のシーケンシングデータエラーに応じて調節された、(i) でカウントされたシーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数と相関する工程と、を含むプロセスを実施させる。工程(i)、(ii)、(iii)、および(iv)は、複数の標的の各々について実施することができる。本方法は、多重化することができる。

【0051】

いくつかの実施形態では、本プロセスは、1つ以上のシーケンシングデータエラーを決定する前に、(b) で得られたシーケンシングデータを折りたたむ工程をさらに含む。(b) で得られたシーケンシングデータを折りたたむ工程は、類似分子標識を有し、かつ、所定の折りたたみ発生数閾値よりも少ない発生数を有する標的のコピーを、複数の標的について同じ分子標識を有するものとして帰属させる工程を含み、ここで、標的の2つのコピーは、標的の2つのコピーの分子標識の配列が、少なくとも1塩基相違する場合、類似の分子標識を有する。

【0052】

いくつかの実施形態では、所定の折りたたみ発生数閾値は、確率バーコードが、識別可能な配列を有する約6561の分子標識を含む場合、7となりうる。確率バーコードが、識別可能な配列を有する約65536の分子標識を含む場合、所定の折りたたみ発生数閾値は17となりうる。標的の2つのコピーは、標的の2つのコピーの分子標識の配列が、少なくとも1塩基相違する場合、類似の分子標識を有する。いくつかの実施形態では、分子標識は、5～20ヌクレオチドを含む。さまざまな確率バーコードの分子標識は、互いに異なっていてよい。複数の確率バーコードは、識別可能な配列を有する約6561の分子標識を含む。複数の確率バーコードは、識別可能な配列を有する約65536の分子標識を含む。

【0053】

いくつかの実施形態では、シーケンシングデータは、50ヌクレオチド以上のリード長を有する複数の標的の配列を含む。シーケンシングデータは、75ヌクレオチド以上のリード長を有する複数の標的の配列を含む。シーケンシングデータは、100ヌクレオチド以上のリード長を有する複数の標的の配列を含む。(b) で得られたシーケンシングデータは、複数の確率バーコード付き標的に対してポリメラーゼ連鎖反応(PCR)増幅を実施することによって生成することができる。

【0054】

いくつかの実施形態では、1つ以上のシーケンシングデータエラーは、PCR導入エラー、シーケンシング導入エラー、バーコード混入に起因するエラー、ライブラリー作製エラー、またはそれらの任意の組合せでありうる。PCR導入エラーは、PCR増幅エラー、PCR増幅バイアス、不十分なPCR増幅、またはそれらの任意の組合せの結果でありうる。シーケンシング導入エラーは、不正確なベースコーリング、不十分なシーケンシング、またはそれらの任意の組合せの結果でありうる。

【0055】

いくつかの実施形態では、シーケンシングデータ中の標的のクオリティステータスは、完全シーケンシング、不完全シーケンシング、または飽和シーケンシングでありうる。シーケンシングデータ中の標的のクオリティステータスは、複数の確率バーコード中に識別可能な配列を有する分子標識の数と、カウントされたシーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数によって決定することができる。シーケンシングデータ中の標的のクオリティステータスは、(b)で得られたシーケンシングデータ中の標的のクオリティステータスが、完全シーケンシングではなく、かつ、飽和シーケンシングではない場合に、不完全シーケンシングとして分類することができる。

【0056】

いくつかの実施形態では、完全シーケンシングクオリティステータスは、所定の完全シーケンシング散布閾値以上のポアソン分布に対する散布指数によって決定され、ここで、所定の完全シーケンシング散布閾値は、0.9、1、または4でありうる。完全シーケンシングクオリティステータスは、さらに、(b)で得られたシーケンシングデータ中の所定の完全シーケンシング発生数閾値以上の発生数を有する分子標識によって決定することもでき、ここで、所定の完全シーケンシング発生数閾値は、10または18でありうる。

【0057】

いくつかの実施形態では、飽和シーケンシングクオリティステータスは、所定の飽和閾値よりも大きい、識別可能な配列を含む分子標識の数を有する標的によって、決定することができる。飽和シーケンシングクオリティステータスは、さらに、所定の飽和閾値よりも大きい、識別可能な配列を含む分子標識の数を有する複数の標的のうちの1つの他の標的によって決定することもできる。所定の飽和閾値は、確率バーコードが、識別可能な配列を有する約6561の分子標識を含む場合、6557でありうる。所定の飽和閾値は、確率バーコードが、識別可能な配列を有する約65536の分子標識を含む場合、65532でありうる。

【0058】

いくつかの実施形態では、(i)でカウントされたシーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数は、(iv)において、標的が完全シーケンシングクオリティステータスを有していれば、1つ以上の親分子標識についてすべての子供分子標識を決定する工程と；少なくとも1つの子供分子標識および親分子標識について第1の統計解析を実施する工程と；第1の統計解析の帰無仮説が容認されれば、子供分子標識の発生数を親分子標識に帰属させる工程と、によって調節される。

【0059】

いくつかの実施形態では、1つ以上の親分子標識は、所定の完全シーケンシング親閾値以上の発生数を有する分子標識を含み、ここで、所定の完全シーケンシング親閾値は、所定の完全シーケンシング発生数閾値と等しい。子供分子標識は、親分子標識と1塩基相違し、かつ、所定の完全シーケンシング子供閾値以下の発生数を有する分子標識を含み、ここで、所定の完全シーケンシング子供閾値は、3または5でありうる。第1の統計解析の帰無仮説は、真であるという帰無仮説の確率が偽発見率を下回れば、容認することができ、ここで、偽発見率は、5%または10%である。第1の統計解析は、多重二項検定であってよい。

【0060】

いくつかの実施形態では、(i)でカウントされたシーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数は、(iv)において、標的が完全シー

10

20

30

40

50

ケンシングクオリティステータスを有していれば、標的の分子標識を閾値化して、(b)で得られたシーケンシングデータ中の標的に関連付けられた真の分子標識および偽の分子標識を決定する工程によって調節される。標的の分子標識を閾値化する工程は、標的の分子標識について第2の統計解析を実施する工程を含む。

【0061】

いくつかの実施形態では、第2の統計解析を実施する工程は、標的の分子標識の分布およびそれらの発生数を2つのポアソン分布に当てはめる工程と；2つのポアソン分布を用いて真の分子標識の数 n を決定する工程と；(b)で得られたシーケンシングデータから偽の分子標識を除去する工程と、を含み、ここで、偽の分子標識は、 n 番目に豊富な分子標識の発生数よりも低い発生数を有する分子標識を含み、また、真の分子標識は、 n 番目に豊富な分子標識の発生数以上の発生数を有する分子標識を含む。2つのポアソン分布は、真の分子標識に対応する第1ポアソン分布と、偽の分子標識に対応する第2ポアソン分布を含む。

【0062】

いくつかの実施形態では、(i)でカウントされたシーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数は、(iv)において、(b)で得られたシーケンシングデータ中の標的のクオリティステータスが、不完全シーケンシングクオリティステータスである場合、標的が、(b)で得られたシーケンシングデータにおいてノイズであるか否かを決定する工程と；(b)で得られたシーケンシングデータからノイズ標的を除去する工程と、によって調節することができる。ノイズ標的の分子標識の発生数が、不完全シーケンシングクノイズ標的閾値以下であれば、標的はノイズである可能性があり、ここで、不完全シーケンシングノイズ遺伝子閾値は5である。不完全シーケンシングノイズ標的閾値は、完全シーケンシングのクオリティステータスを有する複数の標的の分子標識の中央または平均発生数と等しくてもよい。

【0063】

いくつかの実施形態では、(i)でカウントされたシーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数は、(iv)において、(b)で得られたシーケンシングデータ中の標的のクオリティステータスが不完全シーケンシングクオリティステータスである場合、標的の分子標識を閾値化して、(b)で得られたシーケンシングデータ中の真の分子標識および偽の分子標識を決定する工程によって調節することができる。

【0064】

いくつかの実施形態では、標的の分子標識を閾値化する工程は、分子標識について第3の統計解析を実施する工程を含む。分子標識について第3の統計解析を実施する工程は、ゼロ切断ポアソンモデルを用いて、真の分子標識の数 n を決定する工程と；(b)で得られたシーケンシングデータから偽の分子標識を除去する工程と、を含み、ここで、偽の分子標識は、 n 番目に豊富な分子標識の発生数よりも低い発生数を有する分子標識を含み、また、真の分子標識は、 n 番目に豊富な分子標識の発生数以上の発生数を有する分子標識を含む。

【0065】

いくつかの実施形態では、(i)でカウントされたシーケンシングデータが、(iii)で決定された1つ以上のシーケンシングデータエラーに応じて調節された後、(b)で得られたシーケンシングデータ中の分子標識の少なくとも50%または80%を保持することができる。

【0066】

本明細書には、実行可能コードを含む1つ以上の非一過性コンピュータ読取り媒体が開示され、これは、実行されると、1つ以上のコンピュータデバイスに標的の数を決定させる。いくつかの実施形態では、実行可能コードは、実行されると、1つ以上のコンピュータデバイスに、以下：(a)複数の確率バーコードを用いて、複数の標的に確率バーコードを付けて、複数の確率バーコード付き標的を生成する工程であって、複数の確率バーコー

10

20

30

40

50

ドの各々が分子標識を含む工程と；(b) 確率バーコード付き標的のシーケンシングデータを取得する工程と；(c) 複数の標的の1つ以上について：(i) シーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数をカウントする工程と；(ii) 方向近接性を用いて、標的の分子標識のクラスターを同定する工程と；(iii) (ii) で同定された標的の分子標識のクラスターを用いて、(b) で得られたシーケンシングデータを折りたたむ工程と；(iv) 標的の数を推定する工程であって、推定された標的の数が、(ii) でシーケンシングデータを折りたたんだ後、(i) でカウントされたシーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数と相関する工程と、を含むプロセスを実施させる。複数の標的は、細胞の全トランスクリプトームの標的を含む。

10

【0067】

いくつかの実施形態では、クラスター内の標的の分子標識は、互いの所定の方向近接性閾値内にある。方向近接性閾値は、1のハミング距離である。クラスター内の標的の分子標識は、1つ以上の親分子標識と、1つ以上の親分子標識の子供分子標識を含み、ここで、親分子標識の発生数は、所定の方向近接性発生数閾値以上である。所定の方向近接性発生数閾値は、 $2 \times (\text{子供分子標識の発生数}) - 1$ であってよい。

【0068】

いくつかの実施形態では、(ii) で同定された標的の分子標識のクラスターを用いて、(b) で得られたシーケンシングデータを折りたたむ工程は、子供分子標識の発生数を親分子標識に帰属させる工程を含む。

20

【0069】

いくつかの実施形態では、本方法は、さらに、標的のシーケンシング深度を決定する工程を含みうる。標的の数を推定する工程は、標的のシーケンシング深度が所定のシーケンシング深度閾値を超える場合、(i) でカウントされたシーケンシングデータを調節する工程を含む。所定のシーケンシング深度閾値は、15～20であってよい。(i) でカウントされたシーケンシングデータを調節する工程は、標的の分子標識を閾値化して、(b) で得られたシーケンシングデータ中の標的に関連付けられた真の分子標識および偽の分子標識を決定する工程を含む。標的の分子標識を閾値化する工程は、標的の分子標識について統計解析を実施する工程を含む。統計解析を実施する工程は、以下：標的の分子標識の分布およびそれらの発生数を2つのポアソン分布に当てはめる工程と；2つのポアソン分布を用いて真の分子標識の数 n を決定する工程と；(b) で得られたシーケンシングデータから偽の分子標識を除去する工程と、を含み、ここで、偽の分子標識は、 n 番目に豊富な分子標識の発生数よりも低い発生数を有する分子標識を含み、また、真の分子標識は、 n 番目に豊富な分子標識の発生数以上の発生数を有する分子標識を含む。

30

【0070】

本明細書には、PCRまたはシーケンシングエラーを訂正する方法が開示される。いくつかの実施形態では、本方法は、(a) 確率バーコード付き標的のシーケンシングデータを取得する工程と；(b) 複数の標的の1つ以上について：(i) シーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数をカウントする工程と；(ii) 方向近接性を用いて、標的の分子標識のクラスターを同定する工程と；(iii) (ii) で同定された標的の分子標識のクラスターを用いて、(a) で得られたシーケンシングデータを折りたたむ工程と；(iv) 標的の数を推定する工程と、を含むことができ、ここで、推定された標的の数は、(ii) のシーケンシングデータの折りたたみ後に、(i) でカウントされたシーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数と相関する。複数の標的は、細胞の全トランスクリプトームの標的を含む。いくつかの実施形態では、本方法を用いて、標的の数を決定することができる。本方法は、さらに、(c) 複数の確率バーコードを用いて、複数の標的に確率バーコードを付けて、複数の確率バーコード付き標的を生成する工程と；(d) 確率バーコード付き標的をシーケンシングして、受け取った確率バーコード付き標的のシーケンシングデータを生成する工程と、を含みうる。

40

50

【0071】

いくつかの実施形態では、クラスター内の標的の分子標識は、互いの所定の方向近接性閾値内にある。方向近接性閾値は、1のハミング距離である。クラスター内の標的の分子標識は、1つ以上の親分子標識と、1つ以上の親分子標識の子供分子標識を含み、ここで、親分子標識の発生数は、所定の方向近接性発生数閾値以上である。所定の方向近接性発生数閾値は、 $2 \times (\text{子供分子標識の発生数}) - 1$ であってよい。

【0072】

いくつかの実施形態では、(i i)で同定された標的の分子標識のクラスターを用いて、(b)で得られたシーケンシングデータを折りたたむ工程は、子供分子標識の発生数を親分子標識に帰属させる工程を含む。

10

【0073】

いくつかの実施形態では、本方法は、さらに、標的のシーケンシング深度を決定する工程を含む。標的の数を推定する工程は、標的のシーケンシング深度が所定のシーケンシング深度閾値を超える場合、(i)でカウントされたシーケンシングデータを調節する工程を含む。所定のシーケンシング深度閾値は、15～20であってよい。(i)でカウントされたシーケンシングデータを調節する工程は、標的の分子標識を閾値化して、(b)で得られたシーケンシングデータ中の標的に関連付けられた真の分子標識および偽の分子標識を決定する工程を含む。標的の分子標識を閾値化する工程は、標的の分子標識について統計解析を実施する工程を含む。統計解析を実施する工程は、標的の分子標識の分布およびそれらの発生数を2つのネガティブ二項分布に当てはめる工程と；2つのネガティブ二項分布を用いて真の分子標識の数nを決定する工程と；(b)で得られたシーケンシングデータから偽の分子標識を除去する工程と、を含み、ここで、偽の分子標識は、n番目に豊富な分子標識の発生数よりも低い発生数を有する分子標識を含み、また、真の分子標識は、n番目に豊富な分子標識の発生数以上の発生数を有する分子標識を含む。

20

【0074】

本明細書には、標的の数を決定するためのコンピュータシステムが開示される。いくつかの実施形態では、コンピュータシステムは、実行可能命令を記憶するコンピュータ可読メモリと；コンピュータ可読メモリと連絡する1つ以上のコンピュータプロセッサを含み、ここで、1つ以上のコンピュータプロセッサは、(a)複数の確率バーコードを用いて、複数の標的に確率バーコードを付けて、複数の確率バーコード付き標的を生成する工程であって、複数の確率バーコードの各々が分子標識を含む工程と；(b)確率バーコード付き標的のシーケンシングデータを取得する工程と；(c)複数の標的の1つ以上について：(i)シーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数をカウントする工程と；(i i)方向近接性を用いて、標的の分子標識のクラスターを同定する工程と；(i i i)(i i)で同定された標的の分子標識のクラスターを用いて、(b)で得られたシーケンシングデータを折りたたむ工程と；(i v)標的の数を推定する工程であって、推定された標的の数が、(i i)でシーケンシングデータを折りたたんだ後、(i)でカウントされたシーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数と相関する工程と、を実施するように、実行可能命令によりプログラムされる。複数の標的は、細胞の全トランスクリプトームの標的を含む。

30

40

【0075】

いくつかの実施形態では、クラスター内の標的の分子標識は、互いの所定の方向近接性閾値内にある。方向近接性閾値は、1のハミング距離である。クラスター内の標的の分子標識は、1つ以上の親分子標識と、1つ以上の親分子標識の子供分子標識を含み、ここで、親分子標識の発生数は、所定の方向近接性発生数閾値以上である。所定の方向近接性発生数閾値は、 $2 \times (\text{子供分子標識の発生数}) - 1$ であってよい。

【0076】

いくつかの実施形態では、(i i)で同定された標的の分子標識のクラスターを用いて、(b)で得られたシーケンシングデータを折りたたむ工程は、子供分子標識の発生数を親

50

分子標識に帰属させる工程を含む。

【0077】

いくつかの実施形態では、実行可能命令は、さらに、標的のシーケンシング深度を決定するように、1つ以上のコンピュータプロセッサをプログラムすることができる。標的の数
数を推定する工程は、標的のシーケンシング深度が所定のシーケンシング深度閾値を超える場合、(i)でカウントされたシーケンシングデータを調節する工程を含む。所定のシーケンシング深度閾値は、15~20であってよい。(i)でカウントされたシーケンシングデータを調節する工程は、標的の分子標識を閾値化して、(b)で得られたシーケンシングデータ中の標的に関連付けられた真の分子標識および偽の分子標識を決定する工程を含む。標的の分子標識を閾値化する工程は、標的の分子標識について統計解析を実施する工程を含む。統計解析を実施する工程は、以下：標的の分子標識の分布およびそれらの発生数を2つのネガティブ二項分布に当てはめる工程と；2つのネガティブ二項分布を用いて真の分子標識の数nを決定する工程と；(b)で得られたシーケンシングデータから偽の分子標識を除去する工程と、を含み、ここで、偽の分子標識は、n番目に豊富な分子標識の発生数よりも低い発生数を有する分子標識を含み、また、真の分子標識は、n番目に豊富な分子標識の発生数以上の発生数を有する分子標識を含む。

10

【0078】

本明細書には、PCRまたはシーケンシングエラーを訂正する方法が開示される。いくつかの実施形態では、本方法は、以下：(a)確率バーコード付き標的のシーケンシングデータを取得する工程と；(b)複数の標的の1つ以上について：(i)シーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数をカウントする工程と；(ii)シーケンシングデータ中の標的に関連付けられた識別可能な配列を有するノイズ分子標識の数を決定する工程と；(iii)標的の数を推定する工程と、を含み、ここで、推定された標的の数が、(ii)で決定されたノイズ分子標識の数に応じて調節された、(i)でカウントされたシーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数と相関する。いくつかの実施形態では、本方法は、シーケンシングデータ中の標的のシーケンシングステータスを決定する工程もさらに含む。シーケンシングデータ中の標的のシーケンシングステータスは、飽和シーケンシング、過少シーケンシング、または過剰シーケンシングである。いくつかの実施形態では、本方法を用いて、標的の数を決定することができる。本方法は、さらに、(c)複数の確率バーコードを用いて、複数の標的に確率バーコードを付けて、複数の確率バーコード付き標的を生成する工程と；(d)確率バーコード付き標的をシーケンシングして、受け取った確率バーコード付き標的のシーケンシングデータを生成する工程と、を含みうる。

20

30

【0079】

いくつかの実施形態では、飽和シーケンシングステータスは、所定の飽和閾値よりも大きい、識別可能な配列を含む分子標識の数を有する標的によって決定される。所定の飽和閾値は、確率バーコードが、識別可能な配列を有する約6561の分子標識を含む場合、約6557である。所定の飽和閾値は、確率バーコードが、識別可能な配列を有する約65536の分子標識を含む場合、約65532である。シーケンシングデータ中の標的のシーケンシングステータスが、飽和シーケンシングステータスである場合、(ii)で決定されたノイズ分子標識の数は、ゼロである。

40

【0080】

いくつかの実施形態では、過少シーケンシングステータスは、所定の過少シーケンシング閾値より小さい深度(たとえば、平均、最小、または最大深度)を有する標的によって決定することができる。過少シーケンシング閾値は約4である。過少シーケンシング閾値は、識別可能な配列を有する分子標識の数とは無関係でありうる。シーケンシングデータ中の標的のシーケンシングステータスが、過少シーケンシングステータスである場合、(ii)で決定されたノイズ分子標識の数は、ゼロである。

【0081】

いくつかの実施形態では、過剰シーケンシングステータスは、所定の過剰シーケンシング

50

閾値より大きい、識別可能な配列を有する分子標識の数を含む標的によって決定される。たとえば、過剰シーケンシング閾値は、確率バーコードが、識別可能な配列を有する約 6 5 6 1 分子標識を含む場合、約 2 5 0 でありうる。本方法は、シーケンシングデータ中の標的のシーケンシングステータスが、過剰シーケンシングステータスである場合、シーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数を、所定の過剰シーケンシング閾値にサブサンプリングする工程を含む。

【0082】

いくつかの実施形態では、シーケンシングデータ中の標的に関連付けられた識別可能な配列を有するノイズ分子標識の数を決定する工程は、ネガティブ二項分布当てはめ条件が満たされれば、(i v) シグナルネガティブ二項分布を、(i) でカウントされたシーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数に当てはめる工程であって、シグナルネガティブ二項分布が、シグナル分子標識である、(i) でカウントされたシーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数に対応するステップと；(v) ノイズネガティブ二項分布を、(i) でカウントされたシーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数に当てはめる工程であって、ノイズネガティブ二項分布が、ノイズ分子標識である、(i) でカウントされたシーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数に対応する工程と；(v i) (v) で当てはめたシグナルネガティブ二項分布および (v i) で当てはめたノイズネガティブ二項分布を用いて、ノイズ分子標識の数を決定する工程と、を含む。

【0083】

いくつかの実施形態では、ネガティブ二項分布当てはめ条件は、シーケンシングデータ中の標的のシーケンシングステータスが、過少シーケンシングステータスまたは過剰シーケンシングステータスではないことを含む。(v) で当てはめたシグナルネガティブ二項分布および (v i) で当てはめたノイズネガティブ二項分布を用いて、ノイズ分子標識の数を決定する工程は、シーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の各々について、識別可能な配列のシグナル確率が、シグナルネガティブ二項分布であることを決定する工程と；識別可能な配列のノイズ確率が、ノイズネガティブ二項分布であることを決定する工程と；シグナル確率がノイズ確率より小さければ、識別可能な配列がノイズ分子標識であることを決定する工程と、を含む。

【0084】

いくつかの実施形態では、シーケンシングデータ中の標的に関連付けられた識別可能な配列を有するノイズ分子標識の数を決定する工程は、シーケンシングデータ中の標的のシーケンシングステータスが、過少シーケンシングステータスまたは過剰シーケンシングステータスではなく、かつ、(i) でカウントされたシーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数が、擬似点閾値より少ない場合、(i i) でシーケンシングデータ中の標的に関連付けられた識別可能な配列を有するノイズ分子標識の数を決定する前に、シーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数に擬似点を加える工程を含む。擬似点閾値は、10 である。

【0085】

いくつかの実施形態では、シーケンシングデータ中の標的に関連付けられた識別可能な配列を有するノイズ分子標識の数を決定する工程は、以下：シーケンシングデータ中の標的のシーケンシングステータスが、過少シーケンシングステータスまたは過剰シーケンシングステータスではなく、かつ、(i) でカウントされたシーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数が、擬似点閾値以上である場合、(i i) でシーケンシングデータ中の標的に関連付けられた識別可能な配列を有するノイズ分子標識の数を決定する際に、非ユニーク分子標識を除去する工程を含む。

【0086】

いくつかの実施形態では、非ユニーク分子標識を除去する工程は、シーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数が、所定の再使用分子標

識別閾値より大きい場合、(i i) でシーケンシングデータ中の標的に関連付けられた識別可能な配列を有するノイズ分子標識の数を決定する際に、非ユニーク分子標識を除去する工程を含む。たとえば、再使用分子標識閾値は、確率バーコードが、識別可能な配列を有する約 6 5 6 1 の分子標識を含む場合、約 6 5 0 でありうる。

【0087】

いくつかの実施形態では、非ユニーク分子標識を除去する工程は、シーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数について非ユニーク分子標識の理論上の数を決定する工程と；シーケンシングデータ中の標的に関連付けられた識別可能な配列を有する n 番目に豊富な分子標識よりも大きい発生数を有する分子標識を除去する工程と、を含み、ここで、n は、非ユニーク分子標識の理論数である。

10

【0088】

本明細書には、標的の数を決定するためのコンピュータシステムが開示される。いくつかの実施形態では、コンピュータシステムは、実行可能命令を記憶するコンピュータ可読メモリと；コンピュータ可読メモリと連絡する1つ以上のコンピュータプロセッサを含み、ここで、1つ以上のコンピュータプロセッサは、(a) 複数の確率バーコードを用いて、複数の標的に確率バーコードを付けて、複数の確率バーコード付き標的を生成する工程であって、複数の確率バーコードの各々が分子標識を含む工程と；(b) 確率バーコード付き標的のシーケンシングデータを取得する工程と；(c) 複数の標的の1つ以上について：(i) シーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数をカウントする工程と；(i i) シーケンシングデータ中の標的に関連付けられた識別可能な配列を有するノイズ分子標識の数を決定する工程と；(i i i) 標的の数を推定する工程と、を実施するように、実行可能命令によりプログラムされ、ここで、推定された標的の数は、(i i) で決定されたノイズ分子標識の数に応じて調節された、(i) でカウントされたシーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数と相関する。いくつかの実施形態では、本方法は、シーケンシングデータ中の標的のシーケンシングステータスを決定する工程をさらに含む。シーケンシングデータ中の標的のシーケンシングステータスは、飽和シーケンシング、過少シーケンシング、または過剰シーケンシングである。

20

【0089】

いくつかの実施形態では、飽和シーケンシングステータスは、所定の飽和閾値よりも大きい、識別可能な配列を含む分子標識の数を有する標的によって決定される。たとえば、所定の飽和閾値は、確率バーコードが、識別可能な配列を有する約 6 5 6 1 の分子標識を含む場合、約 6 5 5 7 である。所定の飽和閾値は、確率バーコードが、識別可能な配列を有する約 6 5 5 3 6 の分子標識を含む場合、約 6 5 5 3 2 でありうる。シーケンシングデータ中の標的のシーケンシングステータスが、飽和シーケンシングステータスである場合、(i i) で決定されたノイズ分子標識の数は、ゼロである。

30

【0090】

いくつかの実施形態では、過少シーケンシングステータスは、所定の過少シーケンシング閾値より小さい深度（たとえば、平均、最小、または最大深度）を有する標的によって決定することができる。過少シーケンシング閾値は約 4 である。過少シーケンシング閾値は、識別可能な配列を有する分子標識の数とは無関係でありうる。シーケンシングデータ中の標的のシーケンシングステータスが、過少シーケンシングステータスである場合、(i i) で決定されたノイズ分子標識の数は、ゼロである。

40

【0091】

いくつかの実施形態では、過剰シーケンシングステータスは、所定の過剰シーケンシング閾値より大きい、識別可能な配列を有する分子標識の数を有する標的によって決定される。たとえば、過剰シーケンシング閾値は、確率バーコードが、識別可能な配列を有する約 6 5 6 1 の分子標識を含む場合、約 2 5 0 でありうる。本方法は、シーケンシングデータ中の標的のシーケンシングステータスが、過剰シーケンシングステータスである場合、シーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数を、

50

所定の過剰シーケンシング閾値にサブサンプリングする工程を含む。

【0092】

いくつかの実施形態では、シーケンシングデータ中の標的に関連付けられた識別可能な配列を有するノイズ分子標識の数を決定する工程は、ネガティブ二項分布当てはめ条件が満たされれば、(i v) シグナルネガティブ二項分布を、(i) でカウントされたシーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数に当てはめる工程であって、シグナルネガティブ二項分布が、シグナル分子標識である、(i) でカウントされたシーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数に対応するステップと；(v) ノイズネガティブ二項分布を、(i) でカウントされたシーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数に当てはめる工程であって、ノイズネガティブ二項分布が、ノイズ分子標識である、(i) でカウントされたシーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数に対応する工程と；(v i) (v) で当てはめたシグナルネガティブ二項分布および(v i) で当てはめたノイズネガティブ二項分布を用いて、ノイズ分子標識の数を決定する工程と、を含む。

10

【0093】

いくつかの実施形態では、ネガティブ二項分布当てはめ条件は、以下：シーケンシングデータ中の標的のシーケンシングステータスが、過少シーケンシングステータスまたは過剰シーケンシングステータスではないことを含む。(v) で当てはめたシグナルネガティブ二項分布および(v i) で当てはめたノイズネガティブ二項分布を用いて、ノイズ分子標識の数を決定する工程は、シーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の各々について、識別可能な配列のシグナル確率が、シグナルネガティブ二項分布であることを決定する工程と；識別可能な配列のノイズ確率が、ノイズネガティブ二項分布であることを決定する工程と；シグナル確率がノイズ確率より小さければ、識別可能な配列がノイズ分子標識であることを決定する工程と、を含む。

20

【0094】

いくつかの実施形態では、シーケンシングデータ中の標的に関連付けられた識別可能な配列を有するノイズ分子標識の数を決定する工程は、シーケンシングデータ中の標的のシーケンシングステータスが、過少シーケンシングステータスまたは過剰シーケンシングステータスではなく、かつ、(i) でカウントされたシーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数が、擬似点閾値より少ない場合、(i i) でシーケンシングデータ中の標的に関連付けられた識別可能な配列を有するノイズ分子標識の数を決定する前に、シーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数に擬似点を加える工程を含む。擬似点閾値は、10である。

30

【0095】

いくつかの実施形態では、シーケンシングデータ中の標的に関連付けられた識別可能な配列を有するノイズ分子標識の数を決定する工程は、シーケンシングデータ中の標的のシーケンシングステータスが、過少シーケンシングステータスまたは過剰シーケンシングステータスではなく、かつ、(i) でカウントされたシーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数が、擬似点閾値以上である場合、(i i) でシーケンシングデータ中の標的に関連付けられた識別可能な配列を有するノイズ分子標識の数を決定する際に、非ユニーク分子標識を除去する工程を含む。

40

【0096】

いくつかの実施形態では、非ユニーク分子標識を除去する工程は、シーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数が、所定の再使用分子標識閾値より大きい場合、(i i) でシーケンシングデータ中の標的に関連付けられた識別可能な配列を有するノイズ分子標識の数を決定する際に、非ユニーク分子標識を除去する工程を含む。たとえば、再使用分子標識閾値は、確率バーコードが、識別可能な配列を有する約6561の分子標識を含む場合、約650でありうる。

【0097】

50

本明細書に、実行されると、本明細書に開示した方法のいずれかを実施する実行可能コードを含む1つ以上の非一過性コンピュータ読取り媒体が開示される。

【図面の簡単な説明】

【0098】

【図1】非限定的な例示的確率バーコードを示す。

【図2】非限定的な例示的確率バーコーディングおよびデジタルカウンティングを示す。

【図3】複数の標的からの確率バーコード標的のインデックス付きライブラリーを作製するための非限定的な例示のプロセスを示す概略図である。

【図4】分子標識エラー、サンプル標識エラー、および真の分子標識シグナルの非限定的な例示的分布を示す概略図である。

【図5】分子標識を用いて、PCRおよびシーケンシングエラーを訂正する非限定的な例示の実施形態を示すフローチャートである。

【図6】完全シーケンシングおよび不完全シーケンシングによって得られるシーケンシングデータを示す概略図である。

【図7】方向近接性に基づく分子標識を用いて、PCRおよびシーケンシングエラーを訂正する非限定的な例示の実施形態を示すフローチャートである。

【図8】再帰的置換エラー訂正および分子標識深度変化の二次導関数に基づいて、PCRおよびシーケンシングエラーを訂正する非限定的な例示の実施形態を示すフローチャートである。

【図9】再帰的置換エラー訂正および分布ベースのエラー訂正に基づき、PCRおよびシーケンシングエラーを訂正する非限定的な例示の実施形態を示すフローチャートである。

【図10】2つのネガティブ二項分布を用いたエラー訂正の非限定的な例示の実施形態を示すフローチャートである。

【図11】マイクロウェルプレートのサブサンプリングおよび分子標識のマッピングにより、再帰的置換エラー訂正および分布ベースのエラー訂正に基づき、PCRおよびシーケンシングエラーを訂正する非限定的な例示の実施形態を示すフローチャートである。

【図12】遺伝子のサブサンプリングおよび分子標識のマッピングにより、再帰的置換エラー訂正および分布ベースのエラー訂正に基づき、PCRおよびシーケンシングエラーを訂正する非限定的な例示の実施形態を示すフローチャートである。

【図13】再帰により、再帰的置換エラー訂正および分布ベースのエラー訂正に基づき、PCRおよびシーケンシングエラーを訂正する非限定的な例示の実施形態を示すフローチャートである。

【図14】初期パラメータ推定値について2番目に高い分子標識を用いることにより、再帰的置換エラー訂正および分布ベースのエラー訂正に基づき、PCRおよびシーケンシングエラーを訂正する非限定的な例示の実施形態を示すフローチャートである。

【図15】本開示の方法に使用するのに好適な非限定的な例示的機器を示す。

【図16】本開示の実施形態に関連して使用することができるコンピュータシステムの非限定的な例示的構造を示す。

【図17】本開示の方法に使用するのに好適な複数のコンピュータシステムを含むネットワークを示す非限定的な例示的構造を図示する。

【図18】本開示の方法に従う共有仮想アドレスメモリー空間を用いるマルチプロセッサコンピュータシステムの非限定的な例示的構造を示す。

【図19】完全および不完全シーケンシング遺伝子の非限定的な例を示す。

【図20】1塩基シーケンシングエラーならびに真およびエラーバーコードを分離するための閾値についての訂正後のシーケンシングリードとその等級の非限定的な例示のプロットである。

【図21】ゼロ切断ポアソンモデルの非限定的な例示的図である。

【図22】ウェル当たりの総シーケンシングリードの棒グラフを示す。

【図23】完全シーケンシング遺伝子(%)、真のバーコードとして保持された分子標識(ML)(%)および各ウェルについて保持されたそれらのMLにマッピングされた保持

10

20

30

40

50

リード (%) の棒グラフを示す。

【図 2 4】各ウェルの遺伝子により変動する保持リード (%) の箱ひげ図を示す。

【図 2 5】2つのプレートからのアルゴリズム適用後の未補正 M L 対訂正 M I を使用する主成分分析 (P C A) を示す。

【図 2 6】入力分子の増加と共に使用されるユニーク分子標識の理論計算の例示的プロットである。

【図 2 7】高発現遺伝子 - A T C B についてのマイクロウェルプレート全体の各分子標識の分子標識カバー率を示す例示的プロットであり、ここで、エラー分子標識とリアル分子標識との間に明瞭な分布が観察される。

【図 2 8】高発現遺伝子 - A T C B についてのマイクロウェルプレート全体の各分子標識の分子標識カバー率に、2つのネガティブ二項分布を当てはめる工程を示す例示的プロットである。2つのネガティブ二項分布の当てはめによって、より低い分子標識深度を有する分子標識エラーと、より高い分子標識深度を有する真の分子標識を統計的に識別することができることが実証される。x 軸は、分子深度である。

10

【図 2 9】分子標識訂正を示し、ここで、1 のペアワイズハミング距離が大きな比率を占めた。分子標識訂正後、1 のハミング距離相違する分子標識がクラスター化して、同じ親分子標識へ折りたたまれた。

【図 3 0】訂正された分子標識の数対訂正されたリード数カバー率の曲線を示す。

【図 3 1】再帰的置換エラー訂正の一例の概略図を示す。

【図 3 2】パネル (a) ~ (e) は、分子標識深度変化の二次導関数に基づいて、P C R およびシーケンシングエラーを訂正した例示的な結果を示す。

20

【図 3 3 - 1】パネル (a) ~ (c) は、C D 6 9 について2つのネガティブ二項分布に基づいて、P C R およびシーケンシングエラーを訂正した例示的な結果を示す。

【図 3 3 - 2】同上。

【図 3 4 - 1】パネル (a) ~ (c) は、C D 3 E について2つのネガティブ二項分布に基づき、P C R およびシーケンシングエラーを訂正した例示的な結果を示す。

【図 3 4 - 2】同上。

【図 3 5 - 1】パネル (a) ~ (c) は、高発現遺伝子について2つのネガティブ二項分布に基づき、P C R およびシーケンシングエラーを訂正した例示的な結果を示す。

【図 3 5 - 2】同上。

30

【図 3 6】高発現遺伝子の G リッチ分子標識の再使用の例示的な結果を示す。

【図 3 7】パネル (a) ~ (b) は、2つのネガティブ二項分布を当てはめる前に、高発現遺伝子について入力データを調節した例示的な結果を示す。

【図 3 8 - 1】パネル (a) ~ (j) は、2つのネガティブ二項分布を用いて訂正されたデータセットの非限定的な例示的検証を示す。

【図 3 8 - 2】同上。

【図 3 8 - 3】同上。

【図 3 8 - 4】同上。

【図 3 8 - 5】同上。

【図 3 9 - 1】パネル (a) ~ (d) は、混合 J u r k a t および乳癌 (B r C a) 単一細胞 (86 の被検遺伝子) の 96 ウェルからの P r e c i s e (商標) 標的アッセイの例示的な t - 確率的近傍埋込み (t - S N E) 視覚化を示す。

40

【図 3 9 - 2】同上。

【図 4 0 - 1】パネル (a) ~ (b) は、D B S c a n により計算され、かつ各クラスター中の遺伝子マーカーレベルにより決定された、両方の選択クラスターにおいて、> 0 M L の遺伝子に関する細胞クラスター間の差異発現分析を示す非限定的な例示的プロットである。

【図 4 0 - 2】同上。

【図 4 1 - 1】パネル (a) ~ (d) は、86 の被検遺伝子を含む混合 J u r k a t および乳癌 (T 4 7 D) 単一細胞の 96 ウェルプレートからの B D P r e c i s e (商標)

50

標的アッセイのt-SNE確率的近傍埋込み（t-SNE）視覚化を示す、非限定的な例示的プロットである。

【図41-2】同上。

【図42-1】いずれかのエラー訂正工程前（図42、パネル（a）に示す未補正ML）、ならびにRSECおよびDBEC訂正後（図42、パネル（b）に示す調節ML）に、図41で同定されたさまざまな細胞クラスター間の分子標識カウントによる差異遺伝子発現を表示する非限定的な例示的ヒートマップである。

【図42-2】同上。

【発明を実施するための形態】

【0099】

以下の詳細な説明では、その一部を成す添付の図面を参照にする。これら図面において、類似する符号は、文脈から他の解釈が要求されない限り、一般に、類似の構成要素を同一のものとみなす。詳細な説明、図面、および特許請求の範囲に記載される例示的な実施形態は、限定的であることを意味しない。本明細書に提示される主題の精神または範囲から逸脱することなく、他の実施形態を使用してもよく、また他の変更を実施してもよい。本明細書に概略的に記載され、図面に図示されるように、本開示の態様は、非常に多様な異なる構成で配置、代替、組合せ、分離、および設計することができ、それらのすべては、本明細書において明示的に考慮され、本開示の一部を成すものとするを理解されたい。

【0100】

本明細書で参照にされるすべての特許、公開特許出願、他の刊行物、ならびにGenBankおよび他のデータベースからの配列は、関連技術に関してその全体を参照により組み込むものとする。

【0101】

少数の核酸、たとえば、メッセンジャーリボ核酸（mRNA）分子などの定量は、たとえば、さまざまな発生段階またはさまざまな環境条件下で発現される遺伝子を決定するために、臨床上重要である。しかし、特に、分子数が非常に小さい場合、核酸分子（たとえば、mRNA分子）の絶対数を決定するのは極めて困難となりうる。サンプル中の分子の絶対数を決定する方法は、ディジタルポリメラーゼ連鎖反応（PCR）である。理想的には、PCRは、各サイクルで分子の同一コピーを産生する。しかしながら、PCRは、各分子は、推計学的確率で複製し、この確率は、PCRサイクルおよび遺伝子配列によって変動するため、増幅バイアスおよび不正確な遺伝子発現測定値が生じるといった問題を有する。ユニーク分子標識（分子指標（MI）とも呼ばれる）を有する確率バーコードを用いて、分子数をカウントし、増幅バイアスを補正することができる。Precise（商標）アッセイ（Cellular Research, Inc.（Palo Alto, CA））などの確率バーコーディングは、分子標識（ML）を用いて、逆転写（RT）中にmRNAに標識することによって、PCRおよびライブラリー作製工程により誘導されるバイアスを補正することができる。

【0102】

Precise（商標）アッセイは、RT工程中に、サンプル中のすべてのポリ（A）-mRNAとハイブリダイズさせるために、ポリ（T）オリゴヌクレオチド上に多数（たとえば、6561～65536）のユニーク分子標識を有する確率バーコードの非枯渇プールを使用することができる。分子標識に加えて、確率バーコードのサンプル標識（サンプル指標（SI）とも呼ばれる）を用いて、Precise（商標）プレートの各ウェルを識別することができる。確率バーコードは、ユニバーサルPCRプライミング部位を含んでもよい。RTの最中に、標的遺伝子分子は、確率バーコードとランダムに反応する。各標的分子は、得られた確率バーコードとハイブリダイズして、確率バーコード付きの相補的リボヌクレオチド酸（cDNA）分子を生成する。標識した後、マイクロウェルプレートのマイクロウェルからの確率バーコード付きcDNA分子を、PCR増幅およびシーケンシングのために単一チューブ中にプールすることができる。未補正のシーケンシングデータを分析して、ポアソン補正または2つのネガティブ二項分布に基づく補正方法に

10

20

30

40

50

従い、リードの数、ユニーク分子標識を有する確率バーコードの数、mRNA分子の数を取得しうる。

【0103】

バイアス補正以外に、分子標識は、観察されるシーケンシングリードに存在する出発cDNA分子数を明らかにすることによって、結果の統計的品質のよりよい理解をもたらすことができる。たとえば、多数のリードは、統計的に正確な回答を示しうるが、リードが、わずかに少数の出発mRNA分子から得られる場合、測定精度は損なわれうる。

【0104】

PCRおよびライブラリー作製工程により誘導された増幅バイアスは、たとえば、分子標識によって修正することはできるが、分子の絶対数の定量は、いくつかの他の因子のために依然として困難となりうる。第1に、mRNA分子の数の推定は、分子標識の全体的多様性によって制限されうる。確率バーコーディングの最中に、mRNA分子は、利用可能な確率バーコードとランダムに反応することができる。従って、各mRNA分子は、確率バーコードとハイブリダイズすることができるが；その分子標識は、いずれか所与の遺伝子について必ずしもユニークではない場合もある。mRNA分子の数が、確率バーコードの数に比べて小さいとき、各々のmRNA分子は、ユニーク分子標識を有する確率バーコードとハイブリダイズする傾向があり、分子数のカウントは、分子標識数のカウントと同等となりうる。

【0105】

mRNA分子の数が増加するにつれて、多数のmRNA分子は、同じ分子標識を有する確率バーコードとハイブリダイズする傾向が高くなる。故に、ユニーク分子標識のカウントを使用すると、分子数を過少評価する可能性がある。いくつかの事例では、mRNA分子の数は、ポアソン補正または観察されるユニーク分子標識の総数の2つのネガティブ二項分布に基づく補正に従って推定することができる。しかしながら、6561の確率バーコードの全コレクションが観察される極端な事例では、ポアソン補正または2つのネガティブ二項分布に基づく補正はもはや不可能となりうる。たとえば、65000または100000の出発mRNA分子のいずれにかかわらず、いずれの場合でも、6561飽和確率バーコードの最大値が予想される。

【0106】

第2に、PCRエラー（すなわち、PCR増幅の最中に発生したエラー）は、人工的確率バーコードを誘導して、分子標識カウントを任意で増大させうる。第3に、PCR増幅バイアスおよび非効率的PCRは、エラーと識別不可能なバーコード付き分子の少数コピーを生成しうる。第4に、シーケンシングエラー、確率バーコード配列の不正確なコーリングは、人工的確率バーコードを誘導して、分子標識カウントを増大させうる。さらに、シーケンシング深度は、特に、シーケンシングが、浅すぎて、サンプルライブラリー中に存在する確率バーコード付きmRNAのすべてを検出することができない場合に重要となりうる。

【0107】

1つ以上のPCRを有する標的の数、または訂正若しくは調節されたシーケンシングエラーを決定する方法およびシステムが本明細書に開示される。いくつかの実施形態では、本方法は、（a）複数の確率バーコードを用いて、複数の標的に確率バーコードを付けて、複数の確率バーコード付き標的を生成する工程であって、複数の確率バーコードの各々が分子標識を含む工程と；（b）確率バーコード付き標的のシーケンシングデータを取得する工程と；（c）複数の標的の1つ以上について：（i）シーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数をカウントする工程と；（ii）（b）で得られたシーケンシングデータ中の標的のクオリティステータスを決定する工程と；（iii）（b）で得られたシーケンシングデータ中の1つ以上のシーケンシングデータエラーを決定する工程であって、シーケンシングデータ中の1つ以上のシーケンシングデータエラーを決定する工程が、以下：シーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数、シーケンシングデータ中の標的のクオリティス

10

20

30

40

50

テータス、および複数の確率バーコードに識別可能な配列を有する分子標識の数のうち1つ以上を決定することを含む工程と；(i v) 標的の数を推定する工程であって、推定された標的の数が、(i i i) で決定された1つ以上のシーケンシングデータエラーに応じて調節された、(i) でカウントされたシーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数と相関する工程と、を含む。

【0108】

方向近接性に基づいて訂正若しくは調節された1つ以上のPCRまたはシーケンシングエラーを有する標的の数を決定する方法が開示される。いくつかの実施形態では、本方法は、(a) 複数の確率バーコードを用いて、複数の標的に確率バーコードを付けて、複数の確率バーコード付き標的を生成する工程であって、複数の確率バーコードの各々が分子標識を含む工程と；(b) 確率バーコード付き標的のシーケンシングデータを取得する工程と；(c) 複数の標的の1つ以上について：(i) シーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数をカウントする工程と；(i i) 方向近接性を用いて、標的の分子標識のクラスターを同定する工程と；(i i i) (i i) で同定された標的の分子標識のクラスターを用いて、(b) で得られたシーケンシングデータを折りたたむ工程と；(i v) 標的の数を推定する工程であって、推定された標的の数が、(i i) のシーケンシングデータの折りたたみ後に、(i) でカウントされたシーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数と相関する工程と、を含む。

【0109】

訂正若しくは調節された1つ以上のPCRまたはシーケンシングエラーを有する標的の数を決定するためのコンピュータシステムが開示される。実行されると、1つ以上のコンピュータデバイスに、訂正若しくは調節された1つ以上のPCRまたはシーケンシングエラーを有する標的の数を決定させる、実行可能コードを含む非一過性コンピュータ読取り媒体が開示される。

【0110】

定義

特に定義がない限り、本明細書で用いられる技術用語はすべて、本開示が属する分野の当業者により一般に理解されているものと同一の意味を有する。たとえば、Singleton et al., Dictionary of Microbiology and Molecular Biology 2nd ed., J. Wiley & Sons (New York, NY 1994); Sambrook et al., Molecular Cloning, A Laboratory Manual, Cold Spring Harbor Press (Cold Springs Harbor, NY 1989) を参照されたい。本開示の目的のために、下記の用語を以下に定義する。

【0111】

本明細書で用いられる場合、「アダプター」という用語は、関連核酸の増幅またはシーケンシングを促進するための配列を意味する。関連核酸は、標的核酸を含みうる。関連核酸は、空間標識、標的標識、サンプル標識、指標標識、バーコード、確率バーコード、または分子標識の1つ以上を含みうる。アダプターは、線状であってよい。アダプターは、事前にアデニル化されたアダプターであってよい。アダプターは、二本鎖または一本鎖であってよい。1つ以上のアダプターは、核酸の5' または3' 末端に配置することができる。アダプターが5' および3' 末端に既知の配列を含む場合、既知の配列は、同じ配列でも、異なる配列でもよい。ポリヌクレオチドの5' および／または3' 末端に位置するアダプターは、表面上に固定された1つ以上のオリゴヌクレオチドにハイブリダイズする能力を有する。アダプターは、いくつかの実施形態では、ユニバーサル配列を含む。ユニバーサル配列は、2つ以上の核酸分子と共通のヌクレオチド配列の1領域であってよい。2つ以上の核酸分子は、異なる配列の領域を有する。従って、たとえば、5' アダプターは、同一配列および／またはユニバーサル核酸配列を含み、3' アダプターは、同一配列および／またはユニバーサル配列を含みうる。複数の核酸分子の異なるメンバー中に存在する

ユニバーサル配列は、ユニバーサル配列と相補的な単一ユニバーサルプライマーを用いて、複数の異なる配列の複製または増幅を可能にする。同様に、核酸分子のコレクションの異なるメンバー中に存在する少なくとも1つ、2つ（たとえば、ペア）若しくはそれ以上のユニバーサル配列は、ユニバーサル配列と相補的な少なくとも1つ、2つ（たとえば、一対）若しくはそれ以上の単一ユニバーサルプライマーを用いて、複数の異なる配列の複製または増幅を可能にする。従って、ユニバーサルプライマーは、こうしたユニバーサル配列とハイブリダイズすることができる配列を含む。標的核酸配列担持分子を修飾して、ユニバーサルアダプター（たとえば、非標的核酸配列）を異なる標的核酸配列の一端または両端に結合させることができる。標的核酸に結合した1つ以上のユニバーサルプライマーは、ユニバーサルプライマーのハイブリダイゼーションのための部位を提供することができる。標的核酸に結合した1つ以上のユニバーサルプライマーは、同じでも、互いに異なってもよい。

【0112】

本明細書で用いられる場合、「関連付けられる」または「～に関連付けられる」という用語は、ある時点で2つ以上の種が共配置されているとして同定可能であることを意味する。関連付けは、2つ以上の種が類似の容器内にあることを意味する。関連付けは、インフォマティクスの関連付けでありうる。この場合、たとえば、2つ以上の種に関するデジタル情報が記憶され、かつその情報を用いてこれらの種の1つ以上が共配置されたことを決定可能である。関連付けはまた、物理的関連付けでありうる。いくつかの実施形態では、2つ以上の関連付けられる種は、互いにまたは共通の固体もしくは半固体の表面に「テザー連結」、「結合」、または「固定」される。関連付けは、ビーズなどの固体または半固体の支持体に標識を結合するための共有結合手段または非共有結合手段を意味する。関連付けは、標的と標識との共有結合でありうる。

【0113】

本明細書で用いられる場合、「相補的」という用語は、2つのヌクレオチド間の精密なペアリングの能力を意味する。たとえば、核酸の所与の位置のヌクレオチドが他の核酸のヌクレオチドと水素結合可能である場合、2つの核酸はその位置で互いに相補的であるとみなされる。2つの一本鎖核酸分子間の相補性は、ヌクレオチドの一部のみが結合する場合には「部分的」でありうるし、一本鎖分子間のすべてに相補性が存在する場合には完全でありうる。第1のヌクレオチド配列が第2のヌクレオチド配列に相補的である場合、第1のヌクレオチド配列は第2の配列の「相補体」とであるといえる。第1のヌクレオチド配列が第2の配列の逆（すなわち、ヌクレオチドの順序が逆）の配列に相補的である場合、第1のヌクレオチド配列は第2の配列の「逆相補体」とであるといえる。本明細書で用いられる場合、「相補体」、「相補的」、および「逆相補体」という用語は、同義的に用いることが可能である。ある分子が他の分子にハイブリダイズする場合、それはハイブリダイズしている分子の相補体でありうるということが、本開示から理解される。

【0114】

本明細書で用いられる場合、「デジタルカウンティング」という用語は、サンプル中の標的分子の数を推定する方法を意味する。デジタルカウンティングは、サンプル中の標的に関連付けられたユニーク標識の数を決定する工程を含みうる。この確率的方法は、分子をカウントする問題を、同一の分子の位置決定および同定の問題から、所定の標識のセットの検出に関する一連のあり／なしのデジタル問題に変換する。

【0115】

本明細書で用いられる場合、「標識」という用語は、サンプル内の標的に関連付けられる核酸コードを意味する。標識は、たとえば、核酸標識でありうる。標識は、全体または一部が増幅可能な標識でありうる。標識は、全体または一部がシーケンス可能標識でありうる。標識は、個別に同定可能な天然核酸の一部でありうる。標識は、既知の配列でありうる。標識は、核酸配列の接合（たとえば、天然配列と非天然配列との接合）を含みうる。本明細書で用いられる場合、「標識」という用語は、「インデックス」、「タグ」、または「標識タグ」という用語と同義的に用いうる。標識は、情報を伝達可能である。たと

例えば、種々の実施形態では、標識は、サンプル同一性、サンプル源、細胞同一性、および／または標的を決定するために使用可能である。

【0116】

本明細書で用いられる場合、「非枯渇リザーバー」という用語は、多種多様な標識から構成された確率バーコードのプールを意味する。非枯渇リザーバーは、非枯渇リザーバーが標的のプールに関連付けられる場合、各標的がユニーク確率バーコードに関連付けられる可能性が高くなるように、多数の異なる確率バーコードを含みうる。各標識標的分子のユニーク性は、ランダム選択の統計により決定可能であり、標識の多様性と比較してコレクション中の同一の標的分子のコピー数に依存する。得られる標識標的分子のセットのサイズは、バーコーディングプロセスの確率的性質により決定可能であり、次いで、検出された確率バーコードの数の解析は、元のコレクションまたはサンプル中に存在する標的分子の数の計算を可能にする。存在する標的分子のコピー数とユニーク確率バーコードの数との比が低い場合、標識標的分子はきわめてユニークである（すなわち、2つ以上の標的分子が1つの所与の標識で標識される確率は非常に低い）。

【0117】

本明細書で用いられる場合、「核酸」という用語は、ポリヌクレオチド配列またはその断片を意味する。核酸はヌクレオチドを含みうる。核酸は細胞に対して外因性または内因性でありうる。核酸は細胞フリー環境中に存在しうる。核酸は遺伝子またはその断片でありうる。核酸はDNAでありうる。核酸はRNAでありうる。核酸は1つ以上のアナログ（たとえば、修飾された骨格、糖または核酸塩基）を含みうる。アナログのいくつかの例としては、限定されるものではないが、5-ブロモウラシル、ペプチド核酸、ゼノ核酸、モルホリノ体、ロックド核酸、グリコール核酸、トレオース核酸、ジデオキシヌクレオチド、コルジセピン、7-デアザ-GTP、フルオロフォア（たとえば、糖に結合されたローダミンまたはフルオレセイン）、チオール含有ヌクレオチド、ビオチン結合ヌクレオチド、蛍光塩基アナログ、CpGアイランド、メチル-7-グアノシン、メチル化ヌクレオチド、イノシン、チオウリジン、プソイドウリジン、ジヒドロウリジン、キューオシン、およびワイオシンが挙げられる。「核酸」、「ポリヌクレオチド」、「標的ポリヌクレオチド」、および「標的核酸」は、同義的に用いうる。

【0118】

核酸は、新しいまたは向上した特徴（たとえば、向上した安定性）を有する核酸を提供するために1つ以上の修飾（たとえば、塩基修飾、骨格修飾）を含みうる。核酸は核酸アフィニティータグを含みうる。ヌクレオシドは塩基-糖の組合せでありうる。ヌクレオシドの塩基部分はヘテロ環塩基でありうる。かかるヘテロ環塩基の2つの最も一般的なクラスはプリンおよびピリミジンである。ヌクレオチドは、ヌクレオシドの糖部分に共有結合されたリン酸基をさらに含むヌクレオシドでありうる。ペントフラノシル糖を含むヌクレオシドでは、リン酸基は、糖の2'、3'、または5'ヒドロキシル部分に結合可能である。核酸を形成する際、リン酸基は、隣接ヌクレオシドを互いに共有結合して線状高分子化合物を形成可能である。ひいては、この線状高分子化合物のそれぞれの末端をさらに連結して環状化合物を形成可能である。しかしながら、線状化合物が一般に好適である。そのほかに、線状化合物は、内部ヌクレオチド塩基相補性を有しうるので、完全二本鎖または部分二本鎖の化合物を生成するようにフォールディングしうる。核酸内では、リン酸基は、通常、核酸のヌクレオシド間骨格を形成するものとして参照可能である。結合または骨格は、3'→5'ホスホジエステル結合でありうる。

【0119】

核酸は、修飾骨格および／または修飾ヌクレオシド間結合を含みうる。修飾骨格は、骨格中にリン原子を保持するものおよび骨格中にリン原子を有していないものを含みうる。リン原子を中に含有する好適な修飾核酸骨格は、たとえば、ホスホロチオエート、キラルホスホロチオエート、ホスホロジチオエート、ホスホトリエステル、アミノアルキルホスホトリエステル、3'-アルキレンホスホネートや5'-アルキレンホスホネートなどのメチルや他のアルキルのホスホネート、キラルホスホネート、ホスフィネート、3'-アミノホ

10

20

30

40

50

スホルアミデートやアミノアルキルホスホルアミデートなどのホスホルアミデート、ホスホロジアミデート、チオノホスホルアミデート、チオノアルキルホスホネート、チオノアルキルホスホトリエステル、セレノホスフェート、および通常3' - 5' 結合、2' - 5' 結合アナログを有するボラノホスフェート、ならびに1つ以上のヌクレオチド間結合が3' → 3'、5' → 5'、または2' → 2' 結合である逆極性を有するものを含む。

【0120】

核酸は、短鎖アルキルもしくはシクロアルキルのヌクレオシド間結合、混合ヘテロ原子およびアルキルもしくはシクロアルキルのヌクレオシド間結合、または1つ以上の短鎖ヘテロ原子もしくはヘテロ環のヌクレオシド間結合により形成されるポリヌクレオチド骨格を含む。これらは、モルホリノ結合（ヌクレオシドの糖部分から部分的に形成される）、シロキサ骨格、スルフィド、スルホキシド、およびスルホン骨格、ホルムアセチルおよびチオホルムアセチル骨格、メチレンホルムアセチルおよびチオホルムアセチル骨格、リボアセチル骨格、アルケン含有骨格、スルファメート骨格、メチレンイミノおよびメチレンヒドラジノ骨格、スルホネートおよびスルホンアミド骨格、アミド骨格を有するもの、ならびに混合N、O、S、およびCH₂構成部分を有する他のものを含む。

【0121】

核酸は核酸ミメティックを含む。「ミメティック」という用語は、フラノース環のみまたはフラノース環とヌクレオチド間結合の両方が非フラノース基で置き換えられているポリヌクレオチドを含むことを意図し得、フラノース環のみの置換えは、糖サロゲートであるとして参照可能である。ヘテロ環塩基部分または修飾ヘテロ環塩基部分は、適切な標的核酸とのハイブリダイゼーションのために保持可能である。かかる核酸の1つはペプチド核酸（PNA）でありうる。PNAでは、ポリヌクレオチドの糖骨格は、アミド含有骨格特にアミノエチルグリシン骨格で置換え可能である。ヌクレオチドは保持可能であり、かつ骨格のアミド部分のアザ窒素原子に直接的または間接的に結合される。PNA化合物中の骨格は、PNAにアミド含有骨格を与える2つ以上の結合されたアミノエチルグリシン単位を含む。ヘテロ環塩基部分は、骨格のアミド部分のアザ窒素原子に直接的または間接的に結合可能である。

【0122】

核酸はモルホリノ骨格構造を含む。たとえば、核酸は、リボース環の代わりに6員モルホリノ環を含む。これらの実施形態のいくつかでは、ホスホロジアミデートまたは他の非ホスホジエステルのヌクレオシド間結合によりホスホジエステル結合を置換え可能である。

【0123】

核酸は、モルホリノ環に結合されたヘテロ環塩基を有する結合されたモルホリノ単位（すなわちモルホリノ核酸）を含む。結合基は、モルホリノ核酸中のモルホリノモノマー単位を結合可能である。非イオン性モルホリノ系オリゴマー化合物は、細胞タンパク質とのより少ない望ましくない相互作用を有しうる。モルホリノ系ポリヌクレオチドは、核酸の非イオン性ミミックでありうる。モルホリノクラス内のさまざまな化合物は、異なる結合基を用いて連結可能である。ポリヌクレオチドミメティックのさらなるクラスは、シクロヘキセニル核酸（CeNA）として参照可能である。核酸分子中に通常存在するフラノース環は、シクロヘキセニル環で置換え可能である。CeNA DMT保護ホスホロアミダイトモノマーは、ホスホロアミダイト化学を用いたオリゴマー化合物合成のために調製および使用が可能である。核酸鎖中へのCeNAモノマーの取込みは、DNA/RNAハイブリッドの安定性を増加可能である。CeNAオリゴアデニレートは、天然複合体に類似した安定性を有する核酸相補体との複合体を形成可能である。さらなる修飾は、2' - ヒドロキシル基が糖環の4' 炭素原子に結合されて2' - C、4' - C - オキシメチレン結合を形成することにより二環式糖部分を形成するロックド核酸（LNA）を含む。結合は、2' 炭素原子と4' 炭素原子とを架橋するメチレン（-CH₂）、基（式中、nは1または2である）でありうる。LNAおよびLNAアナログは、相補的核酸との非常に高い二本鎖熱安定性（T_m=+3 ~ +10℃）、3' - エキソヌクレアーゼ分解に対する安定性、

および良好な溶解性を示しうる。

【0124】

核酸はまた、核酸塩基（単に「塩基」ということが多い）の修飾または置換を含みうる。本明細書で用いられる場合、「非修飾」または「天然」の核酸塩基は、プリン塩基（たとえば、アデニン（A）およびグアニン（G））、ならびにピリミジン塩基（たとえば、チミン（T）、シトシン（C）およびウラシル（U））を含みうる。修飾核酸塩基は、他の合成および天然の核酸塩基、たとえば、5-メチルシトシン（5-me-C）、5-ヒドロキシメチルシトシン、キサンチン、ヒポキサンチン、2-アミノアデニン、アデニンおよびグアニンの6-メチルおよび他のアルキル誘導体、アデニンおよびグアニンの2-プロピルおよび他のアルキル誘導体、2-チオウラシル、2-チオチミンおよび2-チオシトシン、5-ハロウラシルおよびシトシン、5-プロピニル（ $-C \equiv C-CH_3$ ）ウラシルおよびシトシン、ならびにピリミジン塩基の他のアルキニル誘導体、6-アゾウラシル、シトシンおよびチミン、5-ウラシル（プソイドウラシル）、4-チオウラシル、8-ハロ、8-アミノ、8-チオール、8-チオアルキル、8-ヒドロキシル、ならびに他の8-置換アデニンおよびグアニン、5-ハロ特に5-ブロモ、5-トリフルオロメチルおよび他の5-置換ウラシルおよびシトシン、7-メチルグアニンおよび7-メチルアデニン、2-F-アデニン、2-アミノアデニン、8-アザグアニンおよび8-アザアデニン、7-デアザグアニンおよび7-デアザアデニン、ならびに3-デアザグアニンおよび3-デアアデニンを含みうる。修飾核酸塩基は、三環式ピリミジン、たとえば、フェノキサジンシチジン（1H-ピリミド（5，4-b）（1，4）ベンゾオキサジン-2（3H）-オン）、フェノチアジンシチジン（1H-ピリミド（5，4-b）（1，4）ベンゾチアジン-2（3H）-オン）、置換フェノキサジンシチジン（たとえば、9-（2-アミノエトキシ）-H-ピリミド（5，4-b）（1，4）ベンゾオキサジン-2（3H）-オン）などのG-クランプ、フェノチアジンシチジン（1H-ピリミド（5，4-b）（1，4）ベンゾチアジン-2（3H）-オン）、置換フェノキサジンシチジン（たとえば、9-（2-アミノエトキシ）-H-ピリミド（5，4-b）（1，4）ベンゾオキサジン-2（3H）-オン）などのG-クランプ、カルバゾールシチジン（2H-ピリミド（4，5-b）インドール-2-オン）、ピリドインドールシチジン（H-ピリド（3，'：4，5）ピロロ[2，3-d]ピリミジン-2-オン）を含みうる。

【0125】

本明細書で用いられる場合、「サンプル」という用語は、標的を含む組成物を意味しうる。本開示の方法、デバイス、およびシステムによる分析に好適なサンプルとしては、細胞、組織、器官、または生物が挙げられる。

【0126】

本明細書で用いられる場合、「サンプリングデバイス」または「デバイス」という用語は、サンプルのセクションの採取および／または基材上へのセクションの配置を行うデバイスを含む。サンプルデバイスとは、たとえば、蛍光活性化細胞選別（FACS）機、セルソーター機、生検針、生検デバイス、組織切片化デバイス、マイクロ流体デバイス、ブレードグリッド、および／またはマイクロトームを含む。

【0127】

本明細書で用いられる場合、「固体担体」という用語は、複数の確率バーコードを結合しうる離散した固体または半固体の表面を意味しうる。固体担体は、核酸を（たとえば共有結合または非共有結合で）固定しうるプラスチック、セラミック、金属、または高分子材料（たとえばヒドロゲル）で構成された任意のタイプの中実、多孔性、または中空のスフェア、ボール、ベアリング、シリンダー、または他の類似の構成体を包含しうる。固体担体は、球状（たとえばマイクロスフェア）でありうるかまたは非球状もしくは不規則形状、たとえば、立方体形、直方体形、角錐形、円柱形、円錐形、扁球形、ディスク形などを有する離散粒子を含みうる。アレイ状に離間して配置された複数の固体担体は、基材を含まないこともありうる。固体担体は、「ビーズ」という用語と同義的に用いうる。

【0128】

固体担体は「基材」を意味しうる。基材は固体担体の1種でありうる。基材は、本開示の方法を行いうる連続した固体または半固体の表面を意味しうる。基材は、たとえば、アレイ、カートリッジ、チップ、デバイス、およびスライドを意味しうる。

【0129】

本明細書で用いられる場合、「空間標識」という用語は、空間内の位置と関連させることができる標識を意味しうる。

【0130】

本明細書で用いられる場合、「確率バーコード」という用語は、標識を含むポリヌクレオチド配列を意味しうる。確率バーコードは、確率バーコーディングに使用可能なポリヌクレオチド配列でありうる。確率バーコードは、サンプル中の標的を定量可能である。確率バーコードは、標識を標的に関連付けた後に起こりうるエラーの制御に使用可能である。たとえば、確率バーコードは、増幅またはシーケンシングのエラーを評価可能である。標的に関連付けられた確率バーコードは、確率バーコード標的または確率バーコードタグ標的と呼ぶことが可能である。

【0131】

本明細書で用いられる場合、「遺伝子特異的確率バーコード」という用語は、標識と、遺伝子特異的である標的結合領域とを含むポリヌクレオチド配列を意味しうる。確率バーコードは、確率バーコーディングに使用することができるポリヌクレオチド配列でありうる。確率バーコードは、サンプル中の標的を定量するために使用することができる。確率バーコードは、標識を標的に関連付けた後に起こりうるエラーの制御に使用することができる。たとえば、確率バーコードは、増幅またはシーケンシングのエラーを評価することができる。標的に関連付けられた確率バーコードは、確率バーコード標的または確率バーコードタグ標的と呼ぶことができる。

【0132】

本明細書で用いられる場合、「確率バーコーディング」という用語は、核酸のランダム標識化（たとえばバーコーディング）を意味しうる。確率バーコーディングは、標識を標的に関連付けて、標識に関連付けられた標識を定量するために再帰的ポアソンストラテジーを利用することができる。本明細書で用いられる場合、「確率バーコーディング」という用語は、「遺伝子特異的確率バーコーディング」と置き換え可能に用いられうる。

【0133】

本明細書で用いられる場合、「標的」という用語は、確率バーコードに関連付け可能な組成物を意味しうる。本開示の方法、デバイス、およびシステムによる分析に好適な例示的な標的としては、オリゴヌクレオチド、DNA、RNA、mRNA、マイクロRNA、tRNAなどが挙げられる。標的は一本鎖または二本鎖でありうる。いくつかの実施形態では、標的はタンパク質でありうる。いくつかの実施形態では、標的は脂質である。

【0134】

本明細書で用いられる場合、「逆転写酵素」という用語は、逆転写酵素活性を有する（すなわち、RNA鋳型からのDNAの合成を触媒する）酵素のグループを意味しうる。一般的には、かかる酵素としては、限定されるものではないが、レトロウイルス逆転写酵素、レトロトランスポゾン逆転写酵素、レトロプラスミド逆転写酵素、レトロ逆転写酵素、細菌逆転写酵素、グループIIイントロン由来逆転写酵素、およびそれらの突然変異体、変異体、または誘導体が挙げられる。非レトロウイルス逆転写酵素としては、非LTRレトロトランスポゾン逆転写酵素、レトロプラスミド逆転写酵素、レトロ逆転写酵素、およびグループIIイントロン逆転写酵素が挙げられる。グループIIイントロン逆転写酵素の例としては、ラクトコッカス・ラクティス (*Lactococcus lactis*) L1、LtrBイントロン逆転写酵素、サーモシネコッカス (*Thermosynechococcus*) は、TeI4cイントロン逆転写酵素、またはジオバチルス・ステアロサーモフィラス (*Geobacillus stearothermophilus*) GsI-IIIcイントロン逆転写酵素を伸長する。他のクラスの逆転写酵素としては、多くのクラスの非レトロウイルス逆転写酵素（すなわち、レトロ、グループIIイント

ロン、および特に多様性生成レトロエレメント）が挙げられうる。

【0135】

「ユニバーサルアダプタープライマー」、「ユニバーサルプライマーアダプター」または「ユニバーサルアダプター配列」という用語は、置き換え可能に用いられて、確率バーコードをハイブリダイズして、遺伝子特異的確率バーコードを作製するために使用することができるヌクレオチド配列を指す。ユニバーサルアダプター配列は、たとえば、本開示の方法に用いられるすべての確率バーコードに対してユニバーサルである既知の配列であってよい。たとえば、本明細書に開示する方法を用いて複数の標的が標識される場合、標的特異的配列の各々を同じユニバーサルアダプター配列に連結させてもよい。いくつかの実施形態では、本明細書に開示する方法に、2つ以上のユニバーサルアダプター配列を使用することができる。たとえば、本明細書に開示する方法を用いて複数の標的が標識される場合、標的特異的配列の少なくとも2つを異なるユニバーサルアダプター配列と連結させる。ユニバーサルアダプタープライマーおよびその補体は、2つのオリゴヌクレオチドに含有させてもよく、そのうちの1つは、標的特異的配列を含み、他方は、確率バーコードを含む。たとえば、ユニバーサルアダプター配列は、標的核酸と相補的なヌクレオチド配列を生成するための標的特異的配列を含むオリゴヌクレオチドの一部であってもよい。確率バーコードと、ユニバーサルアダプター配列の相補的配列を含む第2のオリゴヌクレオチドは、ヌクレオチド配列とハイブリダイズして、標的特異的確率バーコードを生成しうる。いくつかの実施形態では、ユニバーサルアダプタープライマーは、本開示の方法で使

10

20

【0136】

本明細書には、PCRおよび／またはシーケンシングの最中に発生したエラーを検出し、および／または訂正する方法およびシステムが開示される。エラーのタイプとしては、限定するものではないが、たとえば、置換エラー（1つ以上の塩基）および非置換エラーがある。置換エラーのうち、1塩基置換エラーは、2塩基以上相違するエラーよりもはるかに頻繁に起こりうる。本方法およびシステムは、たとえば、確率バーコーディングにより分子標的の正確なカウンティングを達成するために使用することができる。

【0137】

確率バーコード

確率バーコーディングは、たとえば、米国特許出願公開第20150299784号明細書、国際公開第2015031691号パンフレット、およびFu et al., Proc Natl Acad Sci U.S.A. 2011 May 31; 108(22): 9026-31に記載されており、これらの刊行物の内容は、その全体を参照により本明細書に組み込む。手短には、確率バーコードは、標的に確率標識（例えば、バーコード、タグ）を付けるために使用することができるポリヌクレオチド配列であってよい。確率バーコードは、1つ以上の標識を含みうる。例示的な標識としては、ユニバーサル標識、細胞標識、分子標識、サンプル標識、プレート標識、空間標識、および／またはプレ空間標識を挙げることができる。図1は、空間標識を有する例示的な確率バーコード104を示す。確率バーコード104は、確率バーコードを固体担体105に連結しうる5'アミンを含んでよい。確率バーコードは、ユニバーサル標識、次元標識、空間標識、細胞標識、および／または分子標識を含みうる。確率バーコード中のさまざまな標識（限定するものではないが、ユニバーサル標識、次元標識、空間標識、細胞標識、および分子標識など）の順序は変動しうる。たとえば、図1に示すように、ユニバーサル標識は、最も5'側の標識であってよく、分子標識は、最も3'側の標識であってもよい。空間標識、次元標識、および細胞標識は、任意の順序であってよい。いくつかの実施形態では、ユニバーサル標識、空間標識、次元標識、細胞標識、および分子標識は、任意の順序であってよい。

30

40

【0138】

標識、たとえば、細胞標識は、規定長さ、たとえば、各々7ヌクレオチド（いくつかのハミングエラー訂正コードに使用されるビット数に相当する）の核酸部分配列の固有のセットを含んでもよく、これらは、エラー訂正能力を賦与するように設計することができる。

50

エラー訂正部分配列のセットは、7つのヌクレオチド配列を含み、これらは、セット内の配列の任意のペア組合せが、規定の「遺伝子距離」（またはミスマッチ塩基の数）を呈示するように、設計することができ、たとえば、3ヌクレオチドの遺伝子距離を呈示するように、1セットのエラー訂正部分配列を設計することができる。この場合、標識化標的核酸分子についてのシーケンシングデータのセット内のエラー訂正配列の見直しによって、増幅若しくはシーケンシングエラーを検出または訂正することが可能になる。いくつかの実施形態では、エラー訂正コードを作製するために用いられる核酸部分配列の長さは、たとえば、約1、2、3、4、5、6、7、8、9、10、15、20、30、31、40、50ヌクレオチド長、またはこれらの値のいずれか2つの間の数もしくは範囲であってよい。いくつかの実施形態では、エラー訂正コードを作製するために、他の長さの核酸部分配列を使用することも可能である。

10

【0139】

確率バーコードは、標的結合領域を含みうる。標的結合領域は、サンプル中の標的と相互作用することができる。標的は、リボ核酸（RNA）、メッセンジャーRNA（mRNA）、microRNA、低分子干渉RNA（siRNA）、RNA分解産物、各々がポリ（A）テールを含有するRNA、またはそれらの任意の組合せであってもよいし、これらを含んでもよい。いくつかの実施形態では、複数の標的は、デオキシリボ核酸（DNA）を含みうる。

【0140】

いくつかの実施形態では、標的結合領域は、mRNAのポリ（A）テールと相互作用することができるオリゴ（dT）配列を含みうる。確率バーコードの標識（たとえば、ユニバーサル標識、次元標識、空間標識、細胞標識、および分子標識）の1つ以上は、確率バーコードの残りの標識の別の1つまたは2つからスペーサによって隔てることができる。スペーサは、たとえば、1、2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19、もしくは20ヌクレオチドまたはそれ以上であってよい。いくつかの実施形態では、確率バーコードの標識のいずれもスペーサによって隔てられない。

20

【0141】

ユニバーサル標識

確率バーコードは1つ以上のユニバーサル標識を含みうる。いくつかの実施形態では、1つ以上のユニバーサル標識は、所与の固体担体に結合される確率バーコードのセット中のすべての確率バーコードで同一でありうる。いくつかの実施形態では、1つ以上のユニバーサル標識は、複数のビーズに結合されるすべての確率バーコードで同一でありうる。いくつかの実施形態では、ユニバーサル標識は、シーケンシングプライマーにハイブリダイズ可能な核酸配列を含みうる。シーケンシングプライマーは、ユニバーサル標識を含む確率バーコードをシーケンスするために使用可能である。シーケンシングプライマー（たとえば、ユニバーサルシーケンシングプライマー）は、高スループットシーケンシングプラットフォームに関連付けられるシーケンシングプライマーを含みうる。いくつかの実施形態では、ユニバーサル標識は、PCRプライマーにハイブリダイズ可能な核酸配列を含みうる。いくつかの実施形態では、ユニバーサル標識は、シーケンシングプライマーおよびPCRプライマーにハイブリダイズ可能な核酸配列を含みうる。シーケンシングプライマーまたはPCRプライマーにハイブリダイズ可能なユニバーサル標識の核酸配列は、プライマー結合部位として参照しうる。ユニバーサル標識は、確率バーコードの転写を開始するために使用しうる配列を含みうる。ユニバーサル標識は、確率バーコードまたは確率バーコード内の領域の伸長のために、使用しうる配列を含みうる。ユニバーサル標識は、約1、2、3、4、5、10、15、20、25、30、35、40、45、50ヌクレオチド長、またはこれらの値のいずれか2つの間の数もしくは範囲であってよい。たとえば、ユニバーサル標識は、少なくとも約10ヌクレオチドを含みうる。ユニバーサル標識は、少なくとも、または多くとも、1、2、3、4、5、10、15、20、25、30、35、40、45、50、100、200、もしくは300ヌクレオチド長

30

40

50

でありうる。いくつかの実施形態では、切断可能なリンカーまたは修飾ヌクレオチドは、担体から確率バーコードを切断して除去することを可能にするユニバーサル標識配列の一部であってよい。

【0142】

次元標識

確率バーコードは1つ以上の次元標識を含みうる。いくつかの実施形態では、次元標識は、確率標識化が行われた次元に関する情報を提供する核酸配列を含みうる。たとえば、次元標識は、標的に確率バーコードが付された時点に関する情報を提供可能である。次元標識は、サンプルの確率バーコーディングの時点に関連付け可能である。次元標識は、確率標識化の時点で活性化可能である。異なる時点で異なる次元標識を活性化可能である。次元標識は、標的、標的のグループ、および／またはサンプルに確率バーコードを付けた順序に関する情報を提供する。たとえば、細胞集団は、細胞周期のG0期に確率バーコードを付けることが可能である。細胞は、細胞周期のG1期に確率バーコードで再びパルスすることが可能である。細胞は、細胞周期のS期に確率バーコードで再びパルスすることが可能であり、他の時期も同様である。各パルス時（たとえば、細胞周期の各期）の確率バーコードは、異なる次元標識を含みうる。こうして、次元標識は、細胞周期のどの期に標的に標識したかに関する情報を提供する。次元標識は、多種多様な生物時間を精査することが可能である。例示的な生物時間としては、限定されるものではないが、細胞周期、転写（たとえば転写開始）、および転写物分解が挙げられうる。他の例として、薬剤治療および／または療法の前および／または後にサンプル（たとえば、細胞、細胞集団）に確率標識を付けることが可能である。識別可能な標的のコピー数の変化は、薬剤および／または療法に対するサンプルの反応の指標でありうる。

【0143】

次元標識は、活性化可能であってよい。活性化可能な次元標識は、特定の時点で活性化可能でありうる。活性化可能な標識は、たとえば、構成的に活性化することができる（たとえば、オフに切り替わらない）。活性化可能な次元標識は、たとえば、可逆的に活性化可能である（たとえば、活性化可能な次元標識は、オン・オフの切替えが可能である）。たとえば、次元標識は、少なくとも1、2、3、4、5、6、7、8、9、もしくは10回またはそれ以上可逆的に活性化可能でありうる。次元標識は、たとえば、少なくとも1、2、3、4、5、6、7、8、9、もしくは10回またはそれ以上可逆的に活性化可能でありうる。いくつかの実施形態では、次元標識は、蛍光、光、化学的イベント（たとえば、切断、他の分子のライゲーション、修飾（たとえば、ペグ化、SUMO化、アセチル化、メチル化、脱アセチル化、脱メチル化）の付加、光化学的イベント（たとえば、光ケーシング）、および非天然ヌクレオチドの導入により活性化可能である。

【0144】

次元標識は、いくつかの実施形態では、所与の固体担体（たとえばビーズ）に結合されるすべての確率バーコードで同一でありうるが、異なる固体担体（たとえばビーズ）では異なりうる。いくつかの実施形態では、同一の固体担体上の確率バーコードの少なくとも60%、70%、80%、85%、90%、95%、97%、99%、または100%は、同一の次元標識を含みうる。いくつかの実施形態では、同一の固体担体上の確率バーコードの少なくとも60%は、同一の次元標識を含みうる。いくつかの実施形態では、同一の固体担体上の確率バーコードの少なくとも95%は、同一の次元標識を含みうる。

【0145】

複数の固体担体（たとえばビーズ）には、10⁶程度またはそれ以上のユニーク次元標識配列が存在可能である。次元標識は、1、2、3、4、5、10、15、20、25、30、35、40、45、50、またはこれらの値のいずれか2つの間の数もしくは範囲、あるいはそうした近似値のヌクレオチド長でありうる。次元標識は、少なくとも、または多くとも、1、2、3、4、5、10、15、20、25、30、35、40、45、50、100、200、もしくは300ヌクレオチド長でありうる。次元標識は、約5～約200ヌクレオチドを含みうる。次元標識は、約10～約150ヌクレオチドを含みうる

。次元標識は、約20～約125ヌクレオチドを含みうる。

【0146】

空間標識

確率バーコードは1つ以上の空間標識を含みうる。いくつかの実施形態では、空間標識は、確率バーコードに関連付けられる標的分子の空間配向に関する情報を提供する核酸配列を含みうる。空間標識は、サンプル中の座標に関連付け可能である。座標は固定座標でありうる。たとえば、座標は基材を基準にして固定可能である。空間標識は二次元または三次元のグリッドを基準にしうる。座標はランドマークを基準にして固定可能である。ランドマークは空間内で同定可能である。ランドマークはイメージング可能な構造体でありうる。ランドマークは生物学的構造体たとえば解剖学的ランドマークでありうる。ランドマークは細胞ランドマーク（たとえばオルガネラ）でありうる。ランドマークは、非天然ランドマーク、たとえば、色コード、バーコード、磁性、蛍光、放射能、またはユニークなサイズもしくは形状のような同定可能な識別子を有する構造体でありうる。空間標識は、物理的パーティション（たとえば、ウェル、容器、またはドロップレット）に関連付け可能である。いくつかの実施形態では、空間内の1つ以上の位置にコードを付けるために複数の空間標識が一緒に使用される。

10

【0147】

空間標識は、所与の固体担体（たとえばビーズ）に結合されるすべての確率バーコードで同一であってよいが、異なる固体担体（たとえばビーズ）については異なってもよい。いくつかの実施形態では、同一の空間標識を含む、同一の固体担体上の確率バーコードのパーセンテージは、60%、70%、80%、85%、90%、95%、97%、99%、100%、またはこれらの値のいずれか2つの間の数もしくは範囲、あるいはそうした近似値でありうる。いくつかの実施形態では、同一の空間標識を含む、同一の固体担体上の確率バーコードのパーセンテージは、少なくとも、または多くとも、60%、70%、80%、85%、90%、95%、97%、99%、もしくは100%でありうる。いくつかの実施形態では、同一の固体担体上の確率バーコードの少なくとも60%が、同一の空間標識を含んでよい。いくつかの実施形態では、同一の固体担体上の確率バーコードの少なくとも95%が、同一の空間標識を含んでよい。

20

【0148】

複数の固体担体（たとえばビーズ）には、10⁶程度またはそれ以上のユニーク空間標識配列が存在可能である。空間標識は、1、2、3、4、5、10、15、20、25、30、35、40、45、50、またはこれらの値のいずれか2つの間の数もしくは範囲、あるいはそうした近似値のヌクレオチド長でありうる。空間標識は、少なくとも、または多くとも1、2、3、4、5、10、15、20、25、30、35、40、45、50、100、200、もしくは300ヌクレオチド長でありうる。空間標識は、約5～約200ヌクレオチドを含みうる。空間標識は、約10～約150ヌクレオチドを含みうる。空間標識は、約20～約125ヌクレオチドを含みうる。

30

【0149】

細胞標識

確率バーコードは、1つ以上の細胞標識を含みうる。いくつかの実施形態では、細胞標識は、どの標的核酸がどの細胞に由来するかを決定するための情報を提供する核酸配列を含みうる。いくつかの実施形態では、細胞標識は、所与の固体担体（たとえばビーズ）に結合されるすべての確率バーコードで同一であるが、異なる固体担体（たとえばビーズ）については異なっている。いくつかの実施形態では、同一の細胞標識を含む、同一の固体担体上の確率バーコードのパーセンテージは、60%、70%、80%、85%、90%、95%、97%、99%、100%、またはこれらの値のいずれか2つの間の数もしくは範囲、あるいはそうした近似値でありうる。いくつかの実施形態では、同一の細胞標識を含む、同一の固体担体上の確率バーコードのパーセンテージは、60%、70%、80%、85%、90%、95%、97%、99%、もしくは100%、またはそうした近似値であってよい。たとえば、同一の固体担体上の確率バーコードの少なくとも60%が、同

40

50

一の細胞標識を含みうる。別の例として、同一の固体担体上の確率バーコードの少なくとも95%が、同一の細胞標識を含んでもよい。

【0150】

複数の固体担体（たとえばビーズ）には、 10^6 程度またはそれ以上のユニーク細胞標識配列が存在可能である。細胞標識は、1、2、3、4、5、10、15、20、25、30、35、40、45、50、またはこれらの値のいずれか2つの間の数もしくは範囲、あるいはそうした近似値のヌクレオチド長でありうる。細胞標識は、少なくとも、または多くとも、1、2、3、4、5、10、15、20、25、30、35、40、45、50、100、200、もしくは300ヌクレオチド長でありうる。たとえば、細胞標識は、約5～約200ヌクレオチドを含みうる。別の例として、細胞標識は、約10～約150ヌクレオチドを含みうる。さらに別の例として、細胞標識は、約20～約125ヌクレオチドを含みうる。

10

【0151】

分子標識

確率バーコードは、1つ以上の分子標識を含みうる。いくつかの実施形態では、分子標識は、確率バーコードにハイブリダイズされた標的核酸種の特定のタイプを同定するための情報を提供する核酸配列を含みうる。分子標識は、確率バーコード（たとえば標的結合領域）にハイブリダイズされた標的核酸種の特定の存在に対するカウンターを提供する核酸配列を含みうる。

【0152】

いくつかの実施形態では、分子標識の多様なセットが所与の固体担体（たとえばビーズ）に結合される。いくつかの実施形態では、 10^2 、 10^3 、 10^4 、 10^5 、 10^6 、 10^7 、 10^8 、 10^9 、またはこれらの値のいずれか2つの間の数もしくは範囲、あるいはそうした近似値のユニーク分子標識配列が存在しうる。たとえば、複数の確率バーコードは、識別可能な配列を有する約6561の分子標識を含みうる。別の例として、複数の確率バーコードは、識別可能な配列を有する約65536の分子標識を含みうる。いくつかの実施形態では、少なくとも、または多くとも、 10^2 、 10^3 、 10^4 、 10^5 、 10^6 、 10^7 、 10^8 、もしくは 10^9 のユニーク分子標識配列が存在しうる。ユニーク分子標識配列は、所与の固体担体（たとえばビーズ）に結合されている。

20

【0153】

分子標識は、1、2、3、4、5、10、15、20、25、30、35、40、45、50ヌクレオチド長、またはこれらの値のいずれか2つの間の数もしくは範囲、あるいはそうした近似値のヌクレオチド長でありうる。分子標識は、少なくとも、または多くとも、1、2、3、4、5、10、15、20、25、30、35、40、45、50、100、200、もしくは300ヌクレオチド長でありうる。

30

【0154】

標的結合領域

確率バーコードは、1つ以上の標的結合領域を含みうる。いくつかの実施形態では、標的結合領域は、対象の標的とハイブリダイズすることができる。いくつかの実施形態では、標的結合領域は、標的（たとえば、標的核酸、標的分子、たとえば、分析される細胞核酸）、たとえば、特定の遺伝子配列に特異的にハイブリダイズする核酸配列を含みうる。いくつかの実施形態では、標的結合領域は、特定の標的核酸の特定の位置に結合（たとえばハイブリダイズ）しうる核酸配列を含みうる。いくつかの実施形態では、標的結合領域は、制限酵素部位オーバーハング（たとえば、EcoRI付着末端オーバーハング）への特異的なハイブリダイゼーションが可能な核酸配列を含みうる。次いで、確率バーコードは、制限部位オーバーハングに相補的な配列を含む任意の核酸分子にライゲートしうる。

40

【0155】

いくつかの実施形態では、標的結合領域は非特異的標的核酸配列を含みうる。非特異的標的核酸配列は、標的核酸の特定の配列に依存せずに複数の標的核酸に結合しうる配列を意味しうる。たとえば、標的結合領域は、ランダムマルチマー配列を含みうるかまたはmR

50

N A分子のポリ（A）テールにハイブリダイズするオリゴ（dT）配列を含みうる。ランダムマルチマー配列は、たとえば、ランダムダイマー、ランダムトリマー、ランダムクアトラマー、ランダムペンタマー、ランダムヘキサマー、ランダムセプタマー、ランダムオクタマー、ランダムノナマー、ランダムデカマー、または任意の長さのより高次のランダムマルチマーの配列でありうる。いくつかの実施形態では、標的結合領域は、所与のビーズに結合されたすべての確率バーコードで同一である。いくつかの実施形態では、所与のビーズに結合された複数の確率バーコードの標的結合領域は、2つ以上の異なる標的結合配列を含む。標的結合領域は、5、10、15、20、25、30、35、40、45、50、またはこれらの値のいずれか2つの間の数もしくは範囲、あるいはそうした近似値のヌクレオチド長でありうる。もしくはそれ以上または概略で少なくともそうしたヌクレオチド長でありうる。標的結合領域は、多くとも約5、10、15、20、25、30、35、40、45、50ヌクレオチド長またはそれ以上でありうる。

10

【0156】

いくつかの実施形態では、標的結合領域は、ポリアデニル化末端を含むmRNAにハイブリダイズすることができるオリゴ（dT）を含みうる。標的結合領域は、遺伝子特異的でありうる。たとえば、標的結合領域は、標的の特定の領域にハイブリダイズするように構成することができる。標的結合領域は、1、2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19、20、21、22、23、24、25、26、27、28、29、30、またはこれらの値のいずれか2つの間の数もしくは範囲、あるいはそうした近似値のヌクレオチド長でありうる。標的結合領域は、少なくとも、または多くとも、1、2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19、20、21、22、23、24、25、26、27、28、29、もしくは30ヌクレオチド長でありうる。標的結合領域は、約5〜30ヌクレオチド長であってもよい。確率バーコードが、遺伝子特異的標的結合領域を含む場合、この確率バーコードは、遺伝子特異的確率バーコードと呼ぶことができる。

20

【0157】

配向性

確率バーコードは、確率バーコードの配向（たとえばアライメント）のために使用することができる1つ以上の配向性を含みうる。確率バーコードは、等電点電気泳動用の部分を含みうる。異なる確率バーコードは、異なる等電点電気泳動点を含みうる。こうした確率バーコードをサンプルに導入した場合、サンプルは、確率バーコードを既知の形態にオリエントするために等電点電気泳動を行うことが可能である。こうして、オリエント性は、サンプルで確率バーコードの既知のマップを作成するために使用可能である。例示的なオリエント性としては、電気泳動移動度（たとえば、確率バーコードのサイズに基づく）、等電点、スピン、伝導率、および／またはセルフアセンブリーが挙げられうる。たとえば、セルフアセンブリーのオリエント性を含む確率バーコードは、活性化時に特定のオリエンテーションにセルフアセンブル可能である（たとえば、核酸ナノ構造）。

30

【0158】

親和性

確率バーコードは、1つ以上の親和性を含みうる。たとえば、空間標識は、親和性を含みうる。親和性は、他のエンティティー（たとえば細胞レセプター）との確率バーコードの結合を促進することができる化学的および／または生物学的部分を含みうる。たとえば、親和性は、抗体、たとえば、サンプル上の特定の部分（たとえばレセプター）に特異的な抗体を含みうる。いくつかの実施形態では、抗体は、確率バーコードを特定の細胞型または分子に誘導することができる。特定の細胞型もしくは分子および／またはその近傍にある標的を確率標識化することができる。抗体は確率バーコードを特定の位置に誘導することができるので、いくつかの実施形態において、親和性は、空間標識のヌクレオチド配列に加え、空間情報も提供することができる。抗体は、治療用抗体、たとえば、モノクローナル抗体またはポリクローナル抗体であってもよい。抗体は、ヒト化されていても、またはキメラであってもよい。抗体は、ネイキッド抗体または融合抗体であってもよい。

40

50

【0159】

抗体は、全長（すなわち、天然に存在するかもしれない）通常の免疫グロブリン遺伝子断片組換えプロセスにより形成される）免疫グロブリン分子（たとえばIgG抗体）または免疫グロブリン分子の免疫活性（すなわち特異的結合）部分たとえば抗体フラグメントでありうる。

【0160】

抗体フラグメントは、たとえば、F(ab')₂、Fab'、Fab、Fv、sFvなどの抗体の一部でありうる。いくつかの実施形態において、抗体フラグメントは、全長抗体により認識される同一の抗原に結合可能である。抗体フラグメントは、抗体の可変領域からなる単離された断片、たとえば、重鎖および軽鎖の可変領域からなる「Fv」フラグメントならびに軽鎖および重鎖の可変領域がペプチドリンカーにより接続された組換え一本鎖ポリペプチド分子（「scFvタンパク質」）を含みうる。例示的な抗体としては、限定されるものではないが、癌細胞に対する抗体、ウイルスに対する抗体、細胞表面レセプター（CD8、CD34、CD45）に結合する抗体、および治療用抗体が挙げられうる。

【0161】

ユニバーサルアダプタープライマー

確率バーコードは、1つ以上のユニバーサルアダプタープライマーを含みうる。たとえば、遺伝子特異的確率バーコードは、ユニバーサルアダプタープライマーを含みうる。ユニバーサルアダプタープライマーは、すべての確率バーコードに対してユニバーサルであるヌクレオチド配列を意味しうる。ユニバーサルアダプタープライマーは、遺伝子特異的確率バーコードを構築するために使用することができる。ユニバーサルアダプタープライマーは、1、2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19、20、21、22、23、24、25、26、27、28、29、30、またはこれらの値のいずれか2つの間の数もしくは範囲、あるいはそうした近似値のヌクレオチド長でありうる。ユニバーサルアダプタープライマーは、少なくとも、または多くとも、1、2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19、20、21、22、23、24、25、26、27、28、29、もしくは30ヌクレオチド長でありうる。ユニバーサルアダプタープライマーは、約5〜30ヌクレオチド長であってもよい。

【0162】

固体担体

本明細書に開示される確率バーコードは、いくつかの実施形態において、固体担体と結合することができる。固体担体は、たとえば、合成粒子であってもよい。いくつかの実施形態では、固体担体上の複数の確率バーコード（たとえば、第1の複数の確率バーコード）の分子標識（たとえば、第1の分子標識）の一部または全部が、少なくとも1ヌクレオチド異なる。同じ固体担体上の確率バーコードの細胞標識は、同じであってもよい。異なる固体担体上の確率バーコードの細胞標識は、少なくとも1ヌクレオチド異なりうる。たとえば、第1の固体担体上の第1の複数の確率バーコードの第1の細胞標識は、同じ配列を有してよく、第2の固体担体上の第2の複数の確率バーコードの第2の細胞標識は、同じ配列を有してよい。第1の固体担体上の第1の複数の確率バーコードの第1の細胞標識と、第2の固体担体上の第2の複数の確率バーコードの第2の細胞標識とは、少なくとも1ヌクレオチド異なりうる。細胞標識は、たとえば、約5〜20ヌクレオチド長でありうる。分子標識は、たとえば、約5〜20ヌクレオチド長でありうる。合成粒子は、たとえば、ビーズであってもよい。

【0163】

ビーズは、たとえば、シリカゲルビーズ、調節多孔性ガラスビーズ、磁気ビーズ、ダイナビーズ、セファデックス／セファロースビーズ、セルロースビーズ、ポリスチレンビーズ、またはそれらの任意の組合せであってもよい。ビーズは、ポリジメチルシロキサン（PDMS）、ポリスチレン、ガラス、ポリプロピレン、アガロース、ゼラチン、ヒドロゲル、常磁性材料、セラミック、プラスチック、ガラス、メチルスチレン、アクリルポリマー、

10

20

30

40

50

チタン、ラテックス、セファロース、セルロース、ナイロン、シリコン、またはそれらの任意の組合せなどの材料を含みうる。

【0164】

いくつかの実施形態では、ビーズは、ポリマービーズ、たとえば、変形性ビーズまたはゲルビーズであってよく、これらは、確率バーコードで官能化されている（たとえば、10 X Genomics (San Francisco, CA) からのゲルビーズなど）。いくつかの実施形態では、ゲルビーズは、ポリマーベースのゲルを含みうる。ゲルビーズは、たとえば、1つ以上のポリマー前駆体を液滴中に封入することによって作製することができる。促進剤（たとえば、テトラメチルエチレンジアミン (TEMED)）にポリマー前駆体を曝露すると、ゲルビーズが作製されうる。

10

【0165】

いくつかの実施形態では、ポリマービーズは、たとえば、所望の条件下で、溶解、溶融、または分解しうる。所望の条件は、環境条件を含みうる。所望の条件は、制御された様式で、ポリマービーズの溶解、溶融、または分解を引き起こしうる。ゲルビーズは、化学的刺激、物理的刺激、生物学的刺激、熱刺激、磁気刺激、電気刺激、光刺激、またはそれらの任意の組合せによって、溶解、溶融、または分解しうる。

【0166】

たとえば、オリゴヌクレオチドバーコードなどの被検物質および／もしくは試薬を、ゲルビーズの内側表面（たとえば、オリゴヌクレオチドバーコードおよび／もしくはオリゴヌクレオチドバーコードを作製するために用いられる材料の拡散を介して進入可能な内部）ならびに／またはゲルビーズの外側表面、あるいは本明細書に記載されるいずれか他のマイクロカプセルにカップリング／固定してもよい。カップリング／固定は、化学結合（たとえば、共有結合、イオン結合）または物理的現象（たとえば、ファンデルワールス力、双極子-双極子相互作用など）の任意の形態を介するものであってよい。いくつかの実施形態では、ゲルビーズまたは本明細書に記載する任意の他のマイクロカプセルに対する試薬のカップリング／固定は、たとえば、不安定部分（たとえば、本明細書に記載の化学架橋剤をはじめとする、化学架橋剤）を介するなど、可逆性であってもよい。刺激を適用すると、不安定部分は、切断されて、固定された試薬が遊離されうる。いくつかの事例では、不安定部分は、ジスルフィド結合である。たとえば、オリゴヌクレオチドバーコードが、ジスルフィド結合を介してゲルビーズに固定されている場合、ジスルフィド結合を還元剤に曝露することにより、ジスルフィド結合を切断して、オリゴヌクレオチドバーコードをビーズから遊離させることができる。不安定部分は、ゲルビーズもしくはマイクロカプセルの一部として、試薬もしくは被検物質をゲルビーズもしくはマイクロカプセルに連結する化学リンカーの一部として、および／または試薬もしくは被検物質の一部として含有させてもよい。

20

30

【0167】

いくつかの実施形態では、ゲルビーズは、限定するものではないが、以下のものをはじめとする、極めて多様なポリマーを含みうる：ポリマー、熱感受性ポリマー、感光性ポリマー、磁気ポリマー、pH感受性ポリマー、塩感受性ポリマー、化学的感受性ポリマー、高分子電解質、多糖、ペプチド、タンパク質、および／またはプラスチック。ポリマーとしては、限定するものではないが、ポリ（N-イソプロピルアクリルアミド）（PNIPAAm）、ポリ（スルホン酸スチレン）（PSS）、ポリ（アリルアミン）（PAAm）、ポリ（アクリル酸）（PAA）、ポリ（エチレンイミン）（PEI）、ポリ（ジアリルジメチル-塩化アンモニウム）（PDADMAC）、ポリ（ピロール）（PPy）、ポリ（ビニルピロリドン）（PVPON）、ポリ（ビニルピリジン）（PVP）、ポリ（メタクリル酸）（PMAA）、ポリ（メチルメタクリレート）（PMMA）、ポリスチレン（PS）、ポリ（テトラヒドロフラン）（PTHF）、ポリ（フタルアルデヒド）（PTHF）、ポリ（ヘキサヒドロピロゲン）（PHV）、ポリ（L-リシン）（PLL）、ポリ（L-アルギニン）（PARG）、乳酸-グリコール酸共重合体（PLGA）などの材料が挙げられる。

40

50

【0168】

多数の化学的刺激を用いて、ビーズの破壊または分解をトリガーすることができる。これらの化学的変化の例として、限定するものではないが、ビーズ壁に対するpH媒介による変化、架橋の化学的切断を介したビーズ壁の崩壊、ビーズ壁の解重合トリガー、およびビーズ壁スイッチング反応が挙げられる。また、バルク変化を用いて、ビーズの破壊をトリガーしてもよい。

【0169】

また、さまざまな刺激を介したマイクロカプセルに対するバルクまたは物理的変化も、試薬を放出するようにカプセルを設計する上で多くの利点をもたらす。バルクまたは物理的変化は、巨視的規模で起こり、その際、ビーズ破断は、刺激により誘導された機械物理的力の結果による。こうしたプロセスとしては、限定するものではないが、圧力誘導破断、ビーズ壁溶融、またはビーズ壁の多孔性変化が挙げられる。

【0170】

生物学的刺激を用いて、ビーズの破壊または分解をトリガーすることもできる。概して、生物学的トリガーは、化学的トリガーと類似しているが、多くの例では、生体分子、または酵素、ペプチド、糖類、核酸などの生存系に一般的に存在する分子が使用される。たとえば、ビーズは、特定のプロテアーゼによる切断に感受性のペプチド架橋を有するポリマーを含んでもよい。さらに具体的には、一例は、GFLGKペプチド架橋を含むマイクロカプセルを含んでもよい。プロテアーゼカテプシンBなどの生物学的トリガーを加えると、シェルウェルのペプチド架橋が切断されて、ビーズの内容物が放出される。他の事例では、プロテアーゼを熱活性化してもよい。別の例では、ビーズは、セルロースを含有するシェル壁を含む。加水分解性酵素キトサンの添加は、セルロース結合の切断、シェル壁の解重合、およびその内部内容物の放出のための生物学的トリガーとして役立つ。

【0171】

さらに、ビーズは、熱刺激の適用時にその内容物を放出するように誘導することもできる。温度の変化は、ビーズにさまざまな変化を引き起こし得る。熱の変化は、ビーズ壁が崩壊するように、ビーズの溶融を引き起こし得る。別の事例では、熱は、ビーズが破断または破裂するように、ビーズの内部成分の内圧を高めうる。また別の事例では、熱は、ビーズを収縮した脱水状態に変形させうる。さらに、熱は、ビーズの壁内の熱感受性ポリマーに作用して、ビーズの破壊を引き起こしうる。

【0172】

マイクロカプセルのビーズ壁に磁気ナノ粒子を含有させると、ビーズの破断トリガー、ならびに多数のビーズの誘導を可能にしうる。本開示のデバイスは、いずれの目的で磁気ビーズを含んでもよい。一例では、高分子電解質含有ビーズにFe₃O₄ナノ粒子を組み込むと、振動磁界刺激の存在下で破断がトリガーされる。

【0173】

ビーズはまた、電気刺激の結果として破壊または分解することもできる。前のセクションに記載した磁気粒子と同様に、電気感受性ビーズも、ビーズの破断トリガー、ならびに電界下でのアラインメント、導電性またはレドックス反応などの他の機能を可能にする。一例では、電気感受性材料を含有するビーズは、内部試薬の放出を制御することができるように、電界下でアラインメントされる。他の例では、電界は、ビーズ壁自体の内部でレドックス反応を誘導することもでき、これにより、多孔性が増加しうる。

【0174】

また、光刺激を用いて、ビーズを破壊することもできる。多数の光トリガーが考えられ、特定の範囲の波長の光子を吸収することができるナノ粒子および発色団などのさまざまな分子を用いるシステムが挙げられる。たとえば、金属酸化物コーティングをカプセルトリガーとして用いることができる。SiO₂でコーティングされた高分子電解質カプセルのUV照射は、ビーズ壁の崩壊を引き起こしうる。また別の例では、アゾベンゼン基などのフォトスイッチ材料をビーズ壁に組み込んでもよい。UVまたは可視光線を適用すると、こうした化学物質は、光子の吸収時に、可逆的シストランス異性化を被る。この態様で

は、光子スイッチの組込みによって、光トリガー適用の際に、崩壊するか、またはより多孔性になりうるビーズ壁が得られる。

【0175】

たとえば、図2に示す確率バーコードの非限定的な例において、ブロック208でのマイクロウェルアレイの複数のマイクロウェルに、単一細胞などの細胞を導入した後、ビーズをブロック212のマイクロウェルアレイの複数のマイクロウェルに導入することができる。各マイクロウェルは、1つのビーズを含みうる。ビーズは、複数の確率バーコードを含みうる。確率バーコードは、ビーズに結合した5'アミン領域を含みうる。確率バーコードは、ユニバーサル標識、分子標識、標的結合領域、またはそれらの任意の組合せを含んでもよい。

10

【0176】

本明細書に開示する確率バーコードは、固体担体（たとえば、ビーズ）に関連（たとえば、結合）させることができる。固体担体と結合した確率バーコードは、各々、ユニーク配列を有する少なくとも100または1000の分子標識を含む群から選択される分子標識を含みうる。いくつかの実施形態では、固体担体と結合した異なる確率バーコードは、異なる配列の分子標識を含んでもよい。いくつかの実施形態では、固体担体と結合した、特定のパーセンテージの確率バーコードが、同じ細胞標識を含む。たとえば、そのパーセンテージは、60%、70%、80%、85%、90%、95%、97%、99%、100%、またはこれらの値のいずれか2つの間の数もしくは範囲、あるいはそうした近似値でありうる。別の例として、パーセンテージは、少なくとも、または多くとも60%、70%、80%、85%、90%、95%、97%、99%、もしくは100%でありうる。いくつかの実施形態では、固体担体と結合した確率バーコードは、同じ細胞標識を含みうる。異なる固体担体と結合した確率バーコードは、ユニーク配列を有する少なくとも100または1000の細胞標識を含む群から選択される、異なる細胞標識を含んでもよい。

20

【0177】

本明細書に開示する確率バーコードは、固体担体（たとえば、ビーズ）に関連（たとえば、結合）させることができる。いくつかの実施形態では、サンプル中の複数の標的に確率バーコードを付ける工程は、複数の確率バーコードと結合した複数の合成粒子を含む固体担体を用いて、実施することができる。いくつかの実施形態では、固体担体は、複数の確率バーコードと結合した複数の合成粒子を含みうる。さまざまな固体担体上の複数の確率バーコードの空間標識は、少なくとも1ヌクレオチド異なりうる。固体担体は、たとえば、2次元または3次元の複数の確率バーコードを含みうる。合成粒子は、ビーズであってよい。ビーズは、シリカゲルビーズ、調節多孔性ガラスビーズ、磁気ビーズ、ダイナビーズ、セファデックス／セファロースビーズ、セルロースビーズ、ポリスチレンビーズ、またはそれらの任意の組合せであってよい。固体担体は、ポリマー、マトリックス、ヒドロゲル、ニードルアレイデバイス、抗体、またはそれらの任意の組合せを含みうる。いくつかの実施形態では、固体担体は、浮動性であってよい。いくつかの実施形態では、固体担体は、半固体または固体アレイに埋め込むことができる。確率バーコードは、固体担体と結合していなくてもよい。確率バーコードは、個別のヌクレオチドであってもよい。確率バーコードは、基材と結合してもよい。

30

40

【0178】

本明細書で使用される場合、「テザー連結」、「結合」、および「固定」という用語は、同義的に用いられて、確率バーコードを固体担体に結合するための共有結合または非共有結合の手段を意味しうる。さまざまな異なるいずれの固体担体も、プレ合成された確率バーコードを結合するための、または確率バーコードを *in situ* 固相合成するための固体担体として使用することができる。

【0179】

いくつかの実施形態では、固体担体はビーズである。ビーズは、核酸を（たとえば共有結合または非共有結合で）固定することができる、固体、多孔性、もしくは中空のスフェア、ボール、ベアリング、シリンダー、または他の類似の構成体の1つ以上のタイプを包含

50

しうる。ビーズは、たとえば、プラスチック、セラミック、金属、もしくは高分子材料、またはそれらの任意の組合せから構成されうる。ビーズは、離散粒子であるか、またはそれを含んでもよく、離散粒子は、球状（たとえばマイクロスフェア）であるか、または非球状もしくは不規則形状、たとえば、立方体形、直方体形、角錐形、円柱形、円錐形、扁球形、ディスク形などを有する。いくつかの実施形態では、ビーズは、非球状の形状でありうる。

【0180】

ビーズは、限定されるものではないが、常磁性材料（たとえば、マグネシウム、モリブデン、リチウム、およびタンタル）、超常磁性材料（たとえば、フェライト（ Fe_3O_4 、マグネタイト）ナノ粒子）、強磁性材料（たとえば、鉄、ニッケル、コバルト、それらのいくつかの合金、およびいくつかの希土類金属化合物）、セラミック、プラスチック、ガラス、ポリスチレン、シリカ、メチルスチレン、アクリルポリマー、チタン、ラテックス、セファロース、アガロース、ヒドロゲル、ポリマー、セルロース、ナイロン、ならびにそれらの任意の組合せなどのさまざまな材料を含みうる。

【0181】

いくつかの実施形態では、ビーズ（たとえば、確率バーコードが結合されたビーズ）は、ヒドロゲルビーズである。いくつかの実施形態では、ビーズは、ヒドロゲルを含む。

【0182】

本明細書に開示するいくつかの実施形態は、1つ以上の粒子（たとえば、ビーズ）を含む。粒子は各々、複数のオリゴヌクレオチド（たとえば、確率バーコード）を含みうる。複数のオリゴヌクレオチドは各々、分子標識配列、細胞標識配列、および標的結合領域（たとえば、オリゴdT配列、遺伝子特異的配列、ランダム多量体、またはそれらの組合せ）を含みうる。複数のオリゴヌクレオチドの各々の細胞標識配列は、同じであってもよい。異なる粒子上のオリゴヌクレオチドの細胞標識配列は、異なる粒子上のオリゴヌクレオチドを同定できるように、相違してもよい。異なる細胞標識配列の数は、異なる実装において相違してもよい。いくつかの実施形態では、細胞標識配列の数は、10、100、200、300、400、500、600、700、800、900、1000、2000、3000、4000、5000、6000、7000、8000、9000、10000、20000、30000、40000、50000、60000、70000、80000、90000、100000、 10^6 、 10^7 、 10^8 、 10^9 、またはこれらの値のいずれか2つの間の数もしくは範囲、またはそれ以上、あるいはそうした近似値でありうる。いくつかの実施形態では、細胞標識配列の数は、少なくとも、または多くとも10、100、200、300、400、500、600、700、800、900、1000、2000、3000、4000、5000、6000、7000、8000、9000、10000、20000、30000、40000、50000、60000、70000、80000、90000、100000、 10^6 、 10^7 、 10^8 、もしくは 10^9 でありうる。いくつかの実施形態では、複数の粒子の1、2、3、4、5、6、7、8、9、10、20、30、40、50、60、70、80、90、100、200、300、400、500、600、700、800、900、1000以下、またはそれ以上が、同じ細胞配列のオリゴヌクレオチドを含む。いくつかの実施形態では、同じ細胞配列のオリゴヌクレオチドを含む複数の粒子は、多くとも0.1%、0.2%、0.3%、0.4%、0.5%、0.7%、0.8%、0.9%、1%、2%、3%、4%、5%、6%、7%、8%、9%、10%またはそれ以上であってもよい。いくつかの実施形態では、複数の粒子のいずれも同じ細胞標識配列を含まない。

【0183】

各粒子の複数のオリゴヌクレオチドは、異なる分子標識配列を含みうる。いくつかの実施形態では、分子標識配列の数は、10、100、200、300、400、500、600、700、800、900、1000、2000、3000、4000、5000、6000、7000、8000、9000、10000、20000、30000、40000、50000、60000、70000、80000、90000、100000、

10⁶、10⁷、10⁸、10⁹、またはこれらの値のいずれか2つの間の数もしくは範囲、あるいはそうした近似値でありうる。分子標識配列の数は、少なくとも、または多くとも10、100、200、300、400、500、600、700、800、900、1000、2000、3000、4000、5000、6000、7000、8000、9000、10000、20000、30000、40000、50000、60000、70000、80000、90000、100000、10⁶、10⁷、10⁸、もしくは10⁹でありうる。たとえば、複数のオリゴヌクレオチドの少なくとも100は、異なる分子標識配列を含む。別の例として、単一粒子において、複数のオリゴヌクレオチドの少なくとも100、500、1000、5000、10000、15000、20000、50000、これらの値のいずれか2つの間の数もしくは範囲、またはそれ以上が、異なる分子標識配列を含む。いくつかの実施形態は、確率バーコードを含む複数の粒子を提供する。いくつかの実施形態では、標的の発生数（またはコピーもしくは数）と異なる分子標識配列の比は、少なくとも、1:1、1:2、1:3、1:4、1:5、1:6、1:7、1:8、1:9、1:10、1:11、1:12、1:13、1:14、1:15、1:16、1:17、1:18、1:19、1:20、1:30、1:40、1:50、1:60、1:70、1:80、1:90、またはそれ以上でありうる。いくつかの実施形態では、複数のオリゴヌクレオチドの各々は、サンプル標識、ユニバーサル標識、またはその両方をさらに含む。粒子は、たとえば、ナノ粒子またはマイクロ粒子であってよい。

【0184】

ビーズのサイズは、変動しうる。たとえば、ビーズの直径は、0.1マイクロメートル〜50マイクロメートルの範囲であってよい。いくつかの実施形態では、ビーズの直径は、0.1、0.5、1、2、3、4、5、6、7、8、9、10、20、30、40、50マイクロメートル、またはこれらの値のいずれか2つの間の数もしくは範囲、あるいはそうした近似値でありうる。

【0185】

ビーズの直径は、基材のウェルの直径と関連させることができる。いくつかの実施形態では、ビーズの直径は、ウェルの直径よりも、10%、20%、30%、40%、50%、60%、70%、80%、90%、100%、またはこれらの値のいずれか2つの間の数もしくは範囲、あるいはそうした近似値だけ長いもしくは短い長さであってよい。ビーズの直径は、細胞（たとえば、基材のウェルに閉じ込められた単一細胞）の直径に関連させることができる。いくつかの実施形態では、ビーズの直径は、細胞の直径よりも、10%、20%、30%、40%、50%、60%、70%、80%、90%、100%、200%、250%、300%、またはこれらの値のいずれか2つの間の数もしくは範囲、あるいはそうした近似値だけ長いもしくは短い長さであってよい。

【0186】

ビーズは、基材への埋込みおよび／または結合が可能である。ビーズは、ゲル、ヒドロゲル、ポリマー、および／またはマトリックスへの埋込みおよび／または結合が可能である。基材（たとえば、ゲル、マトリックス、スキャフォールド、またはポリマー）内のビーズの空間位置は、位置アドレスとして機能可能なビーズ上の確率バーコードに存在する空間標識を用いて同定可能である。

【0187】

ビーズの例としては、限定されるものではないが、ストレプトアビジンビーズ、アガロースビーズ、磁気ビーズ、Dynabead（登録商標）、MACS（登録商標）マイクロビーズ、抗体コンジュゲートビーズ（たとえば、抗免疫グロブリンマイクロビーズ）、プロテインAコンジュゲートビーズ、プロテインGコンジュゲートビーズ、プロテインA／Gコンジュゲートビーズ、プロテインLコンジュゲートビーズ、オリゴ（dT）コンジュゲートビーズ、シリカビーズ、シリカ様ビーズ、抗ビオチンマイクロビーズ、抗蛍光色素マイクロビーズ、およびBcMag（商標）カルボキシル末端磁気ビーズが挙げられうる。

【0188】

ビーズは、1つの蛍光光学チャネルまたは複数の光学チャネルで蛍光を発するように量子

10

20

30

40

50

ドットまたは蛍光色素への関連付け（たとえばそれらによる含浸）が可能である。ビーズは、常磁性または強磁性にするために酸化鉄または酸化クロムへの関連付けが可能である。ビーズは同定可能でありうる。たとえば、ビーズは、カメラを用いてイメージング可能である。ビーズは、ビーズに関連付けられた検出可能なコードを有しうる。たとえば、ビーズは、確率バーコードを含みうる。ビーズは、たとえば、有機または無機の溶液中での膨潤に起因してサイズ変化しうる。ビーズは疎水性でありうる。ビーズは親水性でありうる。ビーズは生体適合性でありうる。

【0189】

固体担体（たとえばビーズ）は可視化可能である。固体担体は可視化タグ（たとえば蛍光色素）を含みうる。固体担体（たとえばビーズ）は識別子（たとえば数）でエッチング可能である。識別子はビーズのイメージングにより可視化可能である。

10

【0190】

基材およびマイクロウェルアレイ

本明細書で使用される場合、基材はあるタイプの固体担体を意味しうる。基材は、本開示の確率バーコードを含みうる固体担体を意味しうる。基材は、たとえば、複数のマイクロウェルを含みうる。たとえば、基材は、2つ以上のマイクロウェルを含むウェルアレイであってよい。いくつかの実施形態では、マイクロウェルは、規定の体積の小さい反応チャンバーを含みうる。いくつかの実施形態では、マイクロウェルは、1つ以上の細胞を閉じ込めることができる。いくつかの実施形態では、マイクロウェルは、1つの細胞のみを閉じ込めることができる。いくつかの実施形態では、マイクロウェルは、1つ以上の固体担体を閉じ込めることができる。いくつかの実施形態では、マイクロウェルは、1つの固体担体のみを閉じ込めることができる。いくつかの実施形態では、マイクロウェルは、単一細胞および単一固体担体（たとえば、ビーズ）を閉じ込める。

20

【0191】

確率バーコーディングの方法

本開示は、身体サンプル（たとえば、組織、器官、腫瘍、細胞）における識別可能な位置の識別可能な標的の数を推定する方法を提供する。本方法は、サンプルと接近させて確率バーコードを配置する工程と、サンプルを溶解させる工程と、識別可能な標的を確率バーコードと関連させる工程と、標的を増幅する工程および／または標的をデジタルカウントする工程と、を含みうる。本方法は、さらに、確率バーコード上の空間標識から得られた情報を分析する工程および／または視覚化する工程をさらに含みうる。いくつかの実施形態では、一方法は、サンプル中の複数の標識を視覚化する工程を含む。サンプルのマップに複数の標的をマッピングする工程は、サンプルの二次元マップまたは三次元マップの作製を含みうる。二次元マップまたは三次元マップは、サンプル中の複数の標的に確率バーコードを付ける前または後に作製することができる。サンプル中の複数の標的を視覚化する工程は、サンプルのマップに複数の標的をマッピングする工程を含みうる。サンプルのマップに複数の標的をマッピングする工程は、サンプルの二次元マップまたは三次元マップを作製するステップを含みうる。二次元マップおよび三次元マップは、サンプル中の複数の標的に確率バーコードを付ける前または後に作製することができる。いくつかの実施形態では、二次元マップおよび三次元マップは、サンプルを溶解させる前または後に作製することができる。二次元マップまたは三次元マップの作製前または後にサンプルを溶解させる工程は、サンプルを加熱する工程と、サンプルを洗剤と接触させる工程と、サンプルのpHを変化させる工程、またはそれらの任意の組合せを含みうる。

30

40

【0192】

いくつかの実施形態では、複数の標的に確率バーコードを付ける工程は、複数の確率バーコードを複数の標的とハイブリダイズさせて、確率バーコード付き標的を作製する工程を含む。複数の標的に確率バーコードを付ける工程は、確率バーコード付き標的のインデックス付きライブラリーを作製する工程を含みうる。確率バーコード付き標的のインデックス付きライブラリーを作製する工程は、複数の確率バーコードを含む固体担体を用いて実施することができる。

50

【0193】

サンプルと確率バーコードの接触

本開示は、サンプル（たとえば、細胞）を本開示の基材と接触させる方法を提供する。たとえば、細胞、器官、または組織薄片を含むサンプルを確率バーコードと接触させることができる。たとえば、重力流によって、細胞を接触させることができ、その場合、細胞は沈殿して単層を形成する。サンプルは、組織薄片であってよい。薄片を基材の上に配置することができる。サンプルは、一次元（たとえば、平面表面を形成する）であってよい。サンプル（たとえば、細胞）は、たとえば、基材上に細胞を増殖させる／培養することによって、基材全体に広げることができる。

【0194】

確率バーコードが標的と近接して位置すると、標的は、確率バーコードとハイブリダイズする。識別可能な標的の各々が、本開示の識別可能な確率バーコードと結合し得るように、確率バーコードを非枯渇的比率で接触させることができる。標的と確率バーコード同士の効率的な結合を確実にするために、標的を確率バーコードと架橋させることができる。

【0195】

細胞溶解

細胞および確率バーコードの分配後、細胞は標的分子を遊離するように溶解可能である。細胞溶解は、さまざまな手段のいずれかにより、たとえば、化学的もしくは生化学的手段により、浸透圧ショックにより、または熱溶解、機械溶解、もしくは光学溶解により達成可能である。細胞は、界面活性剤（たとえば、SDS、Liドデシルスルフェート、Triton X-100、Tween-20、もしくはNP-40）、有機溶媒（たとえば、メタノールもしくはアセトン）、または消化酵素（たとえば、プロテイナーゼK、ペプシンまたはトリプシン）、あるいはそれらの任意の組合せを含む細胞溶解緩衝液の添加により溶解可能である。標的と確率バーコードとの関連付けを向上させるために、たとえば、温度の低下および／またはライセートの粘度の増加により、標的分子の拡散速度を変化させることが可能である。

【0196】

いくつかの実施形態では、サンプルは濾紙を用いて溶解可能である。濾紙は濾紙の上を溶解緩衝液で浸漬可能である。濾紙は、サンプルの溶解および基材へのサンプルの標的のハイブリダイゼーションを促進可能な加圧でサンプルに適用可能である。

【0197】

いくつかの実施形態では、溶解は、機械溶解、熱溶解、光学溶解、および／または化学溶解により行うことが可能である。化学溶解は、プロテイナーゼK、ペプシン、トリプシンなどの消化酵素の使用を含む。溶解は、基材への溶解緩衝液の添加により行うことが可能である。溶解緩衝液はトリスHClを含む。溶解緩衝液は、少なくとも約0.01、0.05、0.1、0.5、もしくは1Mまたはそれ以上のトリスHClを含む。溶解緩衝液は、多くとも約0.01、0.05、0.1、0.5、もしくは1Mまたはそれ以上のトリスHClを含む。溶解緩衝液は約0.1MトリスHClを含む。溶解緩衝液のpHは、少なくとも約1、2、3、4、5、6、7、8、9、もしくは10またはそれ以上でありうる。溶解緩衝液のpHは、多くとも約1、2、3、4、5、6、7、8、9、もしくは10またはそれ以上でありうる。いくつかの実施形態では、溶解緩衝液のpHは約7.5である。溶解緩衝液は塩（たとえばLiCl）を含む。溶解緩衝液中の塩の濃度は、少なくとも約0.1、0.5、もしくは1Mまたはそれ以上でありうる。溶解緩衝液中の塩の濃度は、多くとも約0.1、0.5、もしくは1Mまたはそれ以上でありうる。いくつかの実施形態では、溶解緩衝液中の塩の濃度は約0.5Mである。溶解緩衝液は、界面活性剤（たとえば、SDS、Liドデシルスルフェート、トリトンX、トゥイーン、NP-40）を含む。溶解緩衝液中の界面活性剤の濃度は、少なくとも約0.0001%、0.0005%、0.001%、0.005%、0.01%、0.05%、0.1%、0.5%、1%、2%、3%、4%、5%、6%、もしくは7%またはそれ以上でありうる。溶解緩衝液中の界面活性剤の濃度は、多くとも約0.0001

10

20

30

40

50

％、0.0005％、0.001％、0.005％、0.01％、0.05％、0.1％、0.5％、1％、2％、3％、4％、5％、6％、もしくは7％またはそれ以上でありうる。いくつかの実施形態では、溶解緩衝液中の界面活性剤の濃度は約1％Liドデシルスルフェートである。本方法で溶解に使用される時間は、使用される界面活性剤の量に依存性しうる。いくつかの実施形態では、界面活性剤を多く使用するほど、溶解に必要な時間は短くなる。溶解緩衝液はキレート化剤（たとえば、EDTA、EGTA）を含みうる。溶解緩衝液中のキレート化剤の濃度は、少なくとも約1、5、10、15、20、25、もしくは30mMまたはそれ以上でありうる。溶解緩衝液中のキレート化剤の濃度は、多くとも約1、5、10、15、20、25、もしくは30mMまたはそれ以上でありうる。いくつかの実施形態では、溶解緩衝液中のキレート化剤の濃度は約10mMである。溶解緩衝液は還元試薬（たとえば、βメルカプトエタノール、DTT）を含みうる。溶解緩衝液中の還元試薬の濃度は少なくとも約1、5、10、15、20mMまたはそれ以上でありうる。溶解緩衝液中の還元試薬の濃度は多くとも約1、5、10、15、20mMまたはそれ以上でありうる。いくつかの実施形態では、溶解緩衝液中の還元試薬の濃度は約5mMである。いくつかの実施形態では、溶解緩衝液は、約0.1MのトリスHCl、約pH7.5、約0.5M LiCl、約1％リチウムドデシルスルフェート、約10mMEDTA、および約5mM DTTを含みうる。

【0198】

溶解は、約4、10、15、20、25、または30℃の温度で行うことが可能である。溶解は、約1、5、10、15、もしくは20分間またはそれ以上行うことが可能である。溶解細胞は、少なくとも約100000、200000、300000、400000、500000、600000、もしくは700000標的核酸分子またはそれ以上を含みうる。溶解細胞は、多くとも約100000、200000、300000、400000、500000、600000、もしくは700000標的核酸分子またはそれ以上を含みうる。

【0199】

標的核酸分子への確率バーコードの結合

細胞の溶解およびそれからの核酸分子の放出の後、核酸分子は、共局在化された固体担体の確率バーコードにランダムに関連付けすることができる。関連付けは、標的核酸分子の相補的部分への確率バーコードの標的認識領域のハイブリダイゼーションを含みうる（たとえば、確率バーコードのオリゴ（dT）は、標的のポリ（A）テールと相互作用可能である）。ハイブリダイゼーションに使用されるアッセイ条件（たとえば、緩衝液pH、イオン強度、温度など）は、特定の安定なハイブリッドの形成を促進するように選択可能である。いくつかの実施形態では、溶解した細胞から放出された核酸分子は、基材上の複数のプローブに関連付けする（たとえば、基板上のプローブとハイブリダイズする）ことができる。プローブが、オリゴ（dT）を含むとき、mRNA分子は、プローブにハイブリダイズして、逆転写されうる。オリゴヌクレオチドのオリゴ（dT）部分は、cDNA分子の第1鎖合成のためのプライマーとして作用しうる。たとえば、図2、ブロック216に示す確率バーコードの非限定的な例において、mRNA分子は、ビーズ上の確率バーコードをハイブリダイズすることができる。たとえば、一本鎖ヌクレオチド断片は、確率バーコードの標的結合領域にハイブリダイズすることができる。

【0200】

結合は、確率バーコードの標的認識領域と標的核酸分子の一部とのライゲーションをさらに含みうる。たとえば、標的結合領域は、制限部位オーバーハング（たとえば、EcoRI付着末端オーバーハング）への特異的ハイブリダイゼーションが可能でありうる核酸配列を含みうる。アッセイ手順は、制限部位オーバーハングを生成するために制限酵素（たとえばEcoRI）で標的核酸を処置する工程をさらに含みうる。次いで、確率バーコードは、制限部位オーバーハングに相補的な配列を含む任意の核酸分子にライゲートしうる。リガーゼ（たとえばT4DNAリガーゼ）は2つの断片を連結するために使用しうる。

【0201】

たとえば、図 2、ブロック 220 に図示する確率バーコードの非限定的な例では、複数の細胞（または複数のサンプル）からの標識標的（たとえば、標的バーコード分子）は、続いて、たとえば、チューブ中にプールすることができる。たとえば、確率バーコードおよび／または標的バーコード分子が結合したビーズを回収することにより、標識標的をプールすることができる。

【0202】

結合した標的バーコード分子の固体担体ベースのコレクションの回収は、磁気ビーズおよび外部印加磁界の使用により実現しうる。標的バーコード分子をプールした後、すべてのさらなる処理を単一反応槽内で進行させることができる。さらなる処理は、たとえば、逆転写反応、増幅反応、切断反応、解離反応、および／または核酸伸長反応を含みうる。さらなる処理反応は、マイクロウェル内で、すなわち、複数の細胞の標識標的核酸分子を最初にプールすることなく、実施することができる。

【0203】

逆転写

本開示は、（たとえば、図 2 のブロック 224 で）逆転写を用いて確率標的バーコードコンジュゲートを生成する方法を提供する。確率標的バーコードコンジュゲートは、確率バーコードと標的核酸の全部または一部の相補的配列と（すなわち、確率バーコード付き cDNA 分子）を含みうる。関連付けられた RNA 分子の逆転写は、逆転写酵素と共に逆転写プライマーを添加することによって起こりうる。逆転写プライマーは、オリゴ（dT）プライマー、ランダムヘキサヌクレオチドプライマー、または標的特異的オリゴヌクレオチドプライマーでありうる。オリゴ（dT）プライマーは、12～18ヌクレオチド長、または概ねそうしたヌクレオチド長であってよく、哺乳動物 mRNA の 3' 末端の内因性ポリ（A）テールに結合することができる。ランダムヘキサヌクレオチドプライマーは、さまざまな相補的部位で mRNA と結合しうる。標的特異的オリゴヌクレオチドプライマーは、典型的には対象の mRNA を選択的にプライミングする。

【0204】

いくつかの実施形態では標識 RNA 分子の逆転写は、逆転写プライマーの添加によって起こりうる。いくつかの実施形態では、逆転写プライマーは、オリゴ（dT）プライマー、ランダムヘキサヌクレオチドプライマー、または標的特異的オリゴヌクレオチドプライマーである。一般に、オリゴ（dT）プライマーは、12～18ヌクレオチド長であり、哺乳動物 mRNA の 3' 末端の内因性ポリ（A）+テールに結合する。ランダムヘキサヌクレオチドプライマーは、さまざまな相補的部位で mRNA と結合しうる。標的特異的オリゴヌクレオチドプライマーは、典型的には対象の mRNA を選択的にプライミングする。

【0205】

逆転写は、繰返し行うことにより複数の標識 cDNA 分子を生成可能である。本明細書に開示される方法は、少なくとも約 1、2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19、または 20 回の逆転写反応を行う工程を含みうる。本方法は、少なくとも約 25、30、35、40、45、50、55、60、65、70、75、80、85、90、95、または 100 回の逆転写反応を行う工程を含みうる。

【0206】

増幅

核酸増幅反応（たとえば、図 2 のブロック 228 で）は、標識標的核酸分子の複数のコピーを生成するために 1 回以上実施することができる。増幅は、複数の標的核酸配列が同時に増幅される、多重方式で実施してよい。増幅反応は、核酸分子にシーケンシングアダプターを付加するために使用することができる。増幅反応は、存在するのであれば、サンプル標識の少なくとも一部を増幅する工程を含みうる。増幅反応は、細胞および／または分子標識の少なくとも一部を増幅する工程を含みうる。増幅反応は、サンプルタグ、細胞標識、空間標識、分子標識、標的核酸、またはそれらの組合せの少なくとも一部を増幅する工程を含みうる。増幅反応は、複数の核酸の 0.5%、1%、2%、3%、4%、5%、

10

20

30

40

50

6 %、7 %、8 %、9 %、10 %、15 %、20 %、25 %、30 %、35 %、40 %、45 %、50 %、55 %、60 %、65 %、70 %、75 %、80 %、85 %、90 %、95 %、97 %、100 %、またはこれらの値のいずれか2つの間の範囲もしくは数を増幅する工程を含みうる。本方法は、サンプル標識、細胞標識、空間標識、および／または分子標識を含む標的バーコード分子のcDNAコピーを1つ以上生成するために、cDNA合成反応を1回以上行う工程をさらに含みうる。

【0207】

いくつかの実施形態では、ポリメラーゼ連鎖反応（PCR）を用いて、増幅を実施することができる。本明細書で用いられる場合、PCRとは、DNAの相補鎖の同時プライマー伸長により特定のDNA配列の*in vitro*増幅を行う反応を意味しうる。本明細書で用いられる場合、PCRは、その反応の派生形、たとえば、限定されるものではないが、RT-PCR、リアルタイムPCR、ネステッドPCR、定量PCR、多重PCR、デジタルPCR、およびアセンブリーPCRを包含しうる。

10

【0208】

標識核酸の増幅は、非PCRベースの方法を含みうる。非PCRベースの方法の例としては、限定されるものではないが、多重置換増幅（MDA）、転写媒介増幅（TMA）、核酸配列ベースの増幅（NASBA）、鎖置換増幅（SDA）、リアルタイムSDA、ローリングサークル増幅、またはサークルーサークル増幅が挙げられる。他の非PCRベースの増幅方法としては、DNAもしくはRNA標的を増幅するためのDNA依存性RNAポリメラーゼ駆動RNA転写増幅またはRNA指向DNA合成および転写の多重サイクル、リガーゼ連鎖反応（LCR）、およびQ β レプリカーゼ（Q β ）法、パリンドロームプローブの使用、鎖置換増幅、制限エンドヌクレアーゼを用いたオリゴヌクレオチド駆動増幅、プライマーが核酸配列にハイブリダイズされかつ得られた二本鎖が伸長反応および増幅の前に切断される増幅方法、5'エキソヌクレアーゼ活性の欠如した核酸ポリメラーゼを用いた鎖置換増幅、ローリングサークル増幅、および分岐伸長増幅（RAM）が挙げられる。いくつかの実施形態では、増幅は、環化転写物を生成しうる。

20

【0209】

いくつかの実施形態では、本明細書に開示する方法は、確率標識アンプリコンを生成するために標識核酸（たとえば、標識RNA、標識DNA、標識cDNA）上でポリメラーゼ連鎖反応を実施する工程をさらに含む。標識アンプリコンは、二本鎖分子であってよい。二本鎖分子は、二本鎖RNA分子、二本鎖DNA分子、またはDNA分子にハイブリダイズされたRNA分子を含みうる。二本鎖分子の一方または両方の鎖は、サンプル標識、空間標識、細胞標識、および／または分子標識を含みうる。確率標識アンプリコンは、一本鎖分子でありうる。一本鎖分子は、DNA、RNA、またはそれらの組合せを含みうる。本開示の核酸は、合成核酸または改変核酸を含みうる。

30

【0210】

増幅は、1つ以上の非天然ヌクレオチドの使用を含みうる。非天然ヌクレオチドは、光不安定性またはトリガー性のヌクレオチドを含みうる。非天然ヌクレオチドの例としては、限定されるものではないが、ペプチド核酸（PNA）、モルホリノ核酸、およびロックド核酸（LNA）、さらにはグリコール核酸（GNA）およびトレオース核酸（TNA）が挙げられうる。非天然ヌクレオチドは、増幅反応の1サイクル以上に添加することができる。非天然ヌクレオチドの添加は、増幅反応の特定のサイクルまたは時点で産物を同定するために使用しうる。

40

【0211】

増幅反応を1回以上行う工程は、1つ以上のプライマーの使用を含みうる。1つ以上のプライマーは、たとえば、少なくとも1、2、3、4、5、6、7、8、9、10、11、12、13、14、もしくは15ヌクレオチドまたはそれ以上を含みうる。1つ以上のプライマーは、少なくとも1、2、3、4、5、6、7、8、9、10、11、12、13、14、もしくは15ヌクレオチドまたはそれ以上を含みうる。1つ以上のプライマーは、12～15ヌクレオチド未満を含みうる。1つ以上のプライマーは、複数の確率標識標

50

的の少なくとも一部にアニールしうる。1つ以上のプライマーは、複数の確率標識標的の3'末端または5'末端にアニールしうる。1つ以上のプライマーは、複数の確率標識標的の内部領域にアニールしうる。内部領域は、複数の確率標識標的の3'末端から少なくとも約50、100、150、200、220、230、240、250、260、270、280、290、300、310、320、330、340、350、360、370、380、390、400、410、420、430、440、450、460、470、480、490、500、510、520、530、540、550、560、570、580、590、600、650、700、750、800、850、900、または1000ヌクレオチドでありうる。1つ以上のプライマーは、プライマーの一定パネルを含みうる。1つ以上のプライマーは、少なくとも1つ以上のカスタムプライマーを含みうる。1つ以上のプライマーは、少なくとも1つ以上の対照プライマーを含みうる。1つ以上のプライマーは、少なくとも1つ以上の遺伝子特異的プライマーを含みうる。

【0212】

1つ以上のプライマーは、ユニバーサルプライマーを含みうる。ユニバーサルプライマーは、ユニバーサルプライマー結合部位にアニールしうる。1つ以上のカスタムプライマーは、第1のサンプル標識、第2のサンプル標識、空間標識、細胞標識、分子標識、標的、またはそれらの任意の組合せにアニールしうる。1つ以上のプライマーは、ユニバーサルプライマーおよびカスタムプライマーを含みうる。カスタムプライマーは、1つ以上の標的を増幅するように設計しうる。標的は、1つ以上のサンプル中の全核酸のサブセットを含みうる。標的は、1つ以上のサンプル中の全確率標識標的のサブセットを含みうる。1つ以上のプライマーは、少なくとも96カスタムプライマーまたはそれ以上を含みうる。1つ以上のプライマーは、少なくとも960カスタムプライマーまたはそれ以上を含みうる。1つ以上のプライマーは、少なくとも9600カスタムプライマーまたはそれ以上を含みうる。1つ以上のカスタムプライマーは、2つ以上の異なる標識核酸にアニールしうる。2つ以上の異なる標識核酸は、1つ以上の遺伝子に相当しうる。

【0213】

任意の増幅スキームを本開示の方法で使用することができる。たとえば、一スキームでは、第1ラウンドのPCRは、遺伝子特異的プライマーおよびユニバーサルIlluminaシーケンシングプライマー1配列に対するプライマーを用いて、ビーズに結合された分子を増幅することができる。第2ラウンドのPCRは、Illuminaシーケンシングプライマー2配列がフランキングするネステッド遺伝子特異的プライマーとユニバーサルIlluminaシーケンシングプライマー1配列に対するプライマーとを用いて第1のPCR産物を増幅可能である。第3ラウンドのPCRは、P5およびP7とサンプルインデックスを付加して、PCR産物をIlluminaシーケンシングライブラリーにする。150bp×2シーケンシングを用いたシーケンシングは、リード1上の細胞標識および分子標識、リード2上の遺伝子、ならびにインデックス1リード上のサンプルインデックスを明らかにしうる。

【0214】

いくつかの実施形態では、核酸は、化学切断を用いて基材から除去可能である。たとえば、核酸中に存在する化学基または修飾塩基は、固体担体からのその除去を促進するために使用可能である。たとえば、酵素は、基材から核酸を除去するために使用可能である。たとえば、核酸は、制限エンドヌクレアーゼ消化による基材からの除去が可能である。たとえば、dUTPまたはddUTPを含有する核酸のウラシル-d-グリコシラーゼ(UDG)処理は、基材から核酸を除去するために使用可能である。たとえば、核酸は、ヌクレオチド切除を行う酵素、たとえば、塩基除去修復酵素、たとえば、脱プリン/脱ピリミジン(AP)エンドヌクレアーゼを用いて基材から除去可能である。いくつかの実施形態では、核酸は、光切断性基と光とを用いて基材から除去可能である。いくつかの実施形態では、切断性リンカーは、基材から核酸を除去するために使用可能である。たとえば、切断性リンカーは、ビオチン/アビジン、ビオチン/ストレプトアビジン、ビオチン/ニュートラビジン、Ig-プロテインA、光不安定性リンカー、酸または塩基不安定性リンカー

基、またはアプタマーの少なくとも1つを含みうる。

【0215】

プローブが遺伝子特異的である場合、分子は、プローブにハイブリダイズし、逆転写および／または増幅が可能である。いくつかの実施形態では、核酸が合成された後（たとえば、逆転写された後）、増幅が可能である。増幅は、複数の標的核酸配列が同時に増幅される条件で、多重方式で行いうる。増幅は、核酸にシーケンシングアダプターを付加しうる。

【0216】

いくつかの実施形態では、増幅は、たとえばブリッジ増幅を用いて基材上に行うことが可能である。基材上でオリゴ（dT）プローブを用いてブリッジ増幅するのに適合していた末端を生成するために、cDNAにホモポリマーテールを付加することが可能である。ブリッジ増幅では、テンプレート核酸の3'末端に相補的なプライマーは、固体粒子に共有結合された各ペアの第1のプライマーでありうる。テンプレート核酸を含有するサンプルが粒子に接触して1回の熱サイクルが行われる場合、テンプレート分子は第1のプライマーにアニールし、かつ第1のプライマーはヌクレオチドの付加により順方向に伸長して、テンプレート分子とテンプレートに相補的な新たに形成されたDNA鎖とからなる二本鎖分子を形成する。次のサイクルの加熱工程では、二本鎖分子は変性されて、粒子からテンプレート分子を放出し、第1のプライマーを介して粒子に結合された相補的DNA鎖を残存させる。続くアニーリング・伸長工程のアニーリング段階では、相補鎖は、第1のプライマーから除去された位置の相補鎖のセグメントに相補的な第2のプライマーにハイブリダイズ可能である。このハイブリダイゼーションにより、相補鎖は、共有結合により第1のプライマーにかつハイブリダイゼーションにより第2のプライマーに固定されたブリッジを第1および第2のプライマー間に形成可能である。伸長段階では、第2のプライマーは、同一の反応混合物中にヌクレオチドを添加することにより反対方向に伸長し、それによりブリッジを二本鎖ブリッジに変換可能である。次いで、次のサイクルが開始され、二本鎖ブリッジは変性されて、それぞれ第1および第2のプライマーを介して粒子表面に結合された一方の末端と、それぞれ未結合の状態の他方の末端と、を有する2つの一本鎖核酸分子を与えることが可能である。この第2のサイクルのアニーリング・伸長工程では、各鎖は同一の粒子上のこれまで未使用であったさらなる相補的プライマーにハイブリダイズして新しい一本鎖ブリッジを形成可能である。この時点でハイブリダイズされる2つのこれまで未使用であったプライマーは伸長して2つの新しいブリッジを二本鎖ブリッジに変換可能である。

【0217】

増幅反応は、複数の核酸の少なくとも1%、2%、3%、4%、5%、6%、7%、8%、9%、10%、15%、20%、25%、30%、35%、40%、45%、50%、55%、60%、65%、70%、75%、80%、85%、90%、95%、97%、または100%を増幅する工程を含みうる。

【0218】

標識核酸の増幅は、PCRベースの方法または非PCRベースの方法を含みうる。標識核酸の増幅は、標識核酸の指数関数的増幅を含みうる。標識核酸の増幅は、標識核酸の線形増幅を含みうる。増幅は、ポリメラーゼ連鎖反応（PCR）により行うことが可能である。PCRは、DNAの相補鎖の同時プライマー伸長により特定のDNA配列の*in vitro*増幅を行う反応を意味しうる。PCRは、その反応の派生形、たとえば、限定されるものではないが、RT-PCR、リアルタイムPCR、ネステッドPCR、定量PCR、多重PCR、デジタルPCR、サブレーションPCR、セミサブレーションPCR、およびアセンブリーPCRを包含しうる。

【0219】

いくつかの実施形態では、標識核酸の増幅は非PCRベースの方法を含む。非PCRベースの方法の例としては、限定されるものではないが、多重置換増幅（MDA）、転写媒介増幅（TMA）、核酸配列ベースの増幅（NASBA）、鎖置換増幅（SDA）、リアルタイムSDA、ローリングサークル増幅、またはサークルーサークル増幅が挙げられる。

他の非PCRベースの増幅方法としては、DNAもしくはRNA標的を増幅するためのDNA依存性RNAポリメラーゼ駆動RNA転写増幅またはRNA指向DNA合成および転写の多重サイクル、リガーゼ連鎖反応（LCR）、Q β レプリカーゼ（Q β ）、パンドロームプローブの使用、鎖置換増幅、制限エンドヌクレアーゼを用いたオリゴヌクレオチド駆動増幅、プライマーが核酸配列にハイブリダイズされかつ得られた二本鎖が伸長反応および増幅の前に切断される増幅方法、5'エキソヌクレアーゼ活性の欠如した核酸ポリメラーゼを用いた鎖置換増幅、ローリングサークル増幅、および／または分岐伸長増幅（RAM）が挙げられる。

【0220】

いくつかの実施形態では、本明細書に開示される方法は、増幅アンプリコン（たとえば標的）上でネステッドポリメラーゼ連鎖反応を行う工程をさらに含む。アンプリコンは二本鎖分子でありうる。二本鎖分子は、二本鎖RNA分子、二本鎖DNA分子、またはDNA分子にハイブリダイズされたRNA分子を含みうる。二本鎖分子の一方または両方の鎖は、サンプルタグまたは分子識別子標識を含みうる。代替的に、アンプリコンは一本鎖分子でありうる。一本鎖分子は、DNA、RNA、またはそれらの組合せを含みうる。本発明の核酸は、合成核酸または改変核酸を含みうる。

10

【0221】

いくつかの実施形態では、本方法は、多数のアンプリコンを生成するために標識核酸を繰返し増幅する工程を含む。本明細書に開示される方法は、少なくとも約1、2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19、または20回の増幅反応を行う工程を含みうる。代替的に、本方法は、少なくとも約25、30、35、40、45、50、55、60、65、70、75、80、85、90、95、または100回の増幅反応を行う工程を含む。

20

【0222】

増幅工程は、複数の核酸を含む1つ以上のサンプルに1つ以上の対照核酸を添加する工程をさらに含みうる。増幅工程は、複数の核酸に1つ以上の対照核酸を添加する工程をさらに含みうる。対照核酸は、対照標識を含みうる。

【0223】

増幅は、1つ以上の非天然ヌクレオチドの使用を含みうる。非天然ヌクレオチドは、光不安定性および／またはトリガー性ヌクレオチドを含みうる。非天然ヌクレオチドの例としては、限定されるものではないが、ペプチド核酸（PNA）、モルホリノ核酸およびロックド核酸（LNA）、さらにはグリコール核酸（GNA）およびトレオース核酸（TNA）が挙げられる。非天然ヌクレオチドは、増幅反応の1サイクル以上に添加しうる。非天然ヌクレオチドの添加は、増幅反応の特定のサイクルまたは時点で産物を同定するために使用しうる。

30

【0224】

増幅反応を1回以上行う工程は、1つ以上のプライマーの使用を含みうる。1つ以上のプライマーは1つ以上のオリゴヌクレオチドを含みうる。1つ以上のオリゴヌクレオチドは少なくとも約7～9ヌクレオチドを含みうる。1つ以上のオリゴヌクレオチドは12～15ヌクレオチド未満を含みうる。1つ以上のプライマーは、複数の標識核酸の少なくとも一部にアニールしうる。1つ以上のプライマーは、複数の標識核酸の3'末端および／または5'末端にアニールしうる。1つ以上のプライマーは、複数の標識核酸の内部領域にアニールしうる。内部領域は、複数の標識核酸の3'末端から少なくとも約50、100、150、200、220、230、240、250、260、270、280、290、300、310、320、330、340、350、360、370、380、390、400、410、420、430、440、450、460、470、480、490、500、510、520、530、540、550、560、570、580、590、600、650、700、750、800、850、900、または1000ヌクレオチドでありうる。1つ以上のプライマーは、プライマーの一定パネルを含みうる。1つ以上のプライマーは、少なくとも1つ以上のカスタムプライマーを含みうる。1つ以上のプライマ

40

50

ーは、少なくとも1つ以上の対照プライマーを含みうる。1つ以上のプライマーは、少なくとも1つ以上のハウスキーピング遺伝子プライマーを含みうる。1つ以上のプライマーは、ユニバーサルプライマーを含みうる。ユニバーサルプライマーは、ユニバーサルプライマー結合部位にアニールしうる。1つ以上のカスタムプライマーは、第1のサンプルタグ、第2のサンプルタグ、分子識別子標識、核酸、またはその産物にアニールしうる。1つ以上のプライマーは、ユニバーサルプライマーおよびカスタムプライマーを含みうる。カスタムプライマー、1つ以上の標的核酸を増幅するように設計しうる。標的核酸は、1つ以上のサンプル中の全核酸のサブセットを含みうる。いくつかの実施形態では、プライマーには、本開示のアレイに結合されたプローブである。

【0225】

いくつかの実施形態では、サンプル中の複数の標的に確率バーコードを付ける工程は、確率バーコード付き断片の指標インデックスライブラリーを作製する工程をさらに含む。異なる確率バーコードの分子標識は、互いに異なってもよい。確率バーコード付き標的の指標インデックスライブラリーを作製する工程は、サンプル中の複数の標的から複数の指標インデックスポリヌクレオチドを作製する工程を含む。たとえば、第1の指標インデックス標的と第2の指標インデックス標的を含む確率バーコード標的の指標インデックスライブラリーの場合、第1の指標インデックスポリヌクレオチドの標識領域は、第2の指標インデックスポリヌクレオチドの標識領域と、1、2、3、4、5、6、7、8、9、10、20、30、40、50ヌクレオチド異なって、概ね、少なくとも、もしくは多くともこうした値、またはこれらの値のいずれか2つの間の数もしくは範囲のヌクレオチド異なってもよい。いくつかの実施形態では、確率バーコード付き標的の指標インデックスライブラリーを作製する工程は、ポリ(T)領域および標識領域などの複数のオリゴヌクレオチドと、複数の標識、たとえば、mRNA分子を接触させる工程と；各々がcDNA領域および標識領域を含む一本鎖標識cDNA分子を生成するために、逆転写酵素を用いて、第1鎖合成を実施する工程と、を含み、ここで、複数の標的は、異なる配列の少なくとも2つのmRNA分子を含み、複数のオリゴヌクレオチドは、異なる配列の少なくとも2つのオリゴヌクレオチドを含む。確率バーコード付き標的の指標インデックスライブラリーを作製する工程は、さらに、二本鎖標識cDNA分子を生成するために、一本鎖標識cDNA分子を増幅する工程と；標識アンプリコンを生成するために、二本鎖標識cDNA分子上でネステッドPCRを実施する工程と、を含む。いくつかの実施形態では、本方法は、アダプター標識アンプリコンを作製する工程を含みうる。

【0226】

確率バーコーディングは、個々の核酸（たとえば、DNAまたはRNA）分子を標識するために、核酸バーコードもしくはタグを使用しうる。いくつかの実施形態では、これは、DNAバーコードもしくはタグがmRNAから生成される際に、cDNA分子にこれらを付加する工程を含む。ネステッドPCRは、PCR増幅バイアスの最小限化を実施することができる。アダプターは、たとえば、次世代シーケンシング（NGS）を用いるシーケンシングのために付加することができる。シーケンシング結果を用いて、たとえば、図2のブロック232に位置する標的の1つ以上のコピーの細胞標識、分子標識、およびヌクレオチド断片の配列を決定することができる。

【0227】

図3は、確率バーコード付き標的、たとえば、mRNAの指標インデックスライブラリーを作製する非限定的な例示のプロセスを示す概略図である。ステップ1に示すように、逆転写プロセスは、ユニーク分子標識、細胞標識、およびユニバーサルPCR部位を含む各mRNA分子をコードすることができる。特に、分子識別子標識310のセットとRNA分子302のポリ(A)テール領域308の確率論的ハイブリダイゼーションによって、RNA分子302を逆転写して、cDNA領域306を含む標識cDNA分子304を生成することができる。分子識別子標識310の各々は、標的結合領域、たとえば、ポリ(dT)領域312、標識領域314、およびユニバーサルPCR領域316を含みうる。

【0228】

いくつかの実施形態では、細胞標識は、3～20ヌクレオチドを含みうる。いくつかの実施形態では、分子標識は、3～20ヌクレオチドを含みうる。いくつかの実施形態では、複数の確率バーコードの各々は、1つ以上のユニバーサル標識および細胞標識をさらに含み、ユニバーサル標識は、固体担体上の複数の確率バーコードについて同じであり、細胞標識は、固体担体上の複数の確率バーコードについて同じである。いくつかの実施形態では、ユニバーサル標識は、3～20ヌクレオチドを含みうる。いくつかの実施形態では、細胞標識は、3～20ヌクレオチドを含む。

【0229】

いくつかの実施形態では、標識領域314は、分子標識318および細胞標識320を含みうる。いくつかの実施形態では、標識領域314は、1つ以上のユニバーサル標識、次元標識、および細胞標識を含みうる。分子標識318は、1、2、3、4、5、6、7、8、9、10、20、30、40、50、60、70、80、90、100ヌクレオチド長であっても、概ね、少なくとも、もしくは多くともそうしたヌクレオチド長であってもよいし、またはこれらの値のいずれかの間の数もしくは範囲のヌクレオチド長であってもよい。細胞標識320は、1、2、3、4、5、6、7、8、9、10、20、30、40、50、60、70、80、90、100ヌクレオチド長であっても、概ね、少なくとも、もしくは多くともそうしたヌクレオチド長であってもよいし、またはこれらの値のいずれかの間の数もしくは範囲のヌクレオチド長であってもよい。ユニバーサル標識は、1、2、3、4、5、6、7、8、9、10、20、30、40、50、60、70、80、90、100ヌクレオチド長であっても、概ね、少なくとも、もしくは多くともそうしたヌクレオチド長であってもよいし、またはこれらの値のいずれかの間の数もしくは範囲のヌクレオチド長であってもよい。ユニバーサル標識は、固体担体上の複数の確率バーコードについて同じであってもよく、細胞標識は、固体担体上の複数の確率バーコードについて同じであってもよい。次元標識は、1、2、3、4、5、6、7、8、9、10、20、30、40、50、60、70、80、90、100ヌクレオチド長であっても、概ね、少なくとも、もしくは多くともそうしたヌクレオチド長であってもよいし、またはこれらの値のいずれかの間の数もしくは範囲のヌクレオチド長であってもよい。

【0230】

いくつかの実施形態では、標識領域314は、1、2、3、4、5、6、7、8、9、10、20、30、40、50、60、70、80、90、100、200、300、400、500、600、700、800、900、1000の異なる標識を含むか、概ねそうした値の異なる標識を含むか、少なくとも、もしくは多くともそうした値の異なる標識、またはこれらの値のいずれかの間の数もしくは範囲の異なる標識を含みうる。各標識は、1、2、3、4、5、6、7、8、9、10、20、30、40、50、60、70、80、90、100ヌクレオチド長であっても、概ね、少なくとも、もしくは多くともそうしたヌクレオチド長であってもよいし、またはこれらの値のいずれかの間の数もしくは範囲のヌクレオチド長であってもよい。分子識別子標識310のセットは、10、20、40、50、70、80、90、10²、10³、10⁴、10⁵、10⁶、10⁷、10⁸、10⁹、10¹⁰、10¹¹、10¹²、10¹³、10¹⁴、10¹⁵、10²⁰の分子識別子標識310を含むか、概ねそうした値の識別子標識310を含むか、少なくとも、もしくは多くともそうした値の分子識別子標識310、またはこれらの値のいずれかの間の数もしくは範囲の分子識別子標識310を含みうる。また、分子識別子標識310のセットは、たとえば、各々、ユニーク標識領域314を含みうる。余剰の分子識別子標識310を除去するために、標識cDNA分子304を精製することができる。精製は、Ampureビーズ精製を含みうる。

【0231】

工程2に示すように、工程1の逆転写プロセスからの産物を1チューブ中にプールし、第1PCRプライマープールおよび第1ユニバーサルPCRプライマーを用いてPCR増幅することができる。プールする工程は、ユニーク標識領域314によって可能である。特に、ネステッドPCR標識アンプリコン322を生成するために、標識cDNA分子30

10

20

30

40

50

4を増幅することができる。増幅は、多重PCR増幅を含みうる。増幅は、単一反応量で96多重プライマーを用いる多重PCR増幅を含みうる。いくつかの実施形態では、多重PCR増幅は単一反応量で10、20、40、50、70、80、90、102、103、104、105、106、107、108、109、1010、1011、1012、1013、1014、1015、1020の多重プライマーを使用するか、概ねそうした値の多重プライマー、少なくとも、もしくは多くともそうした値の多重プライマーを使用するか、またはこれらの値のいずれかの間の数もしくは範囲の多重プライマーを使用することができる。増幅は、特定の遺伝子を標的とするカスタムプライマー326A~Cの第1PCRプライマープール324と、ユニバーサルプライマー328とを含みうる。カスタムプライマー326は、標識cDNA分子304のcDNA部分306'内の1領域とハイブリダイズすることができる。ユニバーサルプライマー328は、標識cDNA分子304のユニバーサルPCR領域316とハイブリダイズすることができる。

【0232】

図3の工程3に示すように、工程2のPCR増幅からの産物は、ネステッドPCRプライマープールおよび第2ユニバーサルPCRプライマーを用いて増幅することができる。ネステッドPCRは、PCR増幅バイアスを最小限に抑えることができる。特に、ネステッドPCR標識アンプリコン322は、ネステッドPCRによりさらに増幅することもできる。ネステッドPCRは、単一反応量でネステッドPCRプライマー332a~cのネステッドPCRプライマープール330と、第2ユニバーサルPCRプライマー328'とを含む多重PCRを含みうる。ネステッドPCRプライマープール328は、1、2、3、4、5、6、7、8、9、10、20、30、40、50、60、70、80、90、100、200、300、400、500、600、700、800、900、1000の異なるネステッドPCRプライマー330を含むか、概ねそうした値の異なるネステッドPCRプライマー330を含むか、少なくとも、もしくは多くともそうした値の異なるネステッドPCRプライマー330、またはこれらの値のいずれかの間の数もしくは範囲の異なるネステッドPCRプライマー330を含みうる。ネステッドPCRプライマー332は、アダプター334を含有して、標識アンプリコン322のcDNA部分306'内の1領域とハイブリダイズすることができる。ユニバーサルプライマー328'は、アダプター336を含有して、標識アンプリコン322のユニバーサルPCR領域316とハイブリダイズすることができる。このようにして、工程3は、アダプター標識アンプリコン338を生成する。いくつかの実施形態では、ネステッドPCRプライマー332と第2ユニバーサルPCRプライマー328'は、アダプター334および336を含有しなくてもよい。それに代わり、アダプター334および336は、アダプター標識アンプリコン338を生成するために、ネステッドPCRの産物とライゲートすることができる。

【0233】

工程4に示すように、工程3からのPCR産物は、ライブラリー増幅プライマーを用いたシーケンシングのためにPCR増幅することができる。特に、アダプター334および336を用いて、アダプター標識アンプリコン338に対するアッセイをさらに1回以上実施することができる。アダプター334および336は、プライマー340および342とハイブリダイズすることができる。1つ以上のプライマー340および342は、PCR増幅プライマーであってよい。1つ以上のプライマー340および342は、シーケンシングプライマーであってよい。1つ以上のアダプター334および336は、アダプター標識アンプリコン338のさらなる増幅のために使用することができる。1つ以上のアダプター334および336は、アダプター標識アンプリコン338のシーケンシングのために使用することができる。プライマー342は、プレート指標インデックス344を含有することができ、これによって、分子識別子標識318の同じセットを用いて生成されたアンプリコンを、次世代シーケンシング(NGS)を用いた1回のシーケンシング反応でシーケンシングすることができる。

【0234】

PCRおよびシーケンシングエラーの訂正

本明細書には、標的の数を決定するための方法が開示される。いくつかの実施形態では、本方法は、(a) 複数の確率バーコードを用いて、複数の標的に確率バーコードを付けて、複数の確率バーコード付き標的を生成する工程であって、複数の確率バーコードの各々が分子標識を含む工程と；(b) 確率バーコード付き標的のシーケンシングデータを取得する工程と；(c) 複数の標的の1つ以上について：(i) シーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数をカウントする工程と；(ii) (b) で得られたシーケンシングデータ中の標的のクオリティステータスを決定する工程と；(iii) (b) で得られたシーケンシングデータ中の1つ以上のシーケンシングデータエラーを決定する工程であって、シーケンシングデータ中の1つ以上のシーケンシングデータエラーを決定する工程が、以下：シーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数、シーケンシングデータ中の標的のクオリティステータス、および複数の確率バーコードに識別可能な配列を有する分子標識の数のうち1つ以上を決定することを含む工程と；(iv) 標的の数を推定する工程であって、推定された標的の数が、(iii) で決定された1つ以上のシーケンシングデータエラーに応じて調節された、(i) でカウントされたシーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数と相関する工程と、を含む。工程(i)、(ii)、(iii)、および(iv)は、複数の標的の各々について実施することができる。本方法は、多重化することができる。

【0235】

いくつかの実施形態では、本方法は、1つ以上のシーケンシングデータエラーを決定する前に、シーケンシングデータを折りたたむ工程をさらに含む。シーケンシングデータを折りたたむ工程は、類似分子標識を有し、かつ所定の折りたたみ発生数閾値よりも少ない発生数を有する標的のコピーを、複数の標的について同じ分子標識を有するものとして帰属させる工程を含み、ここで、標的の2つのコピーは、標的の2つのコピーの分子標識が、配列において少なくとも1塩基相違する場合、類似の分子標識を有する。

【0236】

1つ以上のシーケンシングデータエラーに応じてシーケンシングデータを調節した後に保持されるシーケンシングデータ中の分子標識のパーセンテージは、変動しうる。いくつかの実施形態では、1つ以上のシーケンシングデータエラーに応じてシーケンシングデータを調節した後に保持されるシーケンシングデータ中の分子標識のパーセンテージは、50%、60%、70%、80%、90%、95%、99%、もしくは99.9%、または概ねそうしたパーセンテージであるか、あるいはこれらの値のいずれか2つの間の数もしくは範囲でありうる。いくつかの実施形態では、1つ以上のシーケンシングデータエラーに応じてシーケンシングデータを調節した後に保持されるシーケンシングデータ中の分子標識のパーセンテージは、少なくとも、または多くとも、50%、60%、70%、80%、90%、95%、99%、もしくは99.9%でありうる。

【0237】

分子標識カウントの決定

図5は、分子標識を用いてPCRおよびシーケンシングエラーを訂正する非限定的な例示的实施形態500を示すフローチャートである。実施形態500は、複数の確率バーコードを用いて、複数の標的に確率バーコード（複数の確率バーコードの各々は、分子標識を含む）を付けて、複数の確率バーコード付き標的を生成する工程の後、ならびに、確率バーコード付き標的のシーケンシングデータを取得する工程の後、開始ブロック504から開始する。

【0238】

標的、たとえば、マイクロウェルアレイのマイクロウェル内の細胞に由来する遺伝子の場合、シーケンシングデータ中の標的に関連付けられた識別可能な配列を含む分子標識の数をブロック508でカウントすることができる。シーケンシングデータ中で、標的の2つのコピーは、類似の分子標識を有してもよく、たとえば、標的の2つのコピーの分子標識は、配列の1塩基が異なりうる。標的の2つのコピーは、いずれも真であってもよく、標

的の一方のコピーが真で、標的の他方のコピーは、シーケンシングエラーもしくはPCRエラーの結果であってもよいし、または標的の両方のコピーが、シーケンシングエラーもしくはPCRエラーの結果であってもよい。

【0239】

シーケンシングデータの折りたたみ

ブロック512で、シーケンシングデータを折りたたむことができる。シーケンシングデータを折りたたむ工程は、類似分子標識を有し、かつ所定の折りたたみ発生数閾値よりも少ない発生数を有する標的のコピーを、複数の標的について同じ分子標識を有するものとして帰属させる工程を含みうる。所定の折りたたみ発生数閾値は、1～100の範囲で変動しうる。いくつかの実施形態では、所定の折りたたみ発生数閾値は、確率バーコードが、識別可能な配列を有する約6561の分子標識を含む場合、1、2、3、4、5、6、7、8、9、10、17、20、30、40、50、60、70、80、90、100、または概ねそうした値であるか、あるいはこれらの値のいずれか2つの間の数もしくは範囲でありうる。いくつかの実施形態では、所定の折りたたみ発生数閾値は、確率バーコードが、識別可能な配列を有する約6561の分子標識を含む場合、少なくとも、または多くとも、1、2、3、4、5、6、7、8、9、10、17、20、30、40、50、60、70、80、90、もしくは100でありうる。たとえば、分子標識は、8ヌクレオチド長であってよく、各ヌクレオチド位置は、アデニン（A）、シトシン（C）、グアニン（G）；C、G、チミン（T）；A、G、T；またはA、C、Tなどの3つの可能性を有しうるため、 $3^8 = 6561$ のユニーク分子標識を生成しうる。

【0240】

いくつかの実施形態では、所定の折りたたみ発生数閾値は、確率バーコードが、識別可能な配列を有する約65536の分子標識を含む場合、1、2、3、4、5、6、7、8、9、10、20、30、40、50、60、70、80、90、100、または概ねそうした値であるか、あるいはこれらの値のいずれか2つの間の数もしくは範囲でありうる。いくつかの実施形態では、所定の折りたたみ発生数閾値は、確率バーコードが、識別可能な配列を有する約65536の分子標識を含む場合、少なくとも、または多くとも、1、2、3、4、5、6、7、8、9、10、20、30、40、50、60、70、80、90、もしくは100でありうる。たとえば、分子標識は、8ヌクレオチド長であってよく、各ヌクレオチド位置は、4つの可能性：A、C、G、Tを有しうるため、 $3^4 = 65536$ のユニーク分子標識を生成しうる。

【0241】

たとえば、標的の5つのコピーが存在しうる。標的の5つのコピーは、

【化1】

TGTGCGTG, TGTGCGCG, TGTGCGGG, GGTGCGTG, および TGGGCGTG

の分子標識を有するものであってよく、分子標識当たりのリードの数は、それぞれ、261、2、2、1、および1である。分子標識

【化2】

TGTGCGCG, TGTGCGGG, GGTGCGTG, および TGGGCGTG

は、それらが、分子標識TGTGCGTGと1ヌクレオチド（下線部）異なっているため、分子標識TGTGCGTGと類似している。識別可能な配列を有する6561の分子標識があり、かつ所定の折りたたみ発生数閾値が7である場合、分子標識

【化3】

TGTGCGCG, TGTGCGGG, GGTGCGTG, および TGGGCGTG

10

20

30

40

50

の発生数は、分子標識 T G T G C G T G に帰属させることができる。

【0242】

別の例として、標的の7つのコピーが存在しうる。標的の7つのコピーは、

【化4】

CGTGTCTG, GGGGGCGA, GCTGCTGG,

TCGGGCGA, CGCGTTCA, CGCGTTA, および TGGGCTTG

の分子標識を有するものであってよく、分子標識当たりのリードの数は、それぞれ、10、7、5、4、1、1、および1である。分子標識

【化5】

CGCGTTA

は、分子標識 C G C G T T C A と、互いに1ヌクレオチド（下線部）異なっているため、類似している。識別可能な配列を有する6561の分子標識があり、かつ所定の折りたたみ発生数閾値が7である場合、分子標識

【化6】

CGCGTTA

の発生数は、分子標識 C G C G T T C A に帰属させることができる。

【0243】

シーケンシングデータエラー

本明細書に開示する方法は、シーケンシングデータエラー、たとえば、1つ以上の標的核酸をカウントする方法に発生するエラーを同定および／または訂正するために使用することができる。いくつかの実施形態では、シーケンシングデータエラーは、PCR導入エラー、シーケンシング導入エラー、バーコード混入に起因するエラー、ライブラリー作製エラー、またはそれらの任意の組合せを含むか、これらでありうる。PCR導入エラーは、PCR増幅エラー、PCR増幅バイアス、不十分なPCR増幅、またはそれらの任意の組合せの結果を含むか、これらでありうる。シーケンシング導入エラーは、不正確なベースコーリング、不十分なシーケンシング、またはそれらの任意の組合せの結果を含むか、これらでありうる。エラーは、1つ以上のヌクレオチドの欠失、1つ以上のヌクレオチドの置換、1つ以上のヌクレオチドの付加、またはそれらの任意の組合せを含むか、これらでありうる。

【0244】

シーケンシングステータスの決定

前述したように、複数の確率バーコードを用いて、複数の標的に確率バーコードを付けることにより、複数の確率バーコード付き標的を生成することができ、複数の確率バーコードの各々は、分子標識、ならびに確率バーコード付き標的のシーケンシングデータの取得を含みうる。標識、たとえば、マイクロウェルアレイのマイクロウェル内の1細胞に由来する遺伝子の場合、シーケンシングデータ中の標的と関連付けられた識別可能な配列を有する分子標識の数をカウントすることができる。カウントされたシーケンシングデータは、たとえば、類似した分子標識を有し、かつ所定の折りたたみ発生数閾値よりも少ない発生数を有する標的のコピーを、複数の標的について同じ分子標識を有するものとして帰属させる工程によって、折りたたむことができる。シーケンシングデータを折りたたんだ後、標的のクオリティステータスを決定することができる。

【0245】

図5を参照にして、いくつかの実施形態では、ブロック516、シーケンシングデータ中の標的のクオリティステータスは、完全シーケンシング、不完全シーケンシング、または

10

20

30

40

50

飽和シーケンシングであると決定することができる。標的のクオリティステータスは、真の分子標識またはリアル分子標識のすべてがシーケンシングランの深度に観察されたか否かに依存しうる。真の分子標識またはリアル分子標識は、エラーまたは偽の分子標識ではない分子標識を意味しうる。エラーまたは偽の分子標識は、PCRエラー、人工物、またはシーケンシングエラーから生じた配列を有する分子標識を意味しうる。シーケンシングデータ中の標的のクオリティステータスは、複数の確率バーコード中の識別可能な配列を有する分子標識の数と、カウントされたシーケンシングデータ中の標的と関連付けられた識別可能な配列を有する分子標識の数によって決定することができる。

【0246】

いくつかの実施形態では、完全シーケンシングクオリティステータスは、所定の完全シーケンシング散布閾値以上のポアソン分布と比較した散布指数によって決定することができる。散布指数は、標的の分散／平均として定義することができる。図6は、完全シーケンシングと不完全シーケンシングにより得られたシーケンシングデータを示す概略図である。図6は、ライブラリー（左側サークル）中の遺伝子Aの3つのコピーと、遺伝子Bの6つのコピーとを示す。遺伝子Aの3つのコピーが、シーケンシングデータ（右上のサークル）中に6回、5回、および1回のシーケンシングリードを有した場合、分散は7、平均は4、散布指数は1.75である。遺伝子Bの6つのコピーが、シーケンシングデータ（右上のサークル）中に9回、2回、2回、2回、1回、および1回のシーケンシングリードを有した場合、分散は9.36、平均は2.83、散布指数は3.31である。これらのシーケンシングデータを用いて、所定の完全シーケンシング散布閾値が、たとえば、完全シーケンシングについて0.9である場合、遺伝子Aおよび遺伝子Bは、完全シーケンシングステータスを有するとみなすことができる。

【0247】

遺伝子Aの1つのコピーが観察されず、遺伝子Aの他の2つのコピーがシーケンシングデータ（右下のサークル）中に2回および3回のシーケンシングリードを有した場合、分散は0.5、平均は2.5、散布指数は0.2である。遺伝子Bの2つのコピーが観察されず、遺伝子Bの他の4つのコピーがシーケンシングデータ（右下のサークル）中に4回、2回、1回、および1回のシーケンシングリードを有した場合、分散は2、平均は2、散布指数は2である。これらのシーケンシングデータを用いて、所定の完全シーケンシング散布閾値が、たとえば、完全シーケンシングについて1.1である場合、遺伝子Aおよび遺伝子Bは、不完全シーケンシングステータスを有するとみなすことができる。

【0248】

所定の完全シーケンシング散布閾値は、0.5～5の範囲で変動しうる。いくつかの実施形態では、所定の完全シーケンシング散布閾値は、0.5、0.6、0.7、0.8、0.9、1、2、3、4、5、6、または概ねそうした値であるか、あるいはこれらの値のいずれか2つの間の数もしくは範囲でありうる。いくつかの実施形態では、所定の完全シーケンシング散布閾値は、少なくとも、または多くとも0.5、0.6、0.7、0.8、0.9、1、2、3、4、5、もしくは6でありうる。

【0249】

いくつかの実施形態では、完全シーケンシングクオリティステータスは、さらに、シーケンシングデータ中の所定の完全シーケンシング発生数閾値以上の発生数を有する分子標識によっても決定することができる。所定の完全シーケンシング発生数閾値は、8～20の範囲で変動しうる。いくつかの実施形態では、完全シーケンシング発生数閾値は、8、9、10、11、12、13、14、15、16、17、18、19、20、または概ねそうした値であるか、あるいはこれらの値のいずれか2つの間の数もしくは範囲でありうる。いくつかの実施形態では、完全シーケンシング発生数閾値は、少なくとも、または多くとも、8、9、10、11、12、13、14、15、16、17、18、19、もしくは20でありうる。

【0250】

いくつかの実施形態では、飽和シーケンシングクオリティステータスは、所定の飽和閾値

よりも大きい、識別可能な配列を含む分子標識の数を有する標的によって、決定することができる。飽和シーケンシングクオリティステータスは、さらに、所定の飽和閾値よりも大きい、識別可能な配列を含む分子標識の数を有する複数の標的のうちの他の1つの標的によって、決定することもできる。

【0251】

所定の飽和閾値は、変動しうる。いくつかの実施形態では、所定の飽和閾値は、確率バーコードが、識別可能な配列を有する約6561の分子標識を含む場合、6000、6100、6200、6300、6400、6500、6557、6558、6559、6560、6561、または概ねそうした値であるか、あるいはこれらの値のいずれか2つの間の数もしくは範囲でありうる。いくつかの実施形態では、所定の飽和閾値は、確率バーコードが、識別可能な配列を有する約6561の分子標識を含む場合、少なくとも、または多くとも、6000、6100、6200、6300、6400、6500、6557、6558、6559、6560、もしくは6561でありうる。いくつかの実施形態では、所定の飽和閾値は、確率バーコードが、識別可能な配列を有する約65536の分子標識を含む場合、64000、64100、64200、64300、64400、64500、64600、64700、64800、64900、65000、65100、65200、65300、65400、65500、65510、65520、65530、65532、65533、65534、65535、または概ねそうした値であるか、あるいはこれらの値のいずれか2つの間の数もしくは範囲でありうる。いくつかの実施形態では、所定の飽和閾値は、確率バーコードが、識別可能な配列を有する約65536の分子標識を含む場合、少なくとも、または多くとも、64000、64100、64200、64300、64400、64500、64600、64700、64800、64900、65000、65100、65200、65300、65400、65500、65510、65520、65530、65532、65533、65534、もしくは65535でありうる。

【0252】

いくつかの実施形態では、シーケンシングデータ中の標的のクオリティステータスは、シーケンシングデータ中の標的のクオリティステータスが、完全シーケンシングではなく、かつ飽和シーケンシングではない場合に、不完全シーケンシングとして分類することができる。

【0253】

完全シーケンシングクオリティステータス

本明細書に開示する方法は、標的が、完全シーケンシングクオリティステータスを有する場合、シーケンシングライブラリー中の標的の数の推定値を提供することができる。シーケンシングライブラリー中の標的が、完全シーケンシングクオリティステータスを有する場合、真の確率バーコードおよびエラー確率バーコードのシーケンシングリードについて個別のポアソンモデルを介して閾値を確立することができる。標的のクオリティステータスは、真の分子標識またはリアル分子標識のすべてがシーケンシングランの深度で観察されたか否かに依存しうる。真の分子標識またはリアル分子標識は、エラーまたは偽の分子標識ではない分子標識を意味しうる。エラーまたは偽の分子標識は、PCRエラー、人工物、またはシーケンシングエラーから生じた配列を有する分子標識を意味しうる。

【0254】

図5を参照にして、決定状態520で、標的分子が、完全シーケンシングステータスを有する場合、実施形態500は、ブロック524に進む。ブロック524では、1塩基のシーケンシングエラーを次の工程により除去することができる。工程(1)、シーケンシングリードが25より大きい場合、最も豊富なシーケンシングリードに関連付けられた分子標識を第1の親分子標識として選択する。たとえば、シーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数をカウントした後、最も高いシーケンシングリードを有するシーケンシングデータ中の標的に関連付けられた分子標識を選択する。

【0255】

工程（２）、子供分子標識：シーケンシングリード ≤ 3 を有し、第１の親分子標識から１塩基隔てた分子標識を同定し；子供分子標識または１塩基子供分子標識が見出されない場合、工程（５）に進む。工程（３）、すべての子供分子標識および親分子標識に対して複数の二項検定を実施し、その帰無仮説が容認された子供分子標識を除去してから、それらのシーケンシングリードをそれらの親に帰属させる。帰無仮説のいずれも容認されなければ、これは、すべての子供分子標識が、親分子標識の１塩基シーケンシングエラーではないことを意味し、その場合、リード訂正を実施する必要はない。工程（４）、分子標識配列ならびにシーケンシングリードを更新する。たとえば、複数の二項検定の帰無仮説が容認されれば、子供分子標識の発生数を親分子標識に帰属させることができる。工程（５）、次に大きいシーケンシングリードを有する分子標識を親分子標識として選択し、適格の親分子標識または適格の子供分子標識がなくなるまで前述の工程を反復する。

10

【0256】

いくつかの実施形態では、カウントされたシーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数は、標的が、完全シーケンシングクオリティステータスを有していれば、１つ以上の親分子標識についてすべての子供分子標識を決定する工程と；少なくとも１つの子供分子標識および親分子標識について複数の二項検定などの統計解析を実施する工程と；統計解析の帰無仮説が容認されれば、子供分子標識の発生数を親分子標識に帰属させる工程と、によって調節することができる。

20

【0257】

いくつかの実施形態では、子供分子標識は、親分子標識と１塩基相違し、かつ、所定の完全シーケンシング子供閾値以下の発生数を有する分子標識を含みうる。所定の完全シーケンシング子供閾値は、変動しうる。いくつかの実施形態では、所定の完全シーケンシング子供閾値は、１、２、３、４、５、６、７、８、９、１０、または概ねそうした値であるか、あるいはこれらの値のいずれか２つの間の数もしくは範囲でありうる。いくつかの実施形態では、所定の完全シーケンシング子供閾値は、少なくとも、または多くとも、１、２、３、４、５、６、７、８、９、もしくは１０でありうる。

【0258】

いくつかの実施形態では、１つ以上の親分子標識は、所定の完全シーケンシング親閾値以上の発生数を有する分子標識を含み、ここで、所定の完全シーケンシング親閾値は、所定の完全シーケンシング発生数閾値、たとえば、８と等しい。第１の統計解析の帰無仮説は、帰無仮説が真である確率が、偽発見率を下回れば、容認されうる。偽発見率は、変動しうる。いくつかの実施形態では、偽発見率は、１％、２％、３％、４％、５％、６％、７％、８％、９％、１０％、１１％、１２％、１３％、１４％、１５％、１６％、１７％、１８％、１９％、２０％、または概ねそうした値であるか、あるいはこれらの値のいずれか２つの間の数もしくは範囲でありうる。いくつかの実施形態では、偽発見率は、少なくとも、または多くとも、１％、２％、３％、４％、５％、６％、７％、８％、９％、１０％、１１％、１２％、１３％、１４％、１５％、１６％、１７％、１８％、１９％、もしくは２０％でありうる。第１の統計解析は、複数の二項検定であってよい。

30

【0259】

ブロック５２８では、ポアソンモデルを標的の分子標識の閾値化に用いて、シーケンシングデータ中の標的に関連付けられた真の分子標識および偽の分子標識を決定することができる。たとえば、人工物から「真である可能性がある」分子標識を識別するために、ポアソンモデルをシーケンシングリードに適用することができる。

40

【0260】

いくつかの実施形態では、カウントされたシーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数は、標的が完全シーケンシングクオリティステータスを有していれば、標的の分子標識を閾値化して、シーケンシングデータ中の標的に関連付けられた真の分子標識および偽の分子標識を決定する工程によって調節されうる。標的の分子標識を閾値化する工程は、標的の分子標識について統計解析を実施する工程を含

50

みうる。

【0261】

いくつかの実施形態では、統計解析を実施する工程は、以下：標的の分子標識の分布およびそれらの発生数を2つのポアソン分布に当てはめる工程と；2つのポアソン分布を用いて、真の分子標識の数 n を決定する工程と；シーケンシングデータから偽の分子標識を除去する工程と、を含み、ここで、偽の分子標識は、 n 番目に豊富な分子標識の発生数よりも低い発生数を有する分子標識を含み、また、真の分子標識は、 n 番目に豊富な分子標識の発生数以上の発生数を有する分子標識を含む。2つのポアソン分布は、真の分子標識に対応する第1ポアソン分布と、偽の分子標識に対応する第2ポアソン分布とを含みうる。

【0262】

ブロック532では、複数の二項検定または2つのポアソン分布を用いて、シーケンシングデータを訂正または調節した後、標的の数を推定して、出力を生成することができる。実施形態500は、終点ブロック536で終了する。

【0263】

飽和シーケンシングクオリティステータス

本明細書に開示する方法は、分子標識カウントを推定する際の大きな不確実性のために、標的が飽和シーケンシングクオリティステータスを有する場合、シーケンシングライブラリ中の標的の数の推定値を提供することができないこともある。図5を参照にして、いくつかの実施形態では、決定状態520で、シーケンシングステータスが完全シーケンシングステータスではない場合、実施形態500は、決定状態540に進む。決定状態540で、標的が、飽和シーケンシングステータスを有する場合、実施形態500は、終点ブロック536に進む。飽和シーケンシングステータスの場合、分子標識カウントを推定する際の大きな不確実性のために、標的の数が決定されないことがある。

【0264】

不完全シーケンシングクオリティステータス

本明細書に開示する方法は、標的が不完全シーケンシングクオリティステータスを有する場合、シーケンシングライブラリ中の標的の数の推定値を提供することができる。シーケンシングライブラリ中の標的は、不完全シーケンシングクオリティステータスを有するとき、ノイズ標的、たとえば、ノイズ遺伝子は除去することができる。標的は、その増幅速度（分子標識当たりの平均リード）が、標的を含む同じライブラリ中の完全にシーケンシングされた遺伝子から由来するエラーの増幅速度と類似していれば、ノイズでありうる。ライブラリ中に存在する識別可能な分子標識を有する標的を含む確率バーコードの数の推定値を補外するために、不完全シーケンシングのクオリティステータスを有する標的のシーケンシングリードに対して、ゼロ切断ポアソンモデルを適用することができる。

【0265】

実施形態500は、出発標的を標識するために用いられる真の確率バーコードのいくつか、不適切なシーケンシング深度のために観察されなかった場合、シーケンシングライブラリ中の標的の数の推定値を提供することができる。決定状態540で、標的が、飽和シーケンシングステータスを有していなければ、標的は、不完全シーケンシングステータスを有し、実施形態500は、ブロック544に進んで、ノイズ標的、たとえば、ノイズ遺伝子を除去する。

【0266】

標的の散布指数が、 > 4 であり、かつ、その標的の最大シーケンシングリードが、 > 18 である場合、ポアソンモデル化を用いて、真のバーコードとエラーバーコードを区別するための閾値を取得しても、やはり相応しい推定値を提供することができる。シーケンシングデータが、軽度の過剰散布、たとえば、 $1.5 < \text{散布指数} \leq 4$ を示し、かつ、その標的の最大シーケンシングリードが ≤ 18 である場合には、ポアソンモデルを用いて、閾値を得ると、真の分子標識カウントを過少評価する恐れがある。過少評価の理由は、低リードを有する分子標識が、恐らく真の分子標識と偽の分子標識との混合でありうるためであり

10

20

30

40

50

うる。その結果、低シーケンシングリードを有するこれらの真の分子標識は、エラーのポアソンモデルに入ること余儀なくされ、真の分子標識のポアソンモデルが、本来あるべきよりも少ない分子標識を有しうる。例えば、1などの低い分子標識カウントが取り除かれた後の分子標識カウントを使うその場限りの方法を使用することができる。散布指数が1に近い、たとえば、0.9～1.5である場合、観察された分子標識カウントが、相応しい推定値で生成されうる。散布指数が、0.1～0.9であれば、過少散布ポアソンモデルを特徴とするゼロ切断ポアソンモデルが、相応しい推定値を生成しうるが；シーケンシングデータ中にエラーが存在する場合には、このモデルは、過大評価する傾向がありうる。

【0267】

いくつかの実施形態では、シーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数は、シーケンシングデータ中の標的のクオリティステータスが、不完全シーケンシングクオリティステータスである場合、標的が、シーケンシングデータにおいてノイズであるか否かを決定する工程と；シーケンシングデータからノイズ標的を除去する工程と、によって調節することができる。ノイズ標的の分子標識の発生数が、不完全シーケンシングノイズ標的閾値以下であれば、標的はノイズでありうる。不完全シーケンシングノイズ遺伝子閾値は、変動しうる。いくつかの実施形態では、不完全シーケンシングノイズ標的閾値は、完全シーケンシングのクオリティステータスを有する複数の標的の分子標識の中央または平均発生数と等しくてもよい。いくつかの実施形態では、不完全シーケンシングノイズ遺伝子閾値は、1、2、3、4、5、6、7、8、9、10、または概ねそうした値であるか、あるいはこれらの値のいずれか2つの間の数もしくは範囲でありうる。いくつかの実施形態では、不完全シーケンシングノイズ遺伝子閾値は、少なくとも、または多くとも、1、2、3、4、5、6、7、8、9、もしくは10でありうる。

【0268】

ブロック548では、ライブラリー中に存在する識別可能な分子標識を有する標的を含む確率バーコードの数の推定値を補外するために、不完全シーケンシングのクオリティステータスを有する標的のシーケンシングリードに対して、ゼロ切断ポアソンモデルを適用する。いくつかの実施形態では、シーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数は、得られたシーケンシングデータ中の標的のクオリティステータスが、不完全シーケンシングクオリティステータスである場合、標的が、シーケンシングデータにおいてノイズであるか否かを決定する工程と；ノイズ標的を除去する工程と、によって調節する。

【0269】

いくつかの実施形態では、シーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数は、シーケンシングデータ中の標的のクオリティステータスが、不完全シーケンシングクオリティステータスである場合、シーケンシングデータ中の真の分子標識と偽の分子標識とを決定するために、標的の分子標識を閾値化する工程によって調節することができる。標的の分子標識を閾値化する工程は、分子標識に対する統計解析を実施する工程を含みうる。分子標識について統計解析を実施する工程は、ゼロ切断ポアソンモデルを用いて、真の分子標識の数 n を決定する工程と；シーケンシングデータから偽の分子標識を除去する工程と、を含みうる。

【0270】

いくつかの実施形態では、偽の分子標識は、 n 番目に豊富な分子標識の発生数よりも低い発生数を有する分子標識を含みうる。真の分子標識は、 n 番目に豊富な分子標識の発生数以上の発生数を有する分子標識を含みうる。

【0271】

シーケンシングデータエラー

本明細書に開示する方法は、シーケンシングデータエラー、たとえば、1つ以上の標的核酸をカウントする方法に発生するエラーを同定および／または訂正するために使用するこ

10

20

30

40

50

とができる。いくつかの実施形態ではエラーは、1つ以上のヌクレオチドの欠失、1つ以上のヌクレオチドの置換、1つ以上のヌクレオチドの付加、またはそれらの任意の組合せを含むか、そうしたものでありうる。エラーは、分子標識（ML）、サンプル標識（SL）、確率バーコード上の他の標識に存在しうる。いくつかの実施形態では、シーケンシングデータエラーは、PCR導入エラー、シーケンシング導入エラー、逆転写（RT）プライマー混入エラー、またはそれらの任意の組合せを含むか、またはそうしたものでありうる。PCR導入エラーは、PCR増幅エラー、PCR増幅バイアス、不十分なPCR増幅、またはそれらの任意の組合せの結果を含むか、またはそうしたものでありうる。シーケンシング導入エラーは、不正確なベースコーリング、不十分なシーケンシング、またはそれらの任意の組合せの結果を含むか、またはそうしたものでありうる。RTプライマー混入エラーは、PCRに進入した逆転写プライマーに起因するエラーでありうる。

10

【0272】

本明細書で使用される場合、「カバー率」または「シーケンシング深度」という用語は、シーケンシングデータ中の特定のMLおよび特定のSLを有するバーコード付き標的のリードの数を意味しうる。たとえば、バーコード付き標的は複数回シーケンシングされうる。従って、特定のMLおよびSLを有するバーコード付き標的を複数回観察することができる。別の例として、細胞は、標的の複数のコピー（たとえば、遺伝子のmRNA分子の複数のコピー）を含有しうる。こうした標的の複数のコピーにバーコードを付けることができる。PCR増幅の後（たとえば、図のブロック28）、特定のMLおよびSLを有するバーコード付き標的の複数のコピーが存在しうる。シーケンシングに際して、特定のMLおよびSLを有するバーコード付き標的の複数のコピーの一部または全部がシーケンシングされうる。シーケンシングデータ中に観察される同じMLおよびSLを有するバーコード付き標的のリードの数は、「カバー率」または「シーケンシング深度」と呼ばれることがある。

20

【0273】

いくつかの実施形態では、シーケンシングデータエラーは、同定および／または訂正することができる。たとえば、細胞からの標的のコピーは、異なるMLと同じSLとを有するバーコードを付けることができる。MLを有するバーコード付き標的は、シーケンシングデータ中の複数のリードを有しうる。異なるMLを有するバーコード付き標的は、少数のリード（たとえば、1リード）のみを有しうる。前のバーコード付き標的の方が、後のバーコード付き標的と比較して、真のML（またはリアルもしくはシグナルML）を有する傾向が高くなりうる。後のバーコード付き標的は、エラーML（または偽のもしくはノイズML）を含みうる。これは、2つのMLは、類似のカバー率またはシーケンシング深度を有することが予想できるためでありうる。少数のリードしか含まない後のバーコード付き標的は、シーケンシングまたはPCRの最中に生じる人工物もしくはエラーとなり得る。

30

【0274】

別の例として、PCRに進入する確率バーコードは、RTプライマー混入エラーを引き起こしうる。いくつかの実施形態ではcDNA分子にmRNA分子を逆転写した後（たとえば、図の24）、cDNA分子に組み込まれない確率バーコードは、たとえば、Ampureビーズ精製により除去することができる。除去方法、たとえば、Ampureビーズ精製は、確率バーコード付きcDNA分子に組み込まれる逆転写によって伸長されない確率バーコードを完全には除去されない可能性がある。たとえば、確率バーコード付きcDNA分子に組み込まれる逆転写によって伸長されない確率バーコードの15%、10%、9%、8%、7%、6%、5%、4%、3%、2%、1%、0.5%、0.1%、またはこれらのいずれか2つの値の間の範囲が、Ampureビーズ精製により除去されない可能性がある。これらの未除去確率バーコードは、cDNA分子の増幅中（たとえば、図のブロック28）にシーケンシングデータエラーを引き起こしうる。確率バーコードは、サンプルの間で非常に類似しうる。たとえば、確率バーコードのサンプル標識は、同じサンプルの場合、同一でありうる。従って、これらの未除去確率バーコードが、PCRの最中に同じサンプルからの他の核酸分子（たとえば、確率バーコード付きmRNA分子のSL

40

50

領域) にハイブリダイズする可能性があることから、PCR交差が起こり、その結果、SLエラーと呼ばれるシーケンシングデータエラーが生じうる。

【0275】

真のML、エラーML、およびSLエラーは、識別可能な分布を有しうる。図4は、分子標識エラー、サンプル標識エラー、および真の分子標識シグナルの非限定的な例示的分布を示す概略図である。図4に示されるように、エラーMLは、PCRまたはシーケンシングエラーに起因する可能性があるため、エラーMLは、より低いMLカバー率を有する傾向があると考えられる。たとえば、エラーMLは、シーケンシングエラーの大部分およびPCRエラーの一部に起因する可能性がある。SLエラーは、PCRに進入する確率バーコードに大部分起因する可能性があるため、SLエラーは、より低いMLカバー率を有する傾向があると考えられる。

10

【0276】

方向近接性に基づくPCRおよびシーケンシングエラーの訂正

本明細書には、PCRまたはシーケンシングエラーを訂正する方法が開示される。いくつかの実施形態では、本方法は、(a) 確率バーコード付き標的のシーケンシングデータを受け取る工程を含む。確率バーコード付き標的は、複数の確率バーコードを用いて、複数の標的に確率バーコードを付けて、複数の確率バーコード付き標的を生成する工程により取得することができ、ここで、複数の確率バーコードの各々が分子標識を含む。いくつかの実施形態では、本方法は、(b) 複数の標的の1つ以上について：(i) シーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数をカウントする工程と；(ii) 方向近接性を用いて、標的の分子標識のクラスターを同定する工程と；(iii) (ii) で同定された標的の分子標識のクラスターを用いて、(b) で受け取られたシーケンシングデータを折りたたむ工程と；(iv) 標的の数を推定する工程であって、推定された標的の数が、(ii) でシーケンシングデータを折りたたんだ後、(i) でカウントされたシーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数と相関する工程と、を含む。複数の標的は、細胞の全トランスクリプトームの標的を含む。いくつかの実施形態では、本方法は、さらに、(c) 複数の確率バーコードを用いて、複数の標的に確率バーコードを付けて、複数の確率バーコード付き標的を生成する工程と；(d) 確率バーコード付き標的をシーケンシングして、受け取った確率バーコード付き標的のシーケンシングデータを生成する工程と、を含む。

20

30

【0277】

図7は、方向近接性に基づく分子標識を用いて、PCRまたはシーケンシングエラーを訂正する、非限定的な例示的实施形態700を示すフローチャートである。方向近接性に基づく分子標識を用いて、PCRまたはシーケンシングエラーを訂正する工程は、再帰的置換エラー訂正(RSEC)と呼ばれることもある。この方法700は、複数の確率バーコード付き標的のシーケンシングデータを受け取った後、ブロック704で開始する。いくつかの実施形態では、方法700は、さらに、複数の確率バーコードを用いて、複数の標的に確率バーコードを付けて、複数のバーコード付き標的を生成する工程を含み、ここで、複数の確率バーコードの各々は、分子標識を含む。いくつかの実施形態では、方法700は、さらに、複数の確率バーコード付き標的をシーケンシングして、シーケンシングデータを取得する工程も含む。

40

【0278】

ブロック708で、複数の標的の1つ以上について：シーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数をカウントすることができる。ブロック712で、方向近接性を用いて、標的の分子標識のクラスターを同定することができる。クラスター内の標的の分子標識は、互いの所定の方向近接性閾値内に位置しうる。方向近接性閾値は、変動しうる。いくつかの実施形態では、所定の方向近接性閾値は、1もしくは2のハミング距離であるか、概ね、少なくとも、または多くとも、そうした距離でありうる。

【0279】

50

いくつかの実施形態では、クラスター内の標的の分子標識は、1つ以上の親分子標識と1つ以上の親分子標識の1つ以上の子供分子標識とを含みうる。親分子標識の発生数は、所定の方向近接性発生数閾値以上であってよい。いくつかの実施形態では、所定の方向近接性発生数閾値は、 $2 \times (\text{子供分子標識の発生数}) - 1$ であるか、概ね、少なくとも、または多くとも、そうした値でありうる。いくつかの実施形態では、所定の方向近接性発生数閾値は、子供分子標識の発生数の1.5倍、2倍、3倍、4倍、5倍、6倍、7倍、8倍、9倍、10倍、または概ねそうした値であるか、あるいはこれらの値のいずれか2つの間の数もしくは範囲でありうる。いくつかの実施形態では、所定の方向近接性発生数閾値は、子供分子標識の発生数の少なくともまたは多くとも1.5倍、2倍、3倍、4倍、5倍、6倍、7倍、8倍、9倍、10倍でありうる。

10

【0280】

ブロック720で、標的の分子標識のクラスターを用いて、シーケンシングデータを折りたたむ。シーケンシングデータを折りたたむ工程は、子供分子標識の発生数を親分子標識に帰属させる工程を含みうる。ブロック732で、シーケンシングデータを折りたたんだ後、標的の数を推定して、出力を生成することができる。方法700は、ブロック736で終了する。

【0281】

いくつかの実施形態では、本方法は、さらに、標的のシーケンシング深度を決定する工程を含む。標的の数を推定する工程は、標的のシーケンシング深度が、所定のシーケンシング深度閾値を超える場合、(i)でカウントされたシーケンシングデータを調節する工程を含む。所定のシーケンシング深度閾値は、15~20の間であってよい。(i)でカウントされたシーケンシングデータを調節する工程は、標的の分子標識を閾値化して、(b)で得られたシーケンシングデータ中の標的に関連付けられた真の分子標識および偽の分子標識を決定する工程を含む。標的の分子標識を閾値化する工程は、標的の分子標識について統計解析を実施する工程を含む。統計解析を実施する工程は、標的の分子標識の分布およびそれらの発生数を、2つのネガティブ二項分布などの2つの分布に当てはめる工程と；2つのネガティブ二項分布を用いて真の分子標識の数nを決定する工程と；(b)で得られたシーケンシングデータから偽の分子標識を除去する工程と、を含み、ここで、偽の分子標識は、n番目に豊富な分子標識の発生数よりも低い発生数を有する分子標識を含み、また、真の分子標識は、n番目に豊富な分子標識の発生数以上の発生数を有する分子標識を含む。

20

30

【0282】

方向近接性および二次導関数に基づくPCRおよびシーケンシングエラーの訂正
本明細書には、標的の数を決定する方法が開示される。いくつかの実施形態では、一方法は、(a)複数の確率バーコードを用いて、複数の標的に確率バーコードを付けて、複数の確率バーコード付き標的を生成する工程であって、複数の確率バーコードの各々が分子標識を含む工程と；(b)確率バーコード付き標的のシーケンシングデータを取得する工程と；(c)複数の標的の1つ以上について：(i)シーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数をカウントする工程と；(ii)方向近接性を用いて、標的の分子標識のクラスターを同定する工程と；(iii)(ii)で同定された標的の分子標識のクラスターを用いて、(b)で得られたシーケンシングデータを折りたたむ工程と；(iv)標的の数を推定する工程であって、推定された標的の数が、(ii)でシーケンシングデータを折りたたんだ後、(i)でカウントされたシーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数と関連する工程と、を含む。複数の標的は、細胞の全トランスクリプトームの標的を含みうる。

40

【0283】

いくつかの実施形態では、本方法は、シーケンシングデータ中の標的のシーケンシングステータスを決定する工程を含む。シーケンシングデータ中の標的のシーケンシングステータスは、飽和シーケンシングを含むか、または飽和シーケンシングであってもよい。いくつかの実施形態では、シーケンシングデータ中の標的のシーケンシングステータスが、飽

50

和シーケンシングステータスである場合、(i v) で推定された標的の数は、(i) でカウントされたシーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数と相関する。

【0284】

いくつかの実施形態では、推定された標的の数は、S L エラーを訂正した後に (i) でカウントされたシーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数と相関する。S L エラーを訂正する工程は、シーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の累積和プロットを作成する工程と；累積和プロットの二次導関数を決定する工程と；累積和プロットの二次導関数の最小値に基づき、M L リード深度カットオフを決定する工程と、を含む。いくつかの実施形態では S L エラーを訂正する工程は、決定された M L リード深度カットオフより低いリード深度を有する、シーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識を除去する工程を含みうる。

【0285】

図 8 は、方向近接性および二次導関数に基づく分子標識を用いて、P C R およびシーケンシングエラーを訂正する、非限定的な例示の実施形態 800 を示すフローチャートである。方法 800 は、複数の確率バーコード付き標的のシーケンシングデータを受け取った後、ブロック 804 から開始する。いくつかの実施形態では、方法 800 は、さらに、複数の確率バーコードを用いて、複数の標的に確率バーコードを付けて、複数の確率バーコード付き標的を生成する工程を含み、ここで、複数の確率バーコードの各々が分子標識を含む。いくつかの実施形態では、方法 800 は、さらに、複数の確率バーコード付き標的をシーケンシングして、シーケンシングデータを取得する工程も含む。

【0286】

ブロック 808 で、複数の標的の 1 つ以上について：シーケンシング中の標的に関連付けられた識別可能な配列を有する分子標識の数をカウントすることができる。決定ブロック 812 で、シーケンシングデータが、飽和シーケンシングステータスを有するか否かを決定することができる。飽和シーケンシングステータスは、所定の飽和閾値より高い、識別可能な配列を有する分子標識の数を有する標的によって決定することができる。所定の飽和閾値は、さまざまな履行で異なりうる。たとえば、確率バーコードが、識別可能な配列を有する約 6561 の分子標識を有する場合、所定の飽和閾値は、約 6557 となりうる。別の例として、確率バーコードが、識別可能な配列を有する約 65536 の分子標識を有する場合、所定の飽和閾値は、約 65532 となりうる。

【0287】

シーケンシングデータが、決定ブロック 812 で飽和シーケンシングステータスを有していない場合、方法 800 は、ブロック 816 に進むことができ、ここで、分子標識カウントが、方向近接性に基づき調節されうる。標的は、たとえば、それが、1000、2000、3000、4000、5000、6000、7000、8000、9000、10000、20000、30000、40000、50000、60000、70000、80000、90000、100000 を超える、またはこれらのいずれか 2 つの間の数もしくは範囲を超える、識別可能な配列を有する分子標識の数を有する場合、飽和シーケンシングステータスを有するとみなすことができる。別の例として、標的は、識別可能な配列を有する確率バーコードの分子バーコードの 50%、60%、70%、80%、90%、95%、99%、もしくは 99.9% を超える、またはこれらのうちいずれか 2 つの間の数もしくは範囲を超える、識別可能な配列を有する分子標識の数を有する場合、飽和シーケンシングステータスを有するとみなすことができる。いくつかの実施形態では、方向近接性に基づき分子カウントを調節する工程は、図 7 を参照にして説明することができる。たとえば、辞書に基づき分子カウントを調節する工程は、方向近接性を用いて、標的の分子標識のクラスターを同定する工程と；同定された標的の分子標識のクラスターを用いて、シーケンシングデータを折りたたむ工程と；標的の数を推定する工程と、を含むことができ、ここで、推定された標的の数は、シーケンシングデータを折りたたんだ後、カウ

ントされたシーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数と相関する。

【0288】

ブロック820で、累積和プロットの二次導関数を決定することができる。累積和プロットの二次導関数を決定する工程は、シーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の累積和プロットを作成する工程を含みうる。

【0289】

ブロック824で、分子標識は、MLリード深度カットオフに基づき調節することができる。MLリード深度カットオフは、累積和プロットの二次導関数の最小値（たとえば、局所的最小値または大域的最小値）に基づくものでよい。いくつかの実施形態では、SLエラーを訂正する工程は、決定されたMLリード深度カットオフより低いリード深度を有するシーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識を除去する工程を含みうる。

【0290】

ブロック828で、標的の数を推定して、シーケンシングデータを折りたたみ、SLエラーを訂正した後の出力を生成することができる。決定ブロック812で、シーケンシングデータが、飽和シーケンシングステータスを有する場合、方法800は、ブロック828に進んで、シーケンシングデータの折りたたみおよびSLエラーの訂正なしに出力を生成することができる。方法800は、ブロック832で終了する。

【0291】

方向近接性に基づくPCRおよびシーケンシングエラーの訂正ならびに分布に基づくエラーの訂正

本明細書には、PCRまたはシーケンシングエラーを訂正する方法が開示される。本方法を用いて、標的の数を決定することができる。いくつかの実施形態では、本方法は、（a）確率バーコード付き標的のシーケンシングデータを受け取る工程を含む。確率バーコード付き標的は、複数の確率バーコードを用いて、複数の標的に確率バーコードを付けて、複数の確率バーコード付き標的を生成する工程により取得することができ、ここで、複数の確率バーコードの各々が分子標識を含む。いくつかの実施形態では、本方法は、（b）複数の標的の1つ以上について：（i）シーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数をカウントする工程と；（ii）シーケンシングデータ中の標的に関連付けられた識別可能な配列を有するノイズ分子標識の数を決定する工程と；（iii）標的の数を推定する工程と、を含み、ここで、推定された標的の数は、（ii）で決定されたノイズ分子標識の数に従って調節される、（i）でカウントされたシーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数と相関する。いくつかの実施形態では、本方法は、シーケンシングデータ中の標的のシーケンシングステータスを決定する工程を含む。いくつかの実施形態では、本方法は、さらに、（c）複数の確率バーコードを用いて、複数の標的に確率バーコードを付けて、複数の確率バーコード付き標的を生成する工程と；（d）確率バーコード付き標的のシーケンシングして、受け取った確率バーコード付き標的のシーケンシングデータを生成する工程と、を含む。

【0292】

図9は、再帰的置換エラー訂正および分布ベースのエラー訂正に基づいて、PCRおよびシーケンシングエラーを訂正する、非限定的な例示の実施形態900を示すフローチャートである。方法900は、複数の確率バーコード付き標的のシーケンシングデータを受け取った後、ブロック904から開始する。いくつかの実施形態では、方法900は、複数の確率バーコードを用いて、複数の標的に確率バーコードを付けて、複数の確率バーコード付き標的を生成する工程をさらに含み、ここで、複数の確率バーコードの各々が分子標識を含む。いくつかの実施形態では、方法900は、さらに、複数の確率バーコード付き標的のシーケンシングして、シーケンシングデータを取得する工程も含む。

【0293】

ブロック 908 で、複数の標的の 1 つ以上について：シーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数をカウントすることができる。決定ブロック 912 で、シーケンシングデータは、飽和シーケンシングステータスを有するか否かを決定することができる。たとえば、標的は、それが、1000、2000、3000、4000、5000、6000、7000、8000、9000、10000、20000、30000、40000、50000、60000、70000、80000、90000、100000 を超える、またはこれらのいずれか 2 つの間の数もしくは範囲を超える、識別可能な配列を有する分子標識の数を有する場合、飽和シーケンシングステータスを有するとみなすことができる。別の例として、標的は、識別可能な配列を有する確率バーコードの分子バーコードの 50%、60%、70%、80%、90%、95%、99%、もしくは 99.9% を超える、またはこれらのうちいずれか 2 つの間の数もしくは範囲を超える、識別可能な配列を有する分子標識の数を有する場合、飽和シーケンシングステータスを有するとみなすことができる。

10

【0294】

いくつかの実施形態では、飽和シーケンシングステータスは、所定の飽和閾値を超える、識別可能な配列を有する分子標識の数を有する標的によって決定することができる。所定の飽和閾値は、さまざまな履行において異なるものであってよい。たとえば、所定の飽和閾値は、1000、2000、3000、4000、5000、6000、6557、7000、8000、9000、10000、20000、30000、40000、50000、60000、65532、70000、80000、90000、100000、またはこれらの値のいずれか 2 つの間の数もしくは範囲であるか、または概ねそうした値でありうる。別の例として、所定の飽和閾値は、少なくとも、または多くとも、1000、2000、3000、4000、5000、6000、6557、7000、8000、9000、10000、20000、30000、40000、50000、60000、65532、70000、80000、90000、もしくは 100000 でありうる。

20

【0295】

いくつかの実施形態では、飽和シーケンシングステータスは、識別可能な配列を有する確率バーコードの分子標識の数に依存しうる。たとえば、確率バーコードが、識別可能な配列を有する約 6561 の分子標識を有する場合、所定の飽和閾値は、約 6557 となりうる。別の例として、確率バーコードが、識別可能な配列を有する約 65536 の分子標識を有する場合、所定の飽和閾値は、約 65532 となりうる。いくつかの実施形態では、飽和シーケンシングステータスは、識別可能な配列を有する確率バーコードの分子標識の数に依存しない場合もある。

30

【0296】

シーケンシングデータが、決定ブロック 912 で、飽和シーケンシングステータスを有していなければ、方法 900 は、ブロック 916 に進み、ここで、分子標識カウントは、方向近接性に基づいて調節されうる。いくつかの実施形態では、方向近接性に基づき分子カウントを調節する工程は、図 7 を参照にして説明することができる。たとえば、辞書に基づき分子カウントを調節する工程は、方向近接性を用いて、標的の分子標識のクラスターを同定する工程と；同定された標的の分子標識のクラスターを用いて、シーケンシングデータを折りたたむ工程と；標的の数を推定する工程と、を含み、ここで、推定された標的の数は、シーケンシングデータを折りたたんだ後、カウントされたシーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数と相関する。

40

【0297】

ブロック 920 で、シーケンシングデータ中の標的のシーケンシングステータスを決定することができる。シーケンシングデータ中の標的のシーケンシングステータスは、過少シーケンシングを含むか、または過少シーケンシングでありうる。決定ブロック 924 で、シーケンシングデータ中の標的のシーケンシングステータスが、過少シーケンシングステータスであるか否かを決定することができる。たとえば、標的は、その深度（たとえば、

50

平均、最小、もしくは最大深度)が、1、2、3、4、5、6、7、8、9、10、20、30、40、50、60、70、80、90、100、またはこれらの値のいずれか2つの間の数もしくは範囲より小さいか、または概ねそうした値より小さい場合、過少シーケンシングステータスを有するとみなすことができる。別の例として、標的は、その深度が、少なくとも、または多くとも、1、2、3、4、5、6、7、8、9、10、20、30、40、50、60、70、80、90、もしくは100より小さい場合、過少シーケンシングステータスを有するとみなすことができる。

【0298】

いくつかの実施形態では、過少シーケンシングステータスは、所定の過少シーケンシング閾値より小さい深度(たとえば、平均、最小、もしくは最大深度)を有する標的によって決定することができる。過少シーケンシング閾値は、さまざまな履行で異なるものであってよい。たとえば、過少シーケンシング閾値は、1、2、3、4、5、6、7、8、9、10、20、30、40、50、60、70、80、90、100、またはこれらの値のいずれか2つの間の数もしくは範囲、または概ねそうした値でありうる。別の例として、過少シーケンシング閾値は、少なくとも、または多くとも、1、2、3、4、5、6、7、8、9、10、20、30、40、50、60、70、80、90、もしくは100でありうる。

【0299】

いくつかの実施形態では、過少シーケンシングステータスは、識別可能な配列を有する確率バーコードの分子標識の数に依存しうる。たとえば、確率バーコードが、識別可能な配列を有する、1000、2000、3000、4000、5000、6000、6561、7000、8000、9000、10000、20000、30000、40000、50000、60000、65532、70000、80000、90000、100000、またはこれらの値のいずれか2つの間の数もしくは範囲、または概ねそうした値の分子標識を有する場合、過少シーケンシング閾値は、10(または別の閾値数)となりうる。別の例として、確率バーコードが、少なくとも、または多くとも、1000、2000、3000、4000、5000、6000、6561、7000、8000、9000、10000、20000、30000、40000、50000、60000、65532、70000、80000、90000、もしくは100000を含む場合、過少シーケンシング閾値は、10(または別の閾値数)となりうる。いくつかの実施形態では、飽和シーケンシングステータスは、識別可能な配列を有する確率バーコードの分子標識の数に依存しない場合もある。

【0300】

決定ブロック924で、シーケンシングデータ中の標的のシーケンシングステータスが、過少シーケンシングステータスではなければ、方法900は、ブロック928に進んで、分子標識カウントをフィルタリングすることができる。分子標識カウントをフィルタリングする工程は、決定ブロック932で、擬似点閾値より少ない、シーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数を決定する工程を含む。擬似点閾値は、さまざまな履行で異なるものであってよい。たとえば、確率バーコードが、識別可能な配列を有する約6561の分子標識を有する場合、擬似点閾値は、1、2、3、4、5、6、7、8、9、10、20、30、40、50、60、70、80、90、100、またはこれらの値のいずれか2つの間の数もしくは範囲、または概ねそうした値でありうる。別の例として、確率バーコードが、識別可能な配列を有する約6561の分子標識を有する場合、擬似点閾値は、少なくとも、または多くとも、1、2、3、4、5、6、7、8、9、10、20、30、40、50、60、70、80、90、もしくは100でありうる。

【0301】

決定ブロック932で、シーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数が、擬似点閾値より少ない場合、方法900は、任意選択で、ブロック936に進み、そこで、シーケンシングデータ中の標的に関連付けられた識別可能な

10

20

30

40

50

配列を有するノイズ分子標識の数を決定する前に、シーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数に擬似点を追加することができる。擬似点は、さまざまな履行で異なる分子標識カウントを有しうる。たとえば、擬似点の分子標識カウントは、0、0001、0001、001、01、1、2、3、4、5、6、7、8、9、10、20、30、40、50、60、70、80、90、100、またはこれらの値のいずれか2つの間の数もしくは範囲、または概ねそうした値でありうる。別の例として、擬似点の分子標識カウントは、少なくとも、または多くとも、0001、0001、001、01、1、2、3、4、5、6、7、8、9、10、20、30、40、50、60、70、80、90、もしくは100でありうる。いくつかの実施形態では、シーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数が、擬似点閾値より少ない場合、方法900は、ブロックに進むことができ、シーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数が、擬似点閾値より少ない場合、方法944。

【0302】

決定ブロック932で、シーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数が、擬似点閾値以上である場合、非ユニーク分子標識をブロック940で除去することができる。非ユニーク分子標識は、ブロック944でシーケンシングデータ中の標的に関連付けられた識別可能な配列を有するノイズ分子標識の数を決定するために、除去することができる。非ユニーク分子標識は、所定の再使用分子標識閾値より大きい、シーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識を含みうる。再使用分子標識閾値は、さまざまな履行において異なるものであってよい。たとえば、再使用分子標識閾値は、確率バーコードが、識別可能な配列を有する約6561の分子標識を含む場合、100、200、300、400、500、600、650、700、900、1000、2000、またはこれらの値のいずれか2つの間の数もしくは範囲、または概ねそうした値でありうる。別の例として、再使用分子標識閾値は、確率バーコードが、識別可能な配列を有する約6561の分子標識を含む場合、少なくとも、または多くとも、100、200、300、400、500、600、650、700、900、1000、もしくは2000でありうる。

【0303】

いくつかの実施形態では、非ユニーク分子標識を除去する工程は、シーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数に対する非ユニーク分子標識の理論上の数を決定する工程を含む。非ユニーク分子標識を除去する工程は、シーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識のn番目に豊富な分子標識より大きな発生数を有する分子標識を除去する工程を含みうる。数nは、非ユニーク分子標識の理論上の数でありうる。

【0304】

ブロック944で、分布ベースのエラー訂正方法を用いて、分子標識カウントを調節することができる。分布ベースのエラー訂正方法は、シーケンシングデータ中の標的に関連付けられた識別可能な配列を有するノイズ分子標識の数を決定する工程を含みうる。ノイズ分子標識の数を決定する工程は、2つのネガティブ二項分布を、シーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数に当てはめる工程を含みうる。たとえば、ノイズ分子標識の数を決定する工程は、シグナルネガティブ二項分布（2つのネガティブ二項分布の一方）を、カウントされたシーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数に当てはめる工程を含んでよく、ここで、シグナルネガティブ二項分布は、シグナル分子標識である、カウントされたシーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数に対応する。ノイズ分子標識の数を決定する工程は、ノイズネガティブ二項分布（2つのネガティブ二項分布の他方）を、カウントされたシーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数に当てはめる工程を含んでよく、ここでノイズネガティブ二項分布は、ノイズ分子標識である、カウントされたシーケンシングデータ中の標的

に関連付けられた識別可能な配列を有する分子標識の数に対応する。ノイズ分子標識の数を決定する工程は、当てはめたシグナルネガティブ二項分布と、当てはめたノイズネガティブ二項分布を用いて、ノイズ分子標識の数を決定する工程を含んでよい。

【0305】

いくつかの実施形態では、当てはめたシグナルネガティブ二項分布と、当てはめたノイズネガティブ二項分布を用いて、ノイズ分子標識の数を決定する工程は、シーケンシングデータ中の標的に関連付けられた識別可能な配列の各々について：識別可能な配列が、シグナルネガティブ二項分布であるシグナル確率を決定する工程を含む。そして、識別可能な配列のノイズ確率が、ノイズネガティブ二項分布であるノイズ確率を決定することができる。さらには、シグナル確率が、ノイズ確率より小さければ、識別可能な配列は、ノイズ分子標識であると決定することができる。いくつかの実施形態では、ブロック944で分子標識カウントを調節する工程は、2つ未満のピークが見出される（シグナルネガティブ二項分布とノイズネガティブ二項分布を決定するために、2つのピークが必要とされるため）場合、シングルトン（たとえば、単一塩基置換）を除去する工程を含みうる。

10

【0306】

ブロック948で、標的の数を推定して、近接性に基づくエラー訂正および分布ベースのエラー訂正後に出力を生成することができる。決定ブロック912で、シーケンシングデータ中の標的のシーケンシングステータスが、飽和シーケンシングステータスである場合、方法900は、ブロック948に進んで、方向近接性および分布ベースのエラー訂正に基づいて分子標識を調節することなく、出力を生成することができる。たとえば、決定されたノイズ分子標識の数は、ゼロであってもよい。

20

【0307】

決定ブロック924で、シーケンシングデータ中の標的のシーケンシングステータスが、過少シーケンシングステータスである場合、方法900は、ブロック948に進んで、分布ベースのエラー訂正に基づいて分子標識を調節することなく、出力を生成することができる。たとえば、決定されたノイズ分子標識の数は、ゼロであってもよい。方法900は、ブロック952で終了する。

【0308】

図10は、2つのネガティブ二項分布を用いたエラー訂正の非限定的な例示的实施形態1000を示すフローチャートである。方法1000のブロック（たとえば、ブロック904～952）は、図9を参照にして説明されている。手短には、方法1000は、複数の確率バーコード付き標的のシーケンシングデータを受け取った後、ブロック904で開始する。いくつかの実施形態では、方法1000は、複数の確率バーコードを用いて、複数の標的に確率バーコードを付けて、複数の確率バーコード付き標的を生成する工程をさらに含み、ここで、複数の確率バーコードの各々は、分子標識を含む。いくつかの実施形態では、方法1000は、さらに、複数の確率バーコード付き標的をシーケンシングして、シーケンシングデータを取得する工程も含む。

30

【0309】

ブロック908で、複数の標的の1つ以上について：シーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数をカウントすることができる。ブロック916で、方向近接性に基づいて分子標識カウントを調節することができる。いくつかの実施形態では、方向近接性に基づいて分子標識カウントを調節する工程は、図7を参照にして説明することができる。

40

【0310】

ブロック920で、シーケンシングデータ中の標的のシーケンシングステータスを決定することができる。シーケンシングデータ中の標的のシーケンシングステータスは、過少シーケンシングを含むか、または過少シーケンシングであってもよい。決定ブロック924で、シーケンシングデータ中の標的のシーケンシングステータスが、過少シーケンシングステータスであるか否かを決定することができる。

【0311】

50

決定ブロック 924 で、シーケンシングデータ中の標的のシーケンシングステータスが、過少シーケンシングステータスでなければ、方法 1000 は、任意選択で、決定ブロック 1004 に進むことができる。決定ブロック 1004 で、標的のシーケンシング深度を所定のシーケンシング深度閾値と比較することができる。シーケンシング深度閾値は、さまざまな履行において異なるものであってよい。たとえば、標的のシーケンシング深度は、1、2、3、4、5、6、7、8、9、10、20、30、40、50、60、70、80、90、100、またはこれらの値のいずれか 2 つの間の数もしくは範囲、または概ねそうした値でありうる。別の例として、標的のシーケンシング深度は、少なくとも、または多くとも、1、2、3、4、5、6、7、8、9、10、20、30、40、50、60、70、80、90、もしくは 100 でありうる。

10

【0312】

標的のシーケンシング深度が、シーケンシング深度閾値より大きい場合、方法 1000 は、ブロック 928 に進む。標的のシーケンシング深度が、シーケンシング深度閾値以下である場合、方法 1000 は、ブロック 1008 に進む。ブロック 1008 で、ブロック 948 の出力を生成する工程の前に、シングルトン（たとえば、単一塩基置換）を除去することができる。

【0313】

ブロック 928 で、分子標識カウントをフィルタリングすることができる。分子標識カウントをフィルタリングする工程は、決定ブロック 912 で、シーケンシングデータが、飽和シーケンシングステータスを有するか否かを決定する工程を含むことができる。シーケンシングデータが、決定ブロック 912 で、飽和シーケンシングステータスを有していない場合、方法 1000 は、決定ブロック 932 に進むことができる。決定ブロック 932 で、擬似点閾値より少ない、シーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数を決定することができる。

20

【0314】

決定ブロック 932 で、シーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数が、擬似点閾値より少ない場合、方法 900 は、任意選択で、ブロック 936 に進み、そこで、シーケンシングデータ中の標的に関連付けられた識別可能な配列を有するノイズ分子標識の数を決定する前に、シーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数に擬似点を追加することができる。いくつかの実施形態では、シーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数が、擬似点閾値より少ない場合、方法 900 は、ブロックに進むことができ、シーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数が、擬似点閾値より少ない場合、方法 944。

30

【0315】

決定ブロック 932 で、シーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数が、擬似点閾値以上である場合、非ユニーク分子標識をブロック 940 で除去することができる。非ユニーク分子標識は、ブロック 944 でシーケンシングデータ中の標的に関連付けられた識別可能な配列を有するノイズ分子標識の数を決定するために、除去することができる。非ユニーク分子標識は、所定の再使用分子標識閾値より大きい、シーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識を含みうる。

40

【0316】

ブロック 944 で、分布ベースのエラー訂正方法を用いて、分子標識カウントを調節することができる。分布ベースのエラー訂正方法は、シーケンシングデータ中の標的に関連付けられた識別可能な配列を有するノイズ分子標識の数を決定する工程を含みうる。ノイズ分子標識の数を決定する工程は、2 つのネガティブ二項分布、すなわち、シグナルネガティブ二項分布とノイズネガティブ二項分布とを、シーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数に当てはめる工程を含みうる。シグナルネガティブ二項分布は、シグナル分子標識である、カウントされたシーケンシングデータ中

50

の標的に関連付けられた識別可能な配列を有する分子標識の数に対応する。ノイズネガティブ二項分布は、ノイズ分子標識である、カウントされたシーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数に対応する。ノイズ分子標識の数を決定する工程は、当てはめたシグナルネガティブ二項分布と、当てはめたノイズネガティブ二項分布を用いて、ノイズ分子標識の数を決定する工程を含むことができる。

【0317】

ブロック948で、標的の数を推定して、近接性に基づくエラー訂正および分布ベースのエラー訂正後に出力を生成することができる。決定ブロック912で、シーケンシングデータ中の標的のシーケンシングステータスが、飽和シーケンシングステータスである場合、方法1000は、ブロック948に進んで、方向近接性および分布ベースのエラー訂正に基づいて分子標識を調節することなく、出力を生成することができる。

【0318】

決定ブロック924で、シーケンシングデータ中の標的のシーケンシングステータスが、過少シーケンシングステータスである場合、方法1000は、ブロック948に進んで、分布ベースのエラー訂正に基づいて分子標識を調節することなく、出力を生成することができる。たとえば、決定されたノイズ分子標識の数は、ゼロであってもよい。方法1000は、ブロック952で終了する。

【0319】

方向近接性に基づくPCRおよびシーケンシングエラーの訂正、分布に基づくエラーの訂正、ならびにサブサンプリング

図11は、2つのネガティブ二項分布を用いたエラー訂正の非限定的な例示的实施形態1100を示すフローチャートである。方法1100のブロック（たとえば、ブロック904～952）は、図9を参照にして説明されている。手短には、方法1100は、複数の確率バーコード付き標的のシーケンシングデータを受け取った後、ブロック904で開始する。いくつかの実施形態では、方法1100は、複数の確率バーコードを用いて、複数の標的に確率バーコードを付けて、複数の確率バーコード付き標的を生成する工程をさらに含み、ここで、複数の確率バーコードの各々は、分子標識を含む。いくつかの実施形態では、方法1100は、さらに、複数の確率バーコード付き標的をシーケンシングして、シーケンシングデータを取得する工程も含む。

【0320】

ブロック908で、複数の標的の1つ以上について：シーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数をカウントすることができる。決定ブロック912で、シーケンシングデータが、飽和シーケンシングステータスを有するか否かを決定することができる。シーケンシングデータが、決定ブロック912で、飽和シーケンシングステータスを有していなければ、方法1100は、ブロック916に進み、ここで、分子標識カウントは、方向近接性に基づいて調節される。いくつかの実施形態では、方向近接性に基づき分子カウントを調節する工程は、図7を参照にして説明することができる。

【0321】

ブロック920で、シーケンシングデータ中の標的のシーケンシングステータスを決定することができる。シーケンシングデータ中の標的のシーケンシングステータスは、過少シーケンシングを含むか、または過少シーケンシングでありうる。決定ブロック924で、シーケンシングデータ中の標的のシーケンシングステータスが、過少シーケンシングステータスであるか否かを決定することができる。

【0322】

決定ブロック924で、シーケンシングデータ中の標的のシーケンシングステータスが、過少シーケンシングステータスでなければ、方法1100は、任意選択で、決定ブロック1104に進むことができる。決定ブロック1104で、シーケンシングデータ中の標的のシーケンシングステータスが、過剰シーケンシングデータであるか否かを決定することができる。たとえば、標的は、その深度（たとえば、平均、最小、もしくは最大深度）が

、50、100、200、250、300、400、500、600、700、800、900、1000、またはこれらの値のいずれか2つの間の数もしくは範囲より大きい、または概ねそうした値より大きい場合、過剰シーケンシングステータスまたは高度発現標的を有するとみなすことができる。別の例として、標的は、その深度が、少なくとも、または多くとも、50、100、200、250、300、400、500、600、700、800、900、もしくは1000より大きい場合、過少シーケンシングステータスを有するとみなすことができる。

【0323】

いくつかの実施形態では、過剰シーケンシングステータスまたは高度発現標的は、所定の過剰シーケンシング閾値より大きい深度（たとえば、平均、最小、もしくは最大深度）を有する標的によって決定することができる。過剰シーケンシング閾値は、さまざまな履行において異なるものであってよい。たとえば、過剰シーケンシング閾値は、50、100、200、250、300、400、500、600、700、800、900、1000、またはこれらの値のいずれか2つの間の数もしくは範囲、または概ねそうした値でありうる。別の例として、過剰シーケンシング閾値は、少なくとも、または多くとも、50、100、200、250、300、400、500、600、700、800、900、1000でありうる。

【0324】

いくつかの実施形態では、過剰シーケンシングステータスは、識別可能な配列を有する確率バーコードの分子標識の数に依存しうる。たとえば、確率バーコードが、識別可能な配列を有する約6561の分子標識を含む場合、過剰シーケンシング閾値は、50、100、200、250、300、400、500、600、700、800、900、1000、またはこれらの値のいずれか2つの間の数もしくは範囲、または概ねそうした値でありうる。別の例として、確率バーコードが、識別可能な配列を有する約6561の分子標識を含む場合、過剰シーケンシング閾値は、少なくとも、または多くとも、50、100、200、250、300、400、500、600、700、800、900、1000でありうる。いくつかの実施形態では、過少シーケンシングステータスは、識別可能な配列を有する確率バーコードの分子標識の数に依存しない場合もある。

【0325】

決定ブロック1104で、標的が、過剰シーケンシングステータスを有する場合、方法1100は、ブロック1108に進む。ブロック1108で、標的のMLカバー率は、たとえば、全標的のMLカバー率をサブサンプリングすることによって減少されうる。たとえば、シーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数は、全標的についての所定の過剰シーケンシング閾値の近似値までサブサンプリングされうる（たとえば、10）。方法1100は、ブロック1108からブロック928に進む。

【0326】

決定ブロック1104で、標的が、過剰シーケンシングステータスを有していなければ、方法1100は、ブロック928に進んで、分子標識カウントをフィルタリングする。分子標識カウントをフィルタリングする工程は、決定ブロック932で、擬似点閾値より少ない、シーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数を決定する工程を含みうる。

【0327】

決定ブロック932で、シーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数が、擬似点閾値より少ない場合、方法900は、任意選択で、ブロック936に進み、そこで、シーケンシングデータ中の標的に関連付けられた識別可能な配列を有するノイズ分子標識の数を決定する前に、シーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数に擬似点を追加することができる。いくつかの実施形態では、シーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数が、擬似点閾値より少ない場合、方法900は、ブロックに進むこと

10

20

30

40

50

ができ、シーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数が、擬似点閾値より少ない場合、方法 9 4 4。

【0328】

決定ブロック 9 3 2 で、シーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数が、擬似点閾値以上である場合、非ユニーク分子標識をブロック 9 4 0 で除去することができる。非ユニーク分子標識は、ブロック 9 4 4 でシーケンシングデータ中の標的に関連付けられた識別可能な配列を有するノイズ分子標識の数を決定するために、除去することができる。非ユニーク分子標識は、所定の再使用分子標識閾値より大きい、シーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識を含みうる。

10

【0329】

ブロック 9 4 4 で、分布ベースのエラー訂正方法を用いて、分子標識カウントを調節することができる。分布ベースのエラー訂正方法は、シーケンシングデータ中の標的に関連付けられた識別可能な配列を有するノイズ分子標識の数を決定する工程を含みうる。ノイズ分子標識の数を決定する工程は、2つのネガティブ二項分布、すなわち、シグナルネガティブ二項分布とノイズネガティブ二項分布とを、シーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数に当てはめる工程を含みうる。シグナルネガティブ二項分布は、シグナル分子標識である、カウントされたシーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数に対応する。ノイズネガティブ二項分布は、ノイズ分子標識である、カウントされたシーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数に対応する。ノイズ分子標識の数を決定する工程は、当てはめたシグナルネガティブ二項分布と、当てはめたノイズネガティブ二項分布を用いて、ノイズ分子標識の数を決定する工程を含む。

20

【0330】

ブロック 9 4 4 で分布ベースのエラー訂正を用いて分子標識カウントを調節した後、方法 1 1 0 0 は、任意選択で、ブロック 1 1 1 2 に進む。ブロック 1 1 1 2 で、ブロック 9 4 4 からの調節された分子標識カウントを、ブロック 9 1 6 で決定されて、方向近接性に基づき調節された分子標識カウントと合わせることができる。たとえば、非ユニーク分子標識は、ブロック 9 4 0 で除去されるため、ブロック 9 4 4 で分布当てはめには使用されない。しかし、これらの分子標識は、ブロック 9 1 6 で決定されて、方向近接性に基づき調節された分子標識カウント中に依然として存在する。従って、ブロック 9 4 4 からの調節された分子標識カウントと、ブロック 9 4 4 で調節された分子標識カウントを合わせて、ブロック 9 4 8 で出力を生成することができる。

30

【0331】

決定ブロック 9 1 2 で、シーケンシングデータ中の標的のシーケンシングステータスが、飽和シーケンシングステータスであれば、方法 1 1 0 0 は、ブロック 9 4 8 に進んで、方向近接性および分布ベースのエラー訂正に基づいて分子標識を調節することなく、出力を生成することができる。決定ブロック 9 2 4 で、シーケンシングデータ中の標的のシーケンシングステータスが、過少シーケンシングステータスであれば、方法 1 1 0 0 は、ブロック 9 4 8 に進んで、分布ベースのエラー訂正に基づいて分子標識を調節することなく、出力を生成することができる。たとえば、決定されるノイズ分子標識の数は、ゼロでありうる。方法 1 1 0 0 は、たとえば、ブロック 9 5 2 で終了しうる。

40

【0332】

図 1 2 は、2つのネガティブ二項分布を用いたエラー訂正の非限定的な例示的实施形態 1 2 0 0 を示すフローチャートである。方法 1 2 0 0 のブロック（たとえば、ブロック 9 0 4 ~ 9 5 2 およびブロック 1 1 0 4）は、図 9 および 1 1 を参照にして説明されている。手短には、方法 1 2 0 0 は、複数の確率バーコード付き標的のシーケンシングデータを受け取った後、ブロック 9 0 4 で開始する。いくつかの実施形態では、方法 1 2 0 0 は、複数の確率バーコードを用いて、複数の標的に確率バーコードを付けて、複数の確率バーコード付き標的を生成する工程をさらに含み、ここで、複数の確率バーコードの各々は、分

50

子標識を含む。いくつかの実施形態では、方法 1 2 0 0 は、さらに、複数の確率バーコード付き標的をシーケンシングして、シーケンシングデータを取得する工程も含む。

【0 3 3 3】

ブロック 9 0 8 で、複数の標的の 1 つ以上について：シーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数をカウントすることができる。決定ブロック 9 1 2 で、シーケンシングデータが、飽和シーケンシングステータスを有するか否かを決定することができる。シーケンシングデータが、決定ブロック 9 1 2 で、飽和シーケンシングステータスを有していなければ、方法 1 2 0 0 は、ブロック 9 1 6 に進み、ここで、分子標識カウントは、方向近接性に基づいて調節されうる。いくつかの実施形態では、方向近接性に基づき分子カウントを調節する工程は、図 7 を参照にして説明することができる。

10

【0 3 3 4】

ブロック 9 2 0 で、シーケンシングデータ中の標的のシーケンシングステータスを決定することができる。シーケンシングデータ中の標的のシーケンシングステータスは、過少シーケンシングを含むか、または過少シーケンシングでありうる。決定ブロック 9 2 4 で、シーケンシングデータ中の標的のシーケンシングステータスが、過少シーケンシングステータスであるか否かを決定することができる。

【0 3 3 5】

決定ブロック 9 2 4 で、シーケンシングデータ中の標的のシーケンシングステータスが、過少シーケンシングステータスでなければ、方法 1 2 0 0 は、任意選択で、決定ブロック 1 1 0 4 に進むことができる。決定ブロック 1 1 0 4 で、シーケンシングデータ中の標的のシーケンシングステータスが、過剰シーケンシングデータであるか否かを決定することができる。

20

【0 3 3 6】

決定ブロック 1 1 0 4 で、標的が、過剰シーケンシングステータスを有するか、または標的が、高度発現標的である場合、方法 1 2 0 0 は、任意選択で、ブロック 1 2 0 8 に進む。ブロック 1 2 0 8 で、標的の M L カバー率は、たとえば、標的毎に M L カバー率をサブサンプリングすることによって減少されうる。たとえば、シーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数は、標的毎に所定の過剰シーケンシング閾値の近似値までサブサンプリングされうる。方法 1 2 0 0 は、ブロック 1 2 0 8 からブロック 9 2 8 に進む。

30

【0 3 3 7】

決定ブロック 1 1 0 4 で、標的が、過剰シーケンシングステータスを有していなければ、方法 1 2 0 0 は、ブロック 9 2 8 に進んで、分子標識カウントをフィルタリングする。分子標識カウントをフィルタリングする工程は、決定ブロック 9 3 2 で、擬似点閾値より少ない、シーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数を決定する工程を含みうる。

【0 3 3 8】

ブロック 9 3 2 で、シーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数が、擬似点閾値より少ない場合、方法 9 0 0 は、任意選択で、ブロック 9 3 6 に進み、そこで、シーケンシングデータ中の標的に関連付けられた識別可能な配列を有するノイズ分子標識の数を決定する前に、シーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数の擬似点を任意選択で追加することができる。いくつかの実施形態では、シーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数が、擬似点閾値より少ない場合、方法 9 0 0 は、ブロックに進むことができ、シーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数が、擬似点閾値より少ない場合、方法 9 4 4。

40

【0 3 3 9】

決定ブロック 9 3 2 で、シーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数が、擬似点閾値以上である場合、非ユニーク分子標識をブロック 9

50

40で除去することができる。非ユニーク分子標識は、ブロック944でシーケンシングデータ中の標的に関連付けられた識別可能な配列を有するノイズ分子標識の数を決定するために、除去することができる。非ユニーク分子標識は、所定の再使用分子標識閾値より大きい、シーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識を含みうる。

【0340】

ブロック944で、分布ベースのエラー訂正方法を用いて、分子標識カウントを調節することができる。分布ベースのエラー訂正方法は、シーケンシングデータ中の標的に関連付けられた識別可能な配列を有するノイズ分子標識の数を決定する工程を含みうる。ノイズ分子標識の数を決定する工程は、2つのネガティブ二項分布、すなわち、シグナルネガティブ二項分布とノイズネガティブ二項分布とを、シーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数に当てはめる工程を含みうる。シグナルネガティブ二項分布は、シグナル分子標識である、カウントされたシーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数に対応する。ノイズネガティブ二項分布は、ノイズ分子標識である、カウントされたシーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数に対応する。ノイズ分子標識の数を決定する工程は、当てはめたシグナルネガティブ二項分布と、当てはめたノイズネガティブ二項分布とを用いて、ノイズ分子標識の数を決定する工程を含む。

【0341】

ブロック944で分布ベースのエラー訂正を用いて分子標識カウントを調節した後、方法1200は、任意選択で、ブロック1112に進む。ブロック1112で、ブロック944からの調節された分子標識カウントを、ブロック916で決定されて、方向近接性に基づき調節された分子標識カウントと合わせることができる。たとえば、非ユニーク分子標識は、ブロック940で除去されるため、ブロック944で分布当てはめには使用されない。しかし、これらの分子標識は、ブロック916で決定されて、方向近接性に基づき調節された分子標識カウント中に依然として存在する。従って、ブロック944からの調節された分子標識カウントと、ブロック944で調節された分子標識カウントを合わせて、ブロック948で出力を生成することができる。

【0342】

決定ブロック912で、シーケンシングデータ中の標的のシーケンシングステータスが、飽和シーケンシングステータスであれば、方法1200は、ブロック948に進んで、方向近接性および分布ベースのエラー訂正に基づいて分子標識を調節することなく、出力を生成することができる。決定ブロック924で、シーケンシングデータ中の標的のシーケンシングステータスが、過少シーケンシングステータスであれば、方法1200は、ブロック948に進んで、分布ベースのエラー訂正に基づいて分子標識を調節することなく、出力を生成することができる。たとえば、決定されるノイズ分子標識の数は、ゼロでありうる。方法1200は、ブロック952で終了する。

【0343】

分布当てはめのための初期パラメータ推定を用いた、方向近接性および分布ベースのエラー訂正に基づくPCRおよびシーケンシングエラーの訂正

図13は、再帰による再帰的置換エラー訂正および分布ベースのエラー訂正に基づくPCRおよびシーケンシングエラーの訂正の非限定的な例示的实施形態13を示すフローチャートである。本方法1300のブロック（たとえば、ブロック904～952）は、図9を参照にして説明されている。手短には、方法1300は、複数の確率バーコード付き標的のシーケンシングデータを受け取った後、ブロック904で開始する。いくつかの実施形態では、方法1300は、複数の確率バーコードを用いて、複数の標的に確率バーコードを付けて、複数の確率バーコード付き標的を生成する工程をさらに含み、ここで、複数の確率バーコードの各々は、分子標識を含む。いくつかの実施形態では、方法1300は、さらに、複数の確率バーコード付き標的をシーケンシングして、シーケンシングデータを取得する工程も含む。

【0344】

ブロック908で、複数の標的の1つ以上について：シーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数をカウントすることができる。決定ブロック912で、シーケンシングデータが、飽和シーケンシングステータスを有するか否かを決定することができる。シーケンシングデータが、決定ブロック912で、飽和シーケンシングステータスを有していなければ、方法1300は、ブロック916に進み、ここで、方向近接性に基づいて分子標識カウントを調節することができる。いくつかの実施形態では、方向近接性に基づいて分子標識カウントを調節する工程は、図7を参照にして説明することができる。

【0345】

ブロック920で、シーケンシングデータ中の標的のシーケンシングステータスを決定することができる。シーケンシングデータ中の標的のシーケンシングステータスは、過少シーケンシングを含むか、または過少シーケンシングでありうる。決定ブロック924で、シーケンシングデータ中の標的のシーケンシングステータスが、過少シーケンシングステータスであるか否かを決定することができる。

【0346】

決定ブロック924で、シーケンシングデータ中の標的のシーケンシングステータスが、過少シーケンシングステータスでなければ、方法1300は、ブロック928に進んで、分子標識カウントをフィルタリングすることができる。分子標識カウントをフィルタリングする工程は、決定ブロック932で、擬似点閾値より少ない、シーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数を決定する工程を含みうる。

【0347】

決定ブロック932で、シーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数が、擬似点閾値より少ない場合、方法900は、任意選択で、ブロック936に進み、そこで、シーケンシングデータ中の標的に関連付けられた識別可能な配列を有するノイズ分子標識の数を決定する前に、シーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数に擬似点を追加することができる。いくつかの実施形態では、シーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数が、擬似点閾値より少ない場合、方法900は、ブロックに進むことができ、シーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数が、擬似点閾値より少ない場合、方法944。

【0348】

決定ブロック932で、シーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数が、擬似点閾値以上である場合、非ユニーク分子標識をブロック940で除去することができる。

【0349】

ブロック944で分子標識カウントを調節する前に、ブロック1304で、2つのネガティブ二項分布の初期パラメータを任意選択で推定することができる。2つのネガティブ二項分布の初期パラメータは、さまざまな履行において異なるものであってよい。いくつかの実施形態では、2つのネガティブ二項分布の各々の平均および散布度は、1でありうる。いくつかの実施形態では、2つのネガティブ二項分布の平均および散布度は、ブロック928からのフィルタリング済分子標識カウントの空でない部分集合の平均および散布度であると推定されうる。たとえば、サブセットは、ブロック928からのフィルタリング済分子標識カウントの25%~75%分位でありうる。これらの分位の上限または下限は、さまざまな履行において異なるものであってよい。いくつかの実施形態では、上限または下限は、1%、2%、3%、4%、5%、6%、7%、8%、9%、10%、20%、30%、40%、50%、70%、80%、90%、99%、またはこれらの値のいずれか2つの間の数もしくは範囲であるか、またはこれらの近似値でありうる。いくつかの実施形態では、上限または下限は、少なくとも、または多くとも、1%、2%、3%、4%、5%、6%、7%、8%、9%、10%、20%、30%、40%、50%、70%、

10

20

30

40

50

80%、90%、99%、もしくは100%でありうる。

【0350】

ブロック944で、分布ベースのエラー訂正方法を用いて、分子標識カウントを調節することができる。分布ベースのエラー訂正方法は、シーケンシングデータ中の標的に関連付けられた識別可能な配列を有するノイズ分子標識の数を決定する工程を含みうる。ノイズ分子標識の数を決定する工程は、2つのネガティブ二項分布、すなわち、シグナルネガティブ二項分布とノイズネガティブ二項分布とを、シーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数に当てはめる工程を含みうる。シグナルネガティブ二項分布は、シグナル分子標識である、カウントされたシーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数に対応する。ノイズネガティブ二項分布は、ノイズ分子標識である、カウントされたシーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数に対応する。ノイズ分子標識の数を決定する工程は、当てはめたシグナルネガティブ二項分布と、当てはめたノイズネガティブ二項分布を用いて、ノイズ分子標識の数を決定する工程を含むことができる。

10

【0351】

ブロック948で、標的の数を推定して、近接性に基づくエラー訂正および分布ベースのエラー訂正後に出力を生成することができる。決定ブロック912で、シーケンシングデータ中の標的のシーケンシングステータスが、飽和シーケンシングステータスである場合、方法1300は、ブロック948に進んで、方向近接性および分布ベースのエラー訂正に基づいて分子標識を調節することなく、出力を生成することができる。

20

【0352】

決定ブロック924で、シーケンシングデータ中の標的のシーケンシングステータスが、過少シーケンシングステータスである場合、方法1300は、ブロック948に進んで、分布ベースのエラー訂正に基づいて分子標識を調節することなく、出力を生成することができる。たとえば、決定されたノイズ分子標識の数は、ゼロであってもよい。方法1300は、たとえば、ブロック952で終了する。

【0353】

図14は、初期パラメータ推定値のための2番目に高い分子標識を用いることによる、再帰的置換エラー訂正および分布ベースのエラー訂正に基づくPCRおよびシーケンシングエラーの訂正の非限定的な例示的实施形態を示すフローチャートである。本方法1400のブロック（たとえば、ブロック904～952）は、図9を参照にして説明されている。手短には、方法1400は、複数の確率バーコード付き標的のシーケンシングデータを受け取った後、ブロック904で開始する。いくつかの実施形態では、方法1400は、複数の確率バーコードを用いて、複数の標的に確率バーコードを付けて、複数の確率バーコード付き標的を生成する工程をさらに含み、ここで、複数の確率バーコードの各々は、分子標識を含む。いくつかの実施形態では、方法1400は、さらに、複数の確率バーコード付き標的をシーケンシングして、シーケンシングデータを取得する工程も含む。

30

【0354】

ブロック908で、複数の標的の1つ以上について：シーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数をカウントすることができる。決定ブロック912で、シーケンシングデータが、飽和シーケンシングステータスを有するか否かを決定することができる。シーケンシングデータが、決定ブロック912で、飽和シーケンシングステータスを有していなければ、方法1400は、ブロック916に進み、ここで、方向近接性に基づいて分子標識カウントを調節することができる。いくつかの実施形態では、方向近接性に基づいて分子標識カウントを調節する工程は、図7を参照にして説明することができる。

40

【0355】

ブロック920で、シーケンシングデータ中の標的のシーケンシングステータスを決定することができる。シーケンシングデータ中の標的のシーケンシングステータスは、過少シーケンシングを含むか、または過少シーケンシングでありうる。決定ブロック924で、

50

シーケンシングデータ中の標的のシーケンシングステータスが、過少シーケンシングステータスであるか否かを決定することができる。

【0356】

決定ブロック924で、シーケンシングデータ中の標的のシーケンシングステータスが、過少シーケンシングステータスでなければ、方法1400は、ブロック928に進んで、分子標識カウントをフィルタリングすることができる。分子標識カウントをフィルタリングする工程は、決定ブロック932で、擬似点閾値より少ない、シーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数を決定する工程を含みうる。

【0357】

決定ブロック932で、シーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数が、擬似点閾値より少ない場合、方法900は、任意選択で、ブロック936に進み、そこで、シーケンシングデータ中の標的に関連付けられた識別可能な配列を有するノイズ分子標識の数を決定する前に、シーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数に擬似点を追加することができる。

【0358】

決定ブロック932で、シーケンシングデータ中の標的に関連付けられた識別可能な配列を有する分子標識の数が、擬似点閾値以上である場合、非ユニーク分子標識をブロック940で除去することができる。

【0359】

ブロック944で、分布ベースのエラー訂正方法を用いて、分子標識カウントを調節することができる。分布ベースのエラー訂正方法のための初期パラメータは、分子標識のカウントに基づくものであってよい。たとえば、ネガティブ二項分布（たとえば、シグナルネガティブ二項分布もしくはノイズネガティブ二項分布）の一方の初期パラメータ（たとえば、平均および散布度）は、分子標識のカウントまたは分子標識の数の平均もしくはカウントに基づくものであってよい。この分子標識は、2番目に高いカウントの分子標識または任意の等級付け（たとえば、10番目に高いカウント）の分子標識であってもよい。分子標識の等級付けは、さまざまな履行において異なりうる。いくつかの実施形態では、等級付けは、1、2、3、4、5、6、7、8、9、10、20、30、40、50、60、70、80、90、100、またはこれらの値のいずれか2つの間の数もしくは範囲、またはこれらの近似値でありうる。いくつかの実施形態では、等級付けは、少なくとも、または多くとも、1、2、3、4、5、6、7、8、9、10、20、30、40、50、60、70、80、90、もしくは100でありうる。分子標識の数は、さまざまな履行において異なりうる。いくつかの実施形態では、分子標識の数は、1、2、3、4、5、6、7、8、9、10、20、30、40、50、60、70、80、90、100、またはこれらの値のいずれか2つの間の数もしくは範囲、またはこれらの近似値でありうる。いくつかの実施形態では、分子標識の数は、少なくとも、または多くとも、1、2、3、4、5、6、7、8、9、10、20、30、40、50、60、70、80、90、もしくは100でありうる。

【0360】

ブロック948で、標的の数を推定して、近接性に基づくエラー訂正および分布ベースのエラー訂正後に出力を生成することができる。決定ブロック912で、シーケンシングデータ中の標的のシーケンシングステータスが、飽和シーケンシングステータスである場合、方法1400は、ブロック948に進んで、方向近接性および分布ベースのエラー訂正に基づいて分子標識を調節することなく、出力を生成することができる。

【0361】

決定ブロック924で、シーケンシングデータ中の標的のシーケンシングステータスが、過少シーケンシングステータスである場合、方法1400は、ブロック948に進んで、分布ベースのエラー訂正に基づいて分子標識を調節することなく、出力を生成することができる。たとえば、決定されたノイズ分子標識の数は、ゼロであってもよい。方法1400は、ブロック952で終了する。

【0362】

シーケンシング

いくつかの実施形態では、確率バーコード付き標的の数を推定する工程は、標識標的、空間標識、分子標識、サンプル標識、細胞標識、またはその任意の産物（たとえば、標識アンプリコン、もしくは標識cDNA分子）の配列を決定する工程を含みうる。増幅された標的をシーケンシングに付すことができる。確率バーコード付き標的またはその任意の産物の配列を決定する工程は、サンプル標識の少なくとも一部、空間標識、細胞標識、分子標識、確率バーコード付き標的の少なくとも一部、その相補鎖、逆相補鎖、またはその任意の組合せの配列を決定するために、シーケンシング反応を実施する工程を含みうる。

【0363】

確率バーコード付き標的（たとえば、増幅された核酸、標識核酸、標識核酸のcDNAコピーなど）の配列の決定は、さまざまなシーケンシング方法を用いて実施することができ、そうした方法として、限定するものではないが、ハイブリダイゼーションによるシーケンシング（SBH）、ライゲーションによるシーケンシング（SBL）、定量的インクリメンタル蛍光ヌクレオチド付加シーケンシング（QIFNAS）、段階的ライゲーションおよび切断、蛍光共鳴エネルギー移動（FRET）、分子ビーコン、TaqManリポータプローブ消化、パイロシーケンシング、蛍光in situシーケンシング（FISSSEQ）、FISSSEQビーズ、ワブル（wobble）シーケンシング、多重シーケンシング、重合コロニー（POLONY）シーケンシング；ナノグリッドローリングサークルシーケンシング（ROLONY）、対立遺伝子特異的オリゴライゲーションアッセイ（たとえば、オリゴライゲーション（OLA）、ライゲートした線状プローブおよびローリングサークル増幅（RCA）読み出しを用いた単一テンプレート分子OLA、ライゲートした錠型（padlock）プローブ、またはライゲートした環状錠型プローブおよびローリングサークル増幅（RCA）を用いた単一テンプレート分子OLAなどが挙げられる。

【0364】

いくつかの実施形態では、確率バーコード標的またはその任意の産物の配列を決定する工程は、ペアエンドシーケンシング、ナノポアシーケンシング、ハイスループットシーケンシング、ショットガンシーケンシング、ダイターミネータシーケンシング、マルチプルプライマーDNAシーケンシング、プライマーウォーキングを含み、サンガー（Sanger）ジデオキシシーケンシング、マクサム・ギルバート（Maxam Gilbert）シーケンシング、パイロシーケンシング、真の単一分子シーケンシング、またはそれらの任意の組合せを含む。あるいは、確率バーコード付き標的またはその任意の産物の配列は、電子顕微鏡検査または化学感受性電界効果トランジスタ（chemFET）アレイにより決定することができる。

【0365】

Roche 454、Illumina Solexa、ABI-SOLID、ION Torrent、Complete Genomics、Pacific Bioscience、Helicos、またはPolonatorプラットフォームといったプラットフォームを用いた環状アレイシーケンシングなどのハイスループットシーケンシング方法も使用することができる。いくつかの実施形態では、シーケンシングは、MiSeqシーケンシングを含みうる。いくつかの実施形態では、シーケンシングは、HiSeqシーケンシングを含みうる。

【0366】

確率バーコード付き標的は、生物のゲノムの遺伝子の約0.01%～生物のゲノムの遺伝子の約100%を占める核酸を含みうる。たとえば、複数の多量体を含む標的相補領域を用いて、サンプル中の相補配列を含む遺伝子を捕捉することにより、生物のゲノムの遺伝子の約0.01%～生物のゲノムの遺伝子の約100%をシーケンシングすることができる。いくつかの実施形態では、確率バーコード付き標的は、生物のトランスクリプトームの転写物の約0.01%～生物のトランスクリプトームの転写物の約100%を占める核酸を含む。たとえば、ポリ（T）テールを含む標的相補領域を用いて、サンプルからm

10

20

30

40

50

RNAを捕捉することにより、生物のトランスクリプトームの転写物の約0.501%～生物のトランスクリプトームの転写物の約100%をシーケンシングすることができる。

【0367】

複数の確率バーコードの空間標識および分子標識の配列を決定する工程は、複数の確率バーコードの0.00001%、0.0001%、0.001%、0.01%、0.1%、1%、2%、3%、4%、5%、6%、7%、8%、9%、10%、20%、30%、40%、50%、60%、70%、80%、90%、99%、100%、またはこれらの値のいずれか2つの間の数もしくは範囲をシーケンシングする工程を含みうる。複数の確率バーコードの標識、たとえば、サンプル標識、空間標識、および分子標識の配列を決定する工程は、複数の確率バーコードの1、10、20、30、40、50、60、70、80、90、100、10³、10⁴、10⁵、10⁶、10⁷、10⁸、10⁹、10¹⁰、10¹¹、10¹²、10¹³、10¹⁴、10¹⁵、10¹⁶、10¹⁷、10¹⁸、10¹⁹、10²⁰、またはこれらの値のいずれか2つの間の数もしくは範囲をシーケンシングする工程を含みうる。複数の確率バーコードの一部または全部をシーケンシングする工程は、約、少なくとも、または多くとも、10、20、30、40、50、60、70、80、90、100、200、300、400、500、600、700、800、900、1000、2000、3000、4000、5000、6000、7000、8000、9000、10000、またはこれらの値のいずれか2つの間の数もしくは範囲のヌクレオチドまたは塩基のリード長の配列を生成する工程を含みうる。

【0368】

シーケンシング工程は、確率バーコード付き標的の少なくともまたは少なくとも約10、20、30、40、50、60、70、80、90、100以上のヌクレオチドまたは塩基対をシーケンシングする工程を含みうる。たとえば、シーケンシング工程は、複数の確率バーコード付き標的に対するポリメラーゼ鎖反応（PCR）増幅を実施することにより、50、75、もしくは100以上のヌクレオチドのリード長を有するシーケンシングデータを生成する工程を含みうる。シーケンシング工程は、確率バーコード付き標的の少なくともまたは少なくとも約200、300、400、500、600、700、800、900、1,000以上のヌクレオチドまたは塩基対をシーケンシングする工程を含みうる。シーケンシング工程は、確率バーコード付き標的の少なくともまたは少なくとも約1,500、2000、3000、4000、5000、6000、7000、8000、9000、もしくは10000以上のヌクレオチドまたは塩基対をシーケンシングする工程を含みうる。

【0369】

シーケンシング工程は、ラン当たり少なくとも約200、300、400、500、600、700、800、900、1,000以上のシーケンシングリードを含みうる。いくつかの実施形態では、シーケンシング工程は、ラン当たり少なくともまたは少なくとも約1,500、2000、3000、4000、5000、6000、7000、8000、9000、もしくは10000以上のシーケンシングリードを含みうる。シーケンシング工程は、ラン当たり約1,600,000,000以下のシーケンシングリードを含みうる。シーケンシング工程は、ラン当たり約200,000,000以下のリードを含みうる。

【0370】

サンプル

いくつかの実施形態では、1つ以上のサンプル中に複数の標識が含有されうる。1サンプルは、1つ以上の細胞、または1つ以上の細胞由来の核酸を含みうる。1サンプルは、単一細胞、または1細胞由来の核酸であってよい。1つ以上の細胞は、1つ以上の細胞型であってよい。1つ以上の細胞型の少なくとも1つは、脳細胞、心臓細胞、癌細胞、循環腫瘍細胞、器官細胞、上皮細胞、転移性細胞、良性細胞、一次細胞、循環細胞、またはそれらの任意の組合せである。

【0371】

10

20

30

40

50

本開示の方法に使用するためのサンプルは、1つ以上の細胞を含みうる。サンプルは、1つ以上の細胞を意味する。いくつかの実施形態では、複数の細胞は、1つ以上の細胞を含みうる。1つ以上の細胞型の少なくとも1つは、脳細胞、心臓細胞、癌細胞、循環腫瘍細胞、器官細胞、上皮細胞、転移性細胞、良性細胞、一次細胞、循環細胞、またはそれらの任意の組合せであってよい。いくつかの実施形態では、細胞は、癌組織、たとえば、乳癌、肺癌、結腸癌、前立腺癌、卵巣癌、膀胱癌、脳癌、黒色腫および非黒色腫皮膚癌などから切除された癌細胞である。いくつかの場合には、細胞は、癌に由来するが体液から採取される（たとえば循環腫瘍細胞）。癌の非限定的な例としては、腺腫、腺癌、扁平上皮細胞癌、基底細胞癌、小細胞癌、大細胞未分化癌、軟骨肉腫、および線維肉腫が挙げられる。サンプルは、組織、細胞単層、固定細胞、組織片、またはそれらの任意の組合せを含みうる。サンプルは、生体サンプル、臨床サンプル、環境サンプル、生体体液、組織、または被検者からの細胞を含みうる。サンプルは、ヒト、哺乳動物、イヌ、ラット、マウス、魚類、ハエ、蠕虫、植物、真菌、細菌、ウイルス、脊椎動物、または非脊椎動物から取得することができる。

【0372】

いくつかの実施形態では、細胞は、ウイルスに感染していてウイルスオリゴヌクレオチドを含有する細胞である。いくつかの実施形態では、ウイルス感染は、一本鎖（+鎖または「センス」）DNAウイルス（たとえば、パルボウイルス）、または二本鎖RNAウイルス（たとえば、レトロウイルス）などのウイルスにより引き起こされうる。いくつかの実施形態では、細胞は、細菌である。これらは、グラム陽性またはグラム陰性菌のいずれかを含みうる。いくつかの実施形態では、細胞は、真菌である。いくつかの実施形態では、細胞は、原生動物またはその他の寄生体である。

【0373】

本明細書で使用されるとき、「細胞」という用語は、1つ以上の細胞を意味しうる。いくつかの実施形態では、細胞は、正常細胞、たとえば、さまざまな発生段階のヒト細胞、またはさまざまな器官もしくは組織型に由来するヒト細胞である。いくつかの実施形態では、非ヒト細胞、たとえば、他のタイプの哺乳動物細胞（たとえば、マウス、ラット、ブタ、イヌ、ウシ、またはウマ）である。いくつかの実施形態では、細胞は、他のタイプの動物または植物細胞である。他の実施形態では、細胞は、任意の原核細胞または真核細胞でありうる。

【0374】

本明細書で使用されるとき、細胞は、細胞をビーズに関連付ける前にソートされる。たとえば、細胞は、蛍光活性化細胞ソーティングまたは磁気活性化細胞ソーティング、またはより一般的にはフローサイトメトリーによりソートすることができる。細胞はサイズ別に濾過することができる。いくつかの実施形態では、リテンテートは、ビーズに関連付けられる細胞を含有する。いくつかの実施形態では、フロースルーは、ビーズに関連付けられる細胞を含有する。

【0375】

サンプルは、複数の細胞を意味しうる。サンプルは、細胞の単層を意味しうる。サンプルは、薄い切片（たとえば、組織薄片）を意味しうる。サンプルは、一次元のアレイに配置することができる細胞の固体または半固体コレクションを意味しうる。

【0376】

データ解析および表示ソフトウェア

データ解析および標的の空間分解能の可視化

本開示は、確率バーコーディングおよび空間標識を使ってデジタルカウンティングを用いて標的の数および位置を推定する方法を提供する。本開示の方法から得られるデータはマップ上に可視化可能である。サンプルの標的の数および位置のマップは、本明細書に記載の方法を用いて生成された情報を用いて構築可能である。マップは、標的の物理的位置を決定するために使用可能である。マップは、複数の標的の位置を同定するために使用可能である。複数の標的は標的の同一種でありうるか、または複数の標的は複数の異なる標

10

20

30

40

50

的でありうる。たとえば、脳のマップを構築して複数の標的のデジタルカウントおよび位置を示すことが可能である。

【0377】

マップは、単一のサンプルのデータから生成可能である。マップは、複数のサンプルのデータを用いて構築可能であり、それにより組合せマップを生成可能である。マップは、何十、何百、および／または何千ものサンプルのデータで構築可能である。複数のサンプルから構成されるマップは、複数のサンプルに共通する領域に関連付けられる標的のデジタルカウントの分布を示すことが可能である。たとえば、レプリケートアッセイは同一のマップ上に表示可能である。少なくとも1、2、3、4、5、6、7、8、9、もしくは10レプリケートまたはそれ以上を同一のマップ上に表示（たとえばオーバーレイ）しう

10

【0378】

複数のサンプルからのデータを合わせることで、合わせたマップの位置的解像度を高めることができる。複数のサンプルの配向は、共通のランドマークにより登録することができ、そこで、サンプル全域に及ぶ個別の位置測定値は、少なくとも部分的に非連続的である。具体的な例は、ミクロトームを用いて、1つの軸上でサンプルを切断してから、別の軸に沿って第2のサンプルを切断するものである。合わせたデータベースは、標的のデジタルカウントを伴う三次元の空間位置を付与するであろう。前述のアプローチを多重化することにより、デジタルカウント統計学の高解像度三次元マップが可能になるであろう。

20

【0379】

機器システムのいくつかの実施形態では、システムは、単一細胞確率バーコーディングアッセイを行うことにより生成されたシーケンスデータセットのデータ解析を提供するためのコードを含むコンピュータ可読媒体を含むであろう。データ解析ソフトウェアにより提供しうるデータ解析機能の例としては、限定されるものではないが、(i) アッセイの実施時に生成された確率バーコードライブラリーをシーケンスすることにより提供されるサンプル標識、細胞標識、空間標識、分子標識、および標的シーケンスデータのデコーディング／デマルチプレクシングのためのアルゴリズム、(ii) リード数／遺伝子／細胞およびユニーク転写物分子数／遺伝子／細胞を決定するためのアルゴリズム、(iii) たとえば、遺伝子発現データにより細胞をクラスター化するためのまたは転写物分子数／遺伝子／細胞などの決定の信頼区間を予測するためのシーケンスデータの統計解析、(iv) たとえば、主成分分析、階層的クラスタリング、k平均値クラスタリング、自己組織化マップ、神経回路網などを用いて、希少細胞のサブ集団を同定するためのアルゴリズム、(v) 遺伝子配列データを既知の参照配列にアライメントするためのおよび突然変異、多型体マーカー、およびスプライス変異体を検出するための配列アライメント機能、ならびに(vi) 増幅またはシーケンシングエラーを補償するための分子標識の自動クラスタリングが挙げられる。いくつかの実施形態では、データ解析の全部または一部を行うために市販のソフトウェアを使用しうる。たとえば、全細胞コレクションで各細胞に存在する1遺伝子以上のコピー数の表を編集するためにSeven Bridges (<https://www.sbggenomics.com/>) ソフトウェアを使用しうる。いくつかの実施形態では、データ解析ソフトウェアは、有用なグラフ形式のシーケンシング結果、たとえば、細胞集団の各細胞に存在する1遺伝子以上のコピー数を示すヒートマップを出力するためのオプションを含みうる。いくつかの実施形態では、データ解析ソフトウェアは、たとえば、細胞集団の各細胞に存在する1遺伝子以上のコピー数と、あるタイプの細胞、あるタイプの希少細胞型、または特異的疾患もしくは病態を有する被験体に由来する細胞と、を相関付けることにより、シーケンシング結果から生物学的意味を抽出するためのアルゴリズムをさらに含みうる。ある実施形態では、データ解析ソフトウェアは、異なる生物学的サンプル全体にわたり細胞集団を比較するためのアルゴリズムをさらに含みうる。

30

40

50

【0380】

いくつかの実施形態では、データ解析機能のすべてを単一ソフトウェアパッケージ内にパッケージ化する。いくつかの実施形態では、データ解析能力の完全セットは、一式のソフトウェアパッケージを含みうる。いくつかの実施形態では、データ解析ソフトウェアは、アッセイ機器システムに依存せずにユーザーが利用可能なスタンドアロンパッケージでありうる。いくつかの実施形態では、ソフトウェアはウェブベースでありうるとともに、ユーザーによるデータの共有を可能しうる。

【0381】

いくつかの実施形態では、データ解析機能性のすべてを単一のソフトウェアパッケージ内にパッケージすることができる。いくつかの実施形態では、データ解析能力の完全セットは、一式のソフトウェアパッケージを含みうる。いくつかの実施形態では、データ解析ソフトウェアは、アッセイ機器システムとは独立に、ユーザーが利用可能なスタンドアロンパッケージであってよい。いくつかの実施形態では、ソフトウェアは、ウェブベースでありうるとともに、ユーザーによるデータの共有が可能になりうる。

【0382】

システムプロセッサおよびネットワーク

一般的には、本開示の機器システム方法にての使用に適したコンピュータまたはプロセッサは、図15に示すように、固定媒体1512を有するサーバー1509に任意選択的に接続可能な媒体1511またはネットワークポート1505から命令を読み取り可能な論理装置としてさらに理解しうる。システム1500は、図15に示すように、CPU1501、ディスクドライブ1503、キーボード1515やマウス1516などのオプションの入力デバイス、およびオプションのモニター1507を含みうる。データ通信は、ローカル位置またはリモート位置のサーバーに対して指定の通信媒体を介して達成可能である。通信媒体は、データを送受信する任意の手段を含みうる。たとえば、通信媒体は、ネットワーク接続、無線接続、またはインターネット接続でありうる。かかる接続は、World Wide Webによる通信を提供可能である。本開示に関するデータは、図15に示すように、かかるネットワークまたは接続を介してあるパーティー1522による受信または閲覧のために伝送可能である。

【0383】

図16が示すコンピュータシステム1600の第1のアーキテクチャー例の例示的な実施形態は、本開示の実施形態例との関連で使用可能である。図16が示すように、コンピュータシステム例は、処理命令用のプロセッサ1602を含みうる。プロセッサの例としては、限定されるものではないが、Intel Xeon（商標）プロセッサ、AMD Opteron（商標）プロセッサ、Samsung 32ビットRISC ARM 1176JZ（F）-S v1.0（商標）プロセッサ、ARM Cortex-A8 Samsung S5PC100（商標）プロセッサ、ARM Cortex-A8 Apple A4（商標）プロセッサ、Marvell PXA 930（商標）プロセッサ、または機能的に等価なプロセッサが挙げられる。実行のマルチスレッドは並列処理に使用可能である。いくつかの実施形態では、クラスター接続の単一コンピュータシステムであるか、または複数のコンピュータ、携帯電話、もしくは個人用携帯情報端末デバイスを含むネットワーク接続の分散システムであるかにかかわらず、複数のプロセッサまたは複数のコアを備えたプロセッサも使用可能である。

【0384】

図16が示すように、高速キャッシュ1604は、プロセッサ1602が最近使用したまたは頻繁に使用する命令またはデータに対する高速メモリーを提供するために、プロセッサ1602に接続または導入することが可能である。プロセッサ1602は、プロセッサバス1608によりノースブリッジ1606に接続可能である。ノースブリッジ1606は、メモリーバス1612によりランダムアクセスメモリー（RAM）1610に接続され、プロセッサ1602によりRAM1610へのアクセスを管理する。ノースブリッジ1606はまた、チップセットバス1616によりサウスブリッジ1614に

接続可能である。サウスブリッジ1614は、ひいては、周辺機器用バス1618に接続される。周辺機器用バスは、たとえば、PCI、PCI-X、PCI Express、または他の周辺機器用バスでありうる。ノースブリッジおよびサウスブリッジはプロセッサチップセットと呼ばれることが多く、プロセッサとRAMと周辺機器用バス1618上の周辺機器要素との間のデータ転送を管理する。いくつかの代替アーキテクチャーでは、ノースブリッジの機能、個別のノースブリッジチップを使用する代わりにプロセッサ中に組み込み可能である。

【0385】

いくつかの実施形態では、システム1600は、周辺機器用バス1618に結合されたアクセラレーターカード1622を含みうる。アクセラレーターは、ある特定の処理を加速するためにフィールドプログラマブルゲートアレイ(FPGA)または他のハードウェアを含みうる。たとえば、アクセラレーターは、アダプティブデータリストラクチャリングのために、または拡張セット処理で使用される代数式を評価するために、使用可能である。

10

【0386】

ソフトウェアおよびデータは、外部記憶装置1624に記憶され、プロセッサによる使用のためにRAM1610またはキャッシュ1604にロード可能である。システム1600は、管理システムリソース用のオペレーティングシステムを含む。オペレーティングシステムの例は、限定されるものではないが、Linux(登録商標)、Windows(登録商標)、MACOS(商標)、BlackBerry OS(商標)、iOS(商標)、および他の機能的に等価なオペレーティングシステム、さらには本発明の実施形態例に従ってデータ記憶および最適化を管理するためのオペレーティングシステムの上で動作するアプリケーションソフトを含む。

20

【0387】

この例では、システム1600はまた、ネットワークインターフェースカード(NIC)1620および1621を含み、ネットワーク接続記憶装置(NAS)などの外部記憶装置および分散並列処理に使用可能な他のコンピュータシステムへのネットワークインターフェースを提供する周辺機器用バスに接続される。

【0388】

図17は、本開示の方法での使用に好適な、複数のコンピュータシステム1702a、および1702b、複数の携帯電話および個人用携帯情報端末1702c、ならびにネットワーク接続記憶装置(NAS)1704a、および1704bを含むネットワーク1700の例示的な図を示す。実施形態例では、システム1712a、1712b、および1712cは、データ記憶を管理し、ネットワーク接続記憶装置(NAS)に記憶されたデータに対するデータアクセスを最適化することができる。データに数学モデルを使用することができ、分散並列処理コンピュータシステム1712a、および1712b、ならびに携帯電話および個人用携帯情報端末システム1712cを用いて評価することができる。コンピュータシステム1712a、および1712b、ならびに携帯電話および個人用携帯情報端末システム1712cはまた、ネットワーク接続記憶装置(NAS)1714aおよび1714bに記憶されたデータのアダプティブデータリストラクチャリングのために並列処理を提供可能である。図17は、一例を示すに過ぎず、多種多様な他のコンピュータアーキテクチャーおよびシステムが、本発明の種々の実施形態に関連して使用することができる。たとえば、ブレードサーバーを用いて、並列処理を提供することができる。プロセッサブレードは、並列処理を提供するためにバックプレーンを介して接続可能である。記憶装置はまた、バックプレーンに接続してもよいし、または個別ネットワークインターフェースを介してネットワーク接続記憶装置(NAS)として存在してもよい。

30

40

【0389】

いくつかの実施形態例では、プロセッサは、個別メモリー空間を保持可能であるとともに、ネットワークインターフェースを介してバックプレーンにまたは他のプロセッサによる並列処理のために他のコネクタにデータを伝送可能である。他の実施形態では、プロセッサの一部または全部は、共有仮想アドレスメモリー空間を使用可能である。

50

【0390】

図18に示すマルチプロセッサコンピュータシステム1800の例示的なブロック図は、実施形態例に従って共有仮想アドレスメモリ空間を使用する。システムは、共有メモリーサブシステム1804にアクセス可能な複数のプロセッサ1802a-fを含む。システムは、メモリーサブシステム1804中で複数のプログラマブルハードウェアメモリアルゴリズムプロセッサ(MAP)1806a-fを組込む。各MAP1806a-fは、メモリー1808a-fと1つ以上のフィールドプログラマブルゲートアレイ(FPGA)1810a-fとを含みうる。MAPは、設定可能な機能ユニットを提供し、特定のアルゴリズムまたはアルゴリズムの一部は、それぞれのプロセッサと緊密に連携して処理するためにFPGA1810a-fに提供可能である。たとえば、MAPは、データモデルに関する代数式を評価するためにおよび実施形態例でアダプティブデータリストラクチャリングを行うために使用可能である。この例では、各MAPは、こうした目的のためにすべてのプロセッサによりグローバルにアクセス可能である。一構成では、各MAPは、関連付けられたメモリー1808a-fにアクセスするためにダイレクトメモリアクセス(DMA)を使用可能であり、それにより、それぞれのマイクロプロセッサ1802a-fに依存せずにかつ非同期的に課題を実行可能になる。この構成では、MAPは、アルゴリズムのパイプライン実行および並行実行のために他のMAPに結果を直接供給可能である。

10

【0391】

以上のコンピュータアーキテクチャーおよびシステムは、単なる例にすぎず、一般のプロセッサ、共プロセッサ、FPGA、および他のプログラマブルロジックデバイス、システムオンチップ(SOC)、特定用途向け集積回路(ASIC)、および他の処理素子および論理素子の任意の組合せを使用するシステムを含めて、多種多様な他のコンピュータ、携帯電話、および個人用携帯情報端末のアーキテクチャーおよびシステムを実施形態例との関連で使用可能である。いくつかの実施形態では、コンピュータシステムの全部または一部は、ソフトウェアまたはハードウェアで実現可能である。任意のさまざまなデータ記憶媒体は、ランダムアクセスメモリー、ハードドライブ、フラッシュメモリー、テープドライブ、ディスクアレイ、ネットワーク接続記憶装置(NAS)、ならびに他のローカルまたは分散データ記憶デバイスおよびシステムを含めて、実施形態例との関連で、使用可能である。

20

30

【0392】

実施形態例では、本開示のコンピュータサブシステムは、以上のまたは他のコンピュータアーキテクチャーおよびシステムのいずれかで実行されるソフトウェアモジュールを用いて実現可能である。他の実施形態では、システムの機能は、ファームウェア、プログラマブルロジックデバイス、たとえば、フィールドプログラマブルゲートアレイ(FPGA)、システムオンチップ(SOL)、特定用途向け集積回路(ASIC)、または他の処理素子および論理素子で、部分的にまたは完全に実現可能である。たとえば、セットプロセッサおよびオプティマイザーは、アクセラレーターカードなどのハードウェアアクセラレーターカードを用いてハードウェアアクセラレーションで実現可能である。

【0393】

システムプロセッサおよびネットワーク

一般的には、図に示すように、本開示の機器システムに含まれるコンピュータまたはプロセッサは、固定媒体12を有するサーバー09に任意選択的に接続可能な媒体11またはネットワークポート05から命令を読み取ることができる論理装置としてさらに理解しうる。図に示すようなシステム00は、CPU01、ディスクドライブ03、キーボード15もしくはマウス16などのオプションの入力デバイス、およびオプションのモニター07を含みうる。データ通信は、ローカル位置またはリモート位置のサーバーに対して指定の通信媒体を介して達成することができる。通信媒体は、データを送受信する任意の手段を含みうる。たとえば、通信媒体は、ネットワーク接続、無線接続、またはインターネット接続でありうる。かかる接続は、World Wide Webによる通信を提供可能

40

50

である。図示される通り本開示に関するデータは、かかるネットワークまたは接続を介して、あるパーティー２２による受信または閲覧のために伝送することができる。

【０３９４】

図は、本開示の実施形態例との関連で 사용할 ことができるコンピュータシステム ００の第 １のアーキテクチャー例の例示的な実施形態を示す。図に示すように、コンピュータシステム例は、処理命令用のプロセッサ ０２を含みうる。プロセッサの非限定的な例としては、Intel Xeon（商標）プロセッサ、AMD Opteron（商標）プロセッサ、Samsung 32ビットRISC ARM 1176JZ（F）-S v 1.0（商標）プロセッサ、ARM Cortex-A8 Samsung S5PC100（商標）プロセッサ、ARM Cortex-A8 Apple A4（商標）プロセッサ、Marvell PXA 930（商標）プロセッサ、または機能的に同等のプロセッサが挙げられる。実行のマルチスレッドは、並列処理に使用可能である。いくつかの実施形態では、クラスター接続の単一コンピュータシステムであるか、または複数のコンピュータ、携帯電話、もしくは個人用携帯情報端末デバイスを含むネットワーク接続の分散システムであるかにかかわらず、複数のプロセッサまたは複数のコアを備えたプロセッサも使用可能である。

【０３９５】

図に示すように、高速キャッシュ ０４は、プロセッサ ０２が最近使用した、または頻繁に使用する命令またはデータに対する高速メモリを提供するために、プロセッサ ０２に接続または搭載することができる。プロセッサ ０２は、プロセッサバス ０８によりノースブリッジ ０６に接続可能である。ノースブリッジ ０６は、メモリバス １２によりランダムアクセスメモリ（RAM）に接続されて、プロセッサ ０２による RAM １０へのアクセスを管理する。ノースブリッジ ０６はまた、チップセットバス １６によりサウスブリッジ １４にも接続される。次いで、サウスブリッジ １４は、周辺機器用バス １８に接続される。周辺機器用バスは、たとえば、PCI、PCI-X、PCI Express、または他の周辺機器用バスであってよい。ノースブリッジおよびサウスブリッジは、プロセッサチップセットと呼ばれることが多く、プロセッサと、RAMと、周辺機器用バス １８上の周辺機器要素との間のデータ転送を管理する。いくつかの代替アーキテクチャーでは、ノースブリッジの機能を、個別のノースブリッジチップを使用する代わりに、プロセッサ内に搭載することができる。

【０３９６】

いくつかの実施形態では、システム ００は、周辺機器用バス １８に結合されたアクセラレーターカード ２２を含みうる。アクセラレーターは、ある特定の処理を加速するために、フィールドプログラマブルゲートアレイ（FPGA）または他のハードウェアを含みうる。たとえば、アクセラレーターは、アダプティブデータストライクチャリングのために、または拡張セット処理で使用される代数式を評価するために、使用することができる。

【０３９７】

ソフトウェアおよびデータは、外部記憶装置 ２４に記憶して、プロセッサによる使用のために RAM １０またはキャッシュ ０４にロードすることができる。システム ００は、管理システムリソース用のオペレーティングシステムを含み；オペレーティングシステムの例は、限定されるものではないが、Linux（登録商標）、Windows（登録商標）、MACOS（商標）、BlackBerry OS（商標）、iOS（商標）、および他の機能的に同等のオペレーティングシステム、さらには本発明の実施形態例に従ってデータ記憶および最適化を管理するためのオペレーティングシステムの上で動作するアプリケーションソフトを含む。

【０３９８】

この例では、システム ００はまた、ネットワークインターフェースカード（NIC） ２０および ２１を含み、これらは、ネットワーク接続記憶装置（NAS）などの外部記憶装置および分散並列処理に使用可能な他のコンピュータシステムへのネットワークインターフェースを提供する周辺機器用バスに接続されている。

【0399】

図は、複数のコンピュータシステム02a、および02b、複数の携帯電話および個人用携帯情報端末02c、ならびにネットワーク接続記憶装置（NAS）04a、および04bを含むネットワーク00の例示的な図を示す。実施形態例では、システム12a、12b、および12cは、データ記憶を管理すると共に、ネットワーク接続記憶装置（NAS）14aおよび14bに記憶されたデータに対するデータアクセスを最適化することができる。数学モデルをデータに使用し、コンピュータシステム12a、および12b、ならびに携帯電話および個人用携帯情報端末システム12c全体を介した分散並列処理を用いて評価することができる。コンピュータシステム12a、および12b、ならびに携帯電話および個人用携帯情報端末システム12cはまた、ネットワーク接続記憶装置（NAS）に記憶されたデータのアダプティブデータリストラクチャリングのために並列処理を提供することもできる。図は、一例を示すに過ぎず、多種多様な他のコンピュータアーキテクチャーおよびシステムを、本発明の種々の実施形態との関連で使用することができる。たとえば、並列処理を提供するために、ブレードサーバーを使用することができる。プロセッサブレードは、並列処理を提供するためにバックプレーンを介して接続することができる。記憶装置はまた、バックプレーンに接続してもよいし、または個別ネットワークインターフェースを介するネットワーク接続記憶装置（NAS）として存在してもよい。

10

【0400】

いくつかの実施形態例では、プロセッサは、個別メモリー空間を維持するとともに、他のプロセッサによる並列処理のためにネットワークインターフェース、バックプレーンにまたは他のコネクタを介してデータを伝送することができる。他の実施形態では、プロセッサの一部または全部は、共有仮想アドレスメモリー空間を使用することができる。

20

【0401】

図は、実施形態例に従って共有仮想アドレスメモリー空間を用いる、マルチプロセッサコンピュータシステム00の例示的なブロック図を示す。システムは、共有メモリーサブシステム04にアクセス可能な複数のプロセッサ02a～fを含む。システムは、メモリーサブシステム04中で複数のプログラマブルハードウェアメモリアルゴリズムプロセッサ（MAP）06a～fを搭載する。各MAP06a～fは、メモリー08a～fと、1つ以上のフィールドプログラマブルゲートアレイ（FPGA）10a～fとを含みうる。MAPは、設定可能な機能ユニットを提供し、それぞれのプロセッサと緊密に連携して処理するために、特定のアルゴリズムまたはアルゴリズムの一部をFPGA10a～fに提供することができる。たとえば、MAPを用いて、データモデルに関する代数式を評価するとともに、実施形態例でアダプティブデータリストラクチャリングを実施することができる。この例では、各MAPは、こうした目的のために、すべてのプロセッサによりグローバルにアクセス可能である。一構成では、各MAPは、ダイレクトメモリアクセス（DMA）を用いて、関連付けられたメモリー08a～fにアクセスすることができ、これによって、それぞれのマイクロプロセッサ02a～fから独立に、かつ非同期的に課題を実行することが可能になる。この構成では、MAPは、パイプライン処理およびアルゴリズムの並行実行のために他のMAPに結果を直接供給することができる。

30

【0402】

以上のコンピュータアーキテクチャーおよびシステムは、単なる例にすぎず、一般的プロセッサ、共プロセッサ、FPGA、および他のプログラマブルロジックデバイス、システムオンチップ（SOC）、特定用途向け集積回路（ASIC）、および他の処理素子および論理素子の任意の組合せを使用するシステムを含め、多種多様な他のコンピュータ、携帯電話、および個人用携帯情報端末のアーキテクチャーおよびシステムを実施形態例との関連で使用することができる。いくつかの実施形態では、コンピュータシステムの全部または一部は、ソフトウェアまたはハードウェアに実現可能である。ランダムアクセスメモリー、ハードドライブ、フラッシュメモリー、テープドライブ、ディスクアレイ、ネットワーク接続記憶装置（NAS）、ならびに他のローカルまたは分散データ記憶デバイスおよびシステムを含め、任意のさまざまなデータ記憶媒体を、実施形態例に関連して使

40

50

用することができる。

【0403】

実施形態例では、本開示のコンピュータサブシステムは、以上のまたは他のコンピュータアーキテクチャーおよびシステムのいずれかで実行されるソフトウェアモジュールを用いて実現可能である。他の実施形態では、システムの機能は、ファームウェア、プログラマブルロジックデバイス、たとえば、フィールドプログラマブルゲートアレイ（FPGA）、システムオンチップ（SOC）、特定用途向け集積回路（ASIC）、または他の処理素子および論理素子で、部分的にまたは完全に実現可能である。たとえば、セットプロセッサおよびオプティマイザーは、アクセラレーターカードなどのハードウェアアクセラレーターカードを用いてハードウェアアクセレーションで実現可能である。

10

【実施例】

【0404】

以上論述した実施形態のいくつかの態様を以下の実施例でさらに詳しく開示するが、これらの実施例は、本開示の範囲を何ら制限することを意図しない。

【0405】

実施例1

1 塩基置換エラーの訂正

本実施例は、1塩基置換を含むPCRまたはシーケンシングエラーの訂正を示す。1塩基置換を含むPCRまたはシーケンシングエラーは、類似の分子標識と、 3^8 のユニーク確率バーコードが存在した場合（ 4^8 のユニーク確率バーコードが存在した場合、17）、 ≤ 7 の発生数、すなわちシーケンシングリードとを有する標的のコピーを、複数の標的の同じ分子標識を有するものとして帰属させる工程により除去した。

20

【0406】

確率バーコードを付ける工程は、その結合領域としてオリゴ（dT）を有する 3^8 （6561）のユニーク確率バーコードの非枯渇プールを用いて、RT工程の前に、サンプル中にポリ（A）を有するmRNAを標識する工程を含みうる。標識する工程はランダムであってよく、各標的分子は、1つの確率バーコードにハイブリダイズすることができる。いずれの標的についても、標的分子の数が、確率バーコードの数よりはるかに小さければ、各標的分子は、恐らく異なる確率バーコードにハイブリダイズするであろう。従って、少数の標的分子しか存在しない場合、少数の標的分子は、恐らく、ハイブリダイゼーション中に、類似の分子標識（ML）を有する確率バーコードにハイブリダイズするであろう。

30

【0407】

3^8 の非枯渇ユニーク確率バーコードからの類似分子標識を有する少なくとも1対の確率バーコードをサンプリングする確率を計算した。2つの分子標識は、それらが1塩基相違する場合、類似の配列を有しうる。このサンプリングイベントは、確率バーコードが、実際に非枯渇でありうるため、置換を含むサンプリングとみなすことができる。この確率は、複数の標的を含む所与のサンプルについて存在する可能性が最も低い類似の分子標識を有する確率バーコードを推定する上で役立つ。問題は、類似の分子標識を有する少なくとも2つの確率バーコードが特定の確率で選択されるために必要な確率バーコードの数として明確に述べることができる。この問題は、 3^8 の識別可能な分子標識を仮定して、類似の配列を有する2つの確率バーコードの確率が0.5を超えるために必要とされる最小サンプルサイズとして明確に述べることができる。従って、この問題は、古典的な誕生日問題の一般化として考えることができる。古典的な誕生日問題は、365の異なる誕生日を仮定して、誕生日が同じ人が2人いる確率が0.5を超えるために必要な最小サンプルサイズを決定することができる。

40

【0408】

このサンプルサイズ r を得るために、 3^8 のユニーク確率バーコードからサンプリングされた r 個の確率バーコードを仮定し、その補集合事象の確率を用いて、少なくとも1対の類似の分子標識を有する確率を計算した。 3^8 のユニーク確率バーコードからランダムに1つの確率バーコードだけが選択された場合、確率バーコードは1つしかないため、その

50

分子標識が、他の確率バーコードの分子標識と類似していない確率、 $p_1 = 1$ である。第2の確率バーコードも 3^8 のユニーク確率バーコードからランダムに選択された場合、その分子標識が、第1の確率バーコードの分子標識と類似していない確率、 $p_2 = (3^8 - 16 - 1) / 3^8$ である。これは、確率バーコードの各位置に3つの考えられる塩基があると想定して、所与の分子標識について、各塩基位置が、2つの考えられる代替ヌクレオチドを有し、その結果、計 2×8 個の1塩基変異体が得られたからであった。第3の確率バーコードが、ユニーク分子標識を有する 3^8 のユニーク確率バーコードから、連続的にランダムに取り出された場合、その分子標識が、前の2つの分子標識と類似していない確率、 $p_3 = (3^8 - 1 - 16 - 1 - 16) / 3^8 = (3^8 - 2 \times 17) / 3^8$ である。確率バーコードは、 r 番目の確率バーコードまで、 3^8 のユニーク確率バーコードから連続的に取り出すことができる。この最後の確率バーコードが、前の確率バーコードと類似しない確率、 $p_r (3^8 - (r - 1) \times 17) / 3^8$ である。 r 個の確率バーコードはすべて独立に取り出されたため、いずれもが類似した配列を持たない確率バーコードを取り出す確率は、 P （類似の配列を有していない全分子標識） $= p_1 \times p_2 \times p_3 \times \dots \times p_r$ である。従って、ユニーク分子標識を有する 3^8 の確率バーコードからの r 個の確率バーコードの間で少なくとも1対の類似する確率バーコードを有する確率は、 P （類似の配列を有する少なくとも1対の分子標識） $= 1 - P$ （類似の配列を有していない全分子標識）であった。次に、この等式から、 P （類似の配列を有する少なくとも1対の分子標識）について望ましい値 $= 0.01, 0.05, 0.1$ 、または所望の値を設定することにより、サンプルサイズ r を計算した。

【0409】

【表1】

表1. 所与の数のユニーク分子標識に対する類似の分子標識を有する確率バーコードを観察する確率

N = 3^8 (6561)		N = 4^8 (65536)	
p	r	p	r
0.01	4	0.01	8
0.05	7	0.05	17
0.1	10	0.1	25
0.2	14	0.2	35
0.5	24	0.5	61
0.9	43	0.9	111

【0410】

表1は、 3^8 または 4^8 のユニーク分子標識を仮定して、 r 個の分子標識の間で少なくとも1つの類似する対を有する確率を示す。 3^8 のユニーク確率バーコードと、 ≤ 7 （ 4^8 のユニーク確率バーコードがある場合には、17）の確率バーコードが選択された場合、類似の分子標識を有する1対の確率バーコードを観察する確率は、0.05未満であり、これは無視できる。従って、この小さい確率により正当化されるように、類似の分子標識は、類似の確率バーコードのリアル見込み選択よりも人工物である可能性が高かったため、訂正することができる。

【0411】

しかし、7～24超の確率バーコードが存在した場合、類似の分子標識を有する1対以上の確率バーコードを観察する確率は、高くなる（たとえば、0.5）であろう。従って、これらの確率バーコードが、真であり、人工物ではないという確率を確信して除外することはできない。対照的に、一般的直観では、もし6561のユニーク可能性の大きなプールから24の確率バーコードだけが取り出されたら、いずれか1塩基のずれが、偶然ではなくシーケンシングエラーの結果でありうると、誤って結論付けられたであろう。

【0412】

たとえば、115の確率バーコードがランダムにサンプリングされた場合、算出される確率は、1つであるため、類似の分子標識を有する少なくとも1対の確率バーコードが存在することは100%確実である。サンプル中に115の標的があると想定して、ハイブリダイゼーションおよび逆転写プロセスの後に、類似の分子標識を有する2対の確率バーコードと、非類似の分子標識を有する111の確率バーコード（合計115の確率バーコード）が観察可能になる。しかし、シーケンシングデータ中に、類似の分子標識を有する3対の確率バーコードと、非類似の分子標識を有する110の確率バーコード（合計116の確率バーコード）が観察された場合には、類似の分子標識を有する2対の確率バーコードのみが真であり、3つ目の対は何らかのエラーにより生成された可能性。この100%の確率は、類似の分子標識を有する少なくとも1対の確率バーコードを観察するイベントは、確率バーコード付けの工程中に115の確率バーコードがランダムにサンプリングされた場合に起こり得ることを示すものであるが；これは、類似の分子標識の観察されたすべての対が真であることを意味するわけではない。類似の分子標識を有する確率バーコードは、確率バーコード付け工程、リアルもしくは真の分子標識から、またはPCRエラー、人工物、またはシーケンシングエラー、エラーもしくは偽の分子標識から生成されうる。従って、類似の分子標識が観察された場合、分子標識の特定の対が真であるか否かを決定するために、さらなる評価が必要であろう。さらに、総分子標識多様性を38から48に増加する場合、分子標識の類似の対を予測するために、各々の確率について、さらなる確率バーコードが必要となりうる。

【0413】

表2および表3は、ユニーク分子標識を有する ≤ 7 の確率バーコードが観察されたとき、そのような発生の確率は0.05未満であったため、類似の分子標識が発生する可能性は非常に低かったことを示す。従って、そうした類似の分子標識は、PCRエラー、人工物、またはシーケンシングエラーによって起こった可能性があり、これらは、分子標識カウントを訂正または調節するために、分子標識カウントから除去すべきである。従って、表2および表3中の真の分子標識の総数は、5から1に、ならびに7から6にそれぞれ減少させることができる。しかし、表4では、23のユニークバーコードが観察されており、これによって、類似の分子標識を有する少なくとも1対の確率バーコードを有する約50%の見込みが予測される。従って、類似の分子標識を有する16対の確率バーコードがリアルである可能性があり、類似の分子標識の各々の対は、それらがリアルであるか否かを確認するために、さらなる評価を要するであろう。

【0414】

【表2】

表2. ウェルH01中の10pgの入力RNAからのCD69未補正データ(最も豊富なバーコードから類似の分子標識4組が観察された)

遺伝子ID	ML	ML当たりの読取り数
CD69	TGTGCGTG	261
CD69	TGTGCGCG	2
CD69	TGTGCGGG	2
CD69	GGTGCGTG	1
CD69	TGGGCGTG	1

【0415】

【表 3】

表 3. ウェル A01 中の 500pg の入力 RNA からの CD64 未補正データ(類似の分子標識 1 組が観察された)。

遺伝子 ID	ML	ML 当たりの読取り数
CD4	CGTGTCTG	10
CD4	GGGGGCGA	7
CD4	GCTGCTGG	5
CD4	TCGGGCGA	4
CD4	CGCGTTCA	1
CD4	CGCGTTTA	1
CD4	TGGGCTTG	1

10

【 0 4 1 6 】

【表 4】

表 4. ウェル E11 からの 10pg の入力 RNA からの TFRC 未補正データ(最も豊富な 4 つの確率バーコードから、類似の分子標識 16 組が合計で観察された)

遺伝子 ID	ML	ML 当たりの読取り数
TFRC	CCGCTCTG	369
TFRC	CGGTGTTC	363
TFRC	CGGGGGTG	335
TFRC	CCCCCTTG	22
TFRC	CGGGGGGG	6
TFRC	CGGGGGCG	5
TFRC	CCGCGCTG	4
TFRC	CGGCGTTC	3
TFRC	CGGTGTCC	2
TFRC	TGGTGTTC	2
TFRC	CTGCTCTG	2
TFRC	GGCCGTGG	2
TFRC	TCGCTCTG	2
TFRC	CCGCCCTG	2
TFRC	CCCGCTTG	1
TFRC	CCGTTCTG	1
TFRC	GGGTGTTC	1
TFRC	CGTTGTTC	1
TFRC	GGGGGTGG	1
TFRC	GGGTCCGG	1
TFRC	CGGCTCTG	1
TFRC	CGGCGTCC	1
TFRC	CCGCTCCG	1

20

30

40

【 0 4 1 7 】

全体として、これらのデータから、観察された類似の分子標識を有する確率バーコードの数は、PCRエラー、人工物、またはシーケンシングエラーによって、恐らく、類似の分

50

子標識を有するこれらの確率バーコードが発生したため、除去されたことが明らかである。

【0418】

実施例2

シーケンシングデータ中の標的のクオリティステータスの決定

この実施例は、シーケンシングデータ中の標的のクオリティステータスが、完全シーケンシングクオリティステータス、不完全シーケンシングクオリティステータス、または飽和シーケンシングクオリティステータスであることを決定する工程を明らかにする。標的のクオリティステータスは、真またはリアル分子標識のすべてが観察されたか否かに依存した。

【0419】

実施例1に示すように、ライブラリー中に存在するユニーク分子標識を有する確率バーコードの完全なカウンティングは、シーケンシング深度に大きく依存しうる。シーケンシングが深いほど、すべての真の分子標識が観察される可能性が高くなった。浅いシーケンシングは、安価ではあるが、多数の分子標識を捉えそこなう可能性があり、また、多分に遺伝子検出感度も損ないうる。完全シーケンシングとは、標的分子を標識するために用いた確率バーコードの真の分子標識がすべて観察されたことを意味し、不完全シーケンシングは、真の分子標識の一部しか観察されなかったことを意味しうる。さらに、48568超の標的分子が発見サンプル中に存在した（これは、識別可能な確率バーコードの6561〜2×標準偏差に基づくポアソン訂正または調節後の分子数の下限である）ことも可能である。次に、飽和シーケンシングは、全分子標識の多様性に対する制限のために、標的分子の数が、決定しにくい場合に起こりうる。しかし、確率バーコード付けのための入力として少量のRNAを使用した場合、飽和シーケンシングが発生する可能性は低くなる。

【0420】

完全または不完全シーケンシングを数学的に定義するために、各々をエラーが一切ない理論上のモデルと比較した。完全な実験条件下で、出発サンプル中の標的分子の各コピーは、 j PCR サイクルおよび各サイクルで C 効率を仮定して、 $(1+C)^j$ 個のコピーを生成することができる。出発サンプル中の各バーコード付き分子について、イルミナシーケンシングは、オリジナルのバーコード付き分子から増幅された $(1+C)^j$ 個のクローナルコピーからのポアソンサンプリングとみなすことができる。理論上、同じ標的遺伝子の場合、 k 個の確率バーコード付き標的分子のシーケンシングは、すべての確率バーコード付き分子が、PCR 後均等に表現可能となりうることから、 $(1+C)^j$ 個のコピーからの反復ポアソンサンプリングとみなすことができる。ポアソンモデルの重要な想定は、平均値が分散と等しく、かつ、シーケンシングリードは、等散布に従うはずであることであった。散布は、分散／平均として定義することができる。

【0421】

実際に、完全シーケンシングは、通常、はるかに低いリード頻度でクラスター化したエラーを伴うことが多い。真の分子標識とは異なり、エラーは、すべてのPCRサイクルに参加しない可能性が高いため、ポアソンと比較してはるかに大きいリード頻度に変化をもたらすコピーが少なくなる。図19A～19Bは、完全および不完全シーケンシング遺伝子の例を示す。図19Aでは、最大シーケンシングリードは、最小シーケンシングリードの350倍を超えた。従って、完全シーケンシングは、ポアソンに比べて、大きい散布指数(>1)を示す傾向がある。

【0422】

対照的に、不完全シーケンシングの場合、ライブラリー中の真の分子標識を有する確率バーコードの一部だけがシーケンシングされたため、シーケンシングリードの変化は、ポアソンと比較して小さくなる。図19Bでは、最大シーケンシングリードは、最小シーケンシングリードの約3倍にすぎなかった。従って、不完全シーケンシングは、ポアソンよりも小さい散布指数(<1)を示す傾向がある。

【0423】

散布指数を計算する以外に、最も豊富な分子標識のシーケンシングリードを、シーケンシ

10

20

30

40

50

ングが完全であるか否かを決定するのに用いることができる。たとえば、豊富な分子指標のリードが25であり、散布指数が5であった場合、シーケンシングステータスは完全として分類することができ；そうでなければ、不完全として分類することができる。シーケンシングエラーが出現し始めるまで、シーケンシングは不完全である可能性があるため、25リードの閾値を使用することができる。いずれかの分子標識が25回超認められたら、シーケンシングエラーが生成される可能性がある。

【0424】

高度に豊富な遺伝子のシーケンシングデータが、確率バーコード中で飽和した、たとえば、ユニーク分子標識を有する 3^8 確率バーコードの場合、6557を超える状況下で、同じウェル内で他の低発現遺伝子のシーケンシング情報を、その遺伝子の散布指数および最大シーケンシングリードの計算に使用することができる。たとえば、同じウェル内で2番目に豊富な遺伝子が、確率バーコード中で飽和しておらず、かつ、不完全シーケンシングとして分類される場合、第1の遺伝子の飽和をリアルと考えることができ、分子の数を計算することはできない。そして、2番目に豊富な遺伝子が、完全シーケンシングとして分類された場合、第1の遺伝子の飽和は人工的である可能性があり、すべての確率バーコードの出現は、エラーによるものでありうる。次いで、真の分子標識の数を確認するために、ポアソンモデルベースの閾値化アルゴリズムを用いることができる。

【0425】

全体として、これらのデータは、シーケンシングステータスが、完全シーケンシング、不完全シーケンシング、または飽和シーケンシングであることを決定する工程を明らかにする。

【0426】

実施例3

完全にシーケンシングされた遺伝子の1塩基置換によるPCRまたはシーケンシングエラーの訂正

この実施例は、完全にシーケンシングされた遺伝子、すなわちシーケンシングデータ中の完全シーケンシングのクオリティステータスを有する遺伝子について1塩基置換によるPCRまたはシーケンシングエラーを訂正する工程を示す。この実施例はまた、シーケンシングデータ中の標的に関連付けられた真の分子標識および偽の分子標識を決定するために、標的、たとえば、遺伝子の分子標識を閾値化する工程も示す。

【0427】

ヌクレオチド当たりのシーケンシングエラー率は、0.1~1%に変動しうるとともに、通常、低頻度リードとして認めることができる。シーケンシングが深く進行するにつれて、多くのシーケンシングエラーが生成される可能性がある。たとえば、真のヌクレオチドシーケンシングエラーが0.5%であり、かつ、分子標識が100回シーケンシングされた場合、この分子標識に関連するシーケンシングエラーの予測数は、分子標識が8ヌクレオチド長であれば、 $100 * (1 - (1 - 0.5\%)^8)$ から計算して、約4でありうる。分子標識が300回シーケンシングされた場合には、シーケンシングエラーの予測数は、約12でありうる。これらのシーケンシングエラーは、カウントを増大する可能性がある人工的分子標識配列を生成しうる。これらの分子標識は、さらなる分析の前に、除去することができる。

【0428】

すべてのシーケンシングエラーのうち、1塩基エラーは、2塩基以上隔てたものよりもはるかに頻繁に起こりうる。1塩基シーケンシングエラーを有する確率は、サンプルサイズ8を含む二項分布および1塩基シーケンシングエラー率と等しい成功確率から導くことができる。1つの目標は、1塩基シーケンシングエラーを訂正することであった。1塩基シーケンシングエラーは、最も豊富かつ近接した（たとえば、ハミング距離に関して）分子標識、すなわち親分子標識の子供として考えることができる。シーケンシングエラーは、親分子標識の真の子供（すなわち、親分子標識から1塩基隔てた子供分子標識）を見出すことにより検出した。

【0429】

【表5】

表5. 10pgの入力RNA ウェル E11 から一塩基シーケンシングエラーを除去した後の更新された TFRC 配列データ

遺伝子 ID	ML	ML 当たりの読取り数
TFRC	CCGCTCTG	378
TFRC	CGGTGTTC	374
TFRC	CGGGGGTG	335
TFRC	CCCCCTTG	23
TFRC	CGGGGGGG	6
TFRC	CGGGGGCG	5
TFRC	CCGCGCTG	4
TFRC	GGCCGTGG	2
TFRC	GGGGGTGG	1
TFRC	GGGTCCGG	1
TFRC	CGGCGTCC	1

10

20

【0430】

親および子供分子標識の選択

親分子標識は、 >25 シーケンシングリードを有することが要求され、子供分子標識は、 3 以下のシーケンシングリードを有することが要求されうる。これらの要件は、下記の推論に基づくものであった。ヌクレオチドごとのシーケンシングエラーの確率が 0.5% と仮定する。分子標識が、 25 回シーケンシングされて、合計して 200 のヌクレオチドが生成された場合、 $200 \times 0.005 = 1$ であるから、 1 ヌクレオチドがエラーであることが予想された。従って、 25 のシーケンシングリードを有する各分子標識について、少なくとも 1 つの子供分子標識を有することが予想された。親分子標識は、 25 のシーケンシングリードを有するべきであると想定されうる。 4 のシーケンシングリードを有する子供分子標識は、シーケンシングエラーである可能性は低かった。これは、 1 つの分子標識中に同じエラーを 4 回導入する確率が、 $8 \times 0.005^4 = 10^{-9}$ であるためであった。もし、合計 106 シーケンシングエラーリードが存在した場合には、 4 回反復されたシーケンシングエラーの予想数は、 $5 \times 10^9 \times 10^6 = 0.005$ となり、これは無視することができた。従って、子供分子標識は、リード ≤ 3 を有するべきである。

30

【0431】

1 塩基隔てた親分子標識とその関連子供分子標識を仮定し、いかにして親の真のシーケンシングエラーである子供分子標識を決定するか？

親分子標識と、シーケンシングリード (R_{child1} , R_{child2} , ..., R_{childm}) を有する、親分子標識とは 1 塩基相違する子供分子標識のセットとを仮定し、多重二項検定を用いて、真の子供分子標識を同定することができる。帰無仮説の下で、真の子供分子標識の存在量は、 $R_{par} \times p$ 以下になるはずであり (数学的に、 $H_0: p < e/2$) ; そうでなければ、存在量は、 $R_{par} \times p$ より大きい ($H_A: p < e/2$) という別の仮説を支持する結論が下され、分子標識は、真の子供分子標識であったという仮説は拒絶されうる。次に、親分子標識とは 1 塩基相違する子供分子標識が 1 回観察された確率は、 $p = e/2$ となる。次に、数学的に、総存在量 ($R_{child} + R_{par}$) から、この子供分子標識を少なくとも R_{child} 回観察する確率 p_{child} は、以下の通りとなる：

40

【数1】

50

$$p_{child} = p(X \leq R_{child} | R_{child} + R_{par}, p = e/2) \\ = \sum_{j=R_{par}}^{j=R_{child}} \binom{R_{par} + R_{child}}{j} p^j (1-p)^{R_{par} + R_{child} - j}, \text{ 式(1)}$$

【0432】

子供分子標識が、実際に、その親分子標識のシーケンシングエラーであった場合、確率 p_{child} は、5%の臨界値より大きいはずである。複数の仮説を同時に検定するため、帰無仮説を拒絶するのに使用する臨界値は、5%レベルに制御される偽発見率（FDR）によって決定することができ、 p_{child} が、5%レベルのFDRより大きければ、仮説を容認することができる。5%に制御されたFDRを用いて、未調節のp値を、たとえば $p_1 \leq p_2 \leq p_m$ のように、小さい順にソーティングすることができる。次に、その対応する順位 j を含む検定を見出すことができる。 $p_{child} \leq j/m * 5\%$ であれば、この子供分子標識は、親分子標識の1塩基シーケンシングエラーであったという帰無仮説を容認することができる。

【0433】

全体として、これらのデータは、完全にシーケンシングされた遺伝子について1塩基シーケンシングエラーを訂正する工程を論証する：工程（1）、そのシーケンシングリードが25より大きければ、最も豊富なシーケンシングリードを有する分子標識を第1の親分子標識として選択する。工程（2）シーケンシングリード ≤ 3 を有する分子標識を選択し、第1の親分子標識と1塩基相違するこれらの分子標識を同定し、それらの子供分子標識と呼び；子供分子標識または1塩基子供分子標識が見出されなければ、工程（5）に進む。工程（3）、子供分子標識および親分子標識のすべてに対して多重二項検定を実施し、帰無仮説が容認される子供分子標識を除去して、それらのシーケンシングリードをその親分子標識に帰属させる。帰無仮説のいずれも容認されなかった場合、これは、すべての子供分子標識が親分子標識の1塩基シーケンシングエラーではなかったことを意味し、リード訂正を実施する必要はない。工程（4）、分子標識配列ならびにシーケンシングリードを更新する。工程（5）、親分子標識として次に大きいシーケンシングリードを有する分子標識を選択し、適格な親分子標識または適格な子供分子標識が残らなくなるまで、前述の工程を反復する。

【0434】

表5は、前述の分析を用いて、1塩基シーケンシングエラーを除去した後、更新されたTFRCシーケンシングデータを示す。分子標識の固有の数は、23（表4に示す）から11に減少した。

【0435】

閾値化のためのポアソンモデルの使用

シーケンシングエラーは、完全シーケンシングの下で出現する可能性が高くなりうる。1塩基シーケンシングエラーなどのいくつかのタイプのエラーは、訂正可能であるが、人工的分子標識のランダム組込みといった他のエラーは、配列類似性に基づいて訂正することができないであろう。その代わり、これらのタイプのエラーは、モデル化によって同定することができる。前述したように、完全シーケンシングは、ポアソンに対して過剰散布される傾向がある。従って、過剰散布を特徴とする2つの特有のポアソンモデルを作製した：1つは、真の分子標識（すなわち、確率バーコード付け工程中に、標的分子を標識するのに用いられる分子標識配列）のためのシーケンシングリードをモデル化するために使用することができ、2つ目のモデルは、エラー分子標識（すなわち、確率バーコード付け工程中に使用されないが、エラーのためにシーケンシング後に出現した分子標識配列）のために使用することができる。シーケンシングエラー率は、約0.1～1%であり、PCRサイクルエラー率は、約0.001%でありうる。PCRエラーは、PCRの後のサイク

ル中に、より多く起こって、低シーケンシングリードを有するエラー分子標識を生じうるが、すべての観察された分子標識配列の大部分に寄与しうる。従って、PCRおよびシーケンシングによって生じたエラーは、多くの場合、真の分子標識よりも低いシーケンシングリードを有しうる。そのため、真の分子標識のシーケンシングリードのポアソン平均は、エラー分子標識のポアソン平均より大きくなる。

【0436】

合計 k 個の識別可能な分子標識があり、それらのうち t 個が、 $BC_1, BC_2, \dots, BC_t$ のような真の分子標識であり、残りが、 $BC_{t+1}, BC_{t+2}, \dots, BC_k$ のようなエラー分子標識であったと想定する。こうした真の分子標識およびエラー分子標識にマッピングされたシーケンシングリードは、 $R_1, R_2, \dots, R_t$ および $R_{t+1}, R_{t+2}, \dots, R_k$ でありうる。さらに、真の分子標識およびエラー分子標識を用いたポアソン平均が、 μ_t および μ_n ($\mu_t > \mu_n$) であると想定すると、プロセス全体の確率は、以下のようになる：

【数2】

$$L_1(X; \mu_t, \mu_n) = (X; \mu_t) * (X; \mu_n) \\ = \prod_{i=1}^t P(X_i = R_i | \mu_t) \prod_{j=(t+1)}^k P(X_j = R_j | \mu_n), \quad \text{式(2)}$$

(式中、 $P(X_i = R_i | \mu_t)$ は、平均 μ_t を有するポアソン過程の下で、存在量 R_i を有する i 番目の分子標識を観察する確率を示す)。

【0437】

真の分子標識の数を決定するために、次のようにモデルの数を考慮した；すべての分子標識が真であると想定したモデル（従って、 $l = k$ ）から出発して；最も少ない分子標識が、エラーであり、他の分子標識はすべて真である（従って、 $l = k - 1$ ）であると想定した2番目のモデル；最も豊富な分子標識だけが真であり、他はすべてエラー分子標識である（従って、 $l = 1$ ）と想定した最後のモデルまで。最後に、最良のモデルは、考慮されるすべてのモデルの間で最も高い尤度を有するか、または最も小さい赤池情報量基準（AIC）と同等のものであり、AICは、所与のデータについて考えられるモデルの各々の相対量を測定することによって、モデル選択に使用することができる。数学的に、AICは、 $AIC = -\log L + 2p$ として定義され、式中、 p は、モデルで推定されるパラメータの数である。従って、 L_k および L_1 については、 $p = 1$ であり、他の場合、 $p = 2$ である。表6に示す例から、比較した8つの可能なモデルのうち、最も大きい3つのシーケンシングリードを有する3つの分子標識だけが真の分子標識であると考えられることがわかる。また、図20は、選択されたモデル（最も大きい3つ）から導かれた閾値が、明らかに真の分子標識を、エラーの可能性が大きいものから区別したことを示す。

【0438】

【表 6】

表 6. すべての可能なモデルと 10pg の入力 RNA ウェル E11 からの TFRC の対数尤度

遺伝子 ID	ML	ML 当たりの読取り数
TFRC	すべて	-1294
TFRC	最大 8	-952
TFRC	最大 7	-813
TFRC	最大 6	-660
TFRC	最大 5	-481
TFRC	最大 4	-263
TFRC	最大 3	-51
TFRC	最大 2	-656
TFRC	最大 1	-1036

10

【0439】

データは、1塩基シーケンシングエラーを除去し、ポアソンモデルを用いて閾値化することによって訂正された、完全にシーケンシングされた遺伝子のシーケンシングリードを示す。

20

【0440】

実施例 4

不完全にシーケンシングされた遺伝子の調節

この実施例は、ノイジー遺伝子を除去するとともに、ゼロ切断ポアソンモデルを用いて、ライブラリー中に存在することが予想される分子標識の総数を推定することにより不完全にシーケンシングされた遺伝子を調節する工程を示す。

【0441】

ノイジー遺伝子の除去

分子標識およびそのシーケンシングリードの統計学を考慮する以外に、遺伝子レベルの解析も有益となりうる。ある遺伝子について、検出された分子標識が非常に少なく、しかも各分子標識が、完全にシーケンシングされた遺伝子に比べて著しく低いリードを有する場合、その遺伝子をノイジーとみなすことができる。この想定は、同じライブラリー内の確率バーコード付き分子が、概ね同じ頻度で増幅およびシーケンシングされるはずであるという論証に基づくものであった。こうした期待は、各分子のシーケンシングの相違に起因する PCR およびシーケンシングバイアスによって影響されうるが、それらは PCR 中におけるサンプルの汚染や望ましくない分子の再結合などの事象によって発生する「ノイズ」に応じて小さいと想定されていた。遺伝子は、その増幅速度（分子標識当たりの平均リード）が、同じライブラリー中で完全にシーケンシングされた遺伝子に由来するエラーの増幅率と類似であった場合、ノイジーでありうる。

30

【0442】

具体的には、完全にシーケンシングされた遺伝子 g_1 が、全部で t_1 個の真の分子標識と e_1 個のエラー分子標識から構成され、それにより、 $R_{g_1,1}$ 、 $R_{g_1,2}$ 、 $\dots$ 、 R_{g_1,t_1} が、真の分子標識にマッピングされたシーケンシングリードであり、 $R^*_{g_1,1}$ 、 $R^*_{g_1,2}$ 、 $\dots$ 、 $R^*_{g_1,e_1}$ が、エラー分子標識にマッピングされたシーケンシングリードであると想定する。次に、 g_1 のエラー分子標識の増幅速度（EAMP）は、

40

【数 3】

$$EAMP_{g_1} = \sum_{i=1}^{e_1} R^*_{g_1,i} / e_1$$

50

であった。同様にして、他の完全にシーケンシングされた遺伝子すべての g_2 、 g_3 、 $\dots$ 、 g_x について、EAMP を計算することができる。観察された計 5 未満の分子標識を有する潜在的ノイジー遺伝子 g_1 、ならびに各分子標識にマッピングされた $R_{g'_1,1}$ 、 $R_{g'_1,2}$ 、 $\dots$ 、 $R_{g'_1,k}$ シーケンシングリードについて、カットオフを適用することができ、その増幅速度を

【数 4】

$$amp_{g'_1} = \sum_{i=1}^k R_{g'_1,i} / k$$

として決定する。 $amp_{g'_1}$ 中央 ($amp_{g'_1}$ 、 $amp_{g'_2}$ 、 $\dots$ 、 $amp_{g'_k}$) であれば、遺伝子 g_1 をノイジー遺伝子であると考えた。そうでなければ、これは、不完全遺伝子とみなすことができる。同様に、他のノイジー遺伝子も検定し、除去した。5 の分子標識をカットオフとして選択した理由は、低い増幅速度を有する遺伝子を 2 つの個別のケース：人工物（5 未満の分子標識が観察されたもの）と不完全シーケンシング（低 PCR / シーケンシングのプライマー失敗により ≥ 5 の分子標識が観察されたもの）に処理することが望ましいと思われるためである。

【0443】

ゼロ切断ポアソンモデルを用いた推定

シーケンシングが不完全であったとき、エラーはデータ中に依然として存在しうるが、全体として不十分なシーケンシングリードのために同定することが困難となりうる。シーケンシングが浅く、ライブラリー中に存在する分子標識のすべてが観察されていない場合、重要な分析のためにいくつかの想定が必要となりうる。すべての観察された分子標識が真であること、ならびに観察されていない真の分子標識が、ゼロで切断されている、すなわち、ゼロ時間で観察された切断分子標識であると想定することができる。所与の遺伝子について確率バーコード付き転写物のすべてがシーケンシングにサンプリングされているわけではないが、検出された分子標識のリードの頻度を用い、ゼロ切断ポアソンモデルを適用することにより、全ライブラリー中に存在する分子標識の完全な多様性を推定することができる。

【0444】

リード (R_1 、 R_2 、 $\dots$ 、 R_k) を有する k 個の識別可能な分子標識が観察され、($S - k$) 個の分子標識が観察されず、リードはゼロであった。1 つの目標は、 S 、すなわち、ライブラリー中に存在することが予想される分子標識の総数を推定することであった。ポアソン平均 μ によりゼロで切断されたポアソン変量として、シーケンシングリード 1、2、3、もしくはそれ以上とを認める頻度を想定し、すべてのシーケンシングリードの合計が n であった場合、尤度は次のように表すことができる：

$$L(S, \mu) \propto S! / (S - k)! \mu^k \exp(-S\mu) \quad (\text{式 3})$$

【0445】

伝統的な推理方式を μ 、 S およびそれらの標準誤差の推定のために適用することができる。 μ の最大尤度 (MLE) は、 n/S であり、 S の MLE への近似値は、 $k / (1 - e^{-n/S})$ または $k / (1 - (1 - 1/S)^n)$ となりうる。図 21 は、分子標識の数およびそれらの対応するシーケンシングリードに基づく、当てはめゼロ切断ポアソンモデルを示す。図 21 に示すように、33 のユニーク分子標識が、部分的にシーケンシングされたライブラリー中の計 39 のリード全体にわたって観察された。シーケンシングリード 1、2、3、および 4 を有する分子標識の頻度に基づき、ポアソンモデルを適用して、全ライブラリー中の計 113 の分子標識が、完了まで進行したシーケンシングを有することを推定した。推定方式は、 μ 、 S およびそれらの標準エラーの推定のために適用した。 μ の MLE は、 n/S であり、 S の MLE への近似値は、 $k / (1 - e^{-n/S})$ または $k / (1 - (1 - 1/S)^n)$ となりうる。

【0446】

全体として、これらのデータは、ノイズ遺伝子を除去するとともに、ライブラリー中に存在することが予想される分子標識の総数を推定するために、ゼロ切断ポアソンモデルを用いることによって訂正された不完全シーケンシング遺伝子のシーケンシングリードを明らかにする。

【0447】

実施例 5

完全シーケンシング遺伝子および不完全シーケンシング遺伝子

この実施例は、完全シーケンシング遺伝子および不完全シーケンシング遺伝子のシーケンシングリードを調節した後に生成されたアウトプットの一例を示す。

【0448】

表 7 は、完全シーケンシング遺伝子および不完全シーケンシング遺伝子のシーケンシングリードを調節した後に生成されたアウトプットの一例を提供する。列の見出しの説明は次の通りであった：「遺伝子 ID」は、検出された遺伝子の名称を示す。「シーケンシングステータス」は、3つの考えられる結果：完全、不完全および飽和を示し、これによって、解析方法が決定される。分類は、散布指数、および最も豊富な分子標識 (ML) にマッピングされたシーケンシングリードに応じて実施した。「未補正 ML」は、その遺伝子について観察されたユニーク分子標識のカウントを示す (非検出遺伝子の場合には「0」)。

「未補正リード」は、未補正 ML にマッピングされたシーケンシングリードの合計を示す (非検出遺伝子の場合には「0」)。「訂正 ML」は、アルゴリズムを適用した後に真の分子標識とみなされたユニーク分子標識のカウントを示す (完全シーケンシング遺伝子の場合のみ、不完全遺伝子の場合には「NA」、ノイズおよび非検出遺伝子の場合には「0」)。「訂正リード」は、訂正 ML にマッピングされたシーケンシングリードの合計を示す (完全シーケンシング遺伝子の場合のみ、不完全遺伝子の場合には「NA」、ノイズおよび非検出遺伝子の場合には「0」)。「補外 ML」は、ゼロ切断ポアソンモデルによるユニーク分子標識の推定数を示す (不完全シーケンシング遺伝子の場合のみ、完全遺伝子の場合には「NA」、ノイズおよび非検出遺伝子の場合には「0」)。「推定 Mol」は、訂正 ML (完全シーケンシング遺伝子の場合) または補外 ML (不完全シーケンシング遺伝子の場合) に基づいて推定された分子の数を示し、ノイズ遺伝子および非検出遺伝子の場合には「0」である。「推定 Mol LB」は、分子の推定数の下限を示す。「推定 Mol UB」は、分子の推定数の上限を示す。

【0449】

表 7 において、出発分子の推定数である、推定 Mol (n) は、次のように計算した：

$n = -m \log(1 - k/m)$ 、式 (4)

(式中、 m は、分子標識 (38) の全多様性であり、 k は、観察されたユニーク分子標識の総数であった)。 n の分散である $\text{var}(n)$ は、テイラー展開を用いて導かれた： $\text{var}(n) = (m / (m - k))^2 \text{var}(k)$ (式中、 $\text{var}(k)$ は、 $m * (1 - (1 - 1/m)^n) - (1 - 1/m)^{n+1} - m(m-1) * ((1 - 2/m)^n - (1 - 1/m)^{2n})$ として表すことができる)。出発分子の推定数の下限および上限 (推定 Mol LB および推定 Mol UB) は、

【数 5】

$$n \pm 2\sqrt{\text{var}(n)}$$

を用いて計算した。

【0450】

全体として、これらのデータは、完全シーケンシング遺伝子および不完全シーケンシング遺伝子を調節する工程を明らかにする。

【0451】

【表 7】

表 7. 10pg の入力 RNA ウェル A06 からの出力例

遺伝子 ID	シーケンシング グステータス	未補正 リード	未補正 ML	訂正リード	訂正 ML	外挿 ML	推定モル LB	推定モル UB
DAP_PRE LABELED	不完全	1	1	NA	NA	NA	NA	NA
PHE_PRE LABELED	完全	3129	59	3059	11	NA	11.01	10.83
KAN_PRE LABELED	完全	1418	71	1259	8	NA	8	7.87
CD4	不完全	2	2	0	0	0	0	0
FOXP3	不完全	0	0	0	0	0	0	0
STAT5A	完全	96	4	94	1	NA	1	1
FOXO1	不完全	1	1	0	0	0	0	0
FOXO3	不完全	3	3	0	0	0	0	0
CD45RA	不完全	0	0	0	0	0	0	0
CD45RO	不完全	0	0	0	0	0	0	0
MKI67	不完全	1	1	0	0	0	0	0
GAPDH	完全	12059	709	10984	38	NA	38.11	37.45
MYC	完全	2842	63	2765	7	NA	7	6.89
LRRC32	不完全	0	0	0	0	0	0	0
IL1R1	完全	177	1	177	1	NA	1	1
CCR8	不完全	10	1	NA	NA	1	1	1
IL-26	不完全	0	0	0	0	0	0	0

【0452】

実施例 6

完全シーケンシング遺伝子および不完全シーケンシング遺伝子の訂正の性能

この実施例は、完全シーケンシング遺伝子のシーケンシングリードの訂正の性能を示す。

この性能は、除去された未補正分子標識カウントおよび除去されたシーケンシングリードのエラーおよびノイズに基づいた。

【0453】

いくつかの完全シーケンシング遺伝子を選択して、完全シーケンシング遺伝子のシーケンシングリードの訂正の性能を検定した。表 8 は、シーケンシングリードを訂正または調節する前、およびその後に、これらの遺伝子についていくつかの測定値を比較する。未補正 ML、未補正リード、訂正 ML、訂正リードは、出力表から直接導入した。未補正 amp（未補正データを用いた増幅速度）およびフィルタリング amp（訂正後の真の分子標識データを用いた増幅速度）を、（未補正リード／未補正 ML）および（訂正リード／訂正 ML）を用いて計算した。観察された分子標識の総数の訂正後に、真の分子標識の数に対する、保持された ML のパーセンテージは、 $100 \times \text{訂正 ML} / \text{未補正 ML}$ であり、保持された % リードも同様に、 $100 \times \text{訂正リード} / \text{未補正リード}$ として定義された。表 8 は、より多い分子標識および総リードを呈示する GAPDH および ACTB を含む、さまざまな存在量レベルの遺伝子例を示す。訂正を適用した後の真の分子標識の数は、未補正データに認められる総分子標識の 7 % 未満を占めたが、これは、分子標識の 93 % 超が、エラー分子標識であると考えられ、廃棄されたことを意味する。未補正分子標識の 93 % は、ノイズとして除去されたが、真の分子標識は、リードの少なくとも 72 % に寄与し、これは、これらの廃棄されたエラー分子標識が、はるかに低いリードから成ることも意味する。さらに、アルゴリズムを適用した後の増幅速度は、137～413 の範囲であり、これは、未補正データを用いて得られたもの（6.1～29.4）よりはるかに高かった。訂正増幅速度は、はるかに実測的な測定値であり、これは、少なくとも 75 % の PCR 効

率と相関した。

【0454】

【表8】

表8. アルゴリズムを使用する前後のデータセット D704(10pg の RNA 入力)およびカスタマー1(単細胞、未知 RNA 入力)からの選択遺伝子の比較

	データセット	D704			カスタマー1		
	ウェル	H01	H01	H01	D01	H01	H12
	遺伝子 ID	GAPDH	MYC	TFRC	LEFTY1	SOX17	ACTB
前	未補正 ML	760	76	45	50	28	2070
	未補正リード	11107	1710	1324	434	297	12625
	未補正増幅	14.61	22.5	29.4	8.68	10.61	6.1
後	訂正 ML	37	4	3	1	1	66
	訂正リード	9351	1589	1239	387	226	9039
	ML 保持%	5%	5%	7%	2%	4%	3%
	リード保持%	84%	93%	94%	89%	76%	72%
	フィルタリングされた増幅	252.73	397.25	413	387	226	137

【0455】

全体として、これらのデータは、完全シーケンシング遺伝子のシーケンシングリードの訂正が、シーケンシングリードの大部分を使用する能力を依然として維持しながら、未補正分子標識カウンティングデータ中のエラーおよびノイズを有意に低減したことを示す。

【0456】

実施例7

確率バーコード付き標的のカウンティングデータを要約および視覚化するためのツール
この実施例は、前の実施例に示される確率バーコード付き標的のカウンティングデータを要約および視覚化するためのツールを示す。

【0457】

検定データのために、Precise (商標) assay (Cellular Research, Inc. (Palo Alto, CA)) による処理のために、単一細胞の2つのプレートを作製した。この実験では、2つの異なる細胞型を4:1比で使用し、各ウェルに配置した細胞のアイデンティティは、実験を実施する研究員から不明にした。この試験の目標は、確率バーコードカウントからの遺伝子発現プロファイルを用いて、各ウェルの細胞型を同定することであった。

【0458】

ウェルにおける全体的シーケンシングデータクオリティを評価するために、ウェル毎のシーケンシングリードの合計を算出した。そして、訂正方法の性能を評価するために、訂正方法の適用前および適用後のいくつかの統計学的測定値を集計し、比較した。さらに、グラフ図は、データの視覚的表示を提供し、異常またはパターンを容易に検出することが可能である。

【0459】

図9および10は、シーケンシングリード<5000 (イタリック体)を含むプレート1のウェル当たりのシーケンシングリードの合計を示す。リード<5000などのはるかに低いリードを有するウェルは、単一細胞がウェルに割り当てられなかったことを示しうるため、さらなる解析では、これらのウェルを除外すべきである。

【0460】

10

20

30

40

50

【表 9】

表 9. プレート 1 に対するウェル当たりのシーケンシングリードの合計

	1	2	3	4	5	6	7	8	9	10	11	12
A	58346	59687	141814	57269	106511	26894	5908	40547	16783	19896	8993	3885
B	3894	24164	52131	40458	61725	55568	3701	5189	3353	2690	3799	3848
C	98195	175593	3627	4508	22856	17253	3435	49083	29189	28969	4570	25593
D	59960	29342	27304	22649	29341	55575	2861	2298	78385	56112	3357	15843
E	18148	49333	25546	40571	4190	31746	111060	3956	57039	12688	13917	31934
F	37105	34979	88619	5457	3552	6499	2951	2874	3127	3118	166177	2848
G	11930	119412	2179	30913	29445	6002	4260	41497	16535	48453	15058	17844
H	29988	28987	3414	2548	74279	4609	4164	4043	35764	2911	3276	38723

10

【0 4 6 1】

【表 1 0】

表 10. プレート 2 に対するウェル当たりのシーケンシングリードの合計

	1	2	3	4	5	6	7	8	9	10	11	12
A	2291	45737	124283	95919	50637	3822	2770	4147	20528	93362	24443	29416
B	26870	26126	3350	42649	63897	2960	71319	13673	40682	28120	2439	3387
C	52029	51582	232	20538	28556	18871	49769	24187	43879	17136	106123	17861
D	41845	124785	26143	4836	2740	18416	37201	90892	87375	21552	29307	31768
E	22866	29578	8506	76852	15211	18138	4087	19625	17151	22380	15928	2242
F	16868	46605	21848	53195	3391	94041	29196	19468	38771	2801	5703	2735
G	22187	30552	20213	3010	3143	4578	1901	2873	163669	34898	60463	19940
H	30216	25110	112489	34535	3976	32726	17868	9807	58375	22972	38446	18290

20

【0 4 6 2】

表 1 0 および 1 1 は、訂正方法の前および後のいくつかの測定値を比較する。これらの表から、「未補正リード」（ウェル当たりのシーケンシングリードの合計）と「未補正 M L」（ウェル当たりの分子標識カウントの総数）に大きな変動が認められた。この大きな変動は、それらの標準偏差（S D）が平均より大きいことに起因しうるが、これも、低リードウェルの存在を示すものである。この方法を用いた後、ウェル当たり約 4 7 % の遺伝子を、存在する全遺伝子の中で完全シーケンシング遺伝子として分類した。遺伝子の大部分が、不完全シーケンシング遺伝子として分類された（たとえば、0 %）場合、本方法は、データ中のノイズを除去しない可能性がある。各ウェルについて、完全遺伝子の訂正後に約 1 5 % の分子標識が保持されたが、これらの分子標識は、平均 9 5 % のシーケンシングリードにマッピングされた。保持された % リードの値が高いほど、ノイズを除去しながら、訂正方法はシグナル（真の分子標識から賦与されたリード）を効果的に捕捉することができる。また、真の分子標識として保持される各分子標識の増幅速度は、1 6 3 . 3 2 であり、訂正方法を適用する前の 2 2 . 7 6 よりはるかに高かった。

30

40

【0 4 6 3】

50

【表 1 1】

表 11. プレート 1 からのデータを使用してアルゴリズムを適用する前後の異なる測定値を使用した統計の概要

	測定値/統計	平均	SD	中央	最小	最大
前	遺伝子	35.31	11.41	35.5	16	60
	未補正 ML	1184.54	1202.71	837	137	5617
	未補正リード	31110.6	35808.4	21272.5	2179	175593
	増幅	22.76	7.42	24.08	4.45	35.53
後	完全遺伝子%	0.47	0.2	0.54	0.1	0.79
	MI 保持%	0.15	0.03	0.16	0.07	0.22
	リード保持%	0.95	0.02	0.96	0.86	0.98
	訂正増幅	163.32	19.36	165.47	116.96	194.86

10

【0 4 6 4】

【表 1 2】

表 12. プレート 2 からのデータを使用してアルゴリズムを適用する前後の異なる測定値を使用した統計の概要

	測定値/統計	平均	SD	中央	最小	最大
前	遺伝子	36.52	9.92	37.5	16	57
	未補正 ML	1159.13	1050.72	851.5	79	5303
	未補正リード	32602.8	32144.1	22919	232	163669
	増幅	25.25	6.44	27.72	2.94	34.09
後	完全遺伝子%	0.53	0.2	0.6	0.05	0.78
	MI 保持%	0.17	0.04	0.17	0.11	0.5
	リード保持%	0.95	0.01	0.95	0.89	0.98
	訂正増幅	161.63	24.06	163.88	27	195.74

20

【0 4 6 5】

図 2 2 は、ウェル当たりの総シーケンシングリードの棒グラフを示す。図 2 2 は、9 6 ウェル全体の相対入力 of 直接の視覚化を達成する。この図から、ウェル C 0 2 および F 1 1 が、他に比べて高いリードを有することがわかり、これは、これらのウェルについての多細胞を示しうる。ウェル A 1 2、B 0 1、B 0 7 ~ B 1 2、C 0 3、C 0 4、C 0 7、C 1 1、D 0 7、D 0 8、D 1 1、E 0 5、E 0 8、F 0 4 ~ F 1 0、F 1 2、G 0 3、G 0 7、H 0 3、H 0 4、H 0 7 ~ H 0 9、H 1 0 ~ H 1 1 は、他のウェルに比べてはるかに低いリードを有するが、これは、これらのウェルに細胞が配置されなかったことを示しうる。

30

【0 4 6 6】

図 2 3 は、%完全シーケンシング遺伝子、真の分子標識として保持された%分子標識 (ML)、および各ウェルについて保持された ML にマッピングされた%保持リードの棒グラフを示す。図 2 3 は、ノイズ (各ウェルの下段) を除去するために、訂正方法を適用することができる、完全として分類された遺伝子のウェル当たりのパーセンテージ；分子標識を用いたウェル当たりのノイズのレベル (訂正方法の適用の前に観察された分子標識に対して、訂正方法の適用後に真の分子標識とみなされた、分子標識のパーセンテージ、各ウェルの上段)；ならびにシーケンシングリードを用いたウェル当たりのノイズのレベル (全未補正リードに比して、真の分子標識にマッピングされたリードのパーセンテージ、各ウェルの中段) を示す。図示するように、完全シーケンシング遺伝子の%はウェルに応じ

40

50

て変動するが、ウェル A 1 2、B 0 1、B 0 7 ~ B 1 2、C 0 3、C 0 4、C 0 7、D 0 7、D 0 8、D 1 1、E 0 5、E 0 8、F 0 4 ~ F 1 0、F 1 2、G 0 3、G 0 7、H 0 3、H 0 6、H 0 7、H 1 0 ~ H 1 1 はでははるかに低く、これは、はるかに低いリードを有するウェルと一致した。上段により示される % 保持 M L は、すべてのウェルで概して 2 0 % 未満であったが、中段により示される % 保持リードは、すべてのウェルで 9 0 % を超えた。このタイプのプロットは、ノイズを除去する上で、また一方では各ウェルのシグナルを最大化する上でも、訂正方法がどれくらい有効であるかについての概念を提供する。

【 0 4 6 7 】

図 2 4 は、各ウェルについて遺伝子により変動する % 保持リードの箱ひげ図を示す。遺伝子レベルでの箱ひげ図は、ウェル中の各遺伝子について訂正方法がどれくらい良く作用したかなどの詳細な情報を明らかにし、これは、ウェルレベルでの棒グラフでは表すことができない。図 2 4 に示すウェル当たりのすべての完全シーケンシング遺伝子についての % 保持リードの箱ひげ図から、遺伝子間の変動は、たとえば、0.6 を超えるひげを有するウェル D 1 1、F 4、F 8、H 3 および H 8 の場合など、重要となりうることが判明した。しかし、これらの 5 つのウェルは、はるかに低い総シーケンシングリード、3 3 5 7、5 4 5 7、2 8 7 4、3 4 1 4 および 4 0 4 3 に対応した。

【 0 4 6 8 】

遺伝子発現データの解析にクラスター化を使用することができる。多次元性を低減し、恐らく相関する変数を、直交変換によって少数の変数にすることによる次元削減のために主成分分析 (P C A) を使用することができる。データ中のクラスターを検索するのに、P C A からの主要な主成分を用いることができる。

【 0 4 6 9 】

図 2 5 A ~ 2 5 B は、2 つのプレートからの未補正 M L 対アルゴリズム適用後の訂正 M I を用いた P C A プロットを示す。図 2 5 A は、総シーケンシングリード > 5 0 0 0 を有するウェル当たりの遺伝子毎の未補正 M L を用いた P C A プロットを示す。この P C A プロットは、第 1 に、総シーケンシングリード < 5 0 0 0 を有するウェルを除去する (その結果、3 つの制御遺伝子を除いて、1 3 9 のウェルと、1 0 7 の遺伝子が残った) 工程 ; 第 2 に、1 3 9 ウェル全体でゼロ未補正 M L を有する遺伝子を除去する (8 5 の遺伝子が残った) 工程 ; 第 3 に、未補正 M L プラスワンの対数を採用して、データセットにゼロを組み込む工程、次に、センタリングおよびスケーリングの後、ログデータに P C A を適用する工程によって生成された。P C A プロットは、明らかに 2 つのクラスターを示すが、両クラスターからの距離がほぼ等しい D 0 2、D 0 5、および F 0 6 などのウェルについては、細胞型を決定するのは困難であった。クラスター化の結果は、ノイズが付加されたために損なわれる可能性があり、少数のノイズ変数であっても明瞭なクラスター構造を損ないうる。従って、特徴 / 変数選択の前処理工程またはフィルタリングもしくは脱ノイズ工程から利益を受けることができる。完全シーケンシングデータに訂正方法を適用することにより、図 2 5 B に示すように、明瞭なクラスター構造が達成された。図 2 5 B の P C A プロットは、未補正 M L (アルゴリズムの適用前に検出された遺伝子すべての分子標識のカウント) ではなく、訂正 M L (訂正方法を適用後の完全シーケンシング遺伝子の真の分子標識のカウント) を用いた以外は、図 2 5 A に示したように明瞭なクラスター構造が得られ、計 1 3 9 のウェルで 7 5 の遺伝子を使用した。2 つの識別可能なクラスターが観察され、これらは、y 軸によって首尾よく隔てられていた (P C 2)。図 2 5 A と比較して、図 2 5 B のクラスターは、サイズがコンパクトであり、各ウェルの細胞が明瞭にクラスターに割り当てられていた。加えて、図 2 5 B の y 軸の右側の小さなクラスターは、3 1 のウェルから成り、総ウェルの約 2 2 % であり、予想された 2 0 % にかなり近い。

【 0 4 7 0 】

全体として、これらのデータは、確率バーコード付き標的のデータカウンティングを要約および視覚化する上で有用ないくつかのツールを明らかにする。

【 0 4 7 1 】

10

20

30

40

50

実施例 8

高度発現遺伝子-ACTBのプレートにおける各MLのMLカバー率

この実施例は、シーケンシングまたはPCRの最中に生じたMLエラーの識別可能な分布が、一般に、MLからの識別可能な分布を有することを実証する。

【0472】

絶対遺伝子発現カウンティングおよびPCRバイアス訂正に加えて、MLは、ライブラリー作製方法およびシーケンシングデータの統計学的クオリティに関するより良い理解をもたらす。同じ遺伝子MLを示すリードの数（MLカバー率と呼ばれる）に関して、ライブラリー作製中に生成されたシーケンシングエラー塩基コールまたはPCRエラーを検出することが可能である。たとえば、単位のリードのみにより表される所与のSLからの遺伝子MLと比較して、複数のリードにより表される所与のSLからの遺伝子MLは、恐らく、正確な測定値である。同じライブラリー中の高MLカバー率の存在下で低MLカバー率バーコードは、往々にして、ライブラリー作製の際のシーケンシングランまたはPCR工程中に生成された人工物もしくはエラーである。シーケンシングまたはPCRの最中に生じたMLエラーは、一般に、真のMLからの識別可能な分布を有する。図27は、高度発現遺伝子-ACTBのマイクロプレートにおける各分子標識の分子標識カバー率を示す例示的なプロットを示し、ここで、識別可能な分布は、エラー分子標識とリアル分子標識の間に観察された。図28は、高度発現遺伝子-ACTBのマイクロプレートにおける各分子標識の分子標識カバー率への2つのネガティブ二項分布の当てはめを示す例示的なプロットである。2つのネガティブ二項分布の当てはめは、低い分子標識深度を有する分子標識エラーと、より高い分子標識深度を有する真の分子標識が、統計学的に識別可能な分布であることを実証する。x軸は、分子深度である。

【0473】

全体として、これらのデータは、シーケンシングまたはPCRの最中に生じたMLエラーが、一般に、真のMLからの識別可能な分布を有することを実証する。

【0474】

実施例 9

PCRまたはシーケンシングエラーによる分子標識の訂正

この実施例は、PCRおよびシーケンシング置換エラーによる分子標識を訂正する方法を明らかにするものであり、これは、均一カバー率の想定なしに、かつ、完全シーケンシングステータスのために高いシーケンシングカバー率を必要とすることなく、全トランスクリプトームアッセイに適用することができる。

【0475】

各リードの第1のマッピング座標およびユニーク分子標識(UMI)に対して重複排除を実施し、同じ開始座標、UML、および鎖を仮定して、リードは、同一であると想定した。重複排除の後、クラスター当たりの最も高いカウントを有するUMLが保持された(表13)。

【0476】

分子標識(ML)は、遺伝子毎に訂正した。各遺伝子について、方向近接性を用いてMLのクラスターを同定した。MLが、1のハミング距離内にあり、かつ、親MLカウント $\geq 2 * (\text{子供MIカウント}) - 1$ であった場合、方向近接性法は、MLをクラスター化した。同じクラスター内のMLはすべて、同じ親MLに由来すると考え、子供MLカウントは、親MLへ折りたたまれた。図29は、分子標識訂正を示し、ここで、1のペアワイズハミング距離が大きな比率を占めた。分子標識訂正後、1のハミング距離相違する分子標識がクラスター化され、同じ親分子標識へ折りたたまれた。図30は、リード数カバー率に対する訂正されたMLの数の曲線を示す。すべてのリードが保持されたため、この方法は、1塩基PCRまたはシーケンシングエラーを除去するために使用することもできる。

【0477】

10

20

30

40

50

【表 13】

表 13. 分子標識を重複排除した後、全トランスクリプトームアッセイの場合、ユニーク分子標識の不十分な数のみが、エラーとしてみなされた

サンプル	ウェル	未補正カウント		重複排除 UML	
		リード数	ユニーク MI 数	リード数	ユニーク MI 数
プレート 2	A02	234646	3079	28925	2335
プレート 3	A02	773050	14126	95023	11410

10

【0478】

全体として、これらのデータは、すべてのリードが保持されたことから、全トランスクリプトームアッセイのデータを訂正または調節するために適用することができる補正方法を実証する。

【0479】

実施例 10

高入力サンプルのための分子標識カウンティング

この実施例は、入力分子が増加するとき、使用されるユニーク分子標識を説明する。

【0480】

mRNA の確率およびユニーク標識を可能にするために、小さなサンプル入力（たとえば、単一細胞）に使用する場合、BD Precise（商標）Targeted Assay が最も好適であると考えられる。転写物の数が、高 RNA／細胞入力実験におけるバーコードプールに比して増加すると、同じ遺伝子を標識するために最小される ML のパーセンテージが増加し、ポアソン分布を用いて理論上計算された（図 26）。こうした状況下で、統計学的訂正なしに、ML を用いて遺伝子発現を定量する工程は、ポアソン訂正も 2 つのネガティブ二項分布に基づく訂正もなしで、初めに存在する分子の数を過小評価するであろう。

20

【0481】

遺伝子当たりの mRNA の数が 6561 バーコードのコレクション全体を超える極めて高い入力サンプルでは、ポアソン訂正または 2 つのネガティブ二項分布に基づく訂正はもはや不可能である。たとえば、65000 または 100000 入力分子のいずれにかかわらず、いずれの場合も最大 6561 の飽和バーコードが予想される。従って、高サンプル入力を有すると思われる遺伝子およびサンプルを改変することができ、それによって、ML カウントは恐らく過小評価されるであろう。

30

【0482】

全体として、これらのデータは、ML を用いて遺伝子発現を定量する場合、未補正データを調節する必要性を実証する。

【0483】

実施例 11

再帰的置換エラー訂正（RSEC）

この実施例では、再帰的置換エラー訂正を明らかにする。

40

【0484】

ML エラーを除去するために、BD Precise（商標）Targeted Assay 分析パイプラインに、2 つの共同的方法を使用することができる。手短には、シーケンシング塩基コール置換エラーに由来する ML エラーを同定し、再帰的置換エラー訂正（RSEC）を用いて真の ML バーコードに調節する。続いて、ライブラリー作製工程由来の ML エラーまたはシーケンシング塩基欠失エラーを、分布ベースのエラー訂正（DBEC）を用いて調節する。

【0485】

RSEC アルゴリズムは、PCR またはシーケンシング置換に由来する ML エラーを調節

50

することができる。これらの稀なエラーイベントは、MLカバー率を調べる際に認められている。たとえば、エラーMLのMLカバー率は、適切なシーケンシングサンプル中のMLよりも有意に低くなりうる（図27）が；初期Molecular Indexing（商標）（逆転写）工程中に、2つの非常に類似したMLを用いた場合、これらは、概して、類似するMLカバー率を有し、除去する必要がない。シーケンシング深度が増大するにつれて、より多くのMLErrorが出現するため、RSECは、高度シーケンシングバーコード付きライブラリーのMLカウントを調節するために重要となりうる。

【0486】

簡潔に述べると、RSECは、エラー訂正において2つの因子：1）ML配列の類似性；および2）それらのMLカバー率を考慮する。各標的遺伝子について、それらのML配列の両方がある、互いに対して1塩基（ハミング距離＝1）内にあれば、MLは接続される。MLxとyとの間の各接続について、

カバー率(y) > 2 * カバー率(x) + 1 式(5)

（式中、yは、「親ML」を示し、xは、「子供ML」を示す）。

【0487】

この代入に基づき、子供MLは、その親へ折りたたまれうる。この過程は、当該遺伝子について同定可能な親／子供MLがもはや存在しなくなるまで、再帰的である。

【0488】

図31は、上に概説した再帰的置換エラー訂正の一例の概略図を示す。RSEC訂正前の未補正データ中のMLは、9つのユニークML：GTCAAATT、GTCAAAAT、GTCAAAAA、TTCAAAAA、TTCAGAAA、CTCAAAAA、TTCAAACT、TTCAAAAT、およびTTCAAACAを含む。RSECを適用することにより、

【化7】

GTCAAATT

は、

【化8】

GTCAAAAT

へ折りたたまれうる。なぜなら、2つのMLは、1ヌクレオチド（下線部）相違し、MLGTCAAATTは、GTCAAAATより低いMLカウントを有するからである。次に、ML

【化9】

GTCAAAAT

は、GTCAAAATより高いMLカウントを有するML

【化10】

GTCAAAA

（ML配列中の相違を下線で示す）へ折りたたまれうる。同様に、MLTTCAGAAAおよびCTCAAAAAは、MLTTCAAAAへ折りたたまれうる。MLTTCAAACTは、MLTTCAAAAATへ折りたたまれ、これが、今度は、MLTTCAAAAAに折りたたまれうる。MLTTCAAACAは、他のすべてのMLと2ヌクレオチド以上相違するため、他の8つのMLのいずれにも折りたたまれない。RSEC訂正前に、未補正MLカウントは9であった。RSEC訂正後、MLカウントは2つ：ML

T T C A A A A A および T T C A A A A A であった。

【0489】

全体として、これらのデータは、未補正MLカウントを訂正するためにRSECを使用する工程を実証する。

【0490】

実施例12

MLカバー率計算

この実施例は、MLカバー率計算を説明する。

【0491】

RSECの後、ウェル当たりの遺伝子MLカウントを評価して、さらなる訂正についてこれらの適合性を判定する。低MLカバー率（<ML当たり4リード）を有する遺伝子は、次の訂正工程を迂回し、最終MLデータ表に報告されて、バイオインフォマティクスパイプラインに「低深度」とであると記録される。考えられる6561のバーコードのうち少なくとも6557が観察されるといった、極めて高い入力を持つ遺伝子の場合、バーコード多様性のために分子の数を決定するのは困難となり、遺伝子は、「飽和」として表示される。2つの決定地点のいずれも満たさない遺伝子MLについては、次のDBECアルゴリズムに進み、出力ログファイル内で「合格」と表示される。さらに、ウェル当たり平均650MLより高いMLを有する遺伝子は、これらのMLの>5%は、ポアソン分布に基づいて再利用されるため、「高入力」とであると記録される（図27）。

【0492】

全体として、この実施例は、MLカバー率計算を説明する。

【0493】

実施例13

分布ベースのエラー訂正（DBEC）

この実施例は、分布ベースのエラー訂正を説明する。

【0494】

RSECとは異なり、DBECアルゴリズムは、MLが、そのML配列にかかわらず、エラーまたは真のシグナルであるかを識別するための方法である。RSECは、エラーを訂正するために、ML配列およびMLカバー率情報の両方に依存するが、DBECは、非置換エラー訂正について訂正するために、主としてMLカバー率だけに依存する。前述したように、エラーバーコードは、一般に、真のバーコードMLカバー率とは異なる低いMLカバー率を有し；このMLカバー率の差は、異なる分布として、MLカバー率のヒストグラムプロットで認めることができる（図27）。この差を仮定して、DBECは、MLEラー（より低いMLカバー率を有する）と、より高いMLカバー率を有する真のシグナルのものとを統計学的に識別するために、2つのネガティブ二項分布を当てはめる。

【0495】

最適分布当てはめのための再使用MLの除去

所与の遺伝子について、検出されたMLが増加するにつれて、再使用されるML（すなわち、同じ遺伝子由来する2つ以上のmRNAを標識するために同じMLが使用される）のパーセンテージは、増加することから、推定することができる。ポアソン分布（ $\gamma_{non-unique}$ ）を用いて、ウェル*i*の再使用MLの数（ $n_{non-unique,i}$ ）をML再使用率方程式（方程式（6））から推定する。推定再使用MLが、ウェル*i*における所与の遺伝子の総MLの5%より大きければ、ウェル*i*におけるこの遺伝子は、「高入力」と表示される。これらの「高入力」データの場合、より優れた二項分布を取得するために、最大MLカバー率MLは、分布当てはめから除外される（しかし、後のカウント工程のために保存される）。

$P(X > 1 \mid \lambda_{non-unique})$, $\lambda_{non-unique} = \text{Number of ML} / 6561$ 式（6）

【数6】

10

20

30

40

50

$$n_{non-unique} = \sum_{i=1}^{n_{wells}} n_{non-unique,i} \quad \text{式(7)}$$

【0496】

低発現遺伝子のための擬似点の追加

MLの固有の数が10未満である場合、往々にして、データの希薄さのために分布を当てはめるのが難しくなる。この問題を改善するために、DBECは、分布当てはめを補助するために用いられる1%シグナルカウントの擬似点を追加するが、それでもなおデータに影響を与えない。

【0497】

パラメータの推定

2つのネガティブ二項分布を当てはめて、シグナルMLからエラーを区別するために、パラメータ推定のための2組の出発数値を概算する。エラー分布は、平均および1の散布を有するネガティブ二項分布であると想定される。

【0498】

エラー／シグナル確率推定

シグナルおよびエラー分布をそれぞれNegative Binomial (μ_{signal} , $\text{size}_{\text{signal}}$) および Negative Binomial (μ_{error} , $\text{size}_{\text{error}}$) として想定する。シグナルMLの数を小さい順に決定するために、所与のMLからのリードの数が、シグナルおよびエラー分布に由来する確率を、方程式(8)が満たされるまで計算し、ここで、先行するMLはすべて、エラーMLとみなされる。

$P(X=r \mid \mu = \mu_{\text{error}}, \text{size} = \text{size}_{\text{error}}) < P(X=r \mid \mu = \mu_{\text{signal}}, \text{size} = \text{size}_{\text{signal}})$ 式(8)

【0499】

全体として、この実施例は、分布ベースのエラー訂正を実施するための計算を示す。

【0500】

実施例14

二次導関数に基づくSLエラーの調節

この実施例は、二次導関数に基づくSLエラーの調節を示す。

【0501】

図32、パネル(a)～(e)は、分子標識深度変化の二次導関数に基づくPCRおよびシーケンシングエラーの訂正の例示的な結果を示す。図32、パネル(a)は、SLエラーおよびシグナルMLが、十分に分離されうること示す。図32、パネル(b)および(d)は、それぞれ、図32、パネル(c)および(e)に示すMLカウントからの分子標識カウントの累積和を示す。図32、パネル(b)および(d)中の縦線は、二次導関数の最大値の位置を示す。図32、パネル(b)および(d)中の点線は、二次導関数の最大値の位置が、MLカウント対MLリード深度のプロットにおいてMLを分離し得ることを示す。

【0502】

全体として、これらのデータは、MLシグナルからSLエラーを分離するために、分子標識の二次導関数の最大値を用いることができることを明らかにする。

【0503】

実施例15

DBECに基づくPCRおよびシーケンシングエラーの訂正

この実施例は、2つのネガティブ二項分布に基づくPCRおよびシーケンシングエラーの訂正を示す。

【0504】

図33、パネル(a)～(c)は、CD69について2つのネガティブ二項分布に基づくPCRおよびシーケンシングエラーの訂正の例示的な結果を示す。図33、パネル(a)

は、図33、パネル(b)のML深度のヒストグラムに示すMLカウントデータでのCD69について2つのネガティブ二項分布(ノイズネガティブ二項分布の D_n と、シグナル二項分布の D_s)の当てはめを示す。図33、パネル(b)の点線は、図33、パネル(a)に示す2つのネガティブ二項分布により決定されたMLシグナルおよびSLエラーの分離を示す。図33、パネル(c)の縦線は、リードの累積和プロットに基づいて決定される二次導関数の局所的な最大値を示す。図33と同様に、図34、パネル(a)~(c)は、CD3Eについての2つのネガティブ二項分布に基づくPCRおよびシーケンシングエラーの訂正の例示的な結果を示す。

【0505】

全体として、これらのデータは、DBECを用いて、PCRおよびシーケンシングエラーを訂正することができることを明らかにする。

10

【0506】

実施例16

ML再使用

この実施例は、高度発現遺伝子のためのML再使用、ならびに分布当てはめ前に高度発現遺伝子の入力データを調節する必要性を明らかにする。

【0507】

図35、パネル(a)~(c)は、高度発現遺伝子ACTBについての2つのネガティブ二項分布に基づくPCRおよびシーケンシングエラーの訂正の例示的な結果を示す。高度発現遺伝子は、過剰シーケンシングステータス(たとえば、100以上のMLカバー率を有する)を有する。いくつかの実施形態では、高度発現遺伝子は、他の基準を用いて決定してもよい。図35、パネル(a)において、縦線右側の分子標識は、高い深度に基づいて恐らく再使用されたMLに対応する。図35、パネル(b)は、分子標識を3つのカテゴリ(MLエラー以外に): SLエラー、シグナルML、および恐らく再使用されたMLに区分することができることを概略的に示す。図35、パネル(c)は、恐らく再使用されたMLを調節せずに、当てはめられた2つのネガティブ二項分布は、理想的ではなかったことを実証する。

20

【0508】

図36は、高度発現遺伝子についてグリッチ分子標識の再使用の例示的な結果を示す。図36は、高度発現遺伝子GAPDH、ACTB、およびHSP90AB1について上位20の高い深度MLを示す。これらの高い深度MLは、多数のGおよびTを有し、これらは、再使用される可能性が高く、バーコード付けは確率的ではなかった。ML二重項は、確率標識を想定する理論計算値より早く起こった。ACTBについては、ウェル当たり350MLが存在した場合、理論上、2.7%の二重項があるはずであったが、実際の二重項は、4パーセント前後であった。

30

【0509】

図37、パネル(a)~(b)は、2つのネガティブ二項分布を当てはめる前の、高度発現遺伝子についての入力データの調節の例示的な結果を示す。図37、パネル(a)は、高度発現遺伝子について調節された、図35、パネル(a)における入力データを示す。図35、パネル(c)における非理想的な分布当てはめとは対照的に、図37、パネル(b)は、当てはめられた2つのネガティブ二項分布を示す。

40

【0510】

全体として、これらのデータは、2つのネガティブ二項分布の当てはめの前に、高度発現遺伝子についてのシーケンシンデータから、再使用されたMLを除去する必要があることを示す。

【0511】

実施例17

2つのネガティブ二項分布を用いたMLカウントの訂正

この実施例は、2つのネガティブ二項分布を用いて訂正された10の標的のMLカウントを示す。

50

【0512】

図38、パネル(a)～(j)は、2つのネガティブ二項分布を用いて訂正されたデータセットの非限定的な例示的検証を示す。図38に示すように、10の標的のMLカウントが訂正された。図38の各パネルの縦線は、2つのネガティブ二項分布を用いて決定された、標的のMLシグナルおよびSLエラーの分離を示す。

【0513】

全体として、これらのデータは、2つのネガティブ二項分布を用いたMLカウントの訂正を検証するものである。

【0514】

実施例18

混合されたJurkatおよび乳癌(BrCa)単一細胞の96ウェルからのBD Precise (商標) Targeted Assayのt-確率的近傍埋込み視覚化

この実施例は、混合されたJurkatおよび乳癌(BrCa)単一細胞についての再帰的置換エラー訂正および分布ベースのエラー訂正に基づいてPCRおよびシーケンシングエラーを訂正する方法を示す。

【0515】

図39、パネル(a)～(d)は、混合されたJurkatおよび乳癌(BrCa)単一細胞の96ウェルからのBD Precise (商標) Targeted Assayの例示的なt-確率的近傍埋込み(t-SNE)視覚化を示す(86の被検遺伝子)。図39、パネル(a)は、ML調節前および後の同じパラメータを有するDBScanを用いて、細胞クラスターを同定したことを示す。図39、パネル(b)～(d)は、色および点サイズの両方により評価される個々のマーカー発現を示す。図39、パネル(b)は、PSMB4、すなわち、両細胞型中に、およびML調節後に存在するハウスキーピング遺伝子を示し、PSMB4シグナルの欠如は、「低シグナル」クラスター中でさらに強調される。図39、パネル(c)は、CD3E、すなわち、Jurkat細胞クラスターを強調するリンパ球マーカーを示す。図39、パネル(d)は、CDH1、すなわち、BrCaクラスターを強調する上皮細胞マーカーを示す。

【0516】

全体として、これらのデータは、ML調節によってMLノイズが除去され、これにより、細胞クラスター間の遺伝子発現の明瞭な区別が可能になったことを実証するものである。

【0517】

実施例19

細胞クラスター間の差異発現分析

この実施例は、低シグナル細胞および乳癌(BrCa)細胞についての再帰的置換エラー訂正および分布ベースのエラー訂正に基づいてPCRおよびシーケンシングエラーを訂正する方法を示す。

【0518】

図40、パネル(a)～(b)は、各々のクラスターでDBScanにより計算され、かつ遺伝子マーカーレベルによって決定された、両方の選択クラスターにおいて>0MLを有する遺伝子について細胞クラスター間の差異発現分析を示す非限定的な例示的プロットである。図40、パネル(a)は、残りの細胞と比較した「低シグナル」クラスター遺伝子発現を示す。図40、パネル(a)の上部は、未補正ML比較を示し、これによって、他の細胞において高い平均発現を有する遺伝子ほど、MLノイズが概して高いことがわかる。図40、パネル(a)の下部は、RSECおよびDBECを用いたML調節後に、「低シグナル」クラスター中に検出されたMLノイズが低減し、クラスター間の遺伝子発現の明瞭な識別を可能にすることを示す。図40、パネル(b)は、残りの細胞と比較した「BrCa」クラスター遺伝子発現を示す。図40、パネル(b)の上部は、非BrCa細胞中の未補正MLも、KRT1、MUC1などのBrCaマーカーの有意なMLカウントを有したことを示す。図40、パネル(b)の下部は、BrCaマーカーの調節されたMLが、BrCaクラスター中で、残りの細胞よりも極めて豊富であったことを示す。

【0519】

全体として、これらのデータは、低シグナル細胞および乳癌細胞などの細胞の場合、再帰的置換エラー訂正および分布ベースのエラー訂正に基づいてPCRおよびシーケンシングエラーを訂正することができることを示す。

【0520】

実施例20

混合JurkatおよびT47D細胞の分子標識の調節

この実施例は、混合JurkatおよびT47D細胞の分子標識を調節する方法を示す。

【0521】

図41、パネル(a)～(d)は、86の被検遺伝子を含む混合Jurkatおよび乳癌(T47D)単一細胞の96ウェルからのBD Precise (商標) Targeted Assayのトータルの確率的近傍埋込み視覚化を示す非限定的な例示的プロットである。図41、パネル(a)は、ML調節前および後に同じパラメータを有するDBScanを用いて、細胞クラスターを同定したことを示す。図41、パネル(b)～(d)は、色および点サイズの両方によって評価される個々のマーカー発現を示す。図41、パネル(b)は、PSMB4、すなわち、両細胞型中に、およびML調節後に存在するハウスキーピング遺伝子の評価を示す。PSMB4シグナルの欠如は、テンプレートなし対照(NTC)クラスターにおいてさらに強調される。図41、パネル(c)は、CD3E、すなわち、Jurkat細胞クラスターを強調するリンパ球マーカーの評価を示す。図41、パネル(d)は、CDH1、すなわち、T47Dクラスターを強調する上皮細胞マーカーの評価を示す。

10

20

【0522】

図42、パネル(a)～(b)は、エラー訂正工程前(図42、パネル(a)に示す未補正ML)ならびにRSECおよびDBEC訂正後(図42、パネル(b)に示す調節ML)に、図41で同定されたさまざまな細胞クラスター間の分子標識カウントによる差異遺伝子発現を表示する非限定的な例示的ヒートマップである。発現の低かった遺伝子は青色で、発現が高かった遺伝子はオレンジ色である。これらの細胞型の間で遺伝子発現が類似する遺伝子は、互いにクラスター化する。エラー訂正がない場合、NTCは、CD3EおよびKRT18(それぞれ、JurkatおよびT47Dマーカーである)などの高度発現遺伝子に由来するノイズを有した。さらに、エラー訂正は、JurkatとT47Dとの間で識別可能な遺伝子発現パターンを明らかにした。

30

【0523】

全体として、これらのデータは、ML調節が、MEノイズを除去することができ、これによって、細胞クラスター間の遺伝子発現の明瞭な区別を可能になることを実証するものである。

【0524】

以上に記載の実施形態の少なくともいくつかでは、実施形態で使用される1つ以上のエレメントは、他の実施形態で互換的に使用可能である。ただし、かかる交換が技術的に実現可能である場合に限る。特許請求された主題の範囲から逸脱することなく、以上に記載の方法および構造に種々の他の省略、追加、および変更を行うことは、当業者であれば分かるであろう。かかる変更および変化はすべて、添付の特許請求の範囲に規定される主題の範囲内に含まれることが意図される。

40

【0525】

本明細書に記載の実質的に任意の複数形および／または単数形の用語の使用に関連して、文脈上および／または適用上適切であれば、当業者は複数形から単数形へおよび／または単数形から複数形への変換が可能である。明確にするために種々の単数形／複数形の入替えを本明細書に明示的に記述しうる。本明細書および添付の特許請求の範囲で用いられる場合、特に文脈上明確に規定されていない限り、単数形の「a」、「an」、および「the」には、複数の参照語が含まれる。本明細書での「or (または)」の意味はいずれも、特に明記されていない限り、「and/or (および／または)」を包含すること

50

が意図される。

【0526】

一般的には、本明細書特に添付の特許請求の範囲（たとえば添付の特許請求の範囲の本文）で用いられる用語は「オープン」用語であることが一般に意図されることは当業者であれば理解されよう（たとえば、「including（～を含む）」という用語は「～を含むがこれらに限定されるものではない」と解釈すべきであり、「having（～を有する）」という用語は「少なくとも～を有する」と解釈すべきであり、「include s（～を含む）」という用語は「～を含むがこれらに限定されるものではない」と解釈すべきであるなど）。さらに、導入クレームレシテーションの特定数が意図される場合、かかる意図は請求項で明示的にリサイトされ、かかるレシテーションの不在下ではかかる意図は存在しないことは当業者であれば理解されよう。たとえば、理解の一助として、以下の添付の特許請求の範囲は、クレームレシテーションを導入するために導入語句「at least one（少なくとも1つ）」および「one or more（1つ以上）」の使用を含みうる。しかしながら、かかる語句が用いられたとしても、不定冠詞「a」または「an」によるクレームレシテーションの導入が、かかる導入クレームレシテーションを含む任意の特定の請求項を、一方のかかるレシテーションを含む実施形態のみに限定することを意味するものと解釈すべきでない。たとえ同一の請求項が導入語句「one or more（1つ以上）」または「at least one（少なくとも1つ）」と不定冠詞たとえば「a」または「an」とを含む場合でさえも、そのように解釈すべきでない（たとえば、「a」および／または「an」は「at least one（少なくとも1つ）」または「one or more（1つ以上）」を意味するものと解釈すべきである）。定冠詞を用いてクレームレシテーションを導入する場合にも、同じことが当てはまる。そのほかに、たとえ特定数の導入クレームレシテーションが明示的にリサイトされたとしても、かかるレシテーションは少なくともリサイトされた数を意味すると解釈すべきであることは当業者であれば分かるであろう（たとえば、「2つのレシテーション」という他の修飾語を含まないペアのレシテーションは、少なくとも2つのレシテーションまたは2つ以上レシテーションを意味する）。さらに、「A、B、およびCの少なくとも1つ」に類似した条件が用いられる場合、一般的には、かかる構成は当業者がその条件を理解する意味であることが意図される（たとえば、「A、B、およびCの少なくとも1つを有する系」は、限定されるものではないが、A単独、B単独、C単独、AとBの両方、AとCの両方、BとCの両方、および／またはAとBとCの全部などを有する系を含であろう）。「A、B、またはCの少なくとも1つなど」に類似した条件が用いられる場合、一般的には、かかる構成は当業者がその条件を理解する意味であることが意図される（たとえば、「A、B、またはCの少なくとも1つを有する系」は、限定されるものではないが、A単独、B単独、C単独、AとBの両方、AとCの両方、BとCの両方、および／またはAとBとCの全部などを有する系を含であろう）。さらに、2つ以上の代替用語を表す実質上任意の選言的な語および／または語句は、明細書、請求項、または図面にかかわらず、用語の1つ、用語のいずれか、または用語の両方を含む可能性が企図されると理解すべきであることは当業者であれば理解されよう。たとえば、「AまたはB」という語句は「A」または「B」または「AおよびB」の可能性を含むものと理解されよう。

【0527】

そのほかに、本開示の特徴または態様がマーカッシュグループにより記述される場合、それにより、本開示は、マーカッシュグループの任意の個別のメンバーまたはメンバーのサブグループにより記述されることは当業者であれば分かるであろう。

【0528】

当業者であれば理解されるであろうが、あらゆる目的で、たとえば、明細書の提供に関して、本明細書に開示された範囲はすべて、あらゆる可能なサブ範囲およびそのサブ範囲の組合せをも包含する。いずれの列挙された範囲も、十分に記述されたものとしてかつその範囲が少なくとも2等分、3等分、4等分、5等分、10等分などされうるものとして容易に認識可能である。たとえば、限定されるものではないが、本明細書で考察した各範囲

10

20

30

40

50

は、下3分の1、中3分の1、上3分の1に容易に分解可能である。同様に、当業者であれば理解されるであろうが、「～まで」、「少なくとも～」、「～超」、「～未満」などの表現はすべて、リサイトされた数を含み、以上で考察したように後続的にサブ範囲に分解可能な範囲を意味する。最終的に、当業者であれば理解されるであろうが、範囲は各個別のメンバーを含む。したがって、たとえば、1～3個の物品を有するグループは、1、2、または3個の物品を有するグループを意味する。同様に、1～5個の物品を有するグループは、1、2、3、4、または5個の物品を有するグループを意味し、他も同様である。

【0529】

種々の態様および実施形態を本明細書に開示してきたが、他の態様および実施形態は当業者には自明であろう。本明細書に開示される種々の態様および実施形態は、例示を目的としたものであり、限定を意図したものではなく、真の範囲および趣旨は、以下の特許請求の範囲により示される。

なお、本発明としては、以下の態様も好ましい。

〔1〕 標的の数を決定する方法であって、

（a）複数の確率バーコードを用いて、複数の標的に確率バーコードを付けて、複数の確率バーコード付き標的を生成する工程であって、前記複数の確率バーコードの各々が分子標識を含む工程と；

（b）前記確率バーコード付き標的のシーケンシングデータを取得する工程と；

（c）前記複数の標的の1つ以上について：

（i）前記シーケンシングデータ中の前記標的に関連付けられた識別可能な配列を有する分子標識の数をカウントする工程と；

（i i）方向近接性を用いて、前記標的の分子標識のクラスターを同定する工程と；

（i i i）（i i）で同定された前記標的の分子標識の前記クラスターを用いて、（b）で得られた前記シーケンシングデータを折りたたむ工程と；

（i v）前記標的の数を推定する工程であって、推定された前記標的の数が、（i i）の前記シーケンシングデータの折りたたみ後に、（i）でカウントされた前記シーケンシングデータ中の前記標的に関連付けられた識別可能な配列を有する分子標識の数と相関する工程と、

を含む、方法。

〔2〕 前記複数の標的が、細胞の全トランスクリプトームの標的を含む、〔1〕に記載の方法。

〔3〕 クラスター内の前記標的の分子標識が、互いの所定の方向近接性閾値内にある、

〔1〕～〔2〕のいずれか一項に記載の方法。

〔4〕 前記方向近接性閾値が、1のハミング距離である、〔3〕に記載の方法。

〔5〕 前記クラスター内の前記標的の前記分子標識が、1つ以上の親分子標識と、前記1つ以上の親分子標識の子供分子標識とを含み、前記親分子標識の発生数が、所定の方向近接性発生数閾値以上である、〔1〕～〔4〕のいずれか一項に記載の方法。

〔6〕 前記所定の方向近接性発生数閾値が、 $2 \times (\text{子供分子標識の発生数}) - 1$ である、〔5〕に記載の方法。

〔7〕 （i i）で同定された前記標的の分子標識の前記クラスターを用いて、（b）で得られた前記シーケンシングデータを折りたたむ工程が、前記子供分子標識の発生数を前記親分子標識に帰属させる工程を含む、〔1〕～〔6〕のいずれか一項に記載の方法。

〔8〕 前記標的のシーケンシング深度を決定する工程をさらに含む、〔1〕～〔7〕のいずれか一項に記載の方法。

〔9〕 前記標的の前記シーケンシング深度が所定のシーケンシング深度閾値を超える場合、前記標的の数を推定する工程が、（i）でカウントされた前記シーケンシングデータを調節する工程を含む、〔8〕に記載の方法。

〔10〕 前記所定のシーケンシング深度閾値が、15～20である、〔9〕に記載の方

10

20

30

40

50

法。

〔11〕 (i) でカウントされた前記シーケンシングデータを調節する工程が、前記標的の分子標識を閾値化して、(b) で得られた前記シーケンシングデータ中の前記標的に関連付けられた真の分子標識および偽の分子標識を決定する工程を含む、〔9〕～〔10〕のいずれか一項に記載の方法。

〔12〕 前記標的の前記分子標識を閾値化する工程が、前記標的の前記分子標識について統計解析を実施する工程を含む、〔11〕に記載の方法。

〔13〕 前記統計解析を実施する工程が、前記標的の前記分子標識の分布およびそれらの発生数を2つのネガティブ二項分布に当てはめる工程と；

前記2つのネガティブ二項分布を用いて真の分子標識の数 n を決定する工程と；

(b) で得られた前記シーケンシングデータから前記偽の分子標識を除去する工程と、を含み、

前記偽の分子標識が、 n 番目に豊富な分子標識の発生数よりも低い発生数を有する分子標識を含み、前記真の分子標識が、 n 番目に豊富な分子標識の発生数以上の発生数を有する分子標識を含む、〔12〕に記載の方法。

〔14〕 前記ネガティブ二項分布が、前記真の分子標識に対応する第1のネガティブ二項分布と、前記偽の分子標識に対応する第2のネガティブ二項分布を含む、〔13〕に記載の方法。

〔15〕 標的の数を決定する方法であって、

(a) 複数の確率バーコードを用いて、複数の標的に確率バーコードを付けて、複数の確率バーコード付き標的を生成する工程であって、前記複数の確率バーコードの各々が分子標識を含む工程と；

(b) 前記確率バーコード付き標的のシーケンシングデータを取得する工程と；

(c) 前記複数の標的の1つ以上について：

(i) 前記シーケンシングデータ中の前記標的に関連付けられた識別可能な配列を有する分子標識の数をカウントする工程と；

(i i) 前記シーケンシングデータ中の前記標的に関連付けられた識別可能な配列を有するノイズ分子標識の数を決定する工程と；

(i i i) 前記標的の数を推定する工程と、

を含み、

推定された前記標的の数が、(i i) で決定された前記ノイズ分子標識の数に応じて調節された、(i) でカウントされた前記シーケンシングデータ中の前記標的に関連付けられた前記識別可能な配列を有する分子標識の数と相関する、方法。

〔16〕 前記シーケンシングデータ中の前記標的のシーケンシングステータスを決定する工程をさらに含む、〔15〕に記載の方法。

〔17〕 前記シーケンシングデータ中の前記標的の前記シーケンシングステータスが、飽和シーケンシング、過少シーケンシング、または過剰シーケンシングである、〔16〕に記載の方法。

〔18〕 前記飽和シーケンシングステータスが、所定の飽和閾値よりも大きい、識別可能な配列を含む分子標識の数を有する前記標的によって決定される、〔17〕に記載の方法。

〔19〕 前記確率バーコードが、識別可能な配列を有する約6561の分子標識を含む場合、前記所定の飽和閾値が、約6557である、〔18〕に記載の方法。

〔20〕 前記確率バーコードが、識別可能な配列を有する約65536の分子標識を含む場合、前記所定の飽和閾値が、約65532である、〔18〕～〔19〕のいずれか一項に記載の方法。

〔21〕 前記シーケンシングデータ中の前記標的の前記シーケンシングステータスが、前記飽和シーケンシングステータスである場合、(i i) で決定された前記ノイズ分子標識の数が、ゼロである、〔17〕～〔20〕のいずれか一項に記載の方法。

10

20

30

40

50

〔22〕 前記過少シーケンシングステータスが、所定の過少シーケンシング閾値より小さい深度を有する前記標的にによって決定され、前記対象の前記深度が、前記シーケンシングデータ中の前記標的に関連付けられた識別可能な配列を有する前記分子標識の、平均、最小、または最大深度を含む、〔17〕～〔21〕のいずれか一項に記載の方法。

〔23〕 前記過少シーケンシング閾値が約4である、〔22〕に記載の方法。

〔24〕 前記過少シーケンシング閾値は、識別可能な配列を有する前記分子標識の数とは無関係である、〔23〕に記載の方法。

〔25〕 前記シーケンシングデータ中の前記標的の前記シーケンシングステータスが、前記過少シーケンシングステータスである場合、(i i) で決定された前記ノイズ分子標識の数が、ゼロである、〔17〕～〔24〕のいずれか一項に記載の方法。

10

〔26〕 前記過剰シーケンシングステータスが、所定の過剰シーケンシング閾値より大きい深度を有する前記標的にによって決定され、前記対象の前記深度が、前記シーケンシングデータ中の前記標的に関連付けられた識別可能な配列を有する前記分子標識の、平均、最小、または最大深度を含む、〔17〕～〔25〕のいずれか一項に記載の方法。

〔27〕 前記確率バーコードが、識別可能な配列を有する約6561の分子標識を含む場合、前記過剰シーケンシング閾値が、約250である、〔26〕に記載の方法。

〔28〕 前記シーケンシングデータ中の前記標的の前記シーケンシングステータスが、前記過剰シーケンシングステータスである場合、前記シーケンシングデータ中の前記標的に関連付けられた識別可能な配列を有する前記分子標識の数を、前記所定の過剰シーケンシング閾値にサブサンプリングする工程をさらに含む、〔26〕～〔27〕のいずれか一項に記載の方法。

20

〔29〕 前記シーケンシングデータ中の前記標的に関連付けられた識別可能な配列を有する前記ノイズ分子標識の数を決定する工程が、ネガティブ二項分布当てはめ条件が満たされる場合、

(i v) シグナルネガティブ二項分布を、(i) でカウントされた前記シーケンシングデータ中の前記標的に関連付けられた識別可能な配列を有する前記分子標識の数に当てはめる工程であって、前記シグナルネガティブ二項分布が、シグナル分子標識である、(i) でカウントされた前記シーケンシングデータ中の前記標的に関連付けられた識別可能な配列を有する分子標識の数に対応するステップと；

(v) ノイズネガティブ二項分布を、(i) でカウントされた前記シーケンシングデータ中の前記標的に関連付けられた識別可能な配列を有する前記分子標識の数に当てはめる工程であって、前記ノイズネガティブ二項分布が、ノイズ分子標識である、(i) でカウントされた前記シーケンシングデータ中の前記標的に関連付けられた識別可能な配列を有する分子標識の数に対応する工程と；

30

(v i) (v) で当てはめた前記シグナルネガティブ二項分布および(v i) で当てはめた前記ノイズネガティブ二項分布を用いて、前記ノイズ分子標識の数を決定する工程と、を含む、

〔17〕～〔28〕のいずれか一項に記載の方法。

〔30〕 前記ネガティブ二項分布当てはめ条件が、前記シーケンシングデータ中の前記標的の前記シーケンシングステータスが、前記過少シーケンシングステータスまたは前記過剰シーケンシングステータスではないことを含む、〔29〕に記載の方法。

40

〔31〕 (v) で当てはめた前記シグナルネガティブ二項分布および(v i) で当てはめた前記ノイズネガティブ二項分布を用いて、前記ノイズ分子標識の数を決定する工程が、前記シーケンシングデータ中の前記標的に関連付けられた前記識別可能な配列の各々について、

前記識別可能な配列のシグナル確率が、前記シグナルネガティブ二項分布であることを決定する工程と；

前記識別可能な配列のノイズ確率が、前記ノイズネガティブ二項分布であることを決定する工程と；

前記シグナル確率が前記ノイズ確率より小さければ、前記識別可能な配列がノイズ分子標

50

識であることを決定する工程と、

を含む、〔29〕～〔30〕のいずれか一項に記載の方法。

〔32〕 前記シーケンシングデータ中の前記標的に関連付けられた識別可能な配列を有する前記ノイズ分子標識の数を決定する工程が、

前記シーケンシングデータ中の前記標的の前記シーケンシングステータスが、前記過少シーケンシングステータスまたは前記過剰シーケンシングステータスではなく、かつ、(i)でカウントされた前記シーケンシングデータ中の前記標的に関連付けられた識別可能な配列を有する前記分子標識の数が、擬似点閾値より少ない場合、(ii)で前記シーケンシングデータ中の前記標的に関連付けられた識別可能な配列を有する前記ノイズ分子標識の数を決定する前に、前記シーケンシングデータ中の前記標的に関連付けられた識別可能な配列を有する前記分子標識の数に擬似点を加える工程を含む、

10

〔17〕～〔31〕のいずれか一項に記載の方法。

〔33〕 前記擬似点閾値が10である、〔32〕に記載の方法。

〔34〕 前記シーケンシングデータ中の前記標的に関連付けられた識別可能な配列を有する前記ノイズ分子標識の数を決定する工程が、

前記シーケンシングデータ中の前記標的の前記シーケンシングステータスが、前記過少シーケンシングステータスまたは前記過剰シーケンシングステータスではなく、かつ、(i)でカウントされた前記シーケンシングデータ中の前記標的に関連付けられた識別可能な配列を有する前記分子標識の数が、擬似点閾値以上である場合、(ii)で前記シーケンシングデータ中の前記標的に関連付けられた識別可能な配列を有する前記ノイズ分子標識の数を決定する際に、非ユニーク分子標識を除去する工程を含む、

20

〔17〕～〔33〕のいずれか一項に記載の方法。

〔35〕 前記非ユニーク分子標識を除去する工程が、前記シーケンシングデータ中の前記標的に関連付けられた識別可能な配列を有する前記分子標識の数が、所定の再使用分子標識閾値より大きい場合、(ii)で前記シーケンシングデータ中の前記標的に関連付けられた識別可能な配列を有する前記ノイズ分子標識の数を決定する際に、前記非ユニーク分子標識を除去する工程を含む、〔34〕に記載の方法。

〔36〕 前記確率バーコードが、識別可能な配列を有する約6561の分子標識を含む場合、前記再使用分子標識閾値が、約650である、〔35〕に記載の方法。

〔37〕 前記非ユニーク分子標識を除去する工程が、

30

前記シーケンシングデータ中の前記標的に関連付けられた識別可能な配列を有する前記分子標識の数について非ユニーク分子標識の理論上の数を決定する工程と；

前記シーケンシングデータ中の前記標的に関連付けられた識別可能な配列を有するn番目に豊富な前記分子標識よりも大きい発生数を有する分子標識を除去する工程と、

を含み、

nが、非ユニーク分子標識の理論数である、〔34〕～〔36〕のいずれか一項に記載の方法。

〔38〕 ハードウェアプロセッサと、

前記ハードウェアプロセッサによって実行される場合、前記プロセッサに〔1〕～〔37〕のいずれか一項に記載の方法を実行させる命令を記憶した非一過性メモリと、

40

を含む、ターゲットの数を決定するためのコンピュータシステム。

〔39〕 〔1〕～〔37〕のいずれか一項に記載の方法を実行するためのコードを含むソフトウェアプログラムを含む、コンピュータ読取り媒体。

〔40〕 標的の数を決定する方法であって、

(a) 複数の確率バーコードを用いて、複数の標的に確率バーコードを付けて、複数の確率バーコード付き標的を生成する工程であって、前記複数の確率バーコードの各々が分子標識を含む工程と；

(b) 前記確率バーコード付き標的のシーケンシングデータを取得する工程と；

(c) 前記複数の標的の1つ以上について：

(i) 前記シーケンシングデータ中の前記標的に関連付けられた識別可能な配列を有する

50

分子標識の数をカウントする工程と；

〔i i〕（b）で得られた前記シーケンシングデータ中の前記標的のクオリティステータスを決定する工程と；

〔i i i〕（b）で得られた前記シーケンシングデータ中の1つ以上のシーケンシングデータエラーを決定する工程であって、前記シーケンシングデータ中の前記1つ以上のシーケンシングデータエラーを決定する工程が、以下：前記シーケンシングデータ中の前記標的に関連付けられた識別可能な配列を有する前記分子標識の数、前記シーケンシングデータ中の前記標的の前記クオリティステータス、および前記複数の確率バーコード中の識別可能な配列を有する前記分子標識の数のうち1つ以上を決定することを含む工程と；

〔i v〕前記標的の数を推定する工程であって、推定された前記標的の数が、〔i i i〕で決定された前記1つ以上のシーケンシングデータエラーに応じて調節された、〔i〕でカウントされた前記シーケンシングデータ中の前記標的に関連付けられた識別可能な配列を有する前記分子標識の数と相関する工程と、を含む、方法。

〔4 1〕 前記1つ以上のシーケンシングデータエラーを決定する前に、（b）で得られた前記シーケンシングデータを折りたたむ工程をさらに含む、〔4 0〕に記載の方法。

〔4 2〕 （b）で得られた前記シーケンシングデータを折りたたむ工程が、類似した分子標識を有し、かつ、所定の折りたたみ発生数閾値よりも少ない発生数を有する標的のコピーを、前記複数の標的について同じ分子標識を有するものとして帰属させる工程を含み、前記標的の2つのコピーは、前記標的の前記2つのコピーの分子標識の配列が少なくとも1塩基相違する場合、類似の分子標識を有する、

〔4 1〕に記載の方法。

〔4 3〕 前記確率バーコードが、識別可能な配列を有する約6 5 6 1の分子標識を含む場合、前記所定の折りたたみ発生数閾値が7である、〔4 2〕に記載の方法。

〔4 4〕 前記確率バーコードが、識別可能な配列を有する約6 5 5 3 6の分子標識を含む場合、前記所定の折りたたみ発生数閾値が1 7である、〔4 2〕に記載の方法。

〔4 5〕 前記標的の2つのコピーが、前記標的の前記2つのコピーの分子標識の配列が少なくとも1塩基相違する場合、類似の分子標識を有する、〔4 2〕～〔4 4〕のいずれか一項に記載の方法。

〔4 6〕 前記分子標識が、5～2 0個のヌクレオチドを含む、〔4 0〕～〔4 5〕のいずれか一項に記載の方法。

〔4 7〕 異なる確率バーコードの前記分子標識が、互いに異なっている、〔4 0〕～〔4 6〕のいずれか一項に記載の方法。

〔4 8〕 前記複数の確率バーコードが、識別可能な配列を有する約6 5 6 1の分子標識を含む、〔4 0〕～〔4 7〕のいずれか一項に記載の方法。

〔4 9〕 前記複数の確率バーコードが、識別可能な配列を有する約6 5 5 3 6の分子標識を含む、〔4 0〕～〔4 7〕のいずれか一項に記載の方法。

〔5 0〕 前記シーケンシングデータが、5 0ヌクレオチド以上のリード長を有する前記複数の標的の配列を含む、〔4 0〕～〔4 9〕のいずれか一項に記載の方法。

〔5 1〕 前記シーケンシングデータが、7 5ヌクレオチド以上のリード長を有する前記複数の標的の配列を含む、〔4 0〕～〔4 9〕のいずれか一項に記載の方法。

〔5 2〕 前記シーケンシングデータが、1 0 0ヌクレオチド以上のリード長を有する前記複数の標的の配列を含む、〔4 0〕～〔4 9〕のいずれか一項に記載の方法。

〔5 3〕 （b）で得られた前記シーケンシングデータが、前記複数の確率バーコード付き標的に対してポリメラーゼ連鎖反応（PCR）増幅を実施することによって生成することができる、〔4 0〕～〔5 2〕のいずれか一項に記載の方法。

〔5 4〕 前記1つ以上のシーケンシングデータエラーが、PCR導入エラー、シーケンシング導入エラー、バーコード混入に起因するエラー、ライブラリー作製エラー、またはそれらの任意の組合せである、〔4 0〕～〔5 3〕のいずれか一項に記載の方法。

10

20

30

40

50

〔55〕 前記PCR導入エラーが、PCR増幅エラー、PCR増幅バイアス、不十分なPCR増幅、またはそれらの任意の組合せの結果である、〔54〕に記載の方法。

〔56〕 前記シーケンシング導入エラーが、不正確なベースコーリング、不十分なシーケンシング、またはそれらの任意の組合せの結果である、〔54〕～〔55〕のいずれか一項に記載の方法。

〔57〕 工程(i)、(ii)、(iii)、および(iv)が、前記複数の標的の各々について実施される、〔40〕～〔56〕のいずれか一項に記載の方法。

〔58〕 前記シーケンシングデータ中の前記標的の前記クオリティステータスが、完全シーケンシング、不完全シーケンシング、または飽和シーケンシングである、〔40〕～〔57〕のいずれか一項に記載の方法。

10

〔59〕 前記シーケンシングデータ中の標的のクオリティステータスが、前記複数の確率バーコード中に識別可能な配列を有する前記分子標識の数と、カウントされた前記シーケンシングデータ中の前記標的に関連付けられた識別可能な配列を有する前記分子標識の数とによって決定される、〔58〕に記載の方法。

〔60〕 前記完全シーケンシングクオリティステータスが、所定の完全シーケンシング散布閾値以上の前記ポアソン分布と比較した散布指数によって決定され、前記所定の完全シーケンシング散布閾値が、0.9である、〔58〕～〔59〕のいずれか一項に記載の方法。

〔61〕 前記所定の完全シーケンシング散布閾値が、1である、〔60〕に記載の方法。

〔62〕 前記所定の完全シーケンシング散布閾値が、4である、〔60〕に記載の方法。

20

〔63〕 前記完全シーケンシングクオリティステータスが、(b)で得られた前記シーケンシングデータ中の所定の完全シーケンシング発生数閾値以上の発生数を有する分子標識によってさらに決定され、前記所定の完全シーケンシング発生数閾値が、10である、〔60〕～〔62〕のいずれか一項に記載の方法。

〔64〕 前記所定の完全シーケンシング発生数閾値が、18である、〔63〕に記載の方法。

〔65〕 前記飽和シーケンシングクオリティステータスが、所定の飽和閾値よりも大きい、識別可能な配列を含む分子標識の数を有する前記標的によって決定される、〔58〕～〔64〕のいずれか一項に記載の方法。

〔66〕 前記飽和シーケンシングクオリティステータスが、前記所定の飽和閾値よりも大きい、識別可能な配列を含む分子標識の数を有する前記複数の標的のうちの1つの他の標的によって、さらに決定される、〔65〕に記載の方法。

30

〔67〕 前記確率バーコードが、識別可能な配列を有する約6561の分子標識を含む場合、前記所定の飽和閾値が、6557である、〔65〕に記載の方法。

〔68〕 前記確率バーコードが、識別可能な配列を有する約65536の分子標識を含む場合、前記所定の飽和閾値が、65532である、〔65〕に記載の方法。

〔69〕 前記シーケンシングデータ中の前記標的の前記クオリティステータスは、(b)で得られた前記シーケンシングデータ中の前記標的の前記クオリティステータスが、完全シーケンシングではなく、かつ、飽和シーケンシングではない場合に、不完全シーケンシングとして分類される、〔40〕～〔68〕のいずれか一項に記載の方法。

40

〔70〕 (i)でカウントされた前記シーケンシングデータ中の前記標的に関連付けられた識別可能な配列を有する前記分子標識の数が、(iv)において、前記標的が前記完全シーケンシングクオリティステータスを有している場合、

1つ以上の親分子標識についてすべての子供分子標識を決定する工程と；

少なくとも1つの子供分子標識および前記親分子標識について第1の統計解析を実施する工程と；

前記第1の統計解析の帰無仮説が容認される場合、前記子供分子標識の前記発生数を前記親分子標識に帰属させる工程と、

によって調節される、〔50〕～〔69〕のいずれか一項に記載の方法。

〔71〕 前記1つ以上の親分子標識が、所定の完全シーケンシング親閾値以上の発生数

50

を有する分子標識を含み、前記所定の完全シーケンシング親閾値が、前記所定の完全シーケンシング発生数閾値と等しい、〔70〕に記載の方法。

〔72〕 前記子供分子標識が、前記親分子標識と1塩基相違し、かつ、所定の完全シーケンシング子供閾値以下の発生数を有する分子標識を含み、前記所定の完全シーケンシング子供閾値が、3である、〔70〕～〔71〕のいずれか一項に記載の方法。

〔73〕 前記所定の完全シーケンシング子供閾値が、5である、〔72〕に記載の方法。

〔74〕 前記帰無仮説が真である確率が偽発見率を下回る場合、前記第1の統計解析の前記帰無仮説が容認され、前記偽発見率が、5%である、〔70〕～〔73〕のいずれか一項に記載の方法。

〔75〕 前記偽発見率が10%である、〔74〕に記載の方法。

10

〔76〕 前記第1の統計解析が、多重二項検定である、〔70〕～〔75〕のいずれか一項に記載の方法。

〔77〕 (i) でカウントされた前記シーケンシングデータ中の前記標的に関連付けられた識別可能な配列を有する前記分子標識の数は、(iv)において、前記標的が前記完全シーケンシングクオリティステータスを有する場合、前記標的の分子標識を閾値化して、(b) で得られた前記シーケンシングデータ中の前記標的に関連付けられた真の分子標識および偽の分子標識を決定する工程によって調節される、〔50〕～〔76〕のいずれか一項に記載の方法。

〔78〕 前記標的の前記分子標識を閾値化する工程が、前記標的の前記分子標識について第2の統計解析を実施する工程を含む、〔77〕に記載の方法。

20

〔79〕 前記第2の統計解析を実施する工程が、前記標的の前記分子標識の分布およびそれらの発生数を2つのポアソン分布に当てはめる工程と；

前記2つのポアソン分布を用いて真の分子標識の数 n を決定する工程と；

(b) で得られた前記シーケンシングデータから前記偽の分子標識を除去する工程と、を含み、

前記偽の分子標識が、 n 番目に豊富な分子標識の前記発生数よりも低い発生数を有する分子標識を含み、前記真の分子標識が、 n 番目に豊富な分子標識の前記発生数以上の発生数を有する分子標識を含む、〔78〕に記載の方法。

〔80〕 前記2つのポアソン分布が、前記真の分子標識に対応する第1のポアソン分布と、前記偽の分子標識に対応する第2のポアソン分布を含む、〔79〕に記載の方法。

30

〔81〕 (i) でカウントされた前記シーケンシングデータ中の前記標的に関連付けられた識別可能な配列を有する前記分子標識の数が、(iv)において、

(b) で得られた前記シーケンシングデータ中の前記標的の前記クオリティステータスが、前記不完全シーケンシングクオリティステータスである場合、前記標的が、(b) で得られたシーケンシングデータにおいてノイズであるか否かを決定する工程と；

(b) で得られた前記シーケンシングデータから前記ノイズ標的を除去する工程と、によって調節される、〔58〕～〔80〕のいずれか一項に記載の方法。

〔82〕 前記ノイズ標的の前記分子標識の前記発生数が、不完全シーケンシングクオリティ標的閾値以下であれば、前記標的はノイズであり、前記不完全シーケンシングノイズ遺伝子閾値が、5である、〔81〕に記載の方法。

40

〔83〕 前記不完全シーケンシングノイズ標的閾値が、完全シーケンシングのクオリティステータスを有する前記複数の標的の前記分子標識の前記中央発生数と等しい、〔82〕に記載の方法。

〔84〕 前記不完全シーケンシングノイズ標的閾値が、完全シーケンシングのクオリティステータスを有する前記複数の標的の前記分子標識の前記平均発生数と等しい、〔82〕に記載の方法。

〔85〕 (i) でカウントされた前記シーケンシングデータ中の前記標的に関連付けられた識別可能な配列を有する前記分子標識の数が、(iv)において、

50

(b) で得られた前記シーケンシングデータ中の前記標的の前記クオリティステータスが前記不完全シーケンシングクオリティステータスである場合、前記標的の前記分子標識を閾値化して、(b) で得られた前記シーケンシングデータ中の真の分子標識および偽の分子標識を決定する工程によって調節される、〔50〕～〔84〕のいずれか一項に記載の方法。

〔86〕 前記標的の前記分子標識を閾値化する工程が、前記分子標識について第3の統計解析を実施する工程を含む、〔85〕に記載の方法。

〔87〕 前記分子標識について前記第3の統計解析を実施する工程が、ゼロ切断ポアソンモデルを用いて、真の分子標識の数 n を決定する工程と；

(b) で得られた前記シーケンシングデータから前記偽の分子標識を除去する工程と、を含み、前記偽の分子標識が、 n 番目に豊富な分子標識の発生数よりも低い発生数を有する分子標識を含み、前記真の分子標識が、 n 番目に豊富な分子標識の前記発生数以上の発生数を有する分子標識を含む、〔86〕に記載の方法。

〔88〕 (i) でカウントされた前記シーケンシングデータが、(iii) で決定された前記1つ以上のシーケンシングデータエラーに応じて調節された後、(b) で得られた前記シーケンシングデータ中の前記分子標識の少なくとも50%が保持される、〔40〕～〔87〕のいずれか一項に記載の方法。

〔89〕 (i) でカウントされた前記シーケンシングデータが、(iii) で決定された前記1つ以上のシーケンシングデータエラーに応じて調節された後、(b) b) で得られた前記シーケンシングデータ中の前記分子標識の少なくとも80%が保持される、〔40〕～〔87〕のいずれか一項に記載の方法。

〔90〕 前記複数の標的に確率バーコードを付ける工程が、前記複数の確率バーコードを前記複数の標的とハイブリダイズさせて、前記確率バーコード付き標的を生成する工程を含む、〔40〕～〔87〕のいずれか一項に記載の方法。

〔91〕 前記複数の標的に確率バーコードを付ける工程が、前記確率バーコード付き標的のインデックス付きライブラリーを作製する工程を含む、〔89〕に記載の方法。

〔92〕 前記確率バーコード付き標的のインデックス付きライブラリーを作製する工程が、前記複数の確率バーコードを含む固体担体を用いて実施される、〔89〕～〔91〕のいずれか一項に記載の方法。

〔93〕 前記固体担体が、前記複数の確率バーコードと結合した複数の合成粒子を含む、〔92〕に記載の方法。

〔94〕 前記複数の確率バーコードの各々が、サンプル標識、ユニバーサル標識および細胞標識の1つ以上を含み、前記サンプル標識が、前記固体担体上の前記複数の確率バーコードに対するものと同じであり、ユニバーサル標識が、前記固体担体上の前記複数の確率バーコードに対するものと同じであり、細胞標識が、前記固体担体上の前記複数の確率バーコードに対するものと同じである、〔92〕～〔93〕のいずれか一項に記載の方法。

〔95〕 前記サンプル標識が、5～20ヌクレオチドを含む、〔94〕に記載の方法。

〔96〕 前記ユニバーサル標識が、5～20ヌクレオチドを含む、〔94〕～〔95〕のいずれか一項に記載の方法。

〔97〕 前記細胞標識が、5～20ヌクレオチドを含む、〔94〕～〔96〕のいずれか一項に記載の方法。

〔98〕 前記固体担体が、2次元または3次元の前記複数の確率バーコードを含む、〔92〕～〔95〕のいずれか一項に記載の方法。

〔99〕 前記合成粒子がビーズである、〔93〕～〔98〕のいずれか一項に記載の方法。

〔100〕 前記ビーズが、シリカゲルビーズ、調節多孔性ガラスビーズ、磁気ビーズ、ダイナビーズ、セファデックス／セファロースビーズ、セルロースビーズ、ポリスチレンビーズ、またはそれらの任意の組合せである、〔99〕に記載の方法。

〔101〕 前記固体担体が、ポリマー、マトリックス、ヒドロゲル、ニードルアレイデ

10

20

30

40

50

バイス、抗体、またはそれらの任意の組合せを含む、〔40〕～〔100〕に記載の方法。

〔102〕 前記複数の標的がサンプル中に含まれる、〔40〕～〔101〕のいずれか一項に記載の方法。

〔103〕 前記サンプルが、1つ以上の細胞を含む、〔102〕に記載の方法。

〔104〕 前記サンプルが単一細胞である、〔102〕に記載の方法。

〔105〕 前記1つ以上の細胞を溶解する工程をさらに含む、〔102〕に記載の方法。

〔106〕 前記1つ以上の細胞を溶解する工程が、前記サンプルを加熱する工程、前記サンプルを洗剤と接触させる工程、前記サンプルのpHを変える工程、またはそれらの任意の組合せを含む、〔105〕に記載の方法。

〔107〕 前記1つ以上の細胞が、1つ以上の細胞型を含む、〔102〕に記載の方法。

〔108〕 前記1つ以上の細胞型の少なくとも1つが、脳細胞、心細胞、癌細胞、循環腫瘍細胞、臓器細胞、上皮細胞、転移細胞、良性細胞、一次細胞、循環細胞、またはそれらの任意の組合せである、〔107〕に記載の方法。

〔109〕 前記複数の標的が、リボ核酸(RNA)、メッセンジャーRNA(mRNA)、microRNA、低分子干渉RNA(siRNA)、RNA分解産物、ポリ(A)テールを各々含むRNA、またはそれらの任意の組合せを含む、〔40〕～〔108〕のいずれか一項に記載の方法。

〔110〕 前記方法が多重化される、〔40〕～〔109〕のいずれか一項に記載の方法。

〔111〕 標的の数を決定する方法であって、

(a) 複数の確率バーコードを用いて、複数の標的に確率バーコードを付けて、複数の確率バーコード付き標的を生成する工程であって、前記複数の確率バーコードの各々が分子標識を含む工程と；

(b) 前記確率バーコード付き標的のシーケンシングデータを取得する工程と；

(c) 前記複数の標的の1つ以上について：

(i) 前記シーケンシングデータ中の前記標的に関連付けられた識別可能な配列を有する分子標識の数をカウントする工程と；

(ii) 方向近接性を用いて、前記標的の分子標識のクラスターを同定する工程と；

(iii) (ii) で同定された前記標的の分子標識の前記クラスターを用いて、(b) で得られた前記シーケンシングデータを折りたたむ工程と；

(iv) 前記標的の数を推定する工程であって、推定された前記標的の数が、(ii) の前記シーケンシングデータの折りたたみ後に、(i) でカウントされた前記シーケンシングデータ中の前記標的に関連付けられた識別可能な配列を有する分子標識の数と相関する工程と、

を含む、方法。

〔112〕 前記複数の標的が、細胞の全トランスクリプトームの標的を含む、〔111〕に記載の方法。

〔113〕 クラスター内の前記標的の分子標識が、互いの所定の方向近接性閾値内にある、〔111〕～〔112〕のいずれか一項に記載の方法。

〔114〕 前記方向近接性閾値が、1のハミング距離である、〔113〕に記載の方法。

〔115〕 前記クラスター内の前記標的の前記分子標識が、1つ以上の親分子標識と、前記1つ以上の親分子標識の子供分子標識とを含み、前記親分子標識の発生数が、所定の方向近接性発生数閾値以上である、〔112〕～〔114〕のいずれか一項に記載の方法。

〔116〕 前記所定の方向近接性発生数閾値が、 $2 \times (\text{子供分子標識の発生数}) - 1$ である、〔115〕に記載の方法。

〔117〕 (ii) で同定された前記標的の分子標識の前記クラスターを用いて、(b) で得られたシーケンシングデータを折りたたむ工程が、前記子供分子標識の前記発生数を前記親分子標識に帰属させる工程を含む、〔111〕～〔116〕のいずれか一項に記載の方法。

〔118〕 前記標的のシーケンシング深度を決定する工程をさらに含む、〔111〕～

10

20

30

40

50

〔117〕のいずれか一項に記載の方法。

〔119〕 前記標的の前記シーケンシング深度が所定のシーケンシング深度閾値を超える場合、前記標的の数を推定する工程が、(i)でカウントされた前記シーケンシングデータを調節する工程を含む、〔118〕に記載の方法。

〔120〕 前記所定のシーケンシング深度閾値が、15～20である、〔119〕に記載の方法。

〔121〕 (i)でカウントされた前記シーケンシングデータを調節する工程が、前記標的の分子標識を閾値化して、(b)で得られた前記シーケンシングデータ中の前記標的に関連付けられた真の分子標識および偽の分子標識を決定する工程を含む、〔119〕～〔120〕のいずれか一項に記載の方法。

10

〔122〕 前記標的の前記分子標識を閾値化する工程が、前記標的の前記分子標識について統計解析を実施する工程を含む、〔121〕に記載の方法。

〔123〕 前記統計解析を実施する工程が、前記標的の前記分子標識の分布およびそれらの発生数を2つのポアソン分布に当てはめる工程と；

前記2つのポアソン分布を用いて真の分子標識の数nを決定する工程と；

(b)で得られた前記シーケンシングデータから前記偽の分子標識を除去する工程と、を含む、

前記偽の分子標識が、n番目に豊富な分子標識の前記発生数よりも低い発生数を有する分子標識を含み、前記真の分子標識が、n番目に豊富な分子標識の前記発生数以上の発生数を有する分子標識を含む、〔122〕に記載の方法。

20

〔124〕 前記2つのポアソン分布が、前記真の分子標識に対応する第1のポアソン分布と、前記偽の分子標識に対応する第2のポアソン分布を含む、〔123〕に記載の方法。

〔125〕 ハードウェアプロセッサと、

前記ハードウェアプロセッサによって実行される場合、前記プロセッサに〔40〕～〔124〕のいずれか一項に記載の方法を実行させる命令を記憶した非一過性メモリと、を含む、ターゲットの数を決定するためのコンピュータシステム。

〔126〕 〔40〕～〔124〕のいずれか一項に記載の方法を実行するためのコードを含むソフトウェアプログラムを含む、コンピュータ読取り媒体。

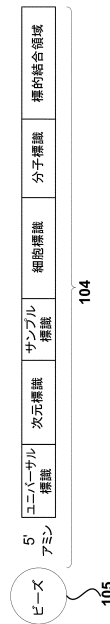
30

40

50

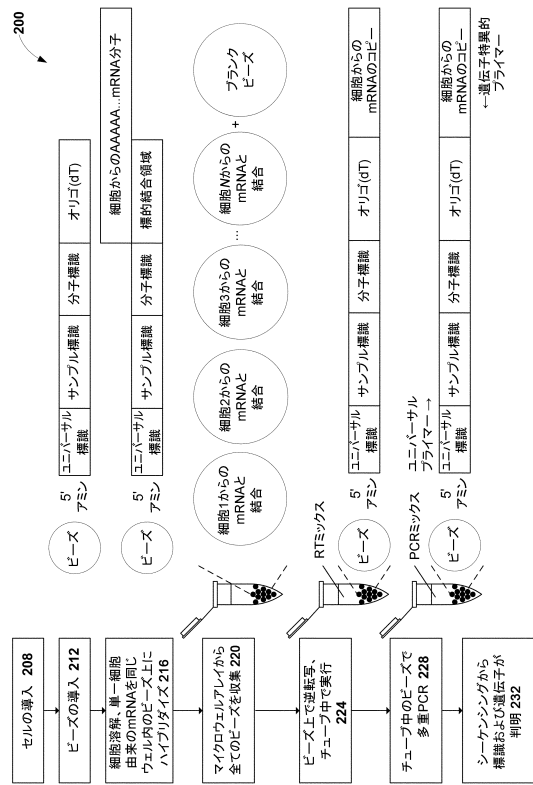
【凶面】

【図 1】



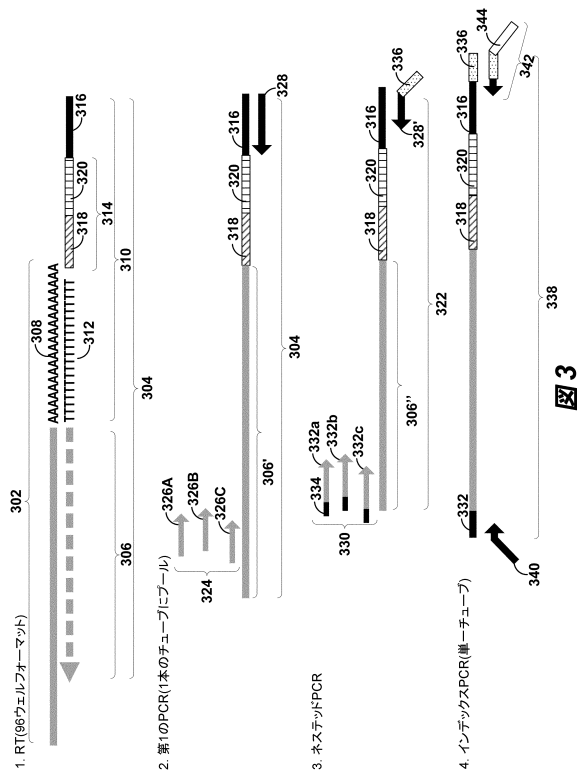
1

【図 2】



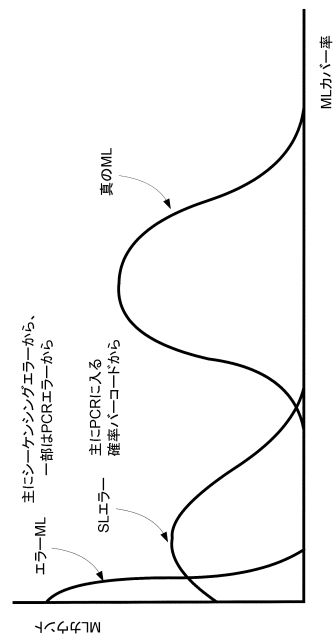
2

【図 3】



3

【図 4】



4.

【図 5】

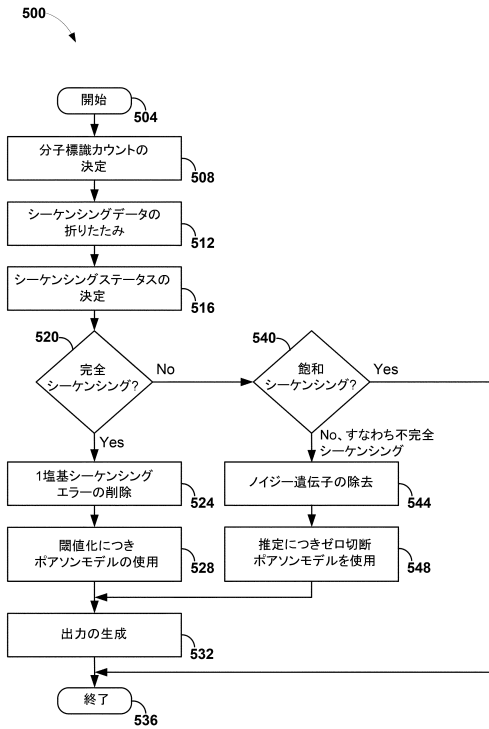
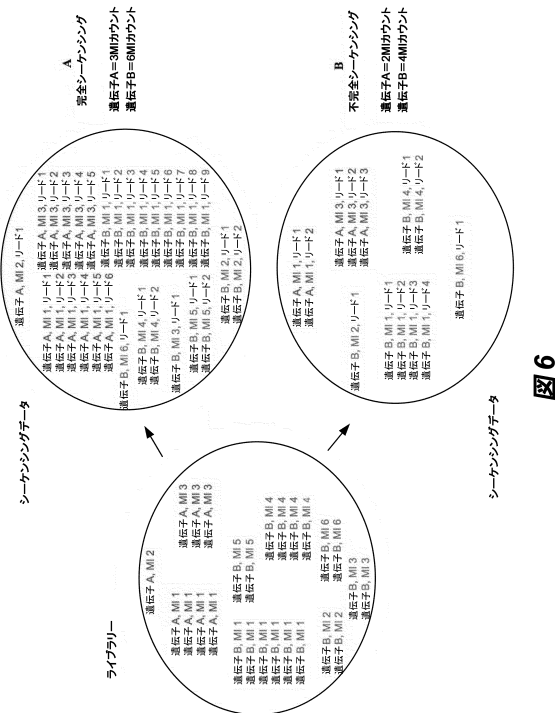


図 5

【図 6】



【図 7】

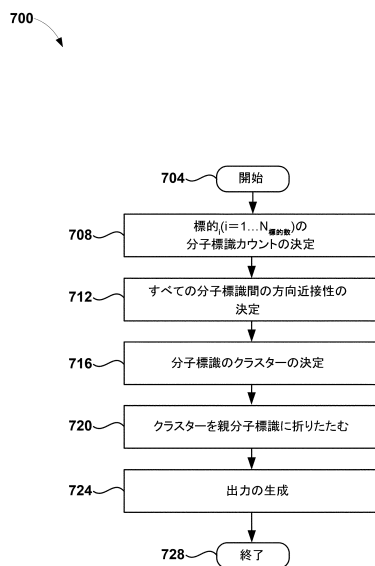


図 7

【図 8】

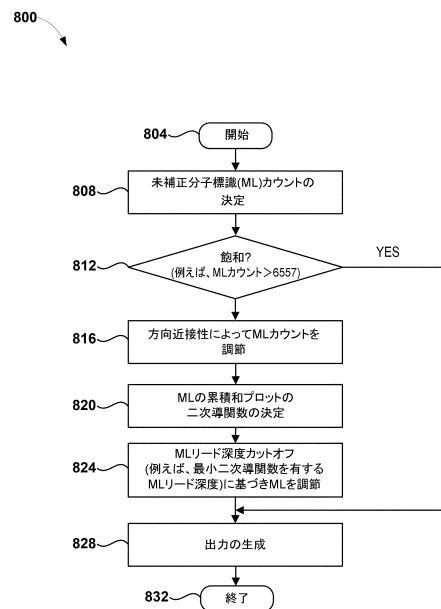


図 8

10

20

30

40

50

【図 9】

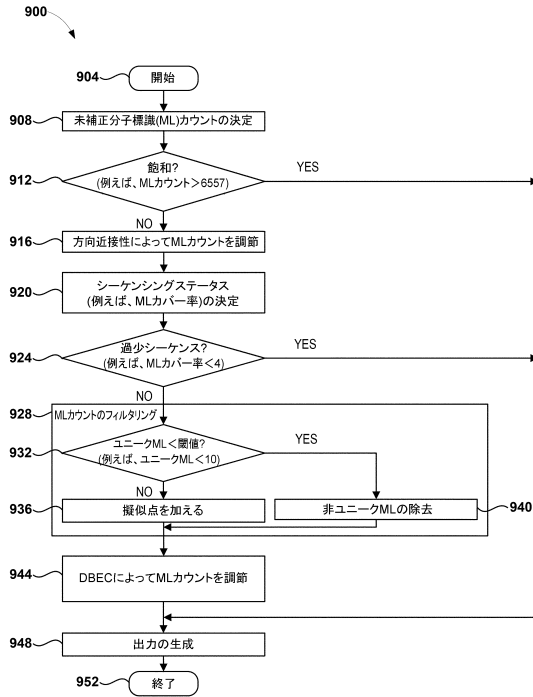


図 9

【図 10】

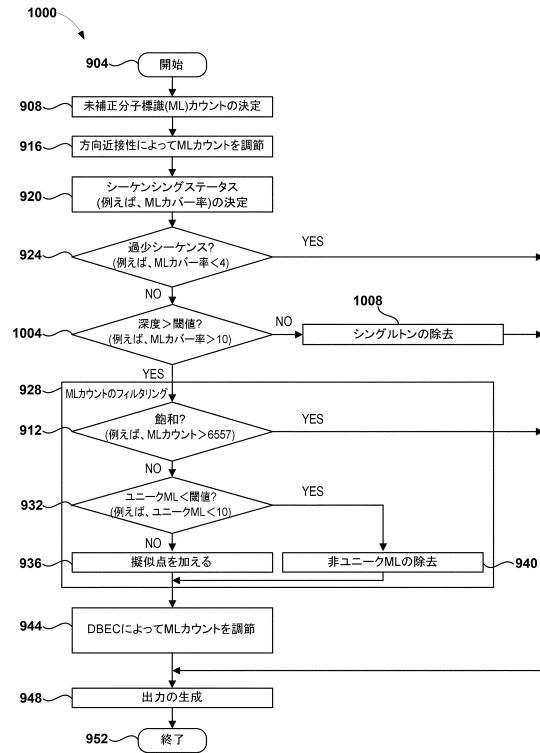


図 10

【図 11】

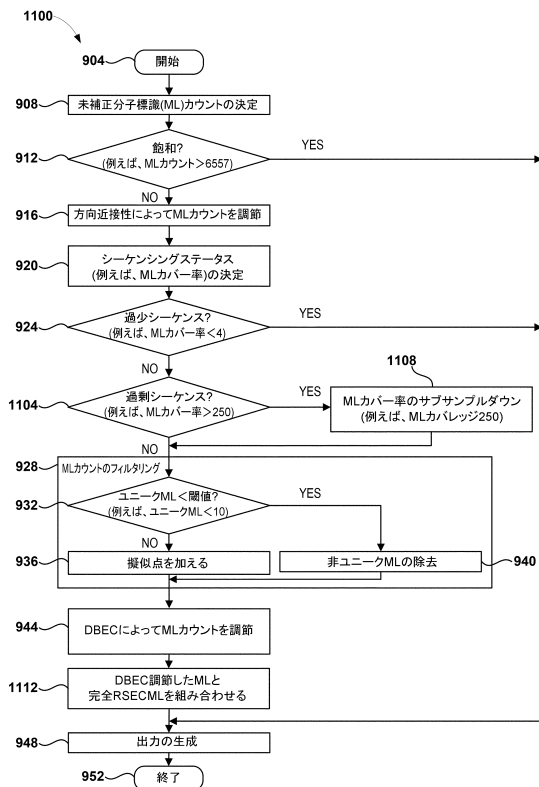


図 11

【図 12】

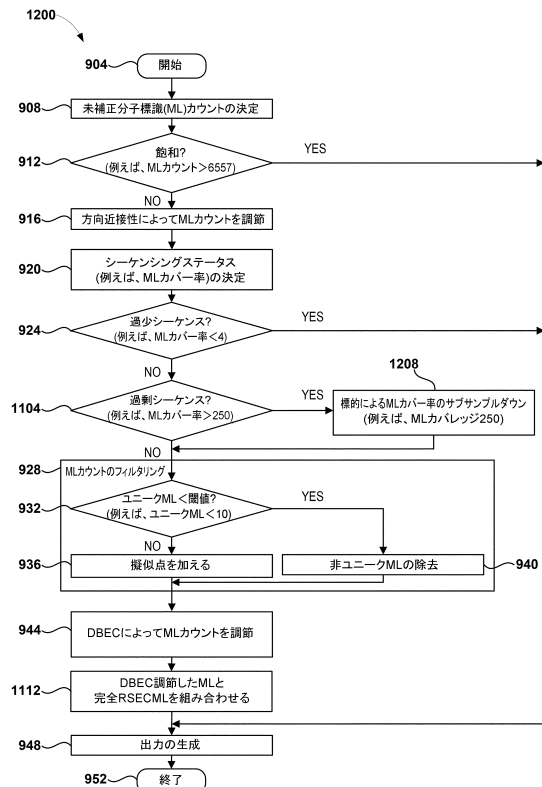


図 12

10

20

30

40

50

【図 1 3】

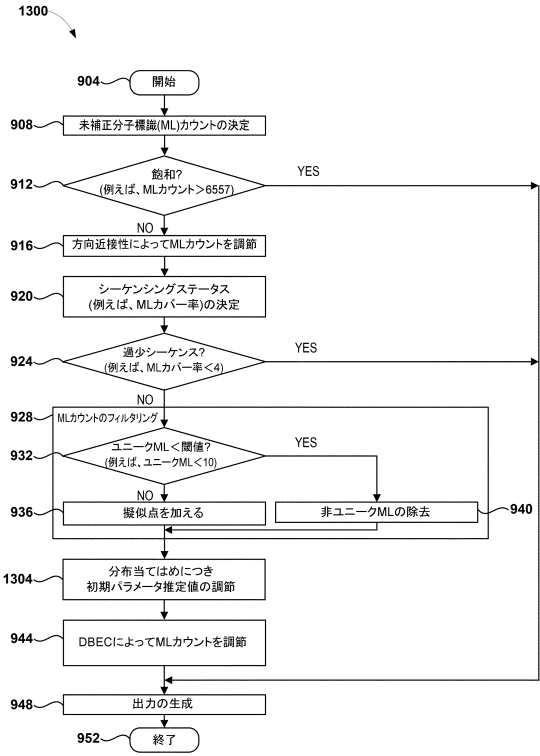


図13

【図 1 4】

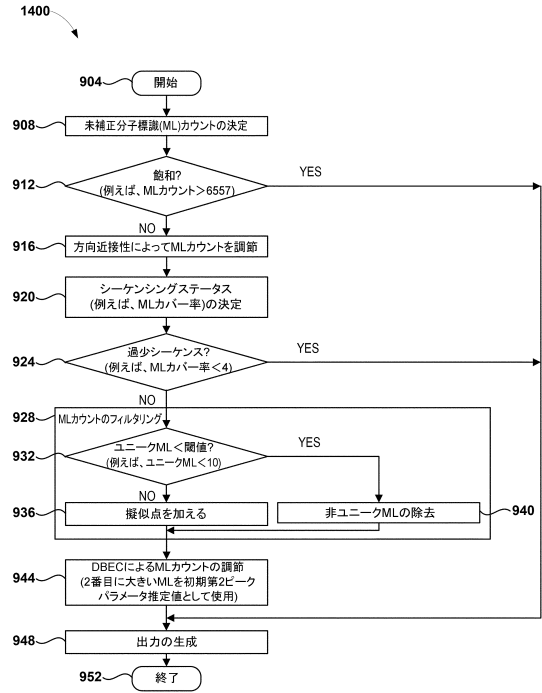


図14

【図 1 5】

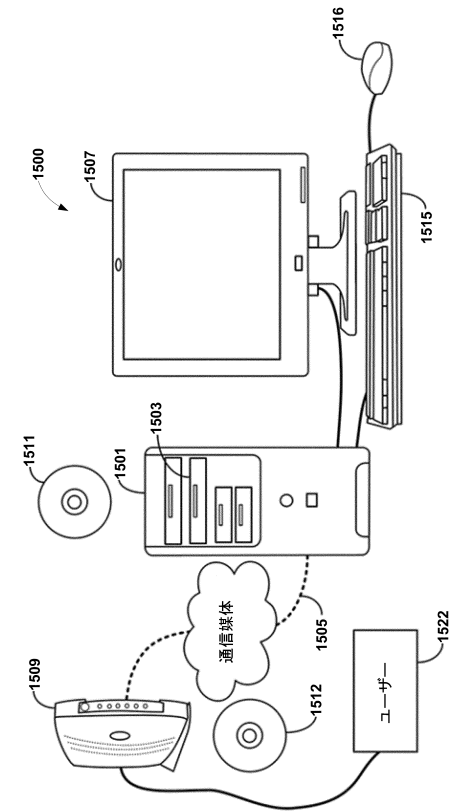


図15

【図 1 6】

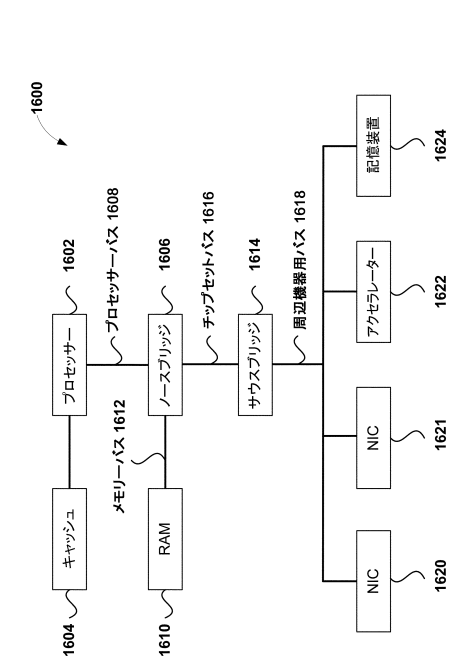


図16

10

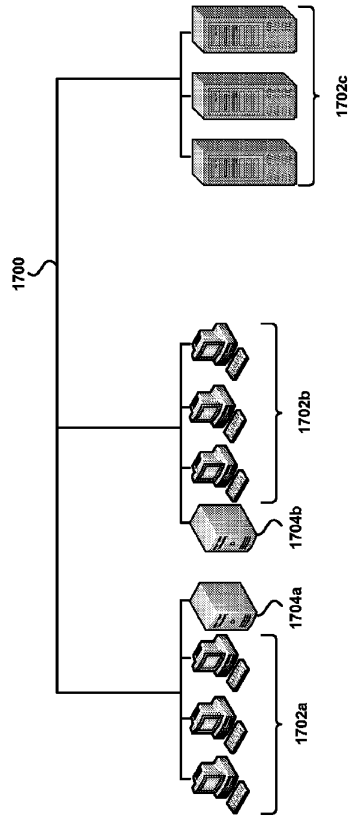
20

30

40

50

【図 17】



【図 19】

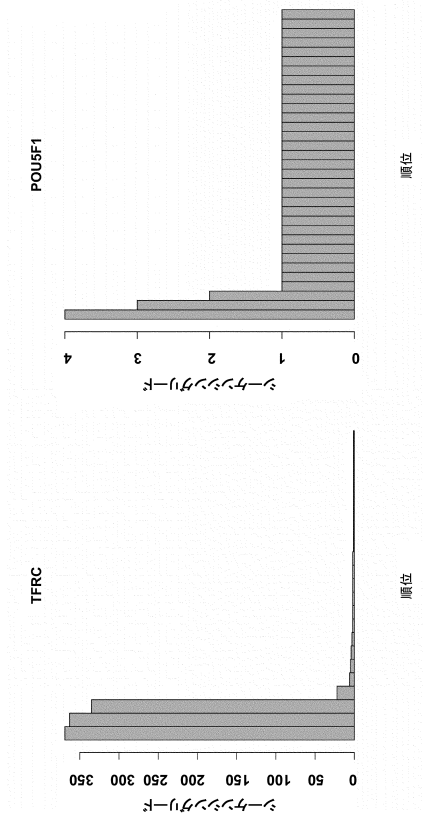


FIG. 17

【図 18】

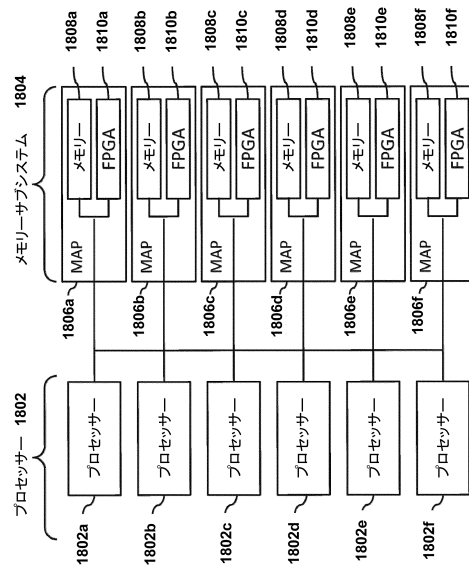


図 18

【図 20】

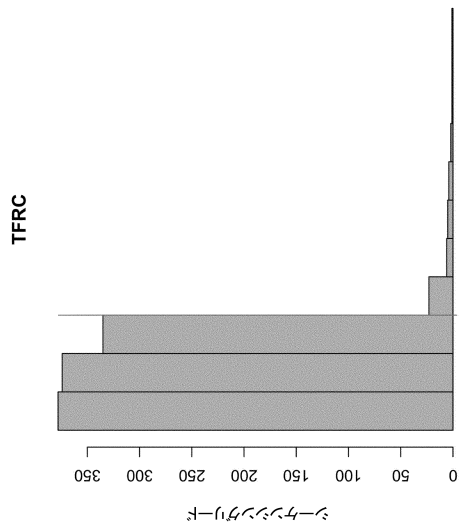


図 20

図 19B

図 19A

10

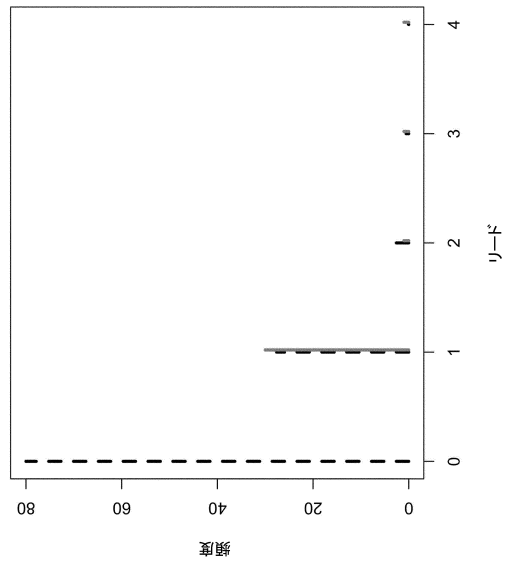
20

30

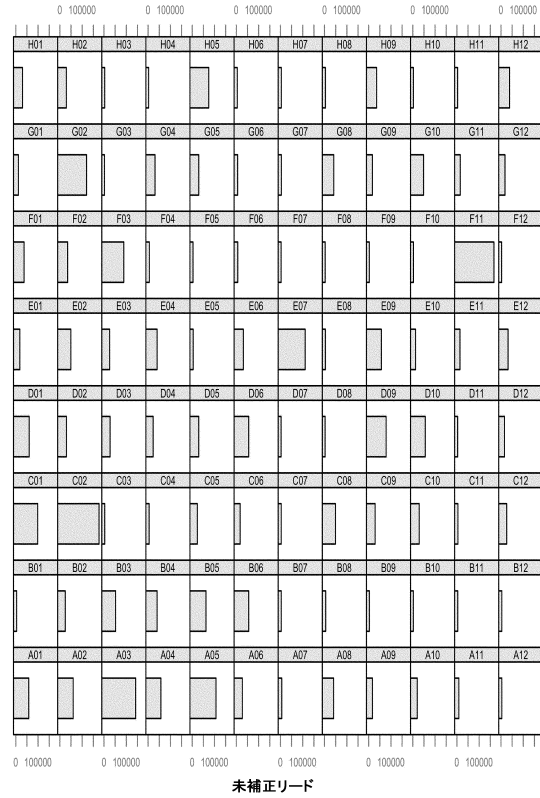
40

50

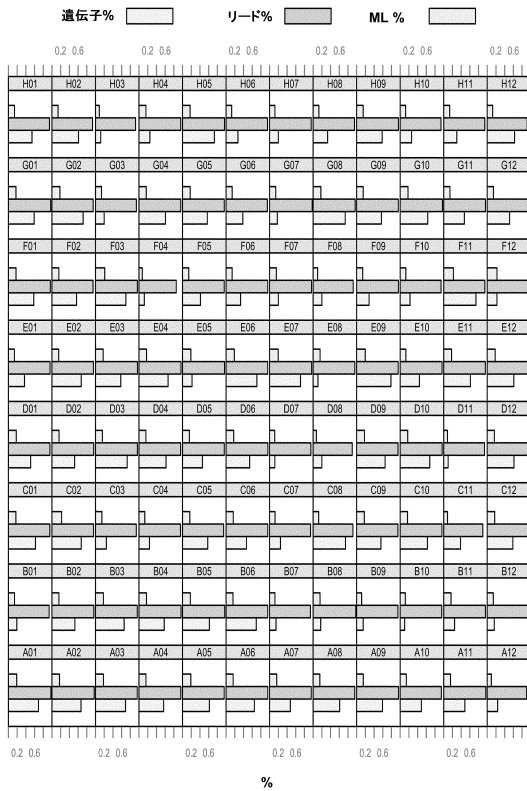
【図 2 1】



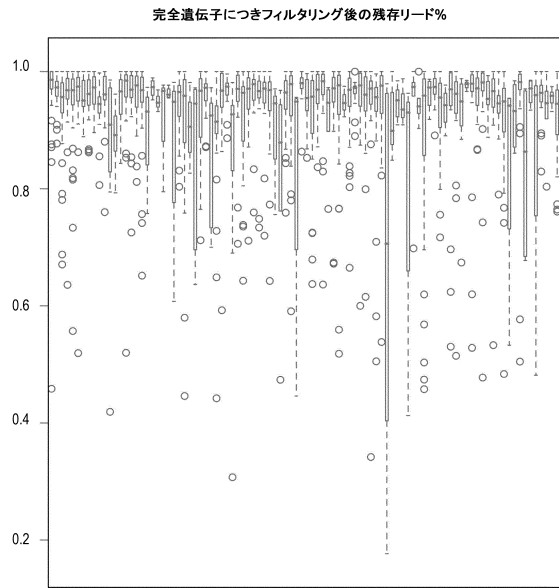
【図 2 2】



【図 2 3】



【図 2 4】



10

20

30

40

50

【図 29】

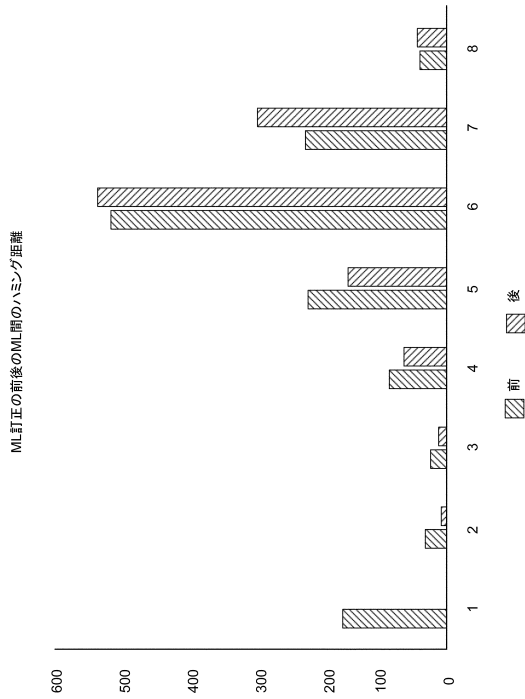


図 29

【図 30】

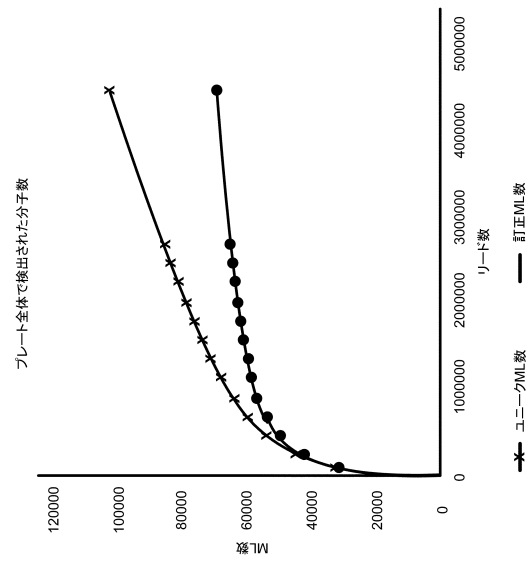


図 30

【図 31】

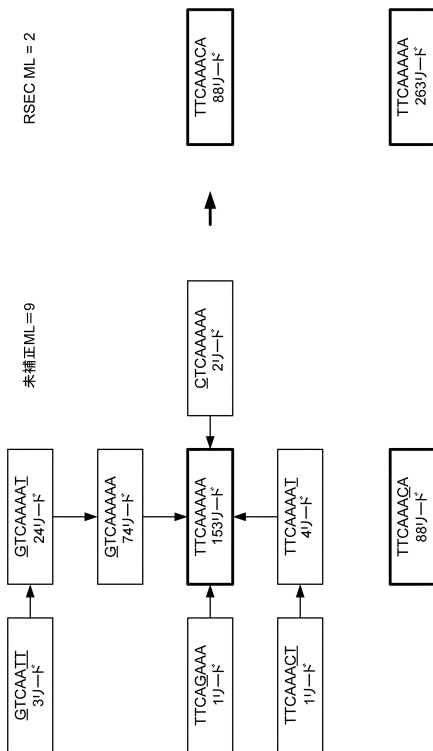


図 31

【図 32】

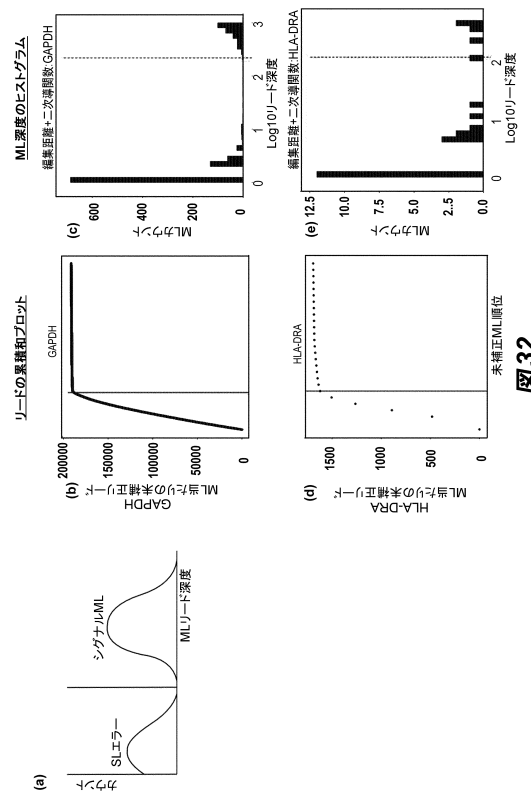
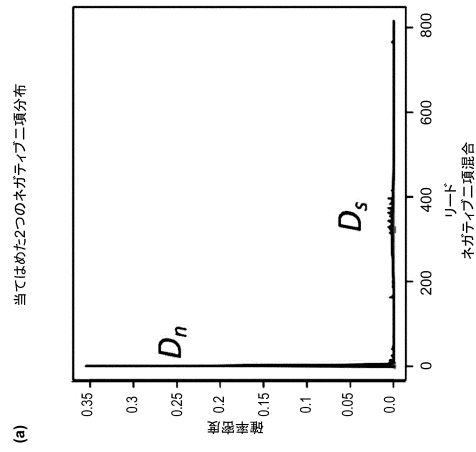


図 32

【図 3 3 - 1】



【図 3 3 - 2】

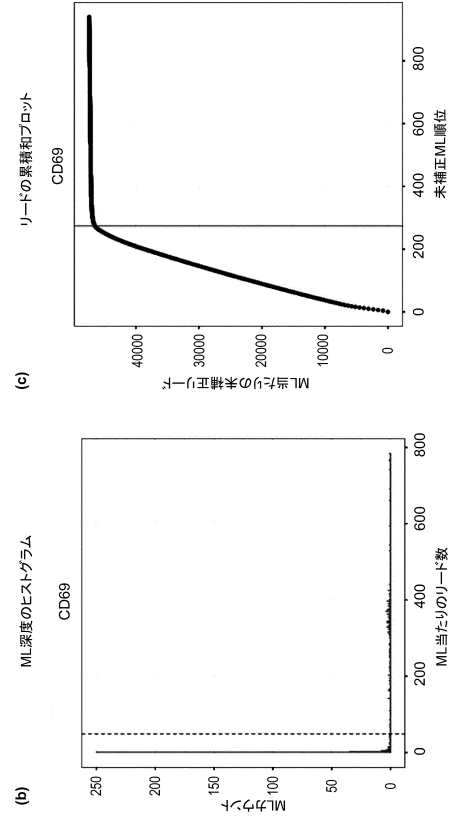
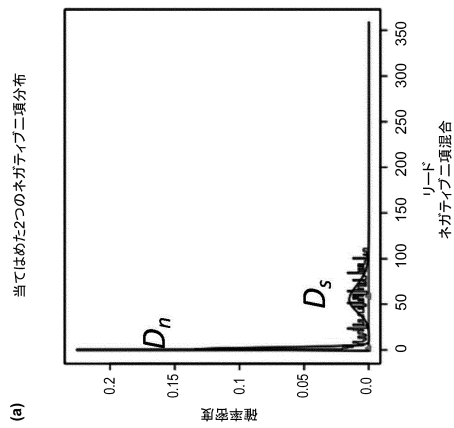


図 33

図 33 (続き)

【図 3 4 - 1】



【図 3 4 - 2】

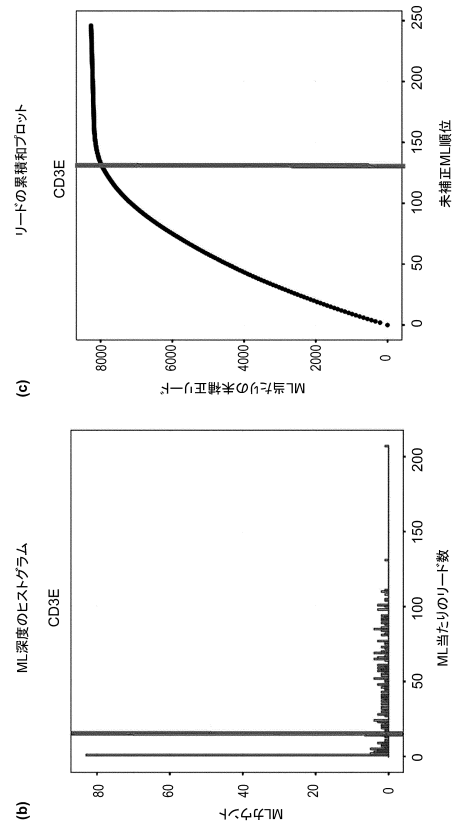


図 34

図 34 (続き)

【図 3 5 - 1】

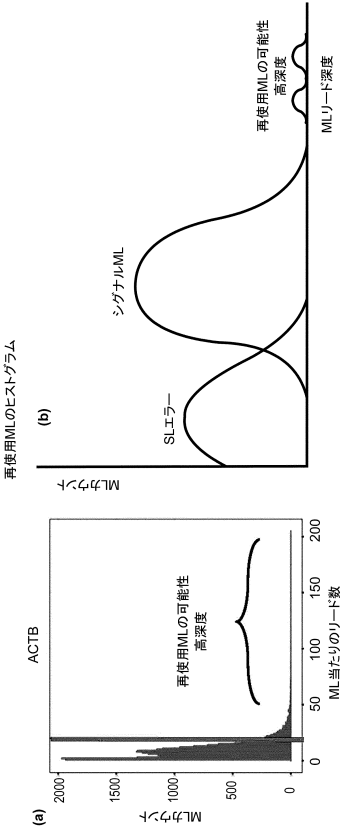


図 35

【図 3 5 - 2】

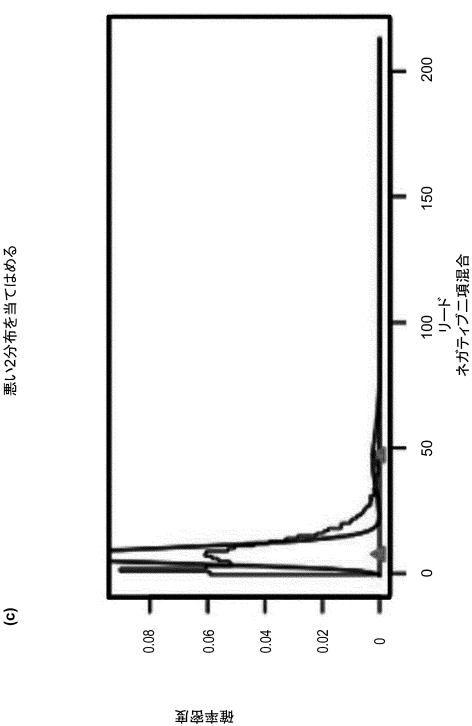


図 35 (続き)

【図 3 6】

遺伝子ID	読取ML	ML 当たりのリード数	遺伝子ID	読取ML	ML 当たりのリード数
GAPDH	GGGGGGGG	1417	ACTB	TTTTTTTA	5395
GAPDH	GGGGGGGG	702	ACTB	GGGGGGGG	4014
GAPDH	GGGGGGGG	630	ACTB	GGGGGGGG	3890
GAPDH	GGGGGGGG	626	ACTB	GGGGGGGG	3749
GAPDH	GGGGGGGG	598	ACTB	GGGGGGGG	3660
GAPDH	GGGGGGGG	598	ACTB	GGGGGGGG	3435
GAPDH	GGGGGGGG	581	ACTB	GGGGGGGG	3347
GAPDH	GGGGGGGG	575	ACTB	GGGGGGGG	3282
GAPDH	GGGGGGGG	563	ACTB	GGGGGGGG	3170
GAPDH	GGGGGGGG	551	ACTB	GGGGGGGG	3070
GAPDH	GGGGGGGG	511	ACTB	GGGGGGGG	2913
GAPDH	GGGGGGGG	479	ACTB	GGGGGGGG	2861
GAPDH	GGGGGGGG	474	ACTB	GGGGGGGG	2817
GAPDH	GGGGGGGG	444	ACTB	GGGGGGGG	2734
GAPDH	GGGGGGGG	442	ACTB	GGGGGGGG	2692
GAPDH	GGGGGGGG	430	ACTB	GGGGGGGG	2667
GAPDH	GGGGGGGG	423	ACTB	GGGGGGGG	2642
GAPDH	GGGGGGGG	419	ACTB	GGGGGGGG	2540
GAPDH	GGGGGGGG	419	ACTB	GGGGGGGG	2508

図 36

【図 3 7】

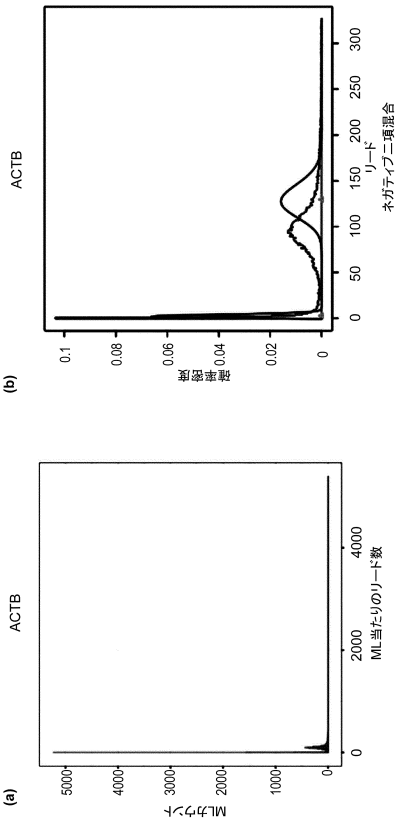


図 37

【図 38 - 1】

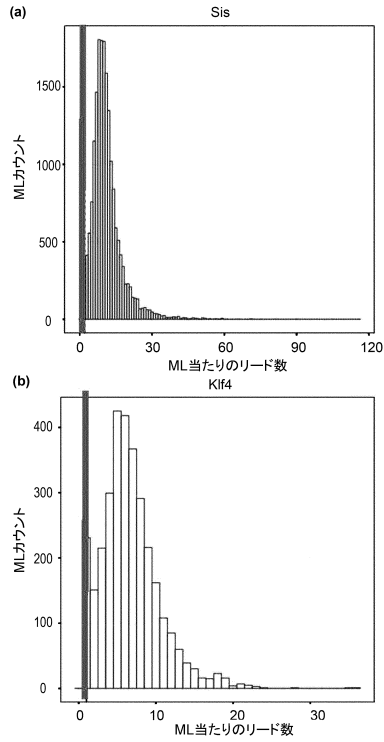


図 38

【図 38 - 2】

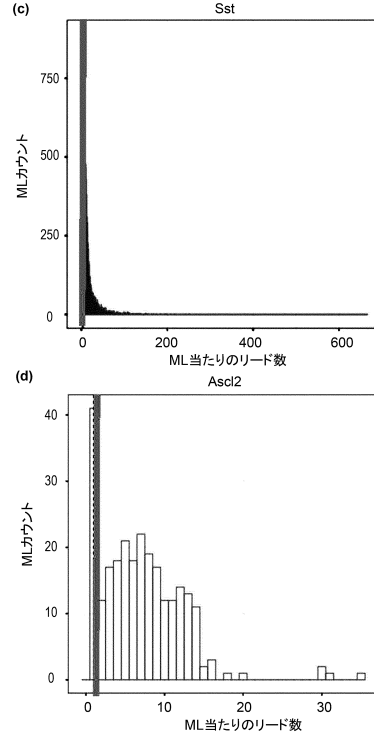


図 38 (続き)

【図 38 - 3】

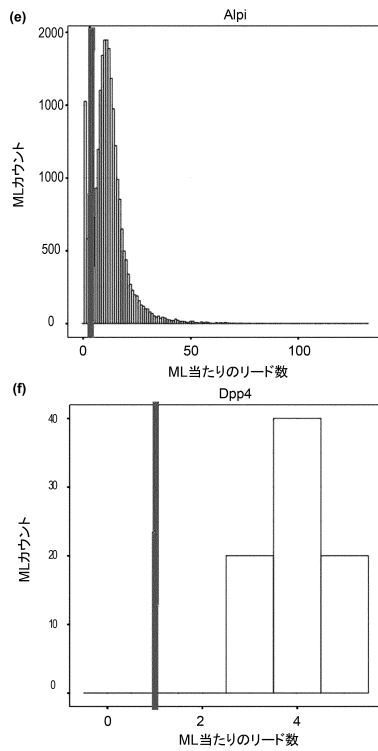


図 38 (続き)

【図 38 - 4】

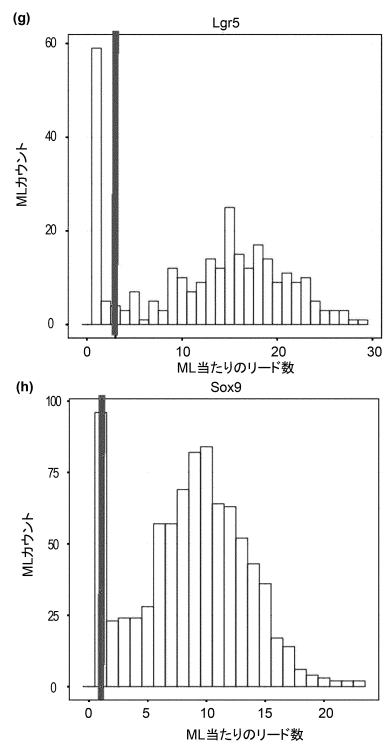


図 38 (続き)

10

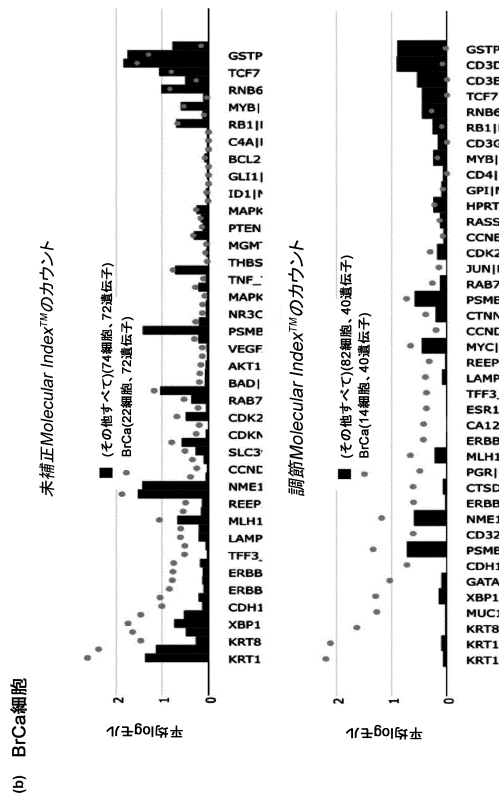
20

30

40

50

【図 40-2】



【圖 4 1 - 1】

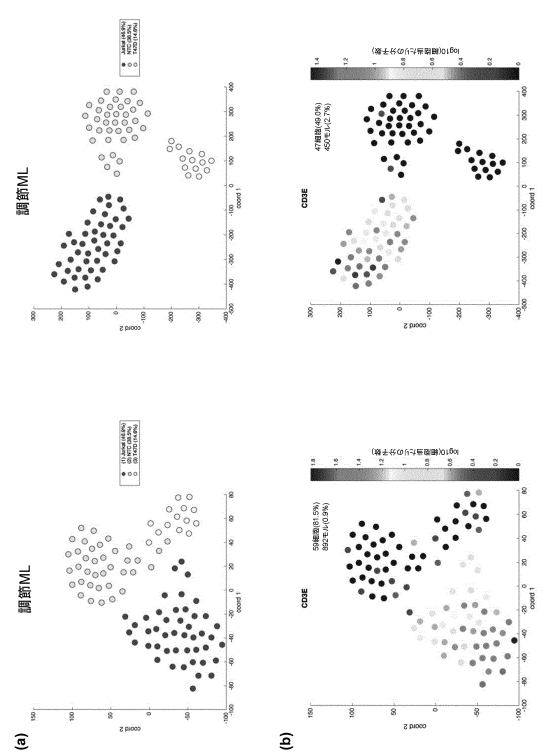


图 41

【図 4 1 - 2】

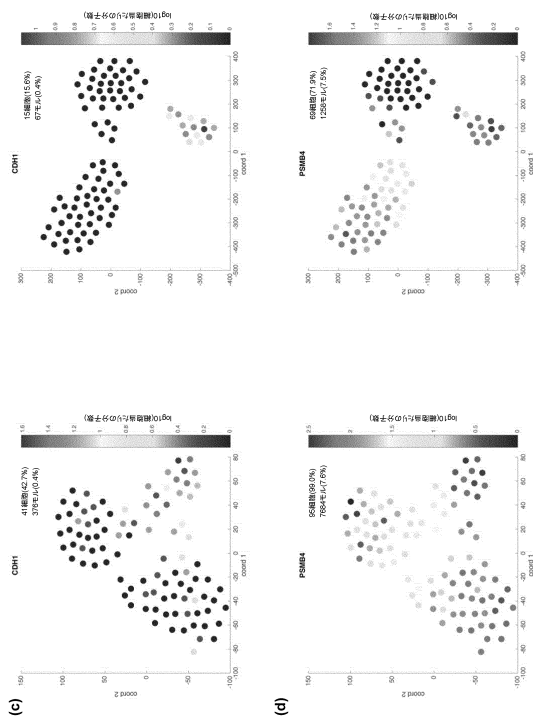


図41 (続き)

【図 4 2 - 1】

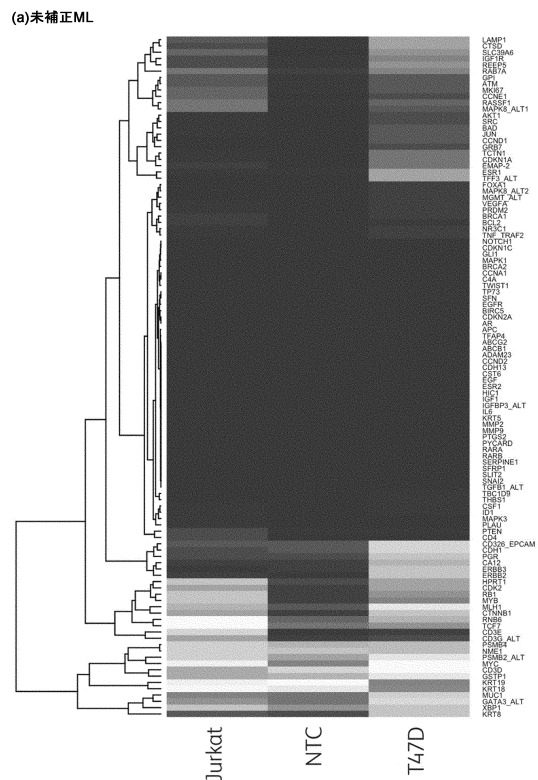


图 42

フロントページの続き

(33)優先権主張国・地域又は機関

米国(US)

(31)優先権主張番号 62/342,137

(32)優先日 平成28年5月26日(2016.5.26)

(33)優先権主張国・地域又は機関

米国(US)

(74)代理人 100103610

弁理士 ▲吉▼田 和彦

(74)代理人 100109070

弁理士 須田 洋之

(74)代理人 100119013

弁理士 山崎 一夫

(74)代理人 100123777

弁理士 市川 さつき

(74)代理人 100111796

弁理士 服部 博信

(74)代理人 100168631

弁理士 佐々木 康匡

(72)発明者 ファン, ジェエ

アメリカ合衆国 カリフォルニア 94025, メンロー パーク, キャンベル アベニュー 4040, スイート 110

(72)発明者 ツァイ, ジェニファー

アメリカ合衆国 カリフォルニア 94025, メンロー パーク, キャンベル アベニュー 4040, スイート 110

(72)発明者 シャム, エリーン

アメリカ合衆国 カリフォルニア 94025, メンロー パーク, キャンベル アベニュー 4040, スイート 110

(72)発明者 デン, リシャ

アメリカ合衆国 カリフォルニア 94025, メンロー パーク, キャンベル アベニュー 4040, スイート 110

(72)発明者 フー, グレン ケー.

アメリカ合衆国 カリフォルニア 94025, メンロー パーク, キャンベル アベニュー 4040, スイート 110

審査官 塩田 徳彦

(56)参考文献 米国特許出願公開第2011/0160078 (US, A1)

国際公開第2015/002908 (WO, A1)

(58)調査した分野 (Int.Cl., DB名)

G16B 5/00 - 99/00