

(19)日本国特許庁(JP)

(12)公表特許公報(A)

(11)公表番号
特表2024-516365
(P2024-516365A)

(43)公表日 令和6年4月15日(2024.4.15)

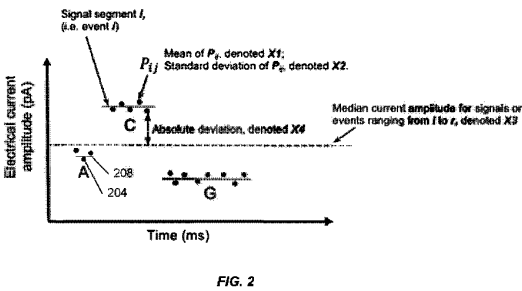
(51)国際特許分類		F I		テーマコード (参考)	
C 1 2 Q	1/6869(2018.01)	C 1 2 Q	1/6869	Z	4 B 0 2 9
C 1 2 M	1/34 (2006.01)	C 1 2 M	1/34	B	4 B 0 6 3
C 1 2 Q	1/6888(2018.01)	C 1 2 Q	1/6888	Z	
C 1 2 Q	1/6883(2018.01)	C 1 2 Q	1/6883	Z	
C 1 2 M	1/00 (2006.01)	C 1 2 M	1/00	A	
審査請求 未請求 予備審査請求 未請求 (全55頁)					
(21)出願番号 特願2023-562533(P2023-562533)		(71)出願人 512037244			
(86)(22)出願日 令和4年4月12日(2022.4.12)		ザ チャイニーズ ユニバーシティ オブ			
(85)翻訳文提出日 令和5年11月15日(2023.11.15)		ホンコン			
(86)国際出願番号 PCT/CN2022/086260		中華人民共和国 ホンコン ニュー テリ			
(87)国際公開番号 WO2022/218290		トリーズ シャーティン ピチウ ビルデ			
(87)国際公開日 令和4年10月20日(2022.10.20)		イング ルーム 3 0 1 オフィス オブ			
(31)優先権主張番号 63/173,728		リサーチ アンド ナレッジ トランスフ			
(32)優先日 令和3年4月12日(2021.4.12)		アー サービスイズ			
(33)優先権主張国・地域又は機関		(74)代理人 100078282			
米国(US)		弁理士 山本 秀策			
(81)指定国・地域 AP(BW,GH,GM,KE,LR,LS,MW,MZ,NA		(74)代理人 100113413			
,RW,SD,SL,ST,SZ,TZ,UG,ZM,ZW),EA(		弁理士 森下 夏樹			
AM,AZ,BY,KG,KZ,RU,TJ,TM),EP(AL,A		(74)代理人 100181674			
T,BE,BG,CH,CY,CZ,DE,DK,EE,ES,FI,FR		弁理士 飯田 貴敏			
,GB,GR,HR,HU,IE,IS,IT,LT,LU,LV,MC,		(74)代理人 100181641			
最終頁に続く		最終頁に続く			

(54)【発明の名称】 電気信号を使用する塩基修飾分析

(57)【要約】

電気信号及び他のデータを使用して塩基修飾を決定するためのシステム及び方法が、本明細書において説明される。実施形態は、様々な塩基修飾によって影響を受ける、ナノポア（1 1 6）を使用することから取得されるもののなどの配列決定に関連する電気信号に由来する特徴、並びにメチル化状態が決定される標的位置周囲のウィンドウにおけるヌクレオチドの同一性を利用することができる。他の特徴は、ヌクレオチドに対応する電気信号のセグメントの統計値のベクトル、及び核酸分子の領域内のウィンドウにおける電気信号の統計値を含み得る。検出された塩基修飾は、生体試料の追加の分析に使用することができる。

【選択図】 図 2



【特許請求の範囲】

【請求項 1】

核酸分子におけるヌクレオチドの修飾を検出するための方法であって、

入力データ構造を受信することであって、前記入力データ構造は、試料核酸分子において配列決定されたヌクレオチドのウィンドウに対応し、前記試料核酸分子は、前記ヌクレオチドに対応する電気信号を測定することによって配列決定され、前記入力データ構造は、以下の特性についての値：

前記ウィンドウ内の各ヌクレオチドについての

前記ヌクレオチドの同一性、

前記それぞれのウィンドウ内の標的位置に対する前記ヌクレオチドの位置、及び

前記ヌクレオチドに対応する前記電気信号のセグメントの第 1 のセグメント統計値を含むベクトルを含む、受信することと、

前記入力データ構造をモデルに入力することであって、前記モデルは、

第 1 の複数の第 1 のデータ構造を受信することであって、前記第 1 の複数の第 1 のデータ構造の各第 1 のデータ構造が、複数の第 1 の核酸分子のそれぞれの核酸分子において配列決定されたヌクレオチドのそれぞれのウィンドウに対応し、前記第 1 の核酸分子の各々は、前記ヌクレオチドに対応する前記電気信号を測定することによって配列決定され、前記修飾は、各第 1 の核酸分子の各ウィンドウにおける標的位置でのヌクレオチドの既知の第 1 の状態を有し、各第 1 のデータ構造が、前記入力データ構造と同じ特性についての値を含む、受信すること、

複数の第 1 の訓練試料を記憶することであって、各々が、前記第 1 の複数の第 1 のデータ構造のうちの 1 つと、前記標的位置での前記ヌクレオチドの前記第 1 の状態を示す第 1 のラベルと、を含む、記憶すること、及び、

前記第 1 の複数の第 1 のデータ構造が前記モデルに入力されたとき、前記複数の第 1 の訓練試料を使用して、前記第 1 のラベルの対応するラベルに一致するか又は一致しない前記モデルの出力に基づいて前記モデルのパラメータを最適化することであって、前記モデルの出力は、前記それぞれのウィンドウにおける前記標的位置での前記ヌクレオチドが前記修飾を有するかどうかを指定する、最適化すること、によって訓練される、入力することと、

前記モデルを使用して、前記入力データ構造の前記ウィンドウ内の前記標的位置でのヌクレオチドに前記修飾が存在するかどうかを決定することと、を含む、方法。

【請求項 2】

前記第 1 のセグメント統計値が、前記ヌクレオチドに対応する前記電気信号の前記セグメントの平均を表す、請求項 1 に記載の方法。

【請求項 3】

前記第 1 のセグメント統計値が、前記ヌクレオチドに対応する前記電気信号の前記セグメントの前記電気信号の変動を表す、請求項 1 に記載の方法。

【請求項 4】

前記第 1 のセグメント統計値が、前記ヌクレオチドに対応する前記電気信号の前記セグメントの平均の正規化された値を表す、請求項 1 に記載の方法。

【請求項 5】

前記ベクトルが、前記ヌクレオチドに対応する前記電気信号の前記セグメントの変動を表す第 2 のセグメント統計値を含む、請求項 1、2、又は 4 のいずれか一項に記載の方法。

【請求項 6】

前記ベクトルが、前記ヌクレオチドに対応する前記電気信号の前記セグメントの平均の正規化された値を表す第 2 のセグメント統計値を含む、請求項 1、2、又は 3 のいずれか一項に記載の方法。

【請求項 7】

前記ベクトルが、前記ヌクレオチドに対応する前記電気信号の前記セグメントの変動を

10

20

30

40

50

表す第 2 のセグメント統計値を含み、

前記ベクトルが、前記第 1 のセグメント統計値の正規化された値を表す第 3 のセグメント統計値を含む、請求項 2 に記載の方法。

【請求項 8】

前記入力データ構造が、前記ウィンドウに等しいか又はそれより大きい前記核酸分子の領域内の前記電気信号の第 1 の領域統計値についての値を含む、先行請求項のいずれか一項に記載の方法。

【請求項 9】

前記第 1 の領域統計値が、前記領域内の前記電気信号の平均又は中央値を表す、請求項 8 に記載の方法。

【請求項 10】

前記第 1 の領域統計値が、前記領域内の前記電気信号の前記平均又は中央値からの前記電気信号の変動の絶対値の中央値又は平均を表す、請求項 8 に記載の方法。

【請求項 11】

前記入力データ構造が、前記領域内の前記電気信号の前記平均又は中央値からの前記電気信号の変動の絶対値の中央値又は平均を表す第 2 の領域統計値を更に含む、請求項 9 に記載の方法。

【請求項 12】

前記領域が、前記試料核酸分子の 1 つの鎖上にある、請求項 8 ～ 11 のいずれか一項に記載の方法。

【請求項 13】

前記領域が、前記試料核酸分子であるか、又は少なくとも 5、10、15、20、25、30、50、100、200、300、400、500、若しくは 1 k、5 k、10 k、50 k、若しくは 1 M ヌクレオチドを含む、請求項 8 ～ 12 のいずれか一項に記載の方法。

【請求項 14】

前記領域が、前記ヌクレオチドの周辺を中心とする、請求項 8 ～ 13 のいずれか一項に記載の方法。

【請求項 15】

前記ウィンドウが、前記試料核酸分子の 2 つの鎖上のヌクレオチドを含む、先行請求項のいずれか一項に記載の方法。

【請求項 16】

前記修飾が、メチル化又は酸化である、先行請求項のいずれか一項に記載の方法。

【請求項 17】

前記電気信号が、電流、電圧、抵抗、インダクタンス、キャパシタンス、又はインピーダンスである、先行請求項のいずれか一項に記載の方法。

【請求項 18】

前記試料核酸分子をナノポアを使用して配列決定することを更に含む、先行請求項のいずれか一項に記載の方法。

【請求項 19】

前記修飾は、メチル化であり、
前記試料核酸分子は、無細胞であり、胎児を妊娠している女性対象の生体試料から取得され、

前記方法が、

前記標的位置での前記ヌクレオチドの修飾状態を使用して、前記試料核酸分子が胎児又は母体起源のものであるかどうかを決定することであって、前記修飾状態は、前記修飾が存在するかどうかであり、任意選択的に、前記試料核酸分子の 1 つ以上の他のヌクレオチドの前記修飾状態である、決定することを更に含む、請求項 1 に記載の方法。

【請求項 20】

前記試料核酸分子が胎児又は母体起源のものであるかどうかを決定することが、

10

20

30

40

50

前記 1 つ以上のヌクレオチドの前記修飾状態を使用して、前記試料核酸分子のメチル化レベルを決定することと、

参照値に対して前記試料核酸分子の前記メチル化レベルを比較することと、を含む、請求項 19 に記載の方法。

【請求項 21】

前記参照値が、1 つ以上の母体核酸分子のメチル化レベルから決定される、請求項 20 に記載の方法。

【請求項 22】

前記参照値に対して前記試料核酸分子の前記メチル化レベルを比較することが、前記試料核酸分子の前記メチル化レベルが前記参照値よりも低いことを決定することを含み、

前記試料核酸分子が胎児又は母体起源のものであるかどうかを決定することが、前記比較を使用して前記試料核酸分子が胎児起源のものであると決定することを含む、請求項 20 に記載の方法。

【請求項 23】

前記試料核酸分子が所定のゲノム領域に整列していることを特定することを更に含む、請求項 19 に記載の方法。

【請求項 24】

前記試料核酸分子は、複数の試料核酸分子のうちの 1 つの試料核酸分子であり、前記方法が、

前記修飾状態を使用して、前記複数の試料核酸分子の各々が胎児又は母体起源であるかどうかを決定することと、

前記複数の試料核酸分子の前記胎児又は母体起源の前記決定を使用して、胎児画分を決定することと、を更に含む、請求項 19 に記載の方法。

【請求項 25】

前記修飾は、メチル化であり、

前記試料核酸分子は、無細胞であり、胎児を妊娠している女性対象の生体試料から取得され、

前記試料核酸分子は、複数の試料核酸分子のうちの 1 つの試料核酸分子であり、前記方法が、

前記複数の試料核酸分子が胎児ゲノムの領域に整列していることを特定することと、

前記複数の試料核酸分子の各試料核酸分子の 1 つ以上のヌクレオチドの修飾状態を決定することと、

前記複数の試料核酸分子の各試料核酸分子についての前記 1 つ以上のヌクレオチドの前記修飾状態を使用して、前記領域のメチル化レベルを決定することと、

前記メチル化レベルを使用して、前記胎児ゲノムの前記領域にコピー数異常が存在するかどうかを決定することと、を更に含む、請求項 1 に記載の方法。

【請求項 26】

核酸分子におけるヌクレオチドの修飾を検出するための方法であって、

第 1 の複数の第 1 のデータ構造を受信することであって、前記第 1 の複数の第 1 のデータ構造の各第 1 のデータ構造は、複数の第 1 の核酸分子のそれぞれの核酸分子において配列決定されたヌクレオチドのそれぞれのウィンドウに対応し、前記第 1 の核酸分子の各々は、前記ヌクレオチドに対応する電気信号を測定することによって配列決定され、前記修飾は、各第 1 の核酸分子の各ウィンドウにおける標的位置でのヌクレオチドの既知の第 1 の状態を有し、各第 1 のデータ構造は、以下の特性についての値：

前記ウィンドウ内の各ヌクレオチドについての

前記ヌクレオチドの同一性、

前記それぞれのウィンドウ内の標的位置に対する前記ヌクレオチドの位置、及び

前記ヌクレオチドに対応する前記電気信号のセグメントの第 1 のセグメント統計値を含むベクトルを含む、受信することと、

複数の第 1 の訓練試料を記憶することであって、各々は、前記第 1 の複数の第 1 のデー

10

20

30

40

50

タ構造のうちの１つと、前記標的位置での前記ヌクレオチドの前記修飾についての前記第１の状態を示す第１のラベルと、を含む、記憶することと、

前記第１の複数の第１のデータ構造がモデルに入力されたとき、前記複数の第１の訓練試料を使用して、前記第１のラベルの対応するラベルに一致するか又は一致しない前記モデルの出力に基づいて前記モデルのパラメータを最適化することによって、前記モデルを訓練することであって、前記モデルの出力は、前記それぞれのウィンドウにおける前記標的位置での前記ヌクレオチドが前記修飾を有するかどうかを指定する、訓練することと、を含む、方法。

【請求項２７】

第２の複数の第２のデータ構造を受信することであって、前記第２の複数の第２のデータ構造の各第２のデータ構造は、複数の第２の核酸分子のそれぞれの核酸分子において配列決定されたヌクレオチドのそれぞれのウィンドウに対応し、前記修飾は、各第２の核酸分子の各ウィンドウ内の標的位置でのヌクレオチドの既知の第２の状態を有し、各第２のデータ構造は、前記第１の複数の第１のデータ構造と同じ特性についての値を含む、受信することと、

複数の第２の訓練試料を記憶することであって、各々は、前記第２の複数の第２のデータ構造のうちの１つと、前記標的位置での前記ヌクレオチドの前記第２の状態を示す第２のラベルと、を含む、記憶することと、を更に含み、

訓練において、

前記第１の状態又は前記第２の状態は、前記修飾が存在することであり、他の状態は、前記修飾が存在しないことであり、

前記モデルは、前記第２の複数の第２のデータ構造が前記モデルに入力されたとき、前記第２のラベルの対応するラベルに一致するか又は一致しない前記モデルの出力に基づいて前記モデルのパラメータを最適化することによって、前記複数の第２の訓練試料を使用することを更に含み、請求項２６に記載の方法。

【請求項２８】

前記複数の第１の核酸分子が、前記複数の前記第２の核酸分子と同じである、請求項２７に記載の方法。

【請求項２９】

前記第１の複数の第１のデータ構造と関連付けられる各ウィンドウは、前記第１の核酸分子の第１の鎖上のヌクレオチド及び前記第１の核酸分子の第２の鎖上のヌクレオチドを含み、

各第１のデータ構造は、前記ウィンドウ内の各ヌクレオチドについて、鎖特性の値を更に含み、前記鎖特性は、前記ヌクレオチドが前記第１の鎖又は前記第２の鎖のいずれかに存在することを示す、請求項２６に記載の方法。

【請求項３０】

前記修飾が、前記標的位置での前記ヌクレオチドのメチル化を含む、請求項２６に記載の方法。

【請求項３１】

前記既知の第１の状態が、前記第１のデータ構造の第１の箇所についてのメチル化状態と、前記第１のデータ構造の第２の箇所についての非メチル化状態と、を含む、請求項３０に記載の方法。

【請求項３２】

前記第１のセグメント統計値が、前記ヌクレオチドに対応する前記電気信号の前記セグメントの平均を表す、請求項２６に記載の方法。

【請求項３３】

前記第１のセグメント統計値が、前記ヌクレオチドに対応する前記電気信号の前記セグメントの前記電気信号の変動を表す、請求項２６に記載の方法。

【請求項３４】

前記第１のセグメント統計値が、前記ヌクレオチドに対応する前記電気信号の前記セグ

10

20

30

40

50

メントの平均の正規化された値を表す、請求項 26 に記載の方法。

【請求項 35】

前記ベクトルが、前記ヌクレオチドに対応する前記電気信号の前記セグメントの変動を表す第 2 のセグメント統計値を含む、請求項 26、32、又は 34 のいずれか一項に記載の方法。

【請求項 36】

前記ベクトルが、前記ヌクレオチドに対応する前記電気信号の前記セグメントの平均の正規化された値を表す第 2 のセグメント統計値を含む、請求項 26、32、又は 33 のいずれか一項に記載の方法。

【請求項 37】

前記ベクトルが、前記ヌクレオチドに対応する前記電気信号の前記セグメントの変動を表す第 2 のセグメント統計値を含み、

前記ベクトルが、前記第 1 のセグメント統計値の正規化された値を表す第 3 のセグメント統計値を含む、請求項 32 に記載の方法。

【請求項 38】

前記ウィンドウに等しいか又はそれより大きい前記それぞれの核酸分子の領域内の前記電気信号の第 1 の領域統計値についての値を含む各第 1 のデータ構造、請求項 26～37 のいずれか一項に記載の方法。

【請求項 39】

前記第 1 の領域統計値が、前記領域内の前記電気信号の平均又は中央値を表す、請求項 38 に記載の方法。

【請求項 40】

前記第 1 の領域統計値が、前記領域内の前記電気信号の前記平均又は中央値からの前記電気信号の変動の絶対値の中央値又は平均を表す、請求項 38 に記載の方法。

【請求項 41】

前記第 1 のデータ構造が、前記領域内の前記電気信号の前記平均又は中央値からの前記電気信号の変動の絶対値の中央値又は平均を表す第 2 の領域統計値を更に含む、請求項 39 に記載の方法。

【請求項 42】

前記領域が、前記それぞれの核酸分子の 1 つの鎖上にある、請求項 38～41 のいずれか一項に記載の方法。

【請求項 43】

前記領域が、前記それぞれの核酸分子であるか、又は少なくとも 5、10、15、20、25、30、50、100、200、300、400、500、若しくは 1k、5k、10k、50k、若しくは 1Mヌクレオチドを含む、請求項 38～45 のいずれか一項に記載の方法。

【請求項 44】

前記領域が、前記ヌクレオチドの周辺を中心とする、請求項 38～43 のいずれか一項に記載の方法。

【請求項 45】

前記ウィンドウが、前記それぞれの核酸分子の 2 つの鎖上のヌクレオチドを含む、請求項 26～44 のいずれか一項に記載の方法。

【請求項 46】

コンピュータ製品であって、複数の命令を記憶する非一時的コンピュータ可読媒体を備え、実行されると、コンピュータシステムを制御して、先行請求項のいずれか一項に記載の方法を実施する、コンピュータ製品。

【請求項 47】

システムであって、

請求項 46 に記載のコンピュータ製品と、

コンピュータ可読媒体に記憶された命令を実行するための 1 つ以上のプロセッサと、を

10

20

30

40

50

備える、システム。

【請求項 48】

先行請求項のいずれか一項に記載の方法を実施するための手段を備える、システム。

【請求項 49】

先行請求項のいずれか一項に記載の方法を実施するように構成された 1 つ以上のプロセスを備える、システム。

【請求項 50】

先行請求項のいずれか一項に記載の方法のステップをそれぞれ実施するモジュールを備える、システム。

【発明の詳細な説明】

【技術分野】

【0001】

関連出願の相互参照

本出願は、2021 年 4 月 12 日に出願された米国仮特許出願第 63 / 173, 728 号に対する優先権の利益を主張し、これは、その全体があらゆる目的で参照により本明細書に組み込まれる。

【背景技術】

【0002】

核酸における塩基修飾の存在は、ウイルス、細菌、植物、真菌、線虫、昆虫、及び脊椎動物（例えば、ヒト）などを含む、異なる生物全体にわたって変動する。最も一般的な塩基修飾は、異なる位置における異なる DNA 塩基へのメチル基の付加、いわゆるメチル化である。メチル化は、5mC（5-メチルシトシン）、4mC（N4-メチルシトシン）、5hmC（5-ヒドロキシメチルシトシン）、5fC（5-ホルミルシトシン）、5caC（5-カルボキシルシトシン）、1mA（N1-メチルアデニン）、3mA（N3-メチルアデニン）、N6-メチルアデニン（6mA）、7mA（N7-メチルアデニン）、3mC（N3-メチルシトシン）、2mG（N2-メチルグアニン）、6mG（O6-メチルグアニン）、7mG（N7-メチルグアニン）、3mT（N3-メチルチミン）、及び 4mT（O4-メチルチミン）などのシトシン、アデニン、チミン、及びグアニンで見出されている。脊椎動物のゲノムでは、5mC が最も一般的なタイプの塩基メチル化であり、グアニンのメチル化がそれに続く（すなわち、CpG の文脈において）。

【0003】

DNA メチル化は哺乳動物の発生に不可欠であり、遺伝子発現及びサイレンシング、胚発生、転写、クロマチン構造、X 染色体不活性化、反復要素の活性に対する保護、有糸分裂中のゲノム安定性の維持、並びに親起源のゲノムインプリンティングの調節において注目すべき役割を果たす。

【0004】

DNA メチル化は、プロモータ及びエンハンサのサイレンシングにおいて、協調的な状態で、多くの重要な役割を果たす（Robertson, 2005、Smith and Meissner, 2013）。多くのヒトの疾患は、DNA メチル化の異常と関連付けられることが見出されており、インプリンティング障害（例えば、ベックウィズ・ウィーデマン症候群及びプラダー・ウィリー症候群）、反復不安定性疾患（例えば、脆弱 X 症候群）、自己免疫障害（例えば、全身性紅斑性狼瘡）、代謝障害（例えば、I 型及び II 型糖尿病）、神経障害、加齢などを含むが、これらに限定されない。

【0005】

DNA 分子のメチロミックな修飾を正確に測定することは、多くの臨床的意味を有する。DNA メチル化を測定するために広く使用されている 1 つの方法は、バイサルファイト配列決定（BS-seq）を使用することである（Lister et al., 2009、Frommer et al., 1992）。このアプローチでは、DNA 試料を、最初にバイサルファイトで処理して、非メチル化シトシン（すなわち、C）をウラシルに変換する。対照的に、メチル化シトシンは、変化せずに残る。次いで、バイサルファイト

10

20

30

40

50

修飾DNAを、DNA配列決定によって分析する。別のアプローチでは、バイサルファイト変換に続いて、修飾DNAは、次いで異なるメチル化プロファイルのバイサルファイト変換DNAを区別できるプライマーを使用して、ポリメラーゼ連鎖反応（PCR）増幅にかけられる（Herman et al., 1996）。この後者のアプローチは、メチル化特異的PCRと呼ばれる。

【0006】

このようなバイサルファイトに基づくアプローチの1つの欠点は、バイサルファイト変換ステップで、処理されたDNAの大部分が著しく分解されることが報告されていることである（Grunau, 2001）。別の欠点は、バイサルファイト変換ステップによって強いCGバイアスが生成され（Ollova et al., 2018）、典型的には、不均一なメチル化状態を有するDNA混合物に対して信号対雑音比が低下することである。更に、バイサルファイト配列決定は、バイサルファイト処理中のDNAの分解により、長鎖DNA分子を配列決定するための理想的な方法ではない。

【0007】

核酸の塩基修飾のバイサルファイトを含まない決定を達成するための多くの継続的な努力がある。しかしながら、バイサルファイト配列決定に匹敵する感度及び特異度レベルを達成している商業的に実行可能なツールは不足している。ナノポア配列決定は、試料の化学ラベリングを必要としないために魅力的である配列決定の一種である。ナノポア配列決定による塩基修飾の検出は、比較的低コストかつ効率的であり得る。

【0008】

したがって、ナノポア配列決定によって塩基修飾を決定する必要がある。本開示において、本発明者らは、塩基修飾決定のための高い感度及び特異度を有するナノポア配列決定によって生成される電流信号を処理するための新しい方法及びシステムを説明する。

【発明の概要】

【課題を解決するための手段】

【0009】

説明される実施形態は、酵素的及び／若しくは化学的変換、又はタンパク質及び／若しくは抗体結合などの鑄型DNAの前処理なしで、核酸中の5mCなどの塩基修飾の判定を可能にする。本開示に存在する実施形態は、例えば、4mC、5hmC、5fC、5caC、1mA、3mA、6mA、7mA、3mC、2mG、6mG、7mG、3mT、4mTなどを含むが、これらに限定されない、異なるタイプの塩基修飾を検出するために使用され得る。そのような実施形態は、様々な塩基修飾によって影響を受ける、ナノポアを使用することから取得されるものなどの配列決定に関連する電気信号に由来する特徴、並びにメチル化状態が決定される標的位置周囲のウィンドウにおけるヌクレオチドの同一性を利用することができる。ヌクレオチドについての生の電気信号はまた、ヌクレオチドの上流又は下流のヌクレオチドに関連し得る。生の電気信号は、好適な技術を使用して異なるヌクレオチドに割り当てられ得る。

【0010】

本発明の実施形態は、ナノポア配列決定とともに使用することができる。ナノポア配列決定システムの一例は、Oxford Nanopore Technologiesによって、商品化されたものである。方法は、ナノポアを使用して測定された電気信号を使用し得る。方法は、ヌクレオチドの同一性、標的位置に対するヌクレオチドの位置、ヌクレオチドに対応する電気信号のセグメントの統計値を含むベクトル、及び核酸分子の領域内のウィンドウにおける電気信号の統計値を使用し得る。

【0011】

本発明者らが開発した方法は、生体試料の塩基修飾を検出して、限定されないが、研究及び診断の目的を含む様々な目的で、試料のメチル化プロファイルを評価するためのツールとして役立ち得る。検出されたメチル化プロファイルは、異なる分析に使用することができる。メチル化プロファイルは、DNAの起源（例えば、母体又は胎児、組織、細菌）を検出するために使用することができる。組織における異常なメチル化プロファイルの検

出は、個体における発達障害及び他の障害の特定に役立つ。

【0012】

本発明の実施形態の性質及び利点に関するより良好な理解は、以下の「発明を実施するための形態」及び添付の図面を参照して得ることができる。

【図面の簡単な説明】

【0013】

【図1】 ナノポア配列決定を例解する。

【図2】 本発明の実施形態による、異なる信号特徴を例解する。

【図3】 本発明の実施形態による、電流信号セグメント化及び信号特徴ベクトルの構築を例解する。

10

【図4】 本発明の実施形態による、ナノポアを通過する各ヌクレオチドについての事象の長さ（すなわち、持続時間）の分布のグラフである。

【図5】 本発明の実施形態による、電流パターン、配列決定位置、及び配列決定文脈を含む統合された表示マトリックスを使用する5mC検出のための原理を例解する。

【図6】 本発明の実施形態による、二本鎖DNAの両方のスタンドに基づいて、電流パターン、配列決定位置、及び配列決定文脈を含む統合された表示マトリックスを使用する塩基修飾検出のための原理を例解する。

【図7】 本発明の実施形態による、塩基修飾分析の性能に対するカーネルサイズの影響を示す。

【図8】 本発明の実施形態による、メチル化検出の観点から、訓練及び試験に使用される配列決定分子の数を示す。

20

【図9A】 本発明の実施形態による、IPM-CNN及びIPM-RNNアプローチを使用する、WGA DNAデータセットとM. S s s I処理DNAデータセットとの間のCpGについてのメチル化される確率の箱ひげ図である。

【図9B】 本発明の実施形態による、IPM-CNN及びIPM-RNNアプローチを使用する、WGA DNAデータセットとM. S s s I処理DNAデータセットとの間のCpGについてのメチル化される確率の箱ひげ図である。

【図9C】 本発明の実施形態による、IPM-CNN及びIPM-RNNアプローチを使用する、WGA DNAデータセットとM. S s s I処理DNAデータセットとの間のCpGについてのメチル化される確率の箱ひげ図である。

30

【図9D】 本発明の実施形態による、IPM-CNN及びIPM-RNNアプローチを使用する、WGA DNAデータセットとM. S s s I処理DNAデータセットとの間のCpGについてのメチル化される確率の箱ひげ図である。

【図10A】 本発明の実施形態による、訓練データセット及び試験データセットに対する受信機オペレータ特徴（ROC）曲線を示す。

【図10B】 本発明の実施形態による、訓練データセット及び試験データセットに対する受信機オペレータ特徴（ROC）曲線を示す。

【図11】 本発明の実施形態による、メチル化分析のための異なるツールの性能の表である。

【図12】 本発明の実施形態による、核酸分子におけるヌクレオチドの修飾を検出するプロセスのフローチャートである。

40

【図13】 本発明の実施形態による、核酸分子におけるヌクレオチドの修飾を検出するプロセスのフローチャートである。

【図14】 本発明の実施形態による、測定システムを例解する。

【図15】 本発明の実施形態による、システム及び方法とともに使用可能な例示的なコンピュータシステムのブロック図を示す。

【図16】 本発明の実施形態による、ROC曲線下面積（AUC）に対するパラメータの異なる組み合わせの影響のグラフを示す。

【図17】 本発明の実施形態による、AUCに対するウィンドウサイズの影響のグラフを示す。

50

【図 18】本発明の実施形態による、電流パターン、配列決定位置、及び配列決定文脈を含む統合された表示マトリックスを使用する 6 mA 検出のための原理を例解する。

【図 19】本発明の実施形態による、6 mA 検出の AUC のグラフを示す。

【図 20】本発明の実施形態による、バフィーコート及びNPC腫瘍試料に由来するDNAについてのIPM-RNNモデルによって決定された単一分子メチル化レベルの比較である。

【図 21】本発明の実施形態による、単一分子メチル化パターンの例を示す。

【図 22】本発明の実施形態による、母体特異的及び胎児特異的無細胞DNA分子の単一分子メチル化レベルのグラフである。

【図 23】本発明の実施形態による、IPM-CNNモデルによって決定されたメチル化パターンを使用して無細胞DNA分子の胎児及び母体起源を決定するためのROC曲線である。

10

【発明を実施するための形態】

【0014】

用語

「組織」は、機能単位としてともに群化する細胞の群に対応する。2つ以上のタイプの細胞が、単一の組織内に見出され得る。異なるタイプの組織は、異なるタイプの細胞（例えば、肝細胞、肺細胞、又は血球）から構成されてもよく、異なる生物（母体対胎児、移植を受けた対象の組織、微生物又はウイルスに感染した生物の組織）由来の組織あるいは健康な細胞対腫瘍細胞に対応してもよい。「参照組織」は、組織特異的メチル化レベルを決定するように使用される組織に対応し得る。異なる個体からの同じ組織タイプの複数の試料を使用して、その組織タイプの組織特異的メチル化レベルを決定することができる。

20

【0015】

「生体試料」は、ヒト対象から採取される任意の細胞試料を指す。生体試料は、組織生検、穿刺吸引物、又は血球であり得る。試料はまた、妊婦から採取された無細胞試料、例えば、血漿又は血清又は尿であり得る。様々な実施形態では、無細胞DNAについて濃縮された妊婦からの生体試料（例えば、遠心分離プロトコルを介して取得された血漿試料）におけるDNAの大部分は、無細胞であり得、例えば、50%超、60%超、70%超、80%超、90%超、95%超、又は99%超のDNAは無細胞であり得る。遠心分離プロトコルは、例えば、3,000g×10分で流体部分を取得することと、残留細胞を除去するために30,000gで更に10分間再遠心分離することと、を含み得る。特定の実施形態では、3,000gの遠心分離ステップに続いて、流体部分の濾過を行うことができる（例えば、直径5µm以下の孔径のフィルターを使用）。

30

【0016】

「配列リード」は、核酸分子の任意の一部又は全部から配列決定されるヌクレオチドの鎖を指す。例えば、配列リードは、核酸断片から配列決定されたヌクレオチドの短鎖（例えば、20～150個）、核酸断片の一方又は両方の末端におけるヌクレオチドの短鎖、又は生体試料中に存在する核酸断片全体の配列決定であり得る。配列リードは、例えば、配列決定技術を使用した、又はプローブを使用した種々の方法で、例えば、ハイブリダイゼーションアレイ若しくは捕捉プローブで、又は単一プライマー若しくは等温増幅を使用した、ポリメラーゼ連鎖反応（PCR）若しくは線形増幅などの増幅技術で、取得され得る。

40

【0017】

「部位」（「ゲノム部位」とも呼ばれる）は、単一の塩基位置、又は相関する塩基位置の群、例えば、CpG部位、又は相関する塩基位置のより大きい群であり得る、単一の部位に対応する。「遺伝子座」は、複数の部位を含む領域に対応し得る。遺伝子座は、遺伝子座をその文脈における部位と等価にするであろうただ1つの部位を含み得る。

【0018】

「メチル化状態」とは、所与の部位でのメチル化の状態を指す。例えば、ある部位は、

50

メチル化されているか、メチル化されていないか、又は場合によっては未決定であるかのいずれかである。

【0019】

各ゲノム部位（例えば、CpG部位）に対する「メチル化指数」は、その部位におけるメチル化を、その部位をカバーするリードの総数にわたって示す、（例えば、配列リード又はプローブから決定されるような）DNA断片の割合を指し得る。「リード」は、DNA断片から取得された情報（例えば、部位におけるメチル化状態）に対応し得る。リードは、1つ以上の部位における特定のメチル化状態のDNA断片と優先的にハイブリダイズする試薬（例えば、プライマー又はプローブ）を使用して、取得することができる。典型的には、このような試薬は、それらのメチル化状態に応じて、DNA分子を示差的に修飾するか又は示差的に認識するプロセス、例えば、バイサルファイト変換、又はメチル化感受性制限酵素、又はメチル化結合タンパク質、又は抗メチルシトシン抗体、又はメチルシトシン及びヒドロキシメチルシトシンを認識する単一分子配列決定技術（例えば、単一分子リアルタイム配列決定（例えば、Pacific Biosciencesからの）、並びにナノポア配列決定（例えば、Oxford Nanopore Technologiesからの））で処理した後に適用される。

10

【0020】

領域の「メチル化密度」は、この領域における部位をカバーするリード数の合計で割ったメチル化を示す、領域内の部位におけるリード数を指し得る。この部位は、具体的な特徴、例えば、CpG部位、を有し得る。したがって、領域の「CpGメチル化密度」は、この領域におけるCpG部位（例えば、特定のCpG部位、CpGアイランド内又はそれより大きい領域内のCpG部位）をカバーするリード数の合計で割ったCpGメチル化を示すリード数を指し得る。例えば、ヒトゲノム中の各100kbビンについてのメチル化密度は、100kb領域にマッピングされた配列リードによってカバーされた全てのCpG部位の割合として、CpG部位においてバイサルファイト処理後に変換されていないシトシン（メチル化されたシトシンに対応する）の総数から決定され得る。この分析はまた、例えば、500bp、5kb、10kb、50kb、若しくは1Mbなどの他のビンサイズに対して実施することができる。領域は、全ゲノム、又は染色体、又は染色体の一部（例えば、染色体腕）であり得る。あるいは、メチル化密度は、本開示において説明される実施形態を使用してナノポア配列決定を使用して、バイサルファイト変換なしで決定することができる。CpG部位のメチル化指数は、領域がそのCpG部位のみを含む場合に、その領域についてのメチル化密度と同じである。「メチル化シトシンの割合」は、領域において、分析されたシトシン残基の総数にわたって、メチル化されている（例えば、バイサルファイト変換後に変換されていない）、すなわち、CpG文脈外のシトシンを含む、ことが示されるシトシン部位の数、「C」、を指し得る。メチル化指数、メチル化密度、1つ以上の部位でメチル化された分子の計数、及び1つ以上の部位でメチル化された分子（例えば、シトシン）の比率が、「メチル化レベル」の例である。バイサルファイト変換とは別に、当業者に既知の他のプロセスは、これらに限定されないが、メチル化状態に感受性のある酵素（例えば、メチル化感受性制限酵素）、メチル化結合タンパク質、メチル化状態に感受性のあるプラットフォームを使用する単一分子配列決定（例えば、ナノポア配列決定（Schreiber et al. Proc Natl Acad Sci 2013; 110: 18910-18915）及び単一分子リアルタイム配列決定（例えば、Pacific Biosciencesからのもの）（Flusberg et al. Nat Methods 2010; 7: 461-465）によるもの）を含み、DNA分子のメチル化状態を調べるために使用することができる。

20

30

40

【0021】

「メチローム」は、ゲノムにおける複数の部位又は遺伝子座のDNAメチル化の量の尺度を提供する。メチロームは、ゲノムの全部、ゲノムの実質的な部分、又はゲノムの比較的わずかな箇所に対応し得る。

【0022】

50

「妊婦血漿メチローム」は、妊娠した動物（例えば、ヒト）の血漿又は血清から決定されたメチロームである。妊婦血漿メチロームは、血漿及び血清が無細胞DNAを含むため、無細胞メチロームの一例である。妊婦血漿メチロームは、体内の異なる器官又は組織又は細胞に由来するDNAの混合物であるため、混合メチロームの一例でもある。一実施形態では、このような細胞は、赤血球（すなわち、赤色細胞）系譜、骨髄系譜（例えば、好中球及びこれらの前駆体）及び巨核球系譜の細胞を含むが、これらに限定されない造血細胞である。妊娠中、血漿メチロームは胎児及び母親からのメチローム情報を含むことがある。「細胞性メチローム」は、患者の細胞（例えば、血球）から決定されるメチロームに対応する。血球のメチロームは、血球メチロームと呼ばれる。

【0023】

「メチル化プロファイル」には、複数の部位又は領域のDNA又はRNAのメチル化に関連する情報が含まれる。DNAメチル化に関連する情報は、CpG部位のメチル化指数、領域中のCpG部位のメチル化密度（略称MD）、連続した領域にわたるCpG部位の分布、2つ以上のCpG部位を含む領域内の各個々のCpG部位のメチル化のパターン又はレベル、及び非CpGメチル化を含み得るが、これらに限定されない。一実施形態では、メチル化プロファイルは、2つ以上のタイプの塩基（例えば、シトシン又はアデニン）のメチル化又は非メチル化のパターンを含み得る。ゲノムの実質的な部分のメチル化プロファイルは、メチロームと等価とみなすことができる。哺乳動物ゲノムにおける「DNAメチル化」とは、典型的には、CpGジヌクレオチド間でシトシン残基の5'炭素へのメチル基の付加（すなわち、5-メチルシトシン）を指す。DNAメチル化は、他の文脈、例えば、CHG及びCHHにおいてシトシンで生じ得、ここで、Hは、アデニン、シトシン、又はチミンである。シトシンのメチル化は、5-ヒドロキシメチルシトシンの形態でもあり得る。N⁶-メチルアデニンなどの非シトシンメチル化もまた、報告されている。

【0024】

「メチル化パターン」とは、メチル化塩基及び非メチル化塩基の順序を指す。例えば、メチル化パターンは、単一のDNA鎖、単一の二本鎖DNA分子、又は別のタイプの核酸分子上のメチル化塩基の順序であり得る。一例として、3つの連続したCpG部位は、以下のメチル化パターン：UUU、MMM、UMM、UMU、UUM、MUM、MUU、又はMMUのうちのいずれかを有し得、ここで、「U」は非メチル化部位を示し、「M」はメチル化部位を示す。限定されないが、この概念をメチル化を含む塩基修飾に拡張する場合、修飾塩基及び非修飾塩基の順序を指す「修飾パターン」という用語を使用するであろう。例えば、修飾パターンは、単一のDNA鎖、単一の二本鎖DNA分子、又は別のタイプの核酸分子上の修飾された塩基の順序であり得る。一例として、3つの連続した潜在的に修飾可能な部位は、以下の修飾パターン：UUU、MMM、UMM、UMU、UUM、MUM、MUU、又はMMUのうちのいずれかを有し得、ここで、「U」は非修飾部位を示し、「M」は修飾部位を示す。メチル化に基づかない塩基修飾の一例は、8-オキソグアニンなどの酸化的変化である。

【0025】

「高メチル化」及び「低メチル化」という用語は、その単一分子のメチル化レベルによって測定される単一のDNA分子のメチル化密度、例えば、その分子内のメチル化された塩基又はヌクレオチドの数を、その分子内のメチル化可能な塩基又はヌクレオチドの総数で割ったものを指し得る。高メチル化分子は、単一分子のメチル化レベルが閾値以上である分子であり、用途ごとに定義され得る。この閾値は、5%、10%、20%、30%、40%、50%、60%、70%、80%、90%、又は95%であり得る。低メチル化分子は、単一分子のメチル化レベルが閾値以下である分子であり、用途ごとに定義され得、用途ごとに変化し得る。この閾値は、5%、10%、20%、30%、40%、50%、60%、70%、80%、90%、又は95%であり得る。

【0026】

「高メチル化」及び「低メチル化」という用語はまた、これらの分子の複数の分子のメ

10

20

30

40

50

チル化レベルによって測定される、DNA分子の集団のメチル化レベルを指してもよい。分子の高メチル化集団は、複数の分子のメチル化レベルが閾値以上である集団であり、用途ごとに定義され得、用途ごとに変化し得る。この閾値は、5%、10%、20%、30%、40%、50%、60%、70%、80%、90%、又は95%であり得る。分子の低メチル化集団は、複数の分子のメチル化レベルが閾値以下である集団であり、用途ごとに定義され得る。この閾値は、5%、10%、20%、30%、40%、50%、60%、70%、80%、90%、又は95%であり得る。一実施形態では、分子の集団は、1つ以上の選択されたゲノム領域に整列され得る。一実施形態では、選択されたゲノム領域は、遺伝性障害、インプリンティング障害、エピジェネティック障害、代謝障害、又は神経障害などの疾患に関連し得る。選択されたゲノム領域は、50ヌクレオチド(nt)、100nt、200nt、300nt、500nt、1000nt、2knt、5knt、10knt、20knt、30knt、40knt、50knt、60knt、70knt、80knt、90knt、100knt、200knt、300knt、400knt、500knt、又は1Mntの鎖長を有し得る。

【0027】

本明細書で使用される「分類」という用語は、試料の特定の特性と関連付けられる任意の数又は他の特徴を指す。例えば、「+」という記号（又は「陽性」という単語）は、試料が欠失又は増幅を有するとして分類されることを意味し得る。分類は、二者択一（例えば、陽性若しくは陰性）であり得、又はより多くのレベルの分類（例えば、1～10若しくは0～1のスケール）を有し得る。

【0028】

「カットオフ」及び「閾値」という用語は、操作に使用される所定の数を指す。例えば、カットオフサイズは、断片が除外されるサイズ超のサイズを指し得る。閾値は、特定の分類が適用されるのを上回る又は下回る値であり得る。これらの用語のいずれかは、これらの文脈のいずれかにおいて使用され得る。カットオフ又は閾値は、「参照値」であり得るか、又は特定の分類を表すか、若しくは2つ以上の分類間を区別する参照値から導出され得る。そのような参照値は、当業者によって理解されるように、様々な方法で決定され得る。例えば、異なる既知の分類を有する対象の2つの異なるコホートについて、メトリックを決定することができ、参照値を1つの分類（例えば、平均）の代表として、又はメトリックの2つのクラスター間の値（例えば、所望の感度及び特異度を取得するために選択される）として選択し得る。別の例として、参照値は、統計分析又は試料のシミュレーションに基づいて決定することができる。

【0029】

「病理のレベル」（又は障害のレベル）は、その細胞の分析を通じて測定することができる、生物と関連付けられる病理の量、程度、又は重症度を指し得る。病理の別の例は、移植された臓器の拒絶反応である。他の例の病理には、ゲノムインプリンティング障害、自己免疫発作（例えば、腎臓を損傷するループス腎炎又は神経系を損傷する多発性硬化症）、炎症性疾患（例えば、肝炎）、線維化プロセス（例えば、肝硬変）、脂肪浸潤（例えば、脂肪肝疾患）、変性プロセス（例えば、アルツハイマー病）、及び虚血性組織損傷（例えば、心筋梗塞又は脳卒中）が含まれ得る。対象の健康な状態は、病理のない分類とみなし得る。

【0030】

「妊娠関連障害」には、母体及び／又は胎児組織における遺伝子の異常な相対的発現レベルを特徴とする任意の障害が含まれる。これらの障害には、子癇前症、子宮内胎児発育遅延、侵襲性胎盤形成、早産、新生児溶血性疾患、胎盤機能不全、胎児水腫、胎児奇形、HELLP（溶血、肝酵素の上昇、及び血小板数の低下）症候群、全身性紅斑性狼瘡（SLE）、及び他の母親の免疫疾患が含まれるが、これらに限定されない。一部の実施形態では、妊娠関連障害は、妊娠期間中の生理学的又は形態学的異常と関連付けられる任意の状態である。

【0031】

10

20

30

40

50

略語「b p」は、塩基対を指す。場合によっては、「b p」は、DNA断片が一本鎖であり、塩基対を含まない場合でも、DNA断片の鎖長を示すために使用され得る。一本鎖DNAの文脈では、「b p」は、ヌクレオチドの鎖長を提供すると解釈される場合がある。

【0032】

略語「n t」は、ヌクレオチドを指す。場合によっては、「n t」は、塩基単位で一本鎖DNAの長さを示すために使用され得る。また、「n t」は、分析される遺伝子座の上流又は下流などの相対位置を示すために使用され得る。技術的概念化、データ表示、処理、及び分析に関する一部の文脈では、「n t」と「b p」は互換的に使用される場合がある。

10

【0033】

「配列文脈」という用語は、DNAのストレッチにおける塩基組成（A、C、G、又はT）及び塩基順序を指し得る。このようなDNAのストレッチは、塩基修飾分析にかけられる塩基又は標的となる塩基を取り巻いている可能性がある。例えば、配列文脈は、塩基修飾分析に供される塩基の上流及び／又は下流の塩基を指し得る。

【0034】

「機械学習モデル」という用語には、試料データ（例えば、訓練データ）を使用して試験データを予測することに基づくモデルが含まれる場合があり、したがって、教師あり学習が含まれ得る。機械学習モデルは、しばしば、コンピュータ又はプロセッサを使用して開発される。機械学習モデルには、統計モデルが含まれ得る。

20

【0035】

「データ分析フレームワーク」という用語は、データを入力として受け取り、次に予測結果を出力することができるアルゴリズム及び／又はモデルを含み得る。「データ分析フレームワーク」の例には、統計モデル、数学的モデル、機械学習モデル、その他の人工知能モデル、及びそれらの組み合わせが含まれる。

【0036】

「リアルタイム配列決定」という用語は、配列決定に関与するプロセス中に、データ収集又は監視を伴う技術を指し得る。例えば、リアルタイム配列決定は、ヌクレオチド鎖がそのナノポアを移行するときにナノポアを通るイオン電流の電気信号監視を伴い得る。

【0037】

「電気信号」という用語は、情報を伝達する電圧又は電流を指し得る。電気信号は、正方形波、長方形波、三角形波、のこぎり波形、又は種々のパルス及びスパイクなどの種々の規則的及び／又は不規則な信号波形タイプ及び／又は形状で表現され得る。電気信号は、経時的な電圧又は電流の変動の視覚的表現を含み得る。電気信号の測定は、特定の時間（例えば、ミリ秒）でサンプリングされ得る。例えば、電流は、1 kHz、2 kHz、3 kHz、4 kHz、5 kHz、10 kHz、20 kHz、30 kHz、40 kHz、50 kHz、100 kHzなどの周波数でサンプリングされる。

30

【0038】

「信号セグメント」又は「セグメント」という用語は、特定のヌクレオチドの配列決定と関連付けられる電気信号のトレースの部分の部分を指し得る。セグメントは、ナノポア配列決定におけるベースコールから決定されたヌクレオチドに対応し得る。セグメントは、トレースの特定の持続時間をカバーし得る。異なるセグメントは、異なる持続時間を有し得る。セグメントは、重複していてもよい。一部の実施形態では、電気信号振幅は、セグメントにおける特定の変動を有し得る。例えば、電気信号振幅は、セグメントにおける電気信号振幅の平均又は中央値の5%、10%、20%、30%、又は40%以内であり得る。

40

【0039】

「約」又は「およそ」という用語は、当業者によって決定される特定の値の許容誤差範囲内を意味し得、これは値の測定又は決定方法、すなわち測定システムの制限について部分的に依存する。例えば、「約」は、当該技術分野の慣例により、1以内又は1を超える

50

標準偏差を意味し得る。あるいは、「約」は、所与の値の最大20%、最大10%、最大5%、又は最大1%の範囲を意味し得る。あるいは、特に生物学的システム又はプロセスに関して、「約」又は「およそ」という用語は、値の1桁以内、5倍以内、より好ましくは2倍以内を意味し得る。本出願及び特許請求の範囲に特定の値が記載されている場合、特に明記しない限り、特定の値の許容誤差範囲内の「約」という用語を想定すべきである。「約」という用語は、当業者によって一般に理解されている意味を有し得る。「約」という用語は、±10%を指し得る。「約」という用語は、±5%を指し得る。

〔発明を実施するための形態〕

〔0040〕

ナノポア配列決定を使用して塩基修飾（例えば、メチル化）を検出する正確かつ効率的な方法が望まれる。調査研究は、DNAメチル化を分析するためにナノポア配列決定によって生成された電気信号を使用する実行可能性を研究している（Simpson et al. Nat Methods. 2017; 14: 407–410、Liu et al. Nat Commun. 2019; 10: 2449、Ni et al. Bioinformatics. 2019; 35: 4586–4595）。5-メチルシトシン（5mC）検出についての報告された性能は、多くの検証研究において最適以下であった。例えば、DeepSignalと呼ばれる計算ツールを使用する5mC検出の感度は、試料NA12878に基づいてH. sapiens R9.4.1Dデータを分析するとき、88%の特異度で79%であることが報告された（Ni et al. Bioinformatics. 2019; 35: 4586–4595）。より高い特異度（例えば>95%）を達成することを目的とする場合、感度は更に低下すると予想される。nanopolish（Liu et al. Nat Commun. 2019; 10: 2449）と呼ばれる別のツールについては、同じデータセットを分析すると、感度は0.46の特異度でわずか0.61であった。nanopolishソフトウェアは、以下の仮定によって隠れマルコフモデルに基づいていた：（1）DNA配列中の6ヌクレオチドオリゴマー（すなわち、6-mer）の電気信号は、ガウス分布に従った、（2）特定の塩基についてのメチル化状態（メチル化又は非メチル化）の確率は、以前の塩基のメチル化状態にのみ依存した、（3）特定の電流レベルを出力する確率は、電流信号を生成するメチル化状態にのみ依存し、任意の他のメチル化状態又は任意の他の電流信号には依存しなかった。それらの仮定は、ナノポア配列決定中に生成された実際の電流信号において不正確であり得、したがって、より低い感度及び特異度につながる。

〔0041〕

Oxford Nanopore配列決定に基づくDNAメチル化分析のためのDeepModと呼ばれる最近の計算ツールは、双方向再帰型ニューラルネットワーク（RNN）を使用することを試みた。しかしながら、そのようなアプローチの設計は、電気信号を用いて配列決定リードからの予測結果を集約することによってゲノム位置におけるメチル化レベルを測定することを目的としており、したがって、単一分子レベルでメチル化パターンを分析する能力を欠いている。加えて、Escherichia coli、Chlamydomonas reinhardtii、及びホモサピエンスを含むデータセットにわたる配列決定深度の中央値は、およそ33xであった。多くの商業用途において、より低い配列決定深度は、経済的コスト及び分析時間を節約するために望ましい。DeepModソフトウェアが、単一分子レベルで実用的に有意義な精度でメチル化パターンを分析できるかどうかは不明である。

〔0042〕

1つの研究では、Yuenらは、ナノポア配列決定からのCpGメチル化検出のためのツールを体系的にベンチマークし、ほとんどのツールが、CpG部位当たりの予想されるメチル化パーセンテージと高い分散及び低い一致を示したと結論付けた（Yuen et al. bioRxiv. 2020; doi: doi.org/10.1101/2020.10.14.340315）。

〔0043〕

Tseらは、Pacific Biosciences (PacBio)からの単一分子リアルタイム配列決定 (SMRT-seq)を使用して、DNA重合中に蛍光体でラベルしたヌクレオチドを組み込むことによって生成されるパルス間持続時間 (IPD) 及びパルス幅 (PW) などの光信号を含むDNAポリメラーゼの動態学的特徴を、畳み込みニューラルネットワークの使用により2つ以上の塩基からなる測定ウィンドウを分析することに基づいて、メチル化CpG部位及び非メチル化CpG部位を区別するために使用することができることを報告した (Tse et al., Proc Natl Acad Sci USA, 2021; 118: e2019768118、米国特許第11,091,794号)。そのような測定ウィンドウは、IPD及びPWを異なる配列決定文脈及び配列決定位置に編成した。しかしながら、ナノポア配列決定は、ナノポアを通過する二本鎖DNAの鎖によって引き起こされる電流信号に応じる、完全に異なる配列決定メカニズムを使用した。そのような生の電気信号は、ナノポアを通過する異なるヌクレオチドによって変動し、特定のヌクレオチドの電気信号は、そのヌクレオチドの近くの上流及び下流のヌクレオチドによって影響を受けるであろう。したがって、異なるヌクレオチドは、検出される電気信号トレースの異なる長さを有し、同一のヌクレオチドでさえ、電気信号トレースの異なる長さを有するであろう。ナノポアを通過する特定のヌクレオチド又は2つ以上のヌクレオチドと関連付けられる電気信号を分析する場合、各塩基上で検出される電気信号トレースの長さは、経時的に固定されない。対照的に、PacBio SMRT-seqを使用する5mC検出についての以前の研究は、各ヌクレオチドについての光信号に関連する2つの固定測定値、すなわち、IPD及びPWに基づいていた (Tse et al., Proc Natl Acad Sci USA, 2021; 118: e2019768118)。したがって、Tseらの研究 (Tse et al., Proc Natl Acad Sci USA, 2021; 118: e2019768118) で提示される訓練されたモデルは、ナノポア配列決定によって生成されるそのような電気信号には適用されない。

10

20

【0044】

本明細書において説明される実施形態は、ヌクレオチド修飾を検出するためにナノポア配列決定から取得される電気信号を使用する。ヌクレオチドメチル化は、本明細書において説明される任意のメチル化を含み得る。ナノポア配列決定から取得される情報は、ヌクレオチドの同一性、標的位置に対するヌクレオチドの位置、ヌクレオチドに対応する電気信号のセグメントの統計値を含むベクトル、及び核酸分子の領域内のウィンドウにおける電気信号の統計値を含み得る。

30

【0045】

本開示に存在する実施形態は、生物から取得された細胞試料 (例えば、細胞株、固形臓器、固形組織、内視鏡検査を介して取得された試料、絨毛膜絨毛試料) から取得されたDNAに使用することができる。本開示における実施形態はまた、環境 (例えば、細菌、細胞汚染物質)、食品 (例えば、肉) から取得された細胞試料にも使用することができる。本開示に存在する実施形態はまた、妊婦から取得された血漿又は血清にも使用することができる。一部の実施形態では、本開示に存在する方法はまた、例えばハイブリダイゼーションプロブ (Albert et al., 2007、Okou et al., 2007、Lee et al., 2011)、又は物理的分離 (サイズなどに基づく) に基づく若しくは制限酵素消化 (例えば、MspI) に続くアプローチ、又はCas9ベースの濃縮 (Watson et al., 2019) を使用して、ゲノムの一部が最初に濃縮されるステップの後で適用され得る。酵素的又は化学的変換は、本発明が機能するのに必要ではないが、特定の実施形態では、そのような変換ステップが、本発明の性能を更に高めるために含まれていてもよい。

40

【0046】

本開示の実施形態は、修飾された塩基を正確かつ効率的に検出することができるようにナノポア配列決定を改善する。塩基修飾は、直接検出され得る。実施形態は、検出のために全ての修飾情報が保たれない可能性がある酵素的又は化学的変換を回避することができ

50

る。更に、特定の酵素的又は化学的変換は、特定のタイプの修飾と互換性がない場合がある。本開示の実施形態はまた、塩基修飾情報をPCR産物に伝達しない可能性があるPCRによる増幅を回避し得る。更に、DNAの両方の鎖を一緒に配列決定することができ、それによって、一方の鎖からの配列と、他方の鎖に相補的な配列との対形成が可能になる。対照的に、PCR増幅は、二本鎖DNAの2つの鎖を分割するため、このような2つの連続する鎖からの配列の組み合わせされた分析は、困難である。

【0047】

更に、ナノポア配列決定は、他の配列決定技術よりも費用効果が高く、ポータブルである。例えば、ナノポア配列決定システムであるOxford Nanopore Technologies MinION（商標）は、およそ5,000USDであるが、一方、光信号ベースの配列決定システムであるPacBio SMRT（商標）Sequel IIシステムは、約500,000~700,000USDのオーダーにある。ナノポア配列決定速度は、毎秒約450ヌクレオチドにあるが、一方、PacBio SMRT（商標）配列決定は、毎秒約5ヌクレオチドである。したがって、同じ期間内に、ナノポア配列決定は、光信号ベースの配列決定システムを用いる場合よりも多くのデータを取得することができる。

【0048】

酵素的又は化学的変換の有無にかかわらず、決定されたメチル化プロファイルは、生体試料の分析に使用することができる。一実施形態では、メチル化プロファイルは、細胞DNAの起源（例えば、母体若しくは胎児、組織、又はウイルス）を検出するために使用することができる。組織における異常なメチル化プロファイルの検出は、個体における発達障害の特定に役立つ。単一分子のメチル化パターンは、キメラDNA（例えば、ウイルスとヒト間）及びハイブリッドDNA（例えば、天然ゲノムでは通常融合されない2つの遺伝子間）又は2つの種間（例えば、遺伝子又はゲノム操作による）を特定することができる。

【0049】

I. ナノポア配列決定の原理

単一分子配列決定技術の例は、ナノポア配列決定（Oxford Nanopore Technologies）である。図1は、DNA分子（例えば、DNA分子104）のナノポア配列決定についての原理を示す。単一のDNA分子がナノメートルサイズの孔を通過する際に、膜を横切るイオン電流の流れによって引き起こされる電気信号パターンは、核酸の配列を決定するために使用された。そのような孔は、例えば、タンパク質（例えば、アルファ溶血素、エロリジン、及びMycobacterium smegmatis ポリンA（MspA））、又はシリコン若しくはグラフェンなどの合成材料によって構築され得るが、これらに限定されない（Magiet al, Brief Bioinform. 2018; 19: 1256-1272）。

【0050】

一実施形態では、二本鎖DNA分子は、末端修復プロセスに供された。そのようなプロセスは、DNAを平滑末端DNAに変換し、続いて、配列決定アダプタライゲーションを容易にするAテールを付加する。各々がモータタンパク質を担持する配列決定アダプタ（すなわち、モータアダプタ）（例えば、モータタンパク質108）は、DNA分子の両端にライゲーションされる。配列決定のプロセスは、モータタンパク質（例えば、モータタンパク質112）が二本鎖DNAをほどくと開始し、第1の鎖がナノポアを通過することを可能にする。DNA鎖がナノポア116を通過するとき、センサ（例えば、電極）は、配列文脈及び関連する塩基修飾（一次元（1D）リードと呼ばれる）に応じて、ピコアンペア（pA）で経時的な（ミリ秒、ms）イオン電流の変化を測定する。グラフ120は、時間に対する例示的な電流信号を示す。別の実施形態では、ヘアピン配列アダプタが、二本鎖DNA分子のために第1の鎖及びその相補鎖と一緒に共有結合的に連結するために使用されるであろう。したがって、配列決定中に、二本鎖DNA分子の鎖が配列決定され、続いて相補鎖（1D²又は二次元（2D）リードと呼ばれる）が配列決定され、これに

より、配列決定の精度が改善する可能性があり得る。なお別の実施形態では、タンパク質によって連結された二本鎖DNA分子の一端は、同じ分子の第1の鎖の配列決定の完了に続く相補鎖の配列決定の可能性を増加させ、1D²リードを生成する。

【0051】

生の信号（例えば、グラフ120における電流）は、ベースコール及び塩基修飾分析に使用される。一部の実施形態では、ベースコール及び塩基修飾分析は、例えば、再帰型ニューラルネットワーク（RNN）、畳み込みニューラルネットワーク（CNN）、隠れマルコフモデル（HMM）、又はそれらの1つ以上の組み合わせであるが、これらに限定されない、機械学習アプローチによって実行される。

【0052】

一実施形態では、本発明者らは、ナノポア配列決定によって生成される電流信号を処理する新しい方法を開発し、処理された信号を、畳み込みニューラルネットワーク（CNN）又は再帰型ニューラルネットワーク（RNN）に基づいて、単一分子レベルでのDNAメチル化の決定のために分析した。

【0053】

II. 電流信号分析

ナノポア配列決定からの電流信号は、塩基修飾を特定するために分析され得る。しかしながら、図1に説明される機械学習アプローチは、ナノポアを使用して取得される生の電流の入力のみを使用しない。本明細書において説明される実施形態は、電流の一部のうちの1つ以上の統計値を使用する。これらの1つ以上の統計値のベクトルを、ヌクレオチドの同一性及びヌクレオチドの位置を含む、ヌクレオチドのウィンドウに対応する他の情報と組み合わせてもよい。ヌクレオチドの位置は、ウィンドウ内の標的位置に対するものであり得、標的位置は、修飾又はその欠如が検出される位置である。ヌクレオチドのウィンドウについての情報は、入力データ構造を形成するために、核酸分子の領域内の電気信号の統計値とともに含まれ得る。これらの入力データ構造上で訓練されたモデルを使用して、塩基修飾を検出することができる。

【0054】

A. 電流ベクトルパラメータ

ナノポアを通過するヌクレオチド鎖について、N個の事象（すなわち、特定された異なるヌクレオチドと関連付けられる信号セグメント）を検出する。一実施形態では、1つの事象は、特定の時間単位（例えば、ミリ秒）でサンプリングされた一連の電気信号を有する、ベースコール中に特定された1つのヌクレオチドに対応する。一例では、電流は、4kHzの周波数でサンプリングされた（Rang et al. Genome Biol. 2018; 19: 90）。別の実施形態では、1つの事象は、特定の時間速度でサンプリングされた一連の電気信号を有する、ベースコール中に特定された2つ以上のヌクレオチドに対応する。

【0055】

図2は、電流信号のグラフを示す。y軸上のピコアンペア単位の電流振幅。ミリ秒単位の時間が、x軸にある。ドット（例えば、ドット204）は、個々の信号測定値を示す。隣接するドットを通る線（例えば、線208）は、ヌクレオチドと関連付けられる信号測定値の信号セグメント（例えば、線208についてのA）を示す。事象iについて、 m_i 電流信号があると仮定して、事象i上の電流信号jの振幅を P_{ij} によって示した。一実施形態では、ヌクレオチドについて、X1、X2、X3、X4、及びX5を含む信号特徴ベクトルを使用して、そのヌクレオチドと関連付けられる電気信号のパターンを特徴付ける。X1、X2、及びX3についての定義を図2に例解する。X1は、 P_{ij} の平均である。X2は、 P_{ij} の標準偏差である。X3は、 P_{ij} の中央値である。。X4は、X3からの電流の絶対偏差の中央値である（図2においてラベルされた絶対偏差は1つのみ）。X5は、標準偏差で除算した電流信号の平均からのX1の差である。X5は、セグメントの電流信号のzスコアとみなすことができる。

【0056】

10

20

30

40

50

一実施形態では、 P_{ij} は、正規化された信号であり得る。正規化は、正規化された信号値が 0 及び 1 の範囲内にあるように、元の範囲からの電流信号を再スケーリングすることを伴い得、ヌクレオチド鎖の一部又は全体に関する最小値及び最大値を使用する。正規化は、正規化された信号値の平均が 0 であり、標準偏差が 1 であるように、電流信号を再スケーリングすることを伴い得る。正規化は、ヌクレオチド鎖の一部又は全体に関する中央値及び偏差の使用により、電流信号を再スケーリングすることを伴い得る。

【0057】

$X1$ 及び $X2$ は、事象 i と関連付けられる P_{ij} の平均及び標準偏差を表す。

【0058】

$X1$ は、以下によって定義される。

10

【数1】

$$X1 = \frac{\sum_{j=1}^{j=m_i} P_{ij}}{m_i}$$

【0059】

$X2$ は、以下によって定義される。

【数2】

$$X2 = \sqrt{\frac{\sum_{j=1}^{j=m_i} (P_{ij} - X1)^2}{m_i - 1}}$$

20

【0060】

$X3$ は、以下によって定義される。

$X3 = \text{中央値}(P_{ij})$ 、

式中、 i は、塩基修飾分析（例えば、CpG 部位でのメチル化）のための調査の塩基の周囲の事象を含む、 $l \sim r$ の範囲である。変数 l 及び r は、（ヌクレオチド配列に対応する）事象の配列のウィンドウの左右を表す。 l と r との間のヌクレオチド配列は、一般に、後述する電流信号パターンの統合された表示マトリックス（IPM と称する）よりも長いはずである。所与の事象 i について、 j は、 $1 \sim m_i$ の範囲である。 $X3$ は、全てのセグメントを決定する際に使用される電流信号の中央値であり得る。 $X3$ は、 $X3$ が単一セグメントだけでなく、それより多くのセグメントについての電流を使用して決定されるため、全てのセグメントについて同じ値であり得る。一部の実施形態では、 $X3$ は、特定のウィンドウのためのものであり得る。他の実施形態では、 $X3$ は、複数のウィンドウにまたがる中央値であり得る。

30

【0061】

$X4$ は、以下によって定義される。

$X4 = \text{中央値}(|P_{ij} - X3|)$ 、

式中、 $|\cdot|$ は、絶対値を表し、 i は、塩基修飾分析（例えば、CpG 部位でのメチル化）のための調査の塩基の周囲の事象を含む、 $l \sim r$ の範囲である。所与の i について、 j は、 $1 \sim m_i$ の範囲である。 $X4$ は、全てのセグメントを決定する際に使用される電流信号の絶対偏差の中央値であり得る。 $X4$ は、単一セグメントだけでなく、それより多くのセグメントについての電流を使用して（例えば、全てのサンプリングされた電流値を使用して）計算され得、したがって、全てのセグメントについて同じ値であり得る。

40

【0062】

$X5$ は、以下によって定義される。

【数3】

50

$$X5 = \frac{X1 - \mu}{\sigma},$$

$$\text{式中、}\mu = \frac{\sum_{i=l}^{l=r} \sum_{j=1}^{j=m_r} P_{ij}}{M-1} \text{ 及び } \sigma = \sqrt{\frac{\sum_{i=l}^{l=r} \sum_{j=1}^{j=m_r} (P_{ij} - \mu)^2}{M-1}},$$

i は、塩基修飾分析（例えば、C p G 部位でのメチル化）のための調査の塩基の周囲の事象を含む、l ~ r の範囲である。所与の i について、j は、1 ~ m_i の範囲である。M は、l ~ r の範囲の事象についてサンプリングされた電流信号の総数である。複数の電流信号と関連付けられ、X 3 を決定するために使用される領域のサイズは、DNA 断片のサイズであり得る。例えば、DNA 断片が 500 bp である場合、次いで、領域のサイズは 500 である。断片が 300 bp である場合、次いで、領域のサイズは 300 である。一部の実施形態では、X 3 を決定するために、DNA 断片をより小さなサブ断片に更に分割することが有用であり得る。X 3 を決定するために使用される領域のサイズは、5 nt、10 nt、20 nt、30 nt、40 nt、50 nt、60 nt、70 nt、90 nt、100 nt、200 nt、300 nt、400 nt、500 nt、600 nt、800 nt、900 nt、1 kb、2 kb、3 kb、4 kb、5 kb、10 kb、50 kb などであり得る。

10

20

【0063】

X 1 及び X 2 は、各ヌクレオチドについての電気信号の局所パターンを表す、事象 i 内の信号変化を反映するために使用することができる。X 3、X 4、及び X 5 は、l ~ r の範囲の他の周囲の事象に対する、事象 i についての信号変化を反映するために使用することができる。一部の実施形態では、周囲の事象は、塩基修飾分析のための調査の塩基の X - nt 上流及び Y - nt 下流であり得る。X は、0、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、31、32、33、34、35、36、37、38、39、40、41、42、43、44、45、46、47、48、49、50、100、150、200、300、400、500、1000、2000、4000、5000、及び 10000 を含み得るが、これらに限定されない。Y は、0、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、31、32、33、34、35、36、37、38、39、40、41、42、43、44、45、46、47、48、49、50、100、150、200、300、400、500、1000、2000、4000、5000、及び 10000 を含み得るが、これらに限定されない。一実施形態では、周囲の事象は、ナノポアを通過するヌクレオチド鎖全体であり得る。

30

【0064】

B. 一本鎖分析

図 3 は、電流信号のグラフを示す。y 軸上のピコアンペア単位の電流振幅。ミリ秒単位の時間が、x 軸にある。トレース 304 は、経時的な電流振幅である。信号セグメント（例えば、セグメント 308）は、ヌクレオチドと関連付けられるトレース 304 の一部である。電流変化は、ナノポアを通過する異なるヌクレオチドに応じて変動するであろう。ナノポア配列決定におけるベースコールは、一般に、電流信号を異なる局所的に静止した状態（すなわち、事象）に変換することに依存する。電流信号を異なる事象に変換するプロセスは、電気信号セグメント化と呼ばれる。イオン電流変化は、信号セグメントにおける 1 つ以上のヌクレオチドに対応する事象の振幅（例えば、ピコアンペア、pA で測定される）、イオン電流の方向、信号セグメントにおける 1 つ以上のヌクレオチドに対応する電流事象の持続時間、イオン電流の変化率、及び異なる信号セグメントにわたる相対的振

40

50

幅を含むが、これらに限定されない。振幅は、電流の強度又は大きさを指し得、交流を意味する必要はない。これらの電流事象は、例えば、Tomb oと呼ばれるソフトウェアを使用して異なる塩基に割り当てられる (Stoiber et al bioRxiv. 2016; doi.org/10.1101/094672)。1つのヌクレオチドは、異なる振幅を有する一連の事象と関連付けられるであろう。そのようなツール (Tomb o) により、マン・ホイットニーのU検定に基づいて、そのような塩基が修飾されたか否かを推測するために2つの試料間のゲノム塩基に割り当てられたナノポア信号における差を検定することを試みた (Stoiber et al bioRxiv. 2016; doi.org/10.1101/094672)。このツール (Tomb o) は、上流及び下流の信号並びに配列文脈を考慮せず、異なる配列リードからの全ての信号がゲノム塩基に集約されたため、単一の分子レベルでメチル化パターンを分析することができなかった。Tomb oの性能は、Nanopolish及びDeepSignalなどの他のツールの性能と比較されている (Yuen et al. bioRxiv. 2020; doi:doi.org/10.1101/2020.10.14.340315)。

【0065】

一実施形態では、ヌクレオチドに関連する信号セグメント内の電流パターンを特徴付けるために、その信号セグメント内の事象のそれらの電流振幅の平均 (X1) 及び標準偏差 (X2) を計算する。分子全体と関連付けられる事象の電流振幅の中央値 (X3) 及び分子全体と関連付けられる事象の電流振幅の絶対偏差の中央値 (X4) が決定される。信号セグメントについての正規化された信号 (X5) は、以下の式によって決定される。

【数4】

$$X5 = \frac{X1 - \mu}{\sigma}$$

式中、X1は、問題のヌクレオチドに関連するその信号セグメント内の事象のそれらの電流振幅の平均であり、 μ は、調査中の分子全体内の事象のそれらの電流振幅の平均であり、 σ は、調査中の分子全体内の事象のそれらの電流振幅の標準偏差である。一実施形態では、平均及び標準偏差は、最大値及び最小値のわずかな指定されたパーセンテージを除いた後に導き出すことができる。

【0066】

ヌクレオチドについて、X1、X2、X3、X4、及びX5を含む信号特徴ベクトルを使用して、そのヌクレオチドと関連付けられる電気信号のパターンを反映させる。例えば、セグメント308は、[X1、X2、X3、X4、X5]の信号特徴ベクトルを有し得る。

【0067】

X1及びX2は、信号セグメントi内の事象の電流振幅の平均及び標準偏差を表す。X3は、分子全体と関連付けられる事象の電流振幅の中央値を表す。X4は、分子全体と関連付けられる事象の電流振幅の絶対偏差の中央値を表す。X5は、信号セグメントiについての正規化された信号を表す。

【0068】

図4は、信号セグメントの長さの頻度のプロットである。ヌクレオチドと関連付けられる電流事象の長さ (すなわち、ミリ秒単位の持続時間) は、x軸にある。長さの頻度は、y軸上に示される。図4は、ヌクレオチドと関連付けられる各信号セグメントの長さが、9の中央値で変数であったことを示す (範囲: 1~3540)。

【0069】

塩基修飾は、その上流及び下流のヌクレオチドと関連付けられる電気信号に影響を与えるであろう。本開示では、本発明者らは、性能を改善するために、塩基修飾分析のためのヌクレオチドに関連する電流信号、目的のヌクレオチドの近くのヌクレオチドと関連付けられる電流信号、及び配列決定文脈を集合的に使用した。CpG部位でのDNAメチル化 (すなわち、シトシンの5番目の炭素でのメチル化) は、脊椎動物のゲノムにおいて最も

一般的なタイプの塩基メチル化である。C p G 部位での D N A メチル化の分析は、本開示のための例示的な例として使用された。

【0070】

図5は、ナノポア配列決定による1つの鎖からの電流信号を使用してメチル化を決定するためのプロセスを示す。ブロック504において、二本鎖D N A 分子を提供する。ブロック508において、二本鎖D N A 分子を、ナノポア配列決定に好適である配列決定アダプタとライゲーションする。ブロック512において、ナノポア配列決定を実施する。単一の二本鎖分子の鎖は、膜に埋め込まれた孔を通して移動し、ナノポアを通して流れるイオン電流信号を変化させる。ブロック516において、電流信号を取得する。イオン電流信号は、例えば、トランス電極によって測定され得る。

10

【0071】

電流信号は、例えば、T o m b o を使用して、セグメント化ステップによって処理される (S t o i b e r e t a l b i o R x i v . 2 0 1 6 ; d o i . o r g / 1 0 . 1 1 0 1 / 0 9 4 6 7 2) 。これらのセグメント化された電気事象は、異なるヌクレオチドに割り当てられる。ブロック520において、統合された表示マトリックス (I P M) を構築する。I P M は、電流信号パターンのマトリックスであり、各塩基についての電流信号、配列決定文脈、及び塩基修飾分析のための遺伝子座の近く又は周囲にある一連のヌクレオチドにまたがる配列決定位置情報を含む。一実施形態では、ヌクレオチドと関連付けられるセグメント化された電気事象は、信号特徴ベクトル、すなわち、[X1、X2、X3、X4、X5]によって説明した。いくつかの信号特徴ベクトルを有する、C p G 部位内のシトシン、及び例えば、そのシトシンの10 n t 上流及び下流 (すなわち、例えば、合計21 n t) を使用して、電流信号パターンのI P M を形成した。説明の目的で、

20

【数5】

5'-T[CCATGC]CATCGTC[GATGCA]G-3'

の21 n t 配列を例として使用して、I P M 524をもたらした。括弧内の塩基は、単純化のために省略した (「・・・」で示される)。アデニン (「A」) の塩基に対応する-2の位置について、「A」と関連付けられる信号特徴ベクトル [X1=1.7、X2=0.29、X3=24.2、X4=436、X5=-0.3] は、「-2」の列と「A」の行との間の対応するセルに充填された。同じ列における他のセルは、「0」で充填された。21 n t 配列文脈に関連する各ヌクレオチドについての残りの信号特徴ベクトルは、同じ規則を使用して充填され、したがって、21 n t I P M が形成された。したがって、そのようなI P M は、電流信号パターン、配列決定文脈、配列決定位置、及び経時的に変化するパターンを同時にコードするであろう。メチル化及び非メチル化D N A データセットに由来するいくつかのI P M を、C N N 又はR N N モデルを訓練するために使用し、続いて、試験試料中のC p G 部位でのメチル化状態を決定するために使用した。

30

【0072】

ブロック528は、C N N 分析を示す。C N N 分析のために、I P M を入力層に供給し、続いて畳み込み層及び出力層のプロセスを行った。C p G についてのメチル化の確率 (すなわち、出力メチル化スコア、0~1の範囲) は、出力層におけるシグモイド関数に基づいて決定した。このアプローチは、I P M - C N N と称する。一実施形態では、メチル化C p G 部位 (M . S s s I 処理D N A) 及び非メチル化C p G 部位 (全ゲノム増幅 (W G A) D N A) についてのI P M を、C N N モデルを訓練するために使用した。M . S s s 処理D N A に由来するデータセットにおけるC p G 部位についてのメチル化の目標値を「1」として定義し、一方、W G A D N A に由来するデータセットにおけるC p G 部位についてのメチル化の目標値を「0」として定義した。シグモイド関数によって計算された出力スコアと所望の目標出力との間の全体的な予測誤差 (2進値: 0又は1) を、反復的にモデルパラメータを更新することによって最小化することにより、I P M - C N N の最適なパラメータを取得した。全体的な予測誤差を、深層学習アルゴリズム (k e r a s

40

50

．i o /) におけるシグモイドクロスエントロピー損失関数によって決定した。訓練データセットから学習したモデルパラメータを、試験データセット内のメチル化状態を分析し、C p G 部位がメチル化されている尤度を示唆する確率スコア（すなわち、メチル化の確率）を出力するために使用した。一実施形態では、C N N モデルは、各々が 25 のカーネルサイズを有する 32、64、128、256 個のフィルターを有する 4 つの二次元（2D）畳み込み層を使用した。正規化線形ユニット（R e L U）の活性化関数をそれらの畳み込み層に使用した。続いて、バッチ正規化層を適用した。平坦化層を更に追加し、続いて 0.5 のドロップアウト率を有するドロップアウト層を追加し、次いで、続いて R e L U 活性化関数の使用により 200 個のニューロンを含む全結合層を追加した。1 個のニューロンを有する出力層を、シグモイド活性化関数を用いて最終的に適用して、メチル化されている C p G 部位についての確率スコア（すなわち、メチル化の確率）を得た。C N N モデルについてのプログラムは、K e r a s 深層学習フレームワーク（<https://keras.io/>）に基づいて実装した。

10

【0073】

ブロック 532 は、R N N 分析を示す。R N N 分析のために、I P M を入力層に供給し、続いて長短期間メモリ（L S T M）層及び出力層のプロセスを行った。C p G についてのメチル化の確率（0～1 の範囲）は、出力層におけるシグモイド関数に基づいて決定した。このアプローチは、I P M - R N N と称する。I P M - R N N において使用されるものと同様の訓練手順を使用して、シグモイド関数によって計算された出力スコアと所望の目標出力との間の全体的な予測誤差（2 進値：0 又は 1）を、反復的にモデルパラメータを更新することによって最小化することにより、I P M - R N N の最適なパラメータを取得した。訓練データセットから学習したモデルパラメータを、試験データセット内のメチル化状態を分析し、C p G 部位がメチル化されている尤度を示唆する確率スコア（すなわち、メチル化の確率）を出力するために使用した。一実施形態では、L S T M ユニットの R N N モデルを、各々が 256 個の隠れノードを有する 2 つの全結合隠れ層とともに使用した。最後の層の後に、ドロップアウト率 0.2 を有するドロップアウト層が続いた。1 個のニューロンを有する出力層を、シグモイド活性化関数を用いて最終的に適用して、メチル化されている C p G 部位についての確率スコア（すなわち、メチル化の確率）を得た。C N N モデルについてのプログラムは、K e r a s 深層学習フレームワーク（keras.io/）に基づいて実装した。

20

30

【0074】

C. 二本鎖分析

図 6 は、ナノポア配列決定による両方の D N A 鎖からの電流信号を使用してメチル化を決定するためのプロセスを示す。一実施形態では、二本鎖 D N A 分子の両方のヌクレオチド鎖からの電流信号は、そのような二本鎖 D N A 分子が、第 2 のヌクレオチド鎖（相補鎖又はクリック鎖と称される）が同じナノポアを通過する第 1 のヌクレオチド鎖（ワトソン鎖と称される）の完了直後に続くように配列決定されるときに得ることができる。同じナノポア内の二本鎖 D N A の両方のヌクレオチド鎖を順次配列決定するためのこの技術は、1 D² 又は 2 D 配列決定と称する。ブロック 604 において、二本鎖 D N A 分子を提供する。ブロック 608 において、二本鎖 D N A 分子を、ナノポア配列決定に好適である配列決定アダプタとライゲーションする。ブロック 612 において、単一の二本鎖分子の鎖が、膜に埋め込まれた孔を通して移動し、続いて相補鎖が移動した。ブロック 616 において、電流信号を、各二本鎖 D N A 分子の両方の鎖について取得する。イオン電流信号は、トランス電極によって測定され得る。得られた電流信号を、G u p p y（O x f o r d N a n o p o r e T e c h n o l o g i e s L t d）を使用して、配列決定された D N A 分子のヌクレオチド情報を推定する（すなわち、ベースコール）ために使用した。一実施形態では、A l b a c o r e（nanoporetech.com/）、W a v e N a n o（W a n g e t a l. Q u a n t i t a t i v e B i o l o g y. 2018; 6: 359-368）、C h i r o n（T e n g e t a l. G i g a S c i e n c e. 2018; 7: g i y 037）、F l a p p i e（github.com/na

40

50

`nanoporetech/flappie`）、`Scrappie (github.com/nanoporetech/scrappie)` などを含むが、これらに限定されない、他のベースコールツールが使用され得る。

【0075】

特定の時間速度（例えば、ミリ秒）でサンプリングされた電流信号は、塩基修飾分析のために異なる検出ヌクレオチドに割り当てられる。電流信号は、例えば、`Tombor` を使用して、セグメント化ステップによって処理される（`Stoiber et al bioRxiv. 2016; doi.org/10.1101/094672`）。これらのセグメント化された電気事象は、異なるヌクレオチドに割り当てられる。ブロック620において、統合された表示マトリックス（IPM）を、各二本鎖DNA分子からの両方の鎖を含むように構築する。一実施形態では、ヌクレオチドと関連付けられるセグメント化された電気事象は、信号特徴ベクトル、すなわち、 $[X_1, X_2, X_3, X_4, X_5]$ によって説明した。相補鎖の対応する塩基からの信号特徴ベクトル、すなわち、 $[X_1', X_2', X_3', X_4', X_5']$ を得た。いくつかの信号特徴ベクトルを有する、CpG部位内のシトシン、及び例えば、そのシトシンの10nt上流及び下流（すなわち、例えば、合計21nt）を使用して、電流信号パターンのIPMを形成した。同じ二本鎖DNA分子の相補鎖内の対応する塩基からのIPMを得た。ワトソン鎖及びクリック鎖に由来するIPMを組み合わせ、塩基修飾分析のために、より高い次元を有する新しいIPMマトリックスを形成した。

【0076】

一部の実施形態では、`NanoMod (Liu et al. BMC Genomics. 2019; 20:78)`、`Albacore (nanoporetech.com/)`、`Chiron (Teng et al. Gigascience. 2018; 7:giy037)`、`Nanopolish (Simpson et al. Nat Methods. 2017; 13:407-410)`、`Scrappie (https://github.com/nanoporetech/scrappie)`、`UNCALLED (Kovaka et al. Nat Biotechnol. 2020; doi:10.1038/s41587-020-0731-9)` などを含む他の計算ツールが、異なるヌクレオチドに電流信号を割り当てるために使用され得る。二本鎖分析のために説明されるこれらの計算ツール及び他の技術は、一本鎖分析のために使用してもよい。

【0077】

説明の目的で、

【数6】

5'-T[CCATGC]CATCGTC[GATGCA]G-3'

の21nt配列を、IPM624のための基礎としての一例として使用した。IPM624は、IPM524と類似していてもよいが、ワトソン鎖及びクリック鎖の両方を含んでもよい。括弧内の塩基は、単純化のために省略した（「・・・」で示される）。ワトソン鎖内のアデニン（「A」）の塩基に対応する-2の位置について、「A」と関連付けられる信号特徴ベクトル、すなわち、 $[X_1=1.7, X_2=0.29, X_3=436, X_4=24.2, X_5=-0.3]$ は、「-2」の列と「ワトソン鎖」によって示される領域内の「A」の行との間の対応するセルに充填された。相補鎖（すなわち、クリック鎖）内のその対応する塩基「T」について、「T」と関連付けられる信号特徴ベクトル、 $[X_1'=-1.9, X_2'=0.23, X_3'=24.2, X_4'=436, X_5'=-1.4]$ は、「-2」の列と「クリック鎖」によって示される領域内の「T」の行との間の対応するセルに充填された。同じ列における他のセルは、「0」で充填された。一部の実施形態では、信号特徴ベクトル内の要素の順序は変更され得る。例えば、 $[X_2, X_1, X_3, X_4, X_5]$ 、 $[X_2, X_3, X_4, X_5, X_1]$ 、 $[X_1, X_3, X_5, X_4, X_2]$ 、又は他の組み合わせを使用することができる。一部の実施形態では、信号特徴ベクト

ルのサイズは、5に制限され得ない。例えば、信号特徴ベクトルのサイズは、より多くの処理された電気信号特徴又は生の電気信号を追加することによって、6、7、8、9、10、15、20、30、40、50、100などを含み得るが、これらに限定されない。信号特徴ベクトルのサイズは、信号特徴ベクトル内のいくつかの特徴を編集又は削除することによって、1、2、3、4を含み得るが、これらに限定されない。

【0078】

21nt配列文脈に関連する各ヌクレオチドについての残りの信号特徴ベクトルは、同じ規則を使用して充填され、したがって、21nt IPMが形成された。したがって、そのようなIPMは、電流信号パターン、配列決定文脈、配列決定位置、及び経時的に変化するパターンを同時にコードするであろう。メチル化及び非メチル化DNAデータセットに由来するいくつかのIPMを、CNN又はRNNモデルを訓練するために使用し、続いて、試験試料中のCpG部位でのメチル化状態を決定するために使用した。

【0079】

ブロック628は、CNN分析を示す。実施形態では、CNNモデルは、各々が 1×25 のカーネルサイズを有する32、64、128、256個のフィルターを有する4つの二次元(2D)畳み込み層を使用した。正規化線形ユニット(ReLU)の活性化関数をそれらの畳み込み層に使用した。続いて、バッチ正規化層を適用した。平坦化層を更に追加し、続いて0.5のドロップアウト率を有するドロップアウト層を追加し、次いで、続いてReLU活性化関数の使用により200個のニューロンを含む全結合層を追加した。1個のニューロンを有する出力層を、シグモイド活性化関数を用いて最終的に適用して、メチル化されているCpG部位についての確率スコア(すなわち、メチル化の確率)を得た。CNNモデルについてのプログラムは、Keras深層学習フレームワーク(keras.io/)に基づいて実装した。一部の実施形態では、カーネルサイズ $n \times m$ は変動され得、「n」は、1、2、3、4、5、10、15、20、30、35、40、45、50、100などを含み得るが、これらに限定されず、「m」は、1、2、3、4、5、10、15、20、30、35、40、45、50、100などを含み得るが、これらに限定されない。

【0080】

図7は、塩基修飾分析の性能に対するカーネルサイズの影響の表である。1列目は、異なるカーネルサイズを示す。2列目は、訓練データセットからのAUC(ROC[受信機オペレータ特徴]曲線下面積)を示す。3列目は、試験データセットからのAUCを示す。図7は、 1×5 、 1×10 、 1×15 、 1×20 、及び 1×25 などのカーネルサイズの範囲が、それぞれ、0.96、0.96、0.97、0.96、及び0.96のAUCによって示されるように、メチル化CpG部位と非メチル化CpG部位との間の区別において同等の性能を与えることを示す。

【0081】

ブロック632は、RNN分析を示す。実施形態では、LSTMユニットを有するRNNモデルを、各々が256個の隠れノードを有する2つの全結合隠れ層とともに使用した。LSTM隠れユニットの電流出力は、電流入力及びLSTMセルに記憶された以前の情報によって決定される。一例として、21nt IPMの第1の行で示す位置と関連付けられた信号特徴ベクトル $[X_1, X_2, X_3, X_4, X_5]$ は、特定の時間ステップでのLSTMユニットについての入力 X_t とみなされた。前方LSTM RNNは、以下のような操作に基づいて時間ステップに従って隠れ層Hを再帰的に計算する(Gers et al. IEEE Transactions on Neural Networks, 2001; 12:1333-1340)。

【数7】

10

20

30

40

50

$$A_{t,F} = \text{シグモイド}(W_{xa,F}X_{t,F} + W_{ha,F}H_{t-1,F} + W_{ca,F} \odot C_{t-1,F} + b_{a,F}),$$

$$F_{t,F} = \text{シグモイド}(W_{xf,F}X_{t,F} + W_{hf,F}H_{t-1,F} + W_{cf,F} \odot C_{t-1,F} + b_{f,F}),$$

$$C_{t,F} = F_{t,F} \odot C_{t-1,F} + A_{t,F} \odot \tanh(W_{xc,F}X_{t,F} + W_{hc,F} \odot C_{t,F} + b_{c,F}),$$

$$O_{t,F} = \text{シグモイド}(W_{xo,F}X_{t,F} + W_{ho,F}H_{t-1,F} + W_{co,F} \odot C_{t,F} + b_{o,F}),$$

$$H_{t,F} = O_{t,F} \odot \tanh(C_{t,F}).$$

【0082】

後方LSTM RNNは、以下のような操作に基づいて時間ステップに従って隠れ層Hを再帰的に計算する(Gers et al. IEEE Transactions on Neural Networks. 2001; 12: 1333-1340)。

10

【数8】

$$A_{t,B} = \text{シグモイド}(W_{xa,B}X_{t,B} + W_{ha,B}H_{t-1,B} + W_{ca,B} \odot C_{t-1,B} + b_{a,B}),$$

$$F_{t,B} = \text{シグモイド}(W_{xf,B}X_{t,B} + W_{hf,B}H_{t-1,B} + W_{cf,B} \odot C_{t-1,B} + b_{f,B}),$$

$$C_{t,B} = F_{t,B} \odot C_{t-1,B} + A_{t,B} \odot \tanh(W_{xc,B}X_{t,B} + W_{hc,B} \odot C_{t,B} + b_{c,B}),$$

$$O_{t,B} = \text{シグモイド}(W_{xo,B}X_{t,B} + W_{ho,B}H_{t-1,B} + W_{co,B} \odot C_{t,B} + b_{o,B}),$$

$$H_{t,B} = O_{t,B} \odot \tanh(C_{t,B}).$$

20

式中、W及びbは重み及びバイアスであり、Xは入力ベクトルであり、Aは入力ゲートの活性化ベクトルであり、Fは忘却ゲートのシグモイド関数であり、Cはセル状態であり、Oは出力ゲートのシグモイド関数であり、HはLSTM隠れユニットの出力である。

【0083】

前方及び後方LSTM RNNユニットの出力が組み合わされる。

【数9】

$$Z_t = H_{t,F} \oplus H_{t,B}.$$

30

【0084】

LSTM RNN出力の最後の層の後に、ドロップアウト率0.2を有するドロップアウト層が続いた。1個のニューロンを有する出力層を、シグモイド活性化関数を用いて最終的に適用して、メチル化されているCpG部位についての確率スコア（すなわち、メチル化の確率）を得た。CNNモデルについてのプログラムは、Keras深層学習フレームワーク(keras.io/)に基づいて実装した。

【0085】

D. パラメータ分析

AUC(ROC[受信機オペレータ特徴]曲線下面積)に対する、異なる電流ベクトルパラメータ及び異なるウィンドウサイズの影響を分析する。本発明者らは、本開示に存在する実施形態に従って、IPM-CNNモデルに基づいて、IPMにおける異なるパラメータの使用による区別力を分析した。このために、WGA DNA及びM. SSSI処理DNAデータセットから、8,282個の分子(38,238個のCpG部位)及び8,247個の分子(39,708個のCpG部位)をそれぞれ分析した。

40

【0086】

図16は、AUCに対するパラメータの異なる組み合わせの影響のグラフを示す。電流ベクトルパラメータの異なる組み合わせは、x軸上にあり、AUCはy軸上にある。図16は、これらに限定されないが、IPMにおけるX1、X2、X3、X4、及びX5のパラメータの異なる組み合わせの使用が、CpGメチル化分析の異なる性能につながったこ

50

とを示す。例えば、IPMにおけるX1の使用は、0.954のAUCをもたらしたが、一方、IPMにおけるX1及びX2の組み合わせは、0.893のAUCを生じさせた。IPMにおけるX1、X2、及びX3の組み合わせは、AUCを0.963に上昇させた。IPMにおけるX1、X2、X3、及びX4の組み合わせは、AUCを0.978に更に上昇させ、続いて、この例では、X1、X2、X3、X4、及びX5の使用による0.977のAUCにおいて性能の横ばいが続いた。したがって、一部の実施形態では、IPMにおけるパラメータの異なる組み合わせは、メチル化CpG部位と非メチル化CpG部位との間の区別における所望の性能を決定することを可能にするであろう。

【0087】

組み合わせではなく個別に、X1、X2、X3、X4、及びX5の使用を試験した。個別に、X1、X2、X3、X4、及びX5を使用する結果は、それぞれ、0.95、0.92、0.98、0.88、及び0.95のAUCをもたらした。X3（すなわち、領域内のP_{ij}の中央値）は、0.98の高いAUCをもたらした。高いAUCは、少なくとも部分的には、完全な断片レベルでのメチル化差の結果であり得る。使用されたデータセットは、WGA（完全に非メチル化）及びM.S.S.L（完全にメチル化）を伴っていた。しかしながら、実際には、断片は、完全にメチル化されていないか、又は完全に非メチル化されていない。完全にメチル化されていない、又は完全に非メチル化されていない試料に対してX3を単独で使用しても、それほど高いAUCをもたらさない場合がある。

【0088】

図17は、AUCに対するウィンドウサイズの影響のグラフを示す。x軸は、ヌクレオチドにおけるウィンドウサイズを示す。y軸は、AUCを示す。IPMで使用されるヌクレオチドの数（ウィンドウサイズとも称される）は、ナノポア配列決定中に生成される電流信号の異なる情報含有量を捕捉し、メチル化分析の性能に影響を与える可能性がある。図17は、IPM-CNNモデルを使用するメチル化CpG部位と非メチル化CpG部位との間の区別における性能が、IPMで使用されるヌクレオチドの数が1ntから10ntに増加するにつれて、0.715のAUCから0.969に徐々に増加するように見えたことを示している。この例では、性能の横ばいは、7ntのウィンドウサイズに達していた。したがって、一部の実施形態では、IPMのウィンドウサイズを調整することは、メチル化CpG部位と非メチル化CpG部位との間の区別における所望の性能を決定することを可能にするであろう。

【0089】

実施形態は、最も高いAUCにつながる電流ベクトルパラメータ又はウィンドウサイズの組み合わせを使用することを必要としない場合がある。より低いAUCは、特定の用途には十分であり得るか、又はより高いAUCは、追加のパラメータに関連する追加の計算コスト及び記憶コストに値しない場合がある。更に、異なるパラメータを調整して、所望のAUC、特異度、及び／又は感度を達成し得る。例えば、より大きなウィンドウサイズを使用して、X1、X2、X3、X4、及びX5の中のより少ないパラメータを使用することを補うことができる。

【0090】

E. 6mA修飾の検出

5mC以外の修飾に対する電流信号分析の適用性を決定するために、電流信号分析を使用して、N6-メチルアデニン（6mA）を検出した。

【0091】

図18は、ナノポア配列決定による1つの鎖からの電流信号を使用して6mAのメチル化を決定するためのプロセスを示す。図18は、5mCのメチル化を決定するためのプロセスを示した図5と同様である。ブロック1804において、二本鎖DNA分子を提供する。ブロック1808において、二本鎖DNA分子を、ナノポア配列決定に好適である配列決定アダプタとライゲーションする。ブロック1812において、ナノポア配列決定を実施する。ブロック1816において、電流信号を取得する。ブロック1820において、統合された表示マトリックス（IPM）を構築する。ブロック1804～1820は、

ブロック504～520と同じであり得る。

【0092】

6mAのメチル化を決定するための説明の目的で、

【数10】

5'-G[TACCCG]GGTACTG[TCTAGA]G-3'

の21nt配列を、メチル化分析のための対象であったヌクレオチドA（例えば、0の位置に対応）を中心に、IPMのための基礎としての一例として使用した。IPM1824は、21nt配列の使用の結果を示す。括弧内の塩基は、単純化のために省略した（「・・」で示される）。1つの鎖におけるアデニン（「A」）の塩基に対応する0の位置について、「A」と関連付けられる信号特徴ベクトル（すなわち、 $[X1=0.39, X2=0.04, X3=389, X4=46.3, X5=0.32]$ ）は、マトリックスの「0」の列と「A」の行との間の対応するセルに充填された。同じ列における他のセルは、「0」で充填された。一部の実施形態では、信号特徴ベクトル内の要素の順序は変更され得る。例えば、 $[X2, X1, X3, X4, X5]$ 、 $[X2, X3, X4, X5, X1]$ 、 $[X1, X3, X5, X4, X2]$ 、又は他の組み合わせを使用してもよい。一部の実施形態では、信号特徴ベクトルのサイズは、5だけではない場合がある。例えば、信号特徴ベクトルのサイズは、より多くの処理された電気信号特徴又は生の電気信号を追加することによって、6、7、8、9、10、15、20、30、40、50、100などを含み得るが、これらに限定されない。信号特徴ベクトルのサイズは、信号特徴ベクトル内のいくつかの特徴を編集又は削除することによって、1、2、3、又は4を含み得るが、これらに限定されない。

【0093】

21nt配列文脈に関連する各ヌクレオチドについての残りの信号特徴ベクトルは、同じ規則を使用して充填され、したがって、21nt IPMが形成された。したがって、そのようなIPMは、電流信号パターン、配列決定文脈、配列決定位置、及び経時的に変化するパターンを同時にコードするであろう。ヌクレオチドAに関連するメチル化及び非メチル化DNAデータセットに由来するいくつかのIPMを、CNN又はRNNモデルを訓練するために使用し、続いて、試験試料中のA部位でのメチル化状態を決定するために使用した。ブロック1828は、CNN分析を示し、ブロック1832は、RNN分析を示す。これらのブロックは、ブロック528及び532と同じであり得る。

【0094】

上の例解された本発明者らのアプローチ（IPM-CNN又はIPM-RNN）がアデニンメチル化（6mA）を決定することができたかどうかを試験するために、本発明者らは、以前の研究（Rand et al. Nat Methods 2017; 14: 411-413）からのpUC19プラスミドDNAのナノポア配列決定結果を含む2つの公開データセットをダウンロードした。第1のデータセット（6mAデータセット）は、全てのGATCモチーフがA部位でメチル化であると想定された、dam及びdcmメチルトランスフェラーゼの両方を含有するE. coliで成長したpUC19プラスミドDNAから生成された。第2のデータセット（uAデータセット）は、全てのA部位が非メチル化であると想定された、非修飾ヌクレオチドを用いるPCR増幅に供されたDNAから生成された。訓練プロセスでは、本発明者らは、IPM-CNNモデルを使用して、6mAデータセットからのGATCモチーフを含有する2052個の分子、及びuAデータセットからの2081個の分子を分析した。

【0095】

図19は、IPM-CNNモデルを使用して得られるAUCを示す。x軸は、特異度を示す。y軸は、感度を示す。線1904は、訓練データセットからの結果を示す。訓練データセットを用いるAUCは、0.94である。訓練プロセスでは、本発明者らは、訓練されたIPM-CNNモデルを、6mAデータセットからのGATCモチーフを含有する

522個の分子、及びuAデータセットからの481個の分子に適用した。試験データセットを用いるAUCは、0.92である。加えて、IPM-RNNモデルを使用する場合、訓練データセット及び試験データセットの両方で0.89のAUCを達成した。これらのデータは、IPM-CNN及びIPM-RNNが、6mA部位を非メチル化A部位から区別することを可能にし得ることを示唆した。

【0096】

実施形態では、ヒト又は非ヒトDNAについての6mAの決定のための訓練データセットは、それぞれ、6mAヌクレオチド及び非メチル化Aヌクレオチドの使用によるPCR増幅に基づいて構築されてもよい。PCRの数サイクルの後、DNA分子の大部分は、6mAヌクレオチドで増幅されたDNAから生成されたデータセットについての6mAヌクレオチドを担持し、一方、DNA分子の大部分は、非メチル化Aヌクレオチドで増幅されたDNAから生成されたデータセットについての非メチル化Aヌクレオチドを担持するであろう。これらの2つのタイプのデータセットは、試験試料中のAヌクレオチドのメチル化状態を決定するためのCNNモデル及び／又はRNNモデルを訓練するために使用され得る。

【0097】

5mCに加えて、6mAを検出するための電流信号分析の使用は、他のメチル化タイプに対するそのような分析の適用性を実証する。したがって、これらの方法は、本明細書において説明される他のメチル化を正確に検出するはずである。

【0098】

F. ヒト対象の非腫瘍組織と腫瘍組織との間のCpGメチル化分析

本明細書において説明される実施形態を使用することによって決定される部位のメチル化は、異なるタイプの組織を識別するために使用することができる。本開示の実施形態によるIPM-RNNモデルを使用して、本発明者らは、上咽頭がん(NPC)腫瘍及びバフィーコート試料に由来する細胞DNA分子についてのメチル化パターンを分析した。このために、本発明者らは、4,406bpの中央値サイズ(四分位範囲(IQR):1,962~8,128bp)、及び1分子当たり32CpGの中央値(IQR:13~61)を有する、NPC腫瘍からの147個の分子を使用した。本発明者らは、6,823bpの中央値サイズ(四分位範囲(IQR):2,515~9,304bp)、及び1分子当たり49CpGの中央値(IQR:23~118)を有する、バフィーコートからの別の147個の分子を分析した。

【0099】

図20は、バフィーコート試料及びNPC腫瘍組織試料からのDNA分子の比較のグラフを示す。x軸は、組織タイプを示す。y軸は、パーセントとしてのメチル化レベルを示す。バフィーコート(中央値:74.8%、IQR:71.1%~80.1%)における単一分子メチル化レベル(すなわち、メチル化されていると決定された分子内のCpG部位のパーセンテージ)は、NPC腫瘍(中央値:50、IQR:45.7~53.1)におけるものよりも有意に高いことが見出された(P値<0.0001、ウィルコクソン順位検定)。腫瘍組織に由来するDNA分子は、低メチル化されているように見え、これは、ショートリードバイサルファイト配列決定に基づく以前の結論と一致していた(Chan et al. Proc Natl Acad Sci USA 2013;110:18761-8)。しかしながら、本明細書において説明される新しいナノポア配列決定技術は、ほぼ全長DNA分子を配列決定し、DNA分子についてのメチル化パターンを分析することを可能にする。例えば、ナノポア配列決定は、ショートリード配列決定プラットフォーム(例えば、Illumina)によって調べることができない、600bpよりも大きいサイズのDNA分子を分析することができる。

【0100】

図21は、腫瘍DNA分子及びバフィーコートDNA分子におけるメチル化パターンを例解する。黒の丸(例えば、丸2104)は、メチル化CpG部位を示す。塗りつぶされていない丸(例えば、丸2108)は、非メチル化CpG部位を示す。丸は、分析される

DNA分子の5'末端に対するCpG部位の相対位置を示す(すなわち、図におけるDNA分子の左側は5'末端に近い)。図21に示されるように、腫瘍組織に由来するDNA分子は、バフィーコート試料に由来するものと比較して、分子においてより多くの非メチル化CpG部位を担持する傾向があった。バフィーコート試料からの分子のうちのわずかに5.4%が、2,091bpの長さの中央値で、50%未満の単一分子メチル化レベルを有した。比較すると、NPC腫瘍組織からの分子のうちの39.5%が、2,924bpの長さの中央値で、50%未満の単一分子メチル化レベルを有した。DNA分子の長さは、897bp~10,424bpの範囲であった。

【0101】

これらのデータは、本明細書において説明されるメチル化を検出するためのナノポア配列決定技術を、単一分子メチル化パターン分析に使用して、各DNA分子(例えば、非腫瘍DNA対腫瘍DNA分子)の起源の組織を組織生検試料から区別できることを示す。組織生検からの単一分子メチル化パターン分析の分析は、腫瘍グレード又はサブタイプの検査、がん又は他の疾患の治療の監視、臓器異常(例えば、腎不全)の評価などを可能にするであろう。

【0102】

G. 胎児DNA分子と母体DNA分子との間の分析

本明細書において説明される実施形態を使用することによって決定される部位のメチル化は、胎児DNA分子と母体DNA分子との間を識別するために使用することができる。IPM-CNNモデルに従って、本発明者らは、母体バフィーコートと胎盤組織との間のSNP情報を利用することによって、第3期トリメスターでの妊婦から得られた、1,262個の胎児特異的無細胞DNA分子(中央値サイズ:530bp、IQR:361~779bp)及び6,108個の母体特異的無細胞DNA分子(中央値サイズ:668bp、IQR:448~1,089bp)について、少なくとも5個のCpG部位を有する単一分子メチル化パターンを決定した。そのような妊婦の血漿DNA中の胎児DNA画分は、26.0%であった。

【0103】

図22は、母体特異的DNA分子と胎児特異的DNA分子との間の単一分子メチル化レベルを示す。x軸は、無細胞DNA分子のカテゴリ、すなわち、母体特異的無細胞DNA分子又は胎児特異的無細胞DNA分子を示す。y軸は、単一分子メチル化レベルをパーセントで示す。単一の血漿DNA分子のメチル化レベルの中央値(すなわち、メチル化されていると決定された分子内のCpG部位のパーセンテージ)は、胎児特異的無細胞DNA分子について66.6%(IQR:28.5~86.6%)であり、これは、母体特異的無細胞DNA分子についてのも(中央値:78.5%、IQR:50~93.7%)よりも有意に低かった(P値:<0.0001、マン・ホイットニーのU検定)。この結果は、無細胞DNA分子のメチル化情報の使用が、各血漿DNA分子の母体及び胎児起源を区別することを可能にしたことを示唆した。

【0104】

加えて、IPM-CNNモデルによって決定されたメチル化パターンを、2021年2月5日に出願された米国特許出願第17/168,950号に記載されているように、バフィーコート及び胎盤組織のそれぞれの参照メチル化パターンと比較することにより、妊婦における胎児起源の血漿DNA分子と母体起源の血漿DNA分子との間の区別について、0.87のAUCを達成することができる。

【0105】

図23は、IPM-CNNモデルによって決定されたメチル化パターンに基づいて、妊婦における無細胞DNA分子の胎児及び母体起源分析のためのROC曲線を示す。x軸は特異度であり、y軸は感度である。

【0106】

III. IPMベースのメチル化決定の評価のためのデータセット

非メチル化データセットは、全ゲノム増幅(WGA)を介して調製された増幅されたD

10

20

30

40

50

NAからの配列決定結果を含んでいた(WGA DNAデータセットとして示される)。WGAにおける非修飾ヌクレオチドの使用は、塩基修飾をほとんど含有しない増幅されたDNAをもたらした(少量の入力ゲノムDNAを除く)。メチル化データセットは、配列決定の前に、M. SssI (Spiroplasma sp. 株MQ1からのメチルトランスフェラーゼ遺伝子を含むEscherichia coliの株から単離されたCpGメチルトランスフェラーゼは、二本鎖DNA中の全てのCpG部位をメチル化することによって処理されたDNAからの配列決定結果を含んでいた(M. SssI処理DNAデータセットとして示される)。M. SssIメチルトランスフェラーゼは、CpG部位をメチル化した。

【0107】

WGA DNAデータセットの調製のために、エキソヌクレアーゼ耐性ランダムプライマーを、反応混合物(phi29反応緩衝液及びdNTPを含む)を加熱ブロック中で95℃で5分間インキュベートし、続いて4℃まで冷却することによって、1ngのDNA鑄型に予めアニリングする。次いで、phi29ポリメラーゼを反応混合物に添加し、30℃で4時間インキュベートした。DNAをAmpure XPビーズで精製し、Qubit蛍光光度計で定量した。典型的には、20μlの反応物から200ngのDNAを取得することができる。

【0108】

M. SssI処理DNAデータセットの調製のために、WGAの後、DNAの半分をM. SssI酵素で処理した。メチルトランスフェラーゼ反応緩衝液、S-アデノシルメチオニン(SAM)、及びM. SssIをDNAと混合し、37℃で2時間インキュベートした。65℃で20分間加熱することによって反応を停止させた。ライゲーション配列決定キット(SQK-LSK109)(Oxford Nanopore)を、ライブラリ調製のために使用した。DNAを、NEBNext Ultra II End Repair/dA-tailing Moduleと一緒に、NEBNext FFPE DNA Repair Mixで処理した。Ampure XPビーズの浄化後、Adapter Mix、Ligation Buffer、及びNEBNext Quick T4 DNA Ligaseを添加することによって、配列決定アダプタを修復されたDNAにライゲーションした。ライゲーションしたDNAを、Ampure XPビーズで浄化し、Short Fragment Bufferで洗浄した。ライブラリをElution Bufferに再懸濁した。R9.4.1フローセルを、WGA(試料__01)及びM. SssI処理(試料__02)ライブラリの各々の配列決定に使用した。フローセルを、最初に、Flush Tether及びFlush Bufferを含むフローセルプライミングミックスでプライミングした。次いで、Sequencing Buffer、Loading Beads、及びDNAライブラリを混合することによって、ライブラリローディングミックスを調製した。ライブラリローディングミックスを、フローセル試料ポートに滴下様式で添加した。ロードされたフローセルをPromethIONにおけるスロットに差し込み、デフォルトパラメータを使用して64時間配列決定した。

【0109】

本発明者らは、試料__01及び試料__02についてそれぞれ1560万個及び1530万個のナノポア配列決定リードを取得し、そのうち1380万個(88.7%)及び1380万個(90.7%)のリードをMinimap2(Li H, Bioinformatics. 2018; 34(18): 3094-3100)を使用することによりヒト参照ゲノム(UCSC hg19)に整列させることができた。リード長の中央値は、試料__01及び試料__02についてそれぞれ、510nt(四分位範囲(IQR): 333~778nt)及び606nt(IQR: 382~911nt)であった。一部の実施形態では、BLASR(Mark J Chaisson et al, BMC Bioinformatics. 2012; 13: 238)、BLAST(Altschul SF et al, J Mol Biol. 1990; 215(3): 403-410)、BL

10

20

30

40

50

AT (Kent WJ, Genome Res. 2002; 12 (4): 656-664)、BWA (Li H et al, Bioinformatics. 2010; 26 (5): 589-595)、NGMLR (Sedlazeck FJ et al, Nat Methods. 2018; 15 (6): 461-468)、及びLAST (Kielbasa SM et al, Genome Res. 2011; 21 (3): 487-493) が、配列決定されたリードを参照ゲノムに整列させるために使用され得る。

【0110】

図8は、IPMに基づいてCNN及びRNNモデルを訓練及び試験するために使用される配列決定分子の数を示す表である。1列目は、データセットである。M. SssI処理DNAは、メチル化DNAデータセットであり、WGA DNAは、非メチル化DNAデータセットである。2列目は、訓練に使用される分子の数及びCpG部位の数である。3列目は、試験に使用される分子の数及びCpG部位の数である。訓練データセットについて、本発明者らは、それぞれ、M. SssI処理DNA (メチル化DNA) 及びWGA DNA (非メチル化DNA) からの7, 989個及び8, 052個の配列決定分子をランダムに使用した。そのような訓練データセットは、38, 470個のメチル化CpG部位及び37, 150個の非メチル化CpG部位を含んでいた。試験データセットについて、本発明者らは、それぞれ、M. SssI処理DNA (メチル化DNA) 及びWGA DNA (非メチル化DNA) からの4, 826個及び5, 041個の配列決定分子をランダムに使用した。そのような訓練データセットは、9, 716個のメチル化CpG部位及び11, 444個の非メチル化CpG部位を含んでいた。

【0111】

図9A～9Dは、IPM-CNN及びIPM-RNNアプローチを使用する、WGA DNAデータセットとM. SssI処理DNAデータセットとの間のCpGについてのメチル化される確率の箱ひげ図である。グラフは、x軸上にデータセットを有する。メチル化の確率は、y軸にある。図9A及び9Bは、IPM-CNN分析の使用の結果を示す。図9Aは、訓練データセットのIPM-CNN分析を示し、M. SssI処理DNAデータセット (中央値: 0.99、IQR: 0.987～0.999) におけるCpGについてのメチル化の確率は、WGA DNAデータセット (中央値: 0.03、IQR: 0.001～0.15) におけるものよりも有意に高かった (P値<0.0001、マン・ホイットニーのU検定)。図9Bは、試験データセットのIPM-CNN分析を示し、WGA DNAデータセット (中央値: 0.4、IQR: 0.002～0.18) とM. SssI処理DNAデータセット (中央値: 0.99、IQR: 0.980～0.999) との間のCpGについてのメチル化される確率における有意差をまた示していた (P値<0.0001、マン・ホイットニーのU検定)。

【0112】

図9C及び9Dは、IPM-RNN分析の使用の結果を示す。図9Cは、訓練データセットのIPM-RNN分析を示し、M. SssI処理DNAデータセット (中央値: 0.994、IQR: 0.92～0.99) におけるCpGについてのメチル化される確率は、WGA DNAデータセット (中央値: 0.079、IQR: 0.059～0.118) におけるものよりも有意に高かった (P値<0.0001、マン・ホイットニーのU検定)。図9Dは、試験データセットのIPM-RNN分析を示し、WGA DNAデータセット (中央値: 0.077、IQR: 0.057～0.115) とM. SssI処理DNAデータセット (中央値: 0.994、IQR: 0.919～0.999) との間のCpGについてのメチル化される確率における有意差をまた示していた (P値<0.0001、マン・ホイットニーのU検定)。これらの結果は、本開示に存在する実施形態による、ナノポア配列決定によって生成された電気信号を使用して、CpG部位でのメチル化状態を決定することが実行可能であることを示した。一実施形態では、0.5のメチル化カットオフの確率を使用して、CpG部位でのメチル化状態を決定することができる。このカットオフの使用により、IPM-CNN分析について、DNAメチル化検出についての特異度及び感度は、訓練データセットではそれぞれ96%及び91%、並びに試験データ

セットではそれぞれ93%及び88%であった。IPM-RNN分析について、DNAメチル化検出についての特異度及び感度は、訓練データセット及び試験データセットの両方について、それぞれ97%及び88%であった。一部の実施形態では、メチル化の確率についてのカットオフは、様々な用途に応じて調整され得る。

【0113】

図10A及び10Bは、受信機オペレータ特徴(ROC)曲線分析を示す。特異度は、x軸上に示される。感度は、y軸上に示される。図10Aは、訓練データセットに対する結果を示す。図10Bは、試験データセットに対する結果を示す。IPM-CNN結果は、線1004及び1008で示される。IPM-RNN結果は、線1012及び1016で示される。DeepMod(Liu et al. Nat Commun. 2019; 10:2449)結果は、線1020及び1024で示される。Nanopolish(Liu et al. Nat Commun. 2019; 10:2449)結果は、線1028及び1032で示される。IPMベースのCNN及びRNN分析は、0.95以上のROC曲線下面積(AUC)で、訓練データセット及び試験データセットの両方で良好な性能をもたらした。IPMベースのCNN及びRNNモデルは、DeepMod(0.83)及びnanopolish(0.91)と比較して、試験データセットにおいて0.95及び0.97のROC曲線下面積(AUC)でより良好な性能をもたらした。DeepMod及びnanopolishを含む他のツールに対する、IPMベースのRNN又はCNNの全ての比較についてのP値(DeLong検定)は、 <0.0001 であることが見出された。これらの結果は、IPM-CNN及びIPM-RNNが、DNAメチル化分析のための他のツールよりも優れていることを示した。

【0114】

図11は、異なる分析についての所与の特異度に対する感度の表である。1列目は、分析のタイプを示す。2列目は、感度を示す。3列目は、特異度を示す。図11は、所与の特異度で、IPM-CNN及びIPM-RNN分析がはるかに高い感度を達成したことを示す。例えば、90%の特異度で、IPM-CNN及びIPM-RNN分析は、それぞれ90%及び93%の感度を達成したが、一方、DeepMod及びnanopolishアプローチは、それぞれわずか53%及び74%の感度を達成した。95%の特異度で、IPM-CNN及びIPM-RNN分析は、それぞれ86%及び90%の感度を達成したが、一方、DeepMod及びnanopolishアプローチは、それぞれわずか38%及び55%の感度を達成した。99%の特異度で、IPM-CNN及びIPM-RNN分析は、それぞれ70%及び83%の感度を達成したが、一方、DeepMod及びnanopolishは、それぞれわずか13%及び16%の感度を達成した。これらの結果は、配列セグメントについての電流信号パターンの統合された表示マトリックスが、DNAメチル化決定の精度を大幅に改善するであろうことを更に実証した。特に、IPM-RNNは、それらのアプローチの中で最高の性能をもたらした。

【0115】

一部の実施形態では、IPMについて、塩基修飾分析にかけられる塩基を取り巻くDNAストレッチの長さは、対称又は非対称であり得る。例えば、その塩基のX-nt上流及びY-nt下流を、塩基修飾分析に使用することができる。Xは、0、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、31、32、33、34、35、36、37、38、39、40、41、42、43、44、45、46、47、48、49、50、100、150、200、300、400、500、1000、2000、4000、5000、及び10000を含み得るが、これらに限定されない。Yは、0、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、31、32、33、34、35、36、37、38、39、40、41、42、43、44、45、46、47、48、49、50、100、150、200、300、400、500、1000、2000、4000、5000、及び10000を

含み得るが、これらに限定されない。X及びYは、同じであっても異なってもよい。

【0116】

一部の実施形態では、核酸における塩基修飾は、ウイルス、細菌、植物、真菌、線虫、昆虫、及び脊椎動物（例えば、ヒト）などを含む異なる生物にわたって、本開示における実施形態に従って分析されるであろう。最も一般的な塩基修飾は、異なる位置における異なるDNA塩基へのメチル基の付加、いわゆるメチル化である。メチル化は、5mC（5-メチルシトシン）、4mC（N4-メチルシトシン）、5hmC（5-ヒドロキシメチルシトシン）、5fC（5-ホルミルシトシン）、5caC（5-カルボキシルシトシン）、1mA（N1-メチルアデニン）、3mA（N3-メチルアデニン）、6mA（N6-メチルアデニン）、7mA（N7-メチルアデニン）、3mC（N3-メチルシトシン）、2mG（N2-メチルグアニン）、6mG（O6-メチルグアニン）、7mG（N7-メチルグアニン）、3mT（N3-メチルチミン）、及び4mT（O4-メチルチミン）などのシトシン、アデニン、チミン、及びグアニンで見出されている。

10

【0117】

一部の実施形態では、電流信号パターンの統合された表示マトリックスは、線形回帰、ロジスティック回帰、深層再帰型ニューラルネットワーク（例えば、長短期間メモリ、LSTM）、ベイズ分類器、隠れマルコフモデル（HMM）、線形判別分析（LDA）、k平均クラスタリング、ノイズを伴う用途の密度ベースの空間クラスタリング（DBSCAN）、ランダムフォレストアルゴリズム、及びサポートベクトルマシン（SVM）を含むが、これらに限定されない、異なる統計及び／又は数学的モデルによって分析され得る。なお別の実施形態では、自然言語処理は、塩基修飾分析のための電気信号分析に適用されるであろう。

20

【0118】

一部の実施形態では、例えば、タンパク質工学技術によるタンパク質α溶血素及びその変異、プログラムされた細菌によって産生される孔タンパク質、合成材料から製造された固体状態のナノポア、グラフェンなどの生物学的ナノポアを含むが、これらに限定されない、異なるタイプのナノポアを使用することができる。

【0119】

実施形態では、これらの方法を使用して、ヒト参照ゲノム（hg19）などの参照ゲノム、例えば、長鎖散在核要素（LINE）反復を参照してガイドRNAを設計することによって、相同配列を共有する多数の長いDNA分子を標的にすることができる。一例では、そのような分析は、胎児の異数性の検出のために、妊婦の母体血漿中の循環無細胞DNAの分析に使用することができる（Kindel et al., PLOS One 2012; 7(7): e41162）。実施形態では、非活性型又は「死んだ」Cas9（dCas9）及びそれに関連する一本鎖ガイドRNA（sgRNA）が、二本鎖DNA分子を切断することなく標的の長いDNAを濃縮するために使用され得る。例えば、sgRNAの3'末端は、余分な普遍的な短い配列を有するように設計され得る。その普遍的な短い配列に相補的なビオチン化一本鎖オリゴヌクレオチドを使用して、dCas9によって結合されたそれらの標的の長いDNA分子を捕捉することができる。別の実施形態では、ビオチン化dCas9タンパク質若しくはsgRNA、又は両方を使用して、濃縮を容易にすることができ

30

40

【0120】

実施形態では、化学的、物理的、酵素的、ゲルベース、及び磁気ビーズベースの方法、又はそのようなアプローチ以上を組み合わせた方法を含むがこれらに限定されないアプローチを使用して、目的の1つ以上の特定のゲノム領域に限定することなく、長いDNA断片を濃縮するためにサイズ選択を実施し得る。

【0121】

IV. 例示的な方法

このセクションでは、塩基修飾を検出するために機械学習モデルを使用する例示的な方法、及び塩基修飾の検出のために機械学習モデルを訓練する例示的な方法を示す。

50

【0122】

A. 修飾の検出

図12は、核酸分子におけるヌクレオチドの修飾を検出することと関連付けられる例示的なプロセス1200のフローチャートである。修飾は、本明細書において説明される任意のメチル化又は任意の酸化を含み得る。酸化は、8-オキソグアニンであり得る。いくつかの実装態様では、図12の1つ以上のプロセスブロックは、システム（例えば、測定システム1400）によって実施され得る。いくつかの実装態様では、図12の1つ以上のプロセスブロックは、システムとは別個の、又はシステムを含む、別のデバイス又はデバイス群によって実施され得る。追加的に、又は代替的に、図12の1つ以上のプロセスブロックは、検出器1420、論理システム1430、ローカルメモリ1435、外部メモリ1440、記憶デバイス1445、及び／又はプロセッサ1450などの測定システム1400のうちの1つ以上の構成要素によって実施され得る。

10

【0123】

ブロック1210では、入力データ構造が受信される。入力データ構造は、試料核酸分子で配列決定されたヌクレオチドのウィンドウに対応し得る。試料核酸分子は、ヌクレオチドに対応する電気信号を測定することによって、配列決定される。電気信号は、電流、電圧、抵抗、インダクタンス、キャパシタンス、又はインピーダンスであり得る。配列決定は、ナノポアを使用することによるものであり得る。プロセス1200は、ナノポアを使用する試料核酸の配列決定を更に含み得る。ナノポアは、本明細書において説明される任意のナノポアであり得る。

20

【0124】

入力データ構造は、いくつかの特性についての値を含み得る。特性は、ウィンドウ内の各ヌクレオチドについて、ヌクレオチドの同一性、それぞれのウィンドウ内の標的位置に対するヌクレオチドの位置、及びヌクレオチドに対応する電気信号のセグメントの第1のセグメント統計値を含むベクトルを含み得る。特性は、ウィンドウに等しいか又はそれより大きい核酸分子の領域内の電気信号の第1の領域統計値を含み得る。例えば、入力データ構造は、統合された表示マトリックス [IPM] を含み得る。

【0125】

ヌクレオチドの同一性は、塩基（例えば、A、T、C、又はG）であり得る。塩基は、ナノポア配列決定を用いるベースコール技術を通して決定され得る。ベースコール技術は、電気信号のセグメントをヌクレオチドと関連付け得る。ヌクレオチドの位置は、標的位置に対するヌクレオチドの距離であり得る。例えば、ヌクレオチドが標的位置からある方向に1ヌクレオチド離れている場合、位置は+1であり得、ヌクレオチドが標的位置から反対方向に1ヌクレオチド離れている場合、位置は-1であり得る。

30

【0126】

第1のセグメント統計値は、ヌクレオチドに対応する電気信号のセグメントの平均を表し得る。一部の実施形態では、第1のセグメント統計値は、ヌクレオチドに対応する電気信号のセグメントの電気信号の変動（例えば、標準偏差）を表し得る。実施形態では、第1のセグメント統計値は、ヌクレオチドに対応する電気信号のセグメントの平均の正規化された値を表し得る。正規化は、第1のセグメント統計値が特定の範囲（例えば、0～1の範囲）にあるように、再スケーリングすることを含み得る。正規化は、ヌクレオチド鎖の一部又は全てについて、中央値、平均値、及び／又は偏差を使用することを含み得る。正規化は、zスコア（例えば、X5）を含む、本明細書において説明される任意のものであり得る。

40

【0127】

ベクトルは、ヌクレオチドに対応する電気信号のセグメントの変動を表す第2のセグメント統計値を含み得る。ベクトルは、第1のセグメント統計値の正規化された値を表す第3のセグメント統計値を含み得る。ベクトルは、本明細書において説明される変数X1、X2、及びX5の任意の組み合わせを含み得る。

【0128】

50

第1の領域統計値は、領域内の電気信号の平均又は中央値を表し得る。例えば、第1の領域統計値は、X3であり得る。実施形態では、第1の領域統計値は、領域内の電気信号の平均又は中央値からの電気信号の変動の絶対値の中央値又は平均を表し得る。変動は、標準偏差であり得る。例えば、第1の領域統計値は、X4であり得る。一部の実施形態では、第1の領域統計値は、任意選択的であり得る。

【0129】

入力データ構造は、領域内の電気信号の平均又は中央値からの電気信号の変動の絶対値の中央値又は平均を表す第2の領域統計値を更に含み得る。例えば、第2の領域統計値は、X4であり得る。

【0130】

第1の領域統計値は、ウィンドウ内の異なるヌクレオチドについて同じ値であり得る。第2の領域統計値は、ウィンドウ内の異なるヌクレオチドについて同じ値であり得る。結果として、第1の領域統計値及び第2の領域統計値は、第1のセグメント統計値及び／又は第2のセグメント統計値を有するベクトルから分離しているとみなされ得る。あるいは、値がヌクレオチドにわたって同じであっても、ベクトルはまた、第1の領域統計値を含み得、かつ／又は第2の領域統計値は、各ヌクレオチドについてのベクトルに含まれ得る。領域統計値を反復するアプローチは、IPM524及びIPM624に例解された。

【0131】

領域は、試料核酸分子の1つの鎖上にあり得る。一部の実施形態では、領域は、試料核酸分子の2つの鎖上にあり得る。ウィンドウは、試料核酸分子の2つの鎖上のヌクレオチドを含み得る。領域は、試料核酸分子であり得る。領域は、少なくとも5、10、15、20、25、30、50、100、200、300、400、500、1k、5k、10k、50k、又は1Mヌクレオチドを含み得る。一部の実施形態では、領域は、50、100、200、300、400、500、1k、5k、10k、50k、又は1Mヌクレオチド未満であり得る。領域は、標的位置でのヌクレオチドの周辺を中心としてもよい。

【0132】

ヌクレオチドのウィンドウは、標的位置でのヌクレオチドの周辺を中心としてもよい。一部の実施形態では、ウィンドウは、標的位置でのヌクレオチドの周辺を中心としない場合がある。ウィンドウは、標的位置でのヌクレオチドからX-nt上流及びY-nt下流を含み得る。Xは、0、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、31、32、33、34、35、36、37、38、39、40、41、42、43、44、45、46、47、48、49、50、100、150、200、300、400、500、1000、2000、4000、5000、及び10000を含み得るが、これらに限定されない。Yは、0、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、31、32、33、34、35、36、37、38、39、40、41、42、43、44、45、46、47、48、49、50、100、150、200、300、400、500、1000、2000、4000、5000、及び10000を含み得るが、これらに限定されない。ウィンドウ内のヌクレオチドの最小数は、2、3、4、5、6、7、8、9、10、20、30、40、50、100、200、又は標的位置の上流及び下流のヌクレオチドの数のうちのいずれかの合計よりも1つ多くてもよい。ウィンドウは、図5で示され、説明されるウィンドウと同様であり得る。

【0133】

ウィンドウは、図6で説明される技術と同様に、核酸分子の2つの鎖を含み得る。

【0134】

ブロック1220において、入力データ構造が、モデルに入力される。モデルは、第1の複数の第1のデータ構造を受信することによって訓練される。第1の複数のデータ構造の各第1のデータ構造は、複数の第1の核酸分子のそれぞれの核酸分子において配列決定

10

20

30

40

50

されたヌクレオチドのそれぞれのウィンドウに対応する。第1の核酸分子の各々は、ヌクレオチドに対応する電気信号を測定することによって、配列決定される。修飾は、各第1の核酸分子の各ウィンドウにおける標的位置でのヌクレオチドの既知の第1の状態を有する。各第1のデータ構造は、入力データ構造と同じ特性に対する値を含む。モデルは、本明細書において説明される任意の機械学習モデルであり得る。

【0135】

モデルは、複数の第1の訓練試料を記憶することによって更に訓練される。各第1の訓練試料は、第1の複数の第1のデータ構造のうちの1つと、標的位置でのヌクレオチドの第1の状態を示す第1のラベルと、を含む。加えて、モデルは、第1の複数の第1のデータ構造がモデルに入力されたとき、複数の第1の訓練試料を使用して、第1のラベルの対応するラベルに一致するか又は一致しないモデルの出力に基づいて、モデルのパラメータを最適化することによって訓練される。モデルの出力は、それぞれのウィンドウにおける標的位置でのヌクレオチドが修飾を有するかどうかを指定する。訓練は、図13で後に説明されるようにしてもよい。

【0136】

ブロック1230において、入力データ構造におけるウィンドウ内の標的位置でのヌクレオチドに修飾が存在するかどうかをモデルを使用して、修飾を決定する。

【0137】

修飾状態は、更なる分析に使用され得る。妊婦から取得された試料において、本開示における実施形態を使用して、メチル化状態に基づいて血漿DNA分子の胎児又は母体起源を決定することができる。母体又は胎児起源は、参照値よりも高い又は低いメチル化レベルを有するゲノム領域によって決定され得る。実施形態では、妊婦から取得された試料は、無細胞、例えば、血漿又は血清であり得る。一部の実施形態では、試料核酸分子は、所定のゲノム領域に整列するものとして特定され得る。所定のゲノム領域は、胎児又は母体ゲノムにおいて高メチル化又は低メチル化であることが既知であり得る。本方法は、標的位置でのヌクレオチドの修飾状態、及び任意選択的に、試料核酸分子の1つ以上の他のヌクレオチドの修飾状態を使用して、試料核酸が胎児又は母体起源のものであると決定することを含み得る。

【0138】

試料核酸分子が胎児又は母体起源のものであるかどうかを決定することは、1つ以上のヌクレオチドのメチル化状態を使用して、試料核酸分子のメチル化レベルを決定することを含み得る。試料核酸分子のメチル化レベルは、参照値と比較され得る。参照値は、1つ以上の母体核酸分子のメチル化レベルから決定され得る。参照値に対して試料核酸分子のメチル化レベルを比較することは、試料核酸分子のメチル化レベルが参照値よりも低いことを決定することを含み得る。試料核酸分子が胎児又は母体起源のものであるかどうかを決定することは、比較を使用して試料核酸分子が胎児起源のものであると決定することを含み得る。

【0139】

一部の実施形態では、試料核酸分子は、複数の試料核酸分子のうちの1つの試料核酸分子であり得る。本方法は、メチル化状態を使用して、複数の試料核酸分子の各々が胎児又は母体起源であるかどうかを決定することを更に含み得る。胎児画分は、複数の試料核酸分子の胎児又は母体起源の決定を使用して決定され得る。

【0140】

一部の実施形態では、修飾状態は、領域にコピー数異常が存在するかどうかを決定するために使用され得る。修飾は、メチル化であり得る。試料核酸分子は、無細胞であり得、胎児を妊娠している女性対象の生体試料から取得され得る。試料核酸分子は、複数の試料核酸分子のうちの1つの試料核酸分子であり得る。本方法は、複数の試料核酸分子が胎児ゲノムの領域に整列していることを特定することを更に含み得る。複数の試料核酸分子の各試料核酸分子の1つ以上のヌクレオチドの修飾状態が決定され得る。領域のメチル化レベルは、複数の試料核酸分子の各試料核酸分子についての1つ以上のヌクレオチドのメチ

10

20

30

40

50

ル化状態を使用して決定され得る。本方法は、メチル化レベルを使用して、胎児ゲノムの領域にコピー数異常が存在するかどうかを決定することを更に含み得る。領域は、染色体であり得、本方法は、コピー数異常が存在することを決定し、胎児が染色体異数性を有することを決定することを更に含み得る。

【0141】

修飾は、1つ以上のヌクレオチドに存在していると決定され得る。障害の分類は、1つ以上のヌクレオチドの修飾の存在を使用して、決定することができる。障害の分類は、修飾の数を使用することを含み得る。修飾の数は、閾値と比較され得る。代替的又は追加的に、分類は、1つ以上の修飾の位置を含み得る。1つ以上の修飾の位置は、核酸分子の配列リードを参照ゲノムに整列することによって、決定することができる。障害と相関していることが知られている特定の位置に修飾があることが示された場合、障害を決定することができる。例えば、メチル化部位のパターンを、障害の参照パターンと比較することができ、その比較に基づいて、障害を決定することができる。参照パターンとの一致又は参照パターンとの実質的な一致（例えば、80%、90%、又は95%以上）は、障害又は障害の可能性が高いことを示している場合がある。障害は、任意の妊娠関連障害（例えば、子癲前症、子宮内胎児発育遅延、侵襲性胎盤形成、及び早産）であり得る。

10

【0142】

統計的に有意な数の核酸分子を分析して、1人以上の妊娠中の対象における障害、組織起源、又は臨床関連DNA画分についての正確な決定を提供することができる。一部の実施形態では、少なくとも1,000個の核酸分子が分析される。他の実施形態では、少なくとも10,000又は50,000又は100,000又は500,000又は1,000,000又は5,000,000、又はそれ以上の核酸分子を分析することができる。更なる例として、少なくとも10,000又は50,000又は100,000又は500,000又は1,000,000の配列リードを生成することができる。

20

【0143】

本方法は、障害の分類は、対象が障害を有すると決定することを含み得る。分類は、修飾の数及び／又は修飾の部位を使用して、障害のレベルを含み得る。

【0144】

1つ以上のヌクレオチドにおける修飾の存在を使用して、胎児のDNA画分、胎児のメチル化プロファイル、母体のメチル化プロファイル、インプリンティング遺伝子領域の存在が決定され得る。

30

【0145】

プロセス1200は、以下に説明される及び／又は本明細書における他の箇所で説明される1つ以上の他のプロセスに関連して、任意の単一の実装態様又は実装態様の任意の組み合わせなどの追加の実施態様を含み得る。

【0146】

図12は、プロセス1200の例示的なブロックを示すが、一部の实装態様では、プロセス1200は、図12に示されているブロックよりも、追加のブロック、より少ないブロック、異なるブロック、又は異なるように配置されたブロックを含み得る。追加的に、又は代替的に、プロセス1200のブロックのうちの2つ以上が、並列に実施され得る。

40

【0147】

B. モデル訓練

図13は、核酸分子におけるヌクレオチドの修飾を検出する例示的な方法1300を示す。例示的な方法1300は、修飾を検出するためにモデルを訓練する方法であり得る。修飾には、メチル化が含まれ得る。メチル化は、本明細書に記載の任意のメチル化を含み得る。修飾は、メチル化及び非メチル化などの個別の状態を有することができ、メチル化の種類を指定する可能性がある。したがって、ヌクレオチドには、3つ以上の状態（分類）が存在してもよい。図13における訓練は、図12の方法1200とともに使用され得る。

50

【0148】

ブロック1310では、複数の第1のデータ構造が受信される。データ構造の様々な例が、本明細書、例えば、図5及び6に説明されている。第1の複数の第1のデータ構造の各第1のデータ構造は、複数の第1の核酸分子のそれぞれの核酸分子において配列決定されたヌクレオチドのそれぞれのウィンドウに対応し得る。第1の複数のデータ構造に関連する各ウィンドウは、4つ以上の連続したヌクレオチドを含んでもよく、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19、20、21又はそれ以上の連続したヌクレオチドが含まれる。各ウィンドウには、同じ数の連続したヌクレオチドが含まれ得る。ウィンドウは、重複している場合がある。各ウィンドウは、第1の核酸分子の第1の鎖上のヌクレオチド及び第1の核酸分子の第2の鎖上のヌクレオチドを含み得る。第1のデータ構造はまた、ウィンドウ内の各ヌクレオチドについて、鎖特性の値を含み得る。鎖特性は、存在するヌクレオチドか、又は第1の鎖若しくは第2の鎖のいずれかを示し得る。ウィンドウは、第1の鎖の対応する位置のヌクレオチドに相補的ではない第2の鎖のヌクレオチドを含み得る。一部の実施形態では、第2の鎖上の全てのヌクレオチドは、第1の鎖上のヌクレオチドに相補的である。一部の実施形態では、各ウィンドウは、第1の核酸分子の1つの鎖のみのヌクレオチドを含み得る。

10

【0149】

第1の複数の第1のデータ構造は、5,000~10,000、10,000~50,000、50,000~100,000、100,000~200,000、200,000~500,000、500,000~1,000,000、又は1,000,000以上の第1のデータ構造を含み得る。複数の第1の核酸分子は、少なくとも1,000、10,000、50,000、100,000、500,000、1,000,000、5,000,000、又はそれ以上の核酸分子を含み得る。更なる例として、少なくとも10,000又は50,000又は100,000又は500,000又は1,000,000又は5,000,000の配列リードを生成することができる。

20

【0150】

第1の核酸分子の各々は、ヌクレオチドに対応する電気信号を測定することによって、配列決定される。電気信号は、ナノポア配列決定からのものであり得る。

【0151】

修飾は、各第1の核酸分子の各ウィンドウにおける標的位置でのヌクレオチドの既知の第1の状態を有する。第1の状態は、修飾がヌクレオチドに存在しないか、又は修飾がヌクレオチドに存在するかであり得る。修飾は、第1の核酸分子に存在しないことが既知の場合があり、又は第1の核酸分子は、修飾が存在しないように処理を受ける場合がある。修飾は、第1の核酸分子に存在することが既知の場合があり、又は第1の核酸分子は、修飾が存在するように処理を受ける場合がある。第1の状態が、修飾が存在しない状態である場合、修飾は、各第1の核酸分子の各ウィンドウに存在せず、標的位置にだけ存在する場合がある。既知の第1の状態は、第1のデータ構造の第1の箇所のメチル化状態と、第1のデータ構造の第2の箇所の非メチル化状態と、を含み得る。メチル化のための既知の第1の状態は、バイサルファイト配列決定を使用する技術を通して、又は単一分子リアルタイム配列決定からの光信号を使用して決定され得る。

30

40

【0152】

標的位置は、それぞれのウィンドウの中心であり得る。偶数のヌクレオチドにまたがるウィンドウの場合、標的位置は、ウィンドウの中心のすぐ上流又はすぐ下流の位置であり得る。一部の実施形態では、標的位置は、第1の位置又は最後の位置を含む、それぞれのウィンドウの他の任意の位置にあってもよい。例えば、ウィンドウが、一方の鎖のnヌクレオチド、1番目の位置からn番目の位置（上流又は下流のいずれか）にまたがる場合、標的位置は、1番目の位置からn番目の位置までの任意の位置にあってもよい。

【0153】

各第1のデータ構造には、ウィンドウ内の特性についての値が含まれる。特性は、ブロック1210で説明される特性のうちのいずれかであり得る。

50

【0154】

ブロック1320では、複数の第1の訓練試料が記憶される。各第1の訓練試料は、第1の複数の第1のデータ構造のうちの1つと、標的位置でのヌクレオチドの修飾についての第1の状態を示す第1のラベルと、を含む。

【0155】

ブロック1330では、第2の複数の第2のデータ構造が受信される。ブロック1330は、任意選択的である。第2の複数の第2のデータ構造の各第2のデータ構造は、複数の第2の核酸分子のそれぞれの核酸分子において配列決定されたヌクレオチドのそれぞれのウィンドウに対応する。第2の複数の核酸分子は、複数の第1の核酸分子と同じであっても異なってもよい。修飾は、各第2の核酸分子の各ウィンドウ内の標的位置でのヌクレオチドの既知の第2の状態を有する。第2の状態は、第1の状態とは異なる状態である。例えば、第1の状態に修飾が存在する場合、第2の状態には修飾が存在せず、その逆も同様である。各第2のデータ構造は、第1の複数の第1のデータ構造と同じ特性についての値を含む。

10

【0156】

ブロック1340では、複数の第2の訓練試料が記憶される。ブロック1340は、任意選択的である。各第2の訓練試料は、第2の複数の第2のデータ構造のうちの1つと、標的位置でのヌクレオチドの修飾についての第2の状態を示す第2のラベルと、を含む。

【0157】

ブロック1350では、モデルは、複数の第1の訓練試料、及び任意選択的に複数の第2の訓練試料を使用して訓練される。訓練は、第1の複数の第1のデータ構造及び任意選択的に第2の複数の第2のデータ構造がモデルに入力されたとき、第1のラベル及び任意選択的に第2のラベルの対応するラベルに一致するか又は一致しないモデルの出力に基づいて、モデルのパラメータを最適化することによって実施される。モデルの出力は、それぞれのウィンドウにおける標的位置でのヌクレオチドが修飾を有するかどうかを指定する。モデルが外れ値を第1の状態とは異なる状態であると特定する可能性があるため、この方法は、複数の第1の訓練試料のみを含み得る。モデルは、機械学習モデルとも呼ばれる、統計モデルであり得る。

20

【0158】

一部の実施形態では、モデルの出力は、複数の状態の各々における確率を含み得る。確率が最も高い状態を、その状態とみなすことができる。

30

【0159】

モデルには、畳み込みニューラルネットワーク(CNN)が含まれ得る。CNNは、第1の複数のデータ構造及び任意選択的に第2の複数のデータ構造をフィルタリングするように構成された畳み込みフィルターのセットを含み得る。フィルターは、本明細書に記載の任意のフィルターであり得る。各層のフィルターの数は、10~20、20~30、30~40、40~50、50~60、60~70、70~80、80~90、90~100、100~150、150~200、又はそれ以上であり得る。フィルターのカーネルサイズは、2、3、4、5、6、7、8、9、10、11、12、13、14、15、15~20、20~30、30~40、又はそれ以上であり得る。CNNは、フィルタリングされた第1の複数のデータ構造、及び任意選択的にフィルタリングされた第2の複数のデータ構造を受信するように構成された入力層を含み得る。CNNはまた、複数のノードを含む複数の隠れ層を含み得る。入力層には、複数の隠れ層の第1の層が結合した。CNNは、複数の隠れ層の最後の層に結合され、出力データ構造を出力するように構成された出力層を更に含み得る。出力データ構造は、特性を含み得る。

40

【0160】

モデルは、再帰型ニューラルネットワーク(RNN)を含み得る。RNNモデルは、測定ウィンドウ内の複数のヌクレオチドと関連付けられるいくつかの長短期間メモリ(LSTM)ユニットを含む。LSTMユニットの数は、測定ウィンドウ内のヌクレオチドの数と等しくてもよい。一部の実施形態では、LSTMユニットの数は、測定ウィンドウ内の

50

ヌクレオチドの数よりも少なくてもよい。LSTMユニットの数は、これらに限定されないが、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、30、40、50、100、200、300、400、500、1,000、2,000、3,000、4,000、5,000、10,000、50,000などであり得る。1つのLSTMユニットは、多くのラウンドの線形又は非線形変換に供されるであろう電流信号特徴に関連する情報を、次のLSTMユニットに伝送することができる。LSTMユニットにわたるそのような情報伝送は、一般に、連続的な様態（例えば、時間ステップに従って）で編成される。LSTMユニットにわたるそのような情報伝送は、双方向であり得る（すなわち、時間的順序及び予約された時間的順序を含む）。各LSTMユニットは、忘却ゲート、入力ゲート、セル状態、及び出力ゲートなどのプログラム可能な操作を含む。それらの操作を通して、1つのLSTMは、以前の時間ステップから来る電流信号情報が記憶されるべきであるかどうか、又は無関係であり、忘れられることができるかどうか（忘却ゲート）を決定することができる。1つのLSTMユニットは、そのようなユニット（入力ゲート）への入力から新しい情報を学習しようとする。ユニットは、現在の時間ステップから次の時間ステップ（出力ゲート）に更新された情報を渡す。本明細書におけるセル状態は、全ての時間ステップとともに情報を運ぶ。いくつかの層のLSTMユニットが使用されてもよい。LSTM層の数は、1、2、3、4、5、6、7、8、9、10、15、20、30などであり得る。層間の全結合が使用されてもよい。シグモイド関数は、一般に、入力ゲート、出力ゲート、及び忘却ゲートについてのゲート関数として使用される。シグモイド関数の出力値は、0～1であり得、ゲート全体の情報の流れがないか、又は完全に流れているかのいずれかを決定する。双曲線正接活性化関数（Tanhとも称される）は、出力ゲートからの情報値を処理して、次のLSTMユニットに渡され得る-1と1との間の値で新しい情報を形成する出力活性化関数として使用され得る。一部の実施形態では、バイナリステップ関数、線形活性化関数、シグモイド関数、正規化線形ユニットなどを含むが、これらに限定されない、他の活性化関数を使用し得る。LSTMの最終層によって生成された値は、各ニューロンが完全に接続されている出力層（すなわち、特定の数のニューロンを有する高密度層）上に渡され得る。高密度層中のニューロンの数は、限定されないが、2、3、4、5、6、7、8、9、10、20、30、40、50、100、200、300、400、500、1000、2000などであり得る。1、2、3、4、5、6、7、8、9、10、20、30、40、50、100、5000、10000などを含むが、これらに限定されない、いくつかの高密度層を使用することができる。出力層は、例えば、メチル化状態を分類するために使用され得るシグモイド活性化関数又はSoftmax活性化関数に基づいて、メチル化スコアを出力し得る。例えば、メチル化スコアが0.5より大きい場合、塩基はメチル化されていると決定される。さもなければ、塩基は非メチル化であると決定される。一部の実施形態では、メチル化状態を分類するために使用される閾値は、限定されないが、少なくとも0.1、0.2、0.3、0.4、0.6、0.7、0.8、0.9などであり得る。一部の実施形態では、モデルにおけるニューロンのいくつかは、過剰適合の問題を最小限に抑えるためにドロップアウトされ得る。ドロップアウトされたニューロンのパーセンテージは、限定されないが、1%、5%、10%、15%、20%、25%、30%、40%、50%、60%、70%などであり得、これらは、異なる層によって異なる場合がある。

【0161】

モデルには、教師あり学習モデルが含まれ得る。教師あり学習モデルは、異なるアプローチ及びアルゴリズムが含まれ得、分析的学習、人工ニューラルネットワーク、誤差逆伝播、ブースティング（メタアルゴリズム）、ベイズ統計、事例ベース推論、決定木学習、帰納論理プログラミング、ガウス過程回帰、遺伝的プログラミング、データ処理のグループ法、カーネル推定器、学習オートマトン、学習分類器システム、最小メッセージ長（決定木、決定グラフなど）、多重線形部分空間学習、ナイーブベイズ分類器、最大エントロピー分類器、条件付き確率場、最近傍アルゴリズム、確率的で近似的に正しい学習（PA

10

20

30

40

50

C) 学習、リップルダウンルール、知識獲得法論、シンボリック機械学習アルゴリズム、サブシンボリック機械学習アルゴリズム、サポートベクトルマシン、最小複雑性マシン (MCM)、ランダムフォレスト、分類器のアンサンブル、通常分類、データ事前処理、不均衡データセットの処理、統計的関係学習、又は Proaftn、多基準分類アルゴリズムが含まれる。モデルは、線形回帰、ロジスティック回帰、深層再帰型ニューラルネットワーク (例えば、長期短期メモリ、LSTM)、ベイズ分類器、隠れマルコフモデル (HMM)、線形判別分析 (LDA)、k 平均クラスタリング、ノイズを伴う用途の密度ベースの空間クラスタリング (DBSCAN)、ランダムフォレストアルゴリズム、サポートベクトルマシン (SVM)、又は本明細書において説明される任意のモデルであり得る。

【0162】

機械学習モデルの訓練の一環として、機械学習モデルのパラメータ (重み、閾値など、例えば、ニューラルネットワークの活性化関数に使用することができるもの) を訓練試料 (訓練セット) に基づいて最適化して、標的位置でのヌクレオチドの修飾を分類する際に最適化された精度を提供する。様々な形式の最適化を行うことができ、例えば、誤差逆伝播、経験的リスク最小化、及び構造的リスク最小化などである。試料の検証セット (データ構造とラベル) を使用して、モデルの精度を検証することができる。交差検証は、訓練と検証のために訓練セットの様々な箇所を使用して行うことができる。モデルは、複数のサブモデルを含むことができ、それによって、アンサンブルモデルを提供する。サブモデルは、より弱いモデルであり得るが、組み合わせると、より正確な最終モデルを提供する。

【0163】

V. 例示的なシステム

図 14 は、本発明の実施形態による、測定システム 1400 を例解する。示されたシステムは、試料ホルダ 1410 内の DNA 分子などの試料 1405 を含み、試料 1405 をアッセイ 1408 と接触させて、物理的特徴 1415 の信号を提供することができる。試料ホルダの例は、アッセイのプロブ及び／又はプライマー、あるいはその中をドロップレットが (アッセイを含むドロップレットとともに) 移動するチューブを含む、フローセルであり得る。試料からの物理的特徴 1415 (例えば、蛍光強度、電圧、又は電流) は、検出器 1420 によって検出される。検出器 1420 は、データ信号を構成するデータポイントを得るために、間隔 (例えば、周期的間隔) を空けて測定を行うことができる。一実施形態では、アナログ-デジタル変換器は、検出器からのアナログ信号をデジタル形態へと、複数回、変換する。試料ホルダ 1410 及び検出器 1420 は、アッセイデバイス、例えば、本明細書に記載される実施形態に従って配列決定を実施する配列決定デバイスを形成し得る。データ信号 1425 は、検出器 1420 から論理システム 1430 に送信される。データ信号 1425 は、ローカルメモリ 1435、外部メモリ 1440、又は記憶デバイス 1445 に保存され得る。

【0164】

論理システム 1430 は、コンピュータシステム、ASIC、マイクロプロセッサなどであり得るか、又はそれらを含み得る。それはまた、ディスプレイ (例えば、モニタ、LED ディスプレイなど)、及びユーザ入力デバイス (例えば、マウス、キーボード、ボタンなど) を含み得るか、又はそれらに連結され得る。論理システム 1430 及び他の構成要素は、スタンドアロン若しくはネットワーク接続されたコンピュータシステムの一部であり得るか、又は検出器 1420 及び／又は試料ホルダ 1410 を含むデバイス (例えば、配列決定デバイス) に直接取り付けられ得るか、又は組み込まれ得る。論理システム 1430 はまた、プロセッサ 1450 で実行するソフトウェアを含み得る。論理システム 1430 は、本明細書に記載される方法のいずれかを実施するようにシステム 1400 を制御するための指示を記憶するコンピュータ可読媒体を含み得る。例えば、論理システム 1430 は、配列決定又は他の物理的動作が実施されるように、試料ホルダ 1410 を含むシステムにコマンドを提供し得る。そのような物理的動作は、特定の順序で実施され得、例えば、試薬は、特定の順序で添加及び除去される。そのような物理的動作は、試料を

取得してアッセイを実施するように使用され得るように、例えば、ロボットアームを含む、ロボットシステムによって実施され得る。

【0165】

本明細書で言及されるコンピュータシステムのうちのいずれも、任意の好適な数のサブシステムを利用し得る。このようなサブシステムの例が、コンピュータシステム10の図15に示される。いくつかの実施形態では、コンピュータシステムは、単一のコンピュータ装置を含み、サブシステムは、コンピュータ装置の構成要素であり得る。他の実施形態では、コンピュータシステムは、各々がサブシステムであり、内部構成要素を備える、複数のコンピュータ装置を含み得る。コンピュータシステムは、デスクトップコンピュータ及びラップトップコンピュータ、タブレット、携帯電話、並びにクラウドベースのシステムを含み得る。

10

【0166】

図15に示されるサブシステムは、システムバス75を介して相互接続される。プリンタ74、キーボード78、記憶デバイス79、ディスプレイアダプタ82に結合されたモニタ76（例えば、LEDなどのディスプレイスクリーン）、及び他などの追加のサブシステムが示されている。I/Oコントローラ71に結合する周辺機器及び入力/出力（I/O）デバイスは、入力/出力（I/O）ポート77（例えば、USB、Lightning、Thunderbolt（商標））などの当該技術分野において既知の任意の数の手段によって、コンピュータシステムに接続され得る。例えば、I/Oポート77又は外部インターフェース81（例えば、Ethernet（登録商標）、Wi-Fiなど）を使用して、Internetなどの広域ネットワーク、マウス入力デバイス、又はスキャナに、コンピュータシステム10を接続し得る。システムバス75を介した相互接続は、中央プロセッサ73が、各サブシステムと通信し、システムメモリ72又は記憶デバイス79（例えば、ハードドライブ又は光ディスクなどの固定ディスク）からの複数の命令の実行、及びサブシステム間の情報交換を制御することを可能にする。システムメモリ72及び/又は記憶デバイス79は、コンピュータ可読媒体を具現化し得る。別のサブシステムは、カメラ、マイクロホン、及び加速度計などのデータ収集デバイス85である。本明細書に言及されるデータのうちのいずれも、1つの構成要素から別の構成要素に出力され得、ユーザに対して出力され得る。

20

【0167】

コンピュータシステムは、例えば、外部インターフェース81によって、内部インターフェースによって、又は1つの構成要素から別の構成要素に接続することができる取り外し可能な記憶デバイスを介してともに接続される、複数の同じ構成要素又はサブシステムを含み得る。いくつかの実施形態では、コンピュータシステム、サブシステム、又は装置は、ネットワーク上で通信することができる。そのような例においては、1つのコンピュータは、クライアント及び別のコンピュータをサーバとみなされ得、各々は、同じコンピュータシステムの一部であり得る。クライアント及びサーバは各々、複数のシステム、サブシステム、又は構成要素を含むことができる。

30

【0168】

実施形態の態様は、制御ロジックの形態で、ハードウェア回路（例えば、特定用途向け集積回路若しくはフィールドプログラマブルゲートアレイ）を使用して、及び/又はモジュール式若しくは集積様態で汎用プログラマブルプロセッサを有するコンピュータソフトウェアを使用して、実装することができる。本明細書で使用される場合、プロセッサは、シングルコアプロセッサ、同じ集積チップ上のマルチコアプロセッサ、又は単一の回路基板上の若しくはネットワーク化された複数の処理ユニット、並びに専用のハードウェアを含み得る。本開示及び本明細書に提供される教示に基づいて、当業者は、ハードウェア、及びハードウェアとソフトウェアとの組み合わせを使用して、本発明の実施形態を実装する他の様式及び/又は方法を認識かつ理解するであろう。

40

【0169】

本出願で記載されるソフトウェア構成要素又は機能のうちのいずれも、例えば、Java

50

a（登録商標）、C、C++、C#、Objective-C、Swiftなどの任意の好適なコンピュータ言語、又は、例えば、従来の技術若しくは物体指向の技術を使用するPerl若しくはPythonなどのスクリプト言語を使用する、プロセッサによって実行されるソフトウェアコードとして実装され得る。ソフトウェアコードは、保存及び／又は伝送のためのコンピュータ可読媒体上に一連の命令又はコマンドとして記憶され得る。好適な非一時的コンピュータ可読媒体は、ランダムアクセスメモリ（random access memory、RAM）、読み出し専用メモリ（read only memory、ROM）、磁気媒体（ハードドライブ若しくはフロッピー（登録商標）ディスクなど）、又は光学媒体（コンパクトディスク（compact disk、CD）若しくはDVD（digital versatile disk、デジタル多用途ディスク）など、又はブルーレイディスク、フラッシュメモリなどを含み得る。コンピュータ可読媒体は、そのような記憶デバイス又は伝送デバイスの任意の組み合わせであり得る。

10

【0170】

そのようなプログラムはまた、インターネットを含む様々なプロトコルに従う有線ネットワーク、光ネットワーク、及び／又は無線ネットワークを介した伝送に適合した搬送波信号を使用して、コード化されかつ伝送され得る。したがって、コンピュータ可読媒体は、そのようなプログラムでコード化されたデータ信号を使用して作成され得る。プログラムコードでコード化されたコンピュータ可読媒体は、互換性のあるデバイスでパッケージ化され得るか、又は他のデバイスとは別個に（例えば、インターネットダウンロードを介して）提供され得る。任意のそのようなコンピュータ可読媒体は、単一のコンピュータ製品（例えば、ハードドライブ、CD、若しくはコンピュータシステム全体）上若しくはその内部に存在し得、システム又はネットワーク内の異なるコンピュータ製品上若しくはその内部に存在し得る。コンピュータシステムは、本明細書に記載の結果のうちのいずれかをユーザに提供するための、モニタ、プリンタ、又は他の好適なディスプレイを含み得る。

20

【0171】

本明細書に記載される方法のいずれも、ステップを実施するように構成することができ、1つ以上のプロセッサを含むコンピュータシステムを用いて全体的に又は部分的に実施され得る。したがって、実施形態は、本明細書に記載される方法のうちのいずれかのステップを実施するように構成されたコンピュータシステムを対象とし得、潜在的には異なる構成要素がそれぞれのステップ又はそれぞれのステップの群を実施する。本明細書の方法のステップは、番号付けされたステップとして提示されているが、同時に若しくは異なる時間に、又は異なる順序で実施され得る。更に、これらのステップの部分は、他の方法からの他のステップの部分と使用され得る。また、あるステップの全て又は部分は、任意選択的であり得る。更に、方法のうちのいずれかのステップのいずれも、これらのステップを実施するためのシステムのモジュール、ユニット、回路、又は他の手段で実施され得る。

30

【0172】

特定の実施形態の具体的な詳細は、本発明の実施形態の趣旨及び範囲から逸脱することなく、任意の好適な様式で組み合わせられ得る。しかしながら、本発明の他の実施形態は、各個々の態様、又はこれらの個々の態様の特定の組み合わせに関する特定の実施形態を対象とし得る。

40

【0173】

本開示の例示的な実施形態の上の説明は、例解及び説明目的のために提示されている。網羅的であること、又は本開示に記載された正確な形態に限定することは意図されず、多くの修正及び変更が、上の教示に鑑みて可能である。

【0174】

「a」、「an」、又は「the」の記述は、それとは反対に具体的に示されない限り、「1つ以上」を意味することが意図される。「又は」の使用は、それとは反対に具体的に示されない限り、「排他的な又は」ではなく「包括的な又は」を意味するように意図さ

50

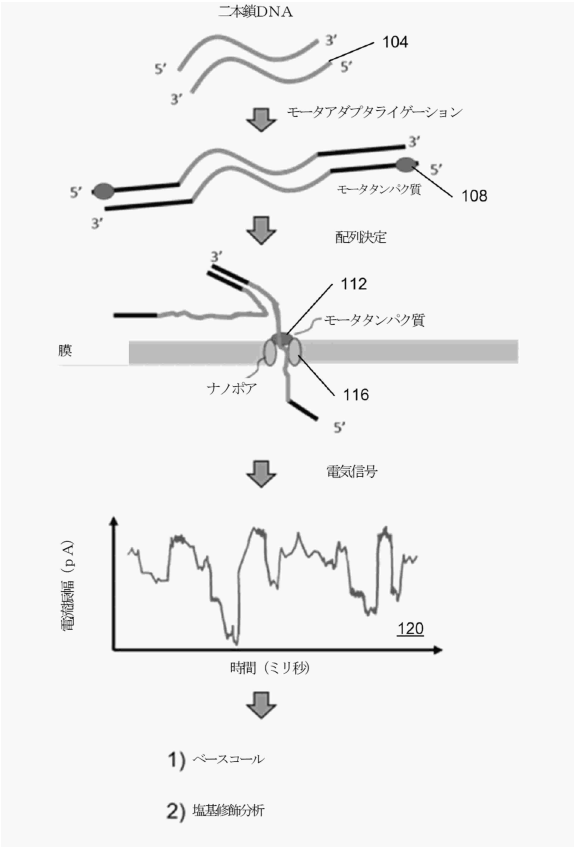
れる。「第1」の構成要素への言及は、第2の構成要素が提供されることを必ずしも必要としない。更に、「第1」又は「第2」の構成要素への言及は、明示的に述べられていない限り、言及される構成要素を特定の場所に限定するものではない。「～に基づいて」という用語は、「少なくとも一部に基づいて」を意味することを意図している。

【0175】

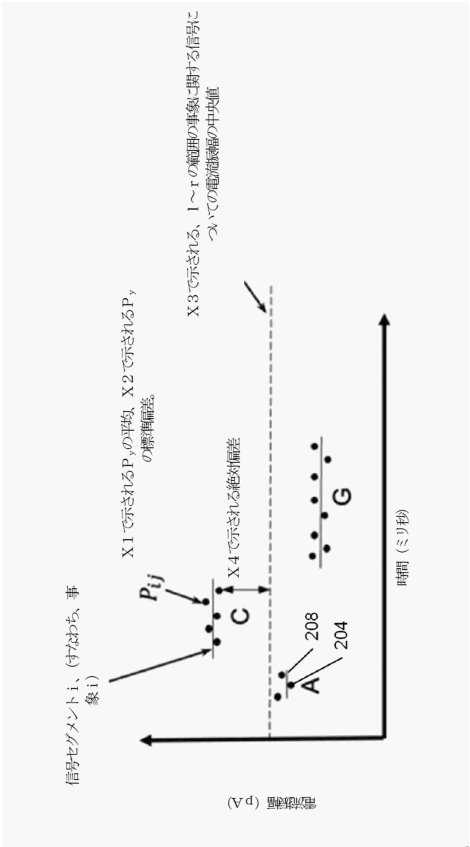
本明細書で言及される全ての特許、特許出願、刊行物、及び明細書は、全ての目的のために参照によりそれらの全体が組み込まれる。いかなるものも、先行技術であるとは認められていない。

【図面】

【図1】



【図2】



10

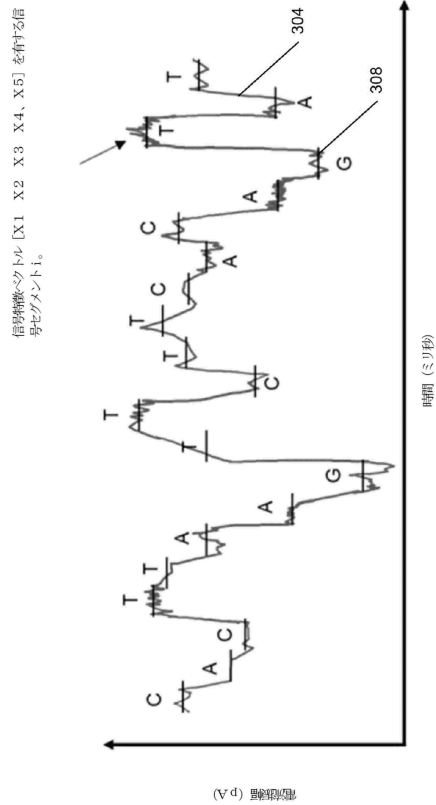
20

30

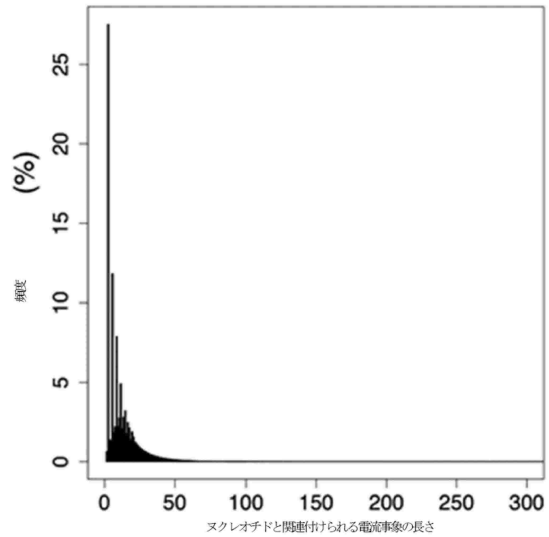
40

50

【図 3】



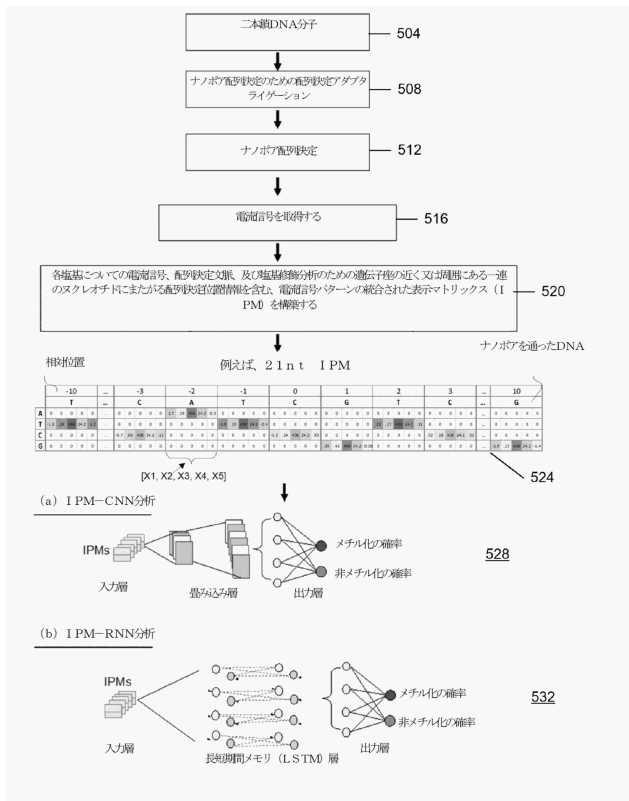
【図 4】



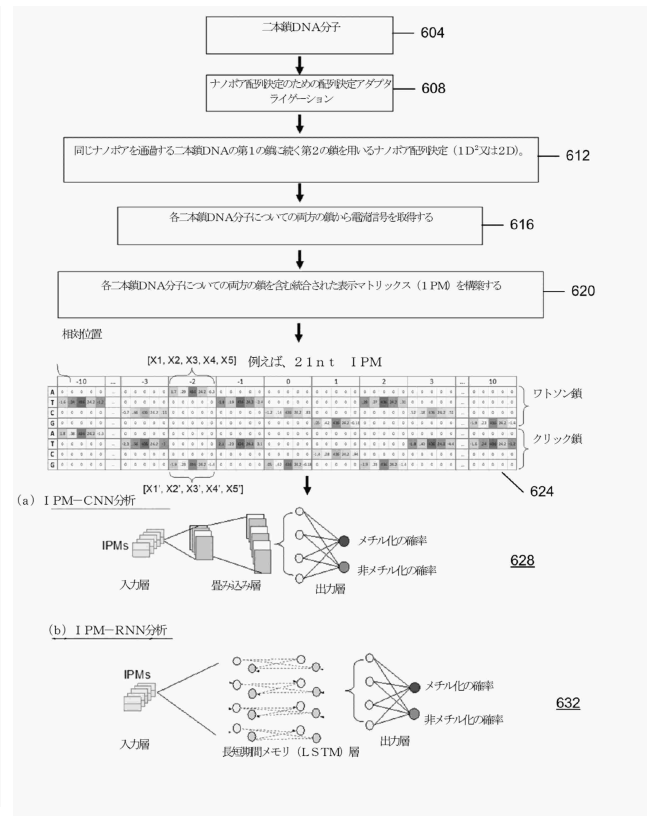
10

20

【図 5】



【図 6】



30

40

50

【 図 7 】

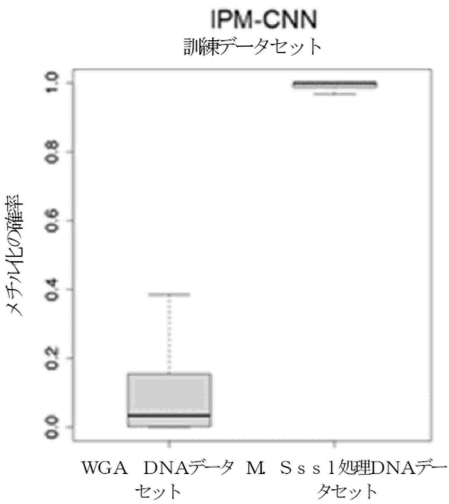
カーネルサイズ	AUC	
	訓練データセット	試験データセット
1x5	0.98	0.96
1x10	0.98	0.96
1x15	0.97	0.97
1x20	0.98	0.96
1x25	0.98	0.96
1x30	0.97	0.94

【 図 8 】

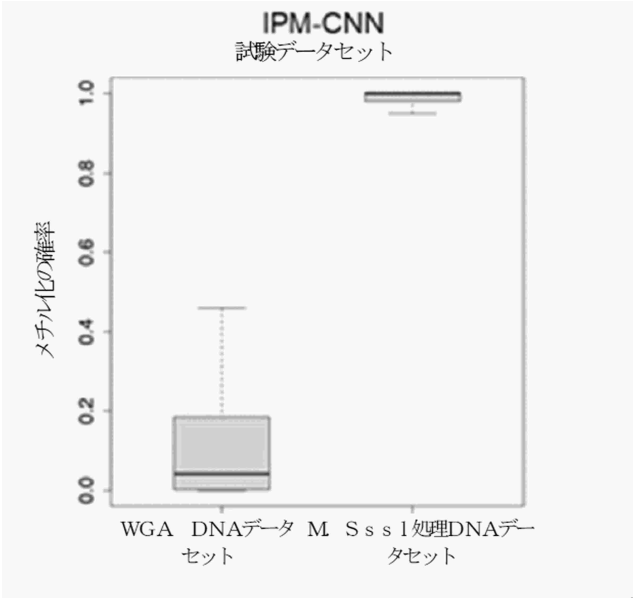
データセット	分子の数 (CpG部位の数)	
	訓練	試験
M S s l 処理DNA	7,989 (38,470)	4,826 (9,716)
WGADNA	8,052 (37,150)	5,041 (11,444)

10

【 図 9 A 】



【 図 9 B 】



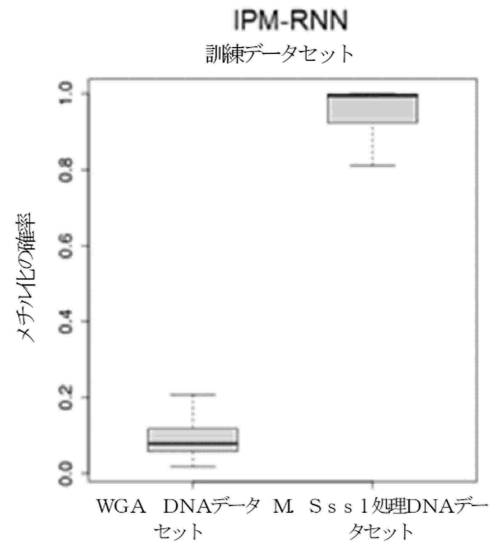
20

30

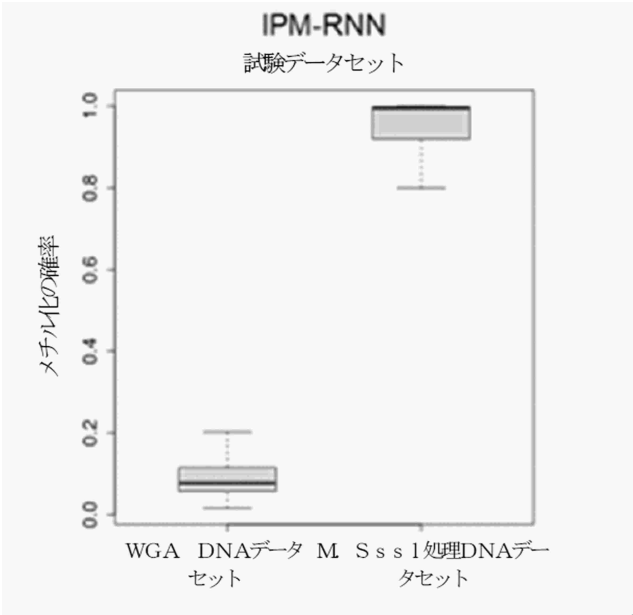
40

50

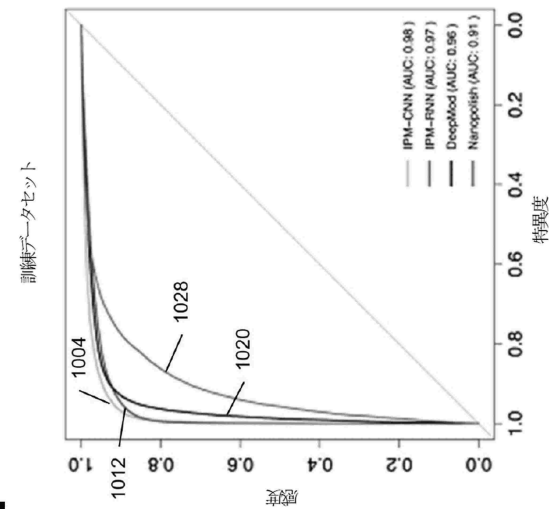
【図 9 C】



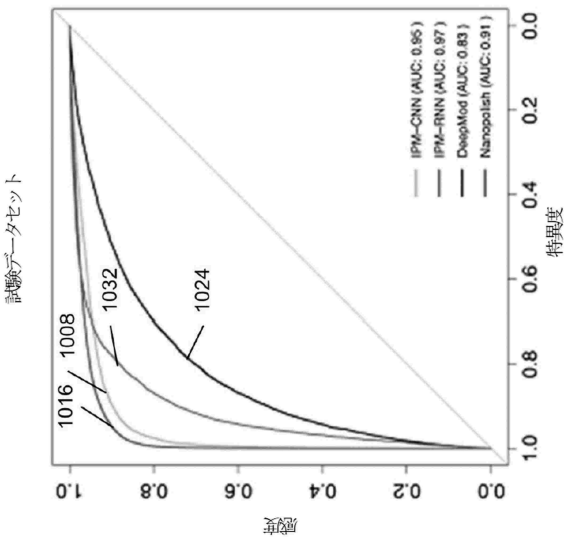
【図 9 D】



【図 10 A】



【図 10 B】



10

20

30

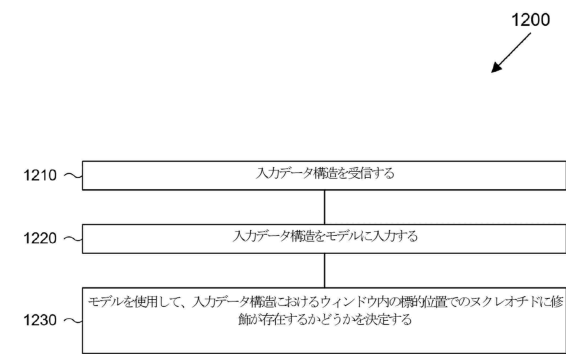
40

50

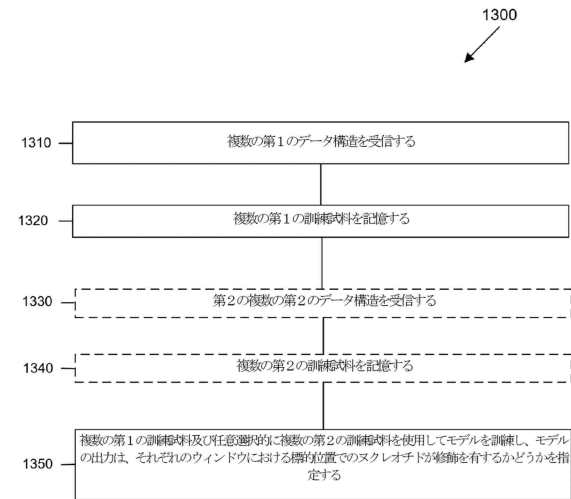
【図 1 1】

	感度	特異度
IPM-CNN	90%	90%
IPM-RNN	93%	
DeepMod	53%	
nanopolish	74%	
IPM-CNN	86%	95%
IPM-RNN	90%	
DeepMod	38%	
nanopolish	55%	
IPM-CNN	70%	99%
IPM-RNN	83%	
DeepMod	13%	
nanopolish	16%	

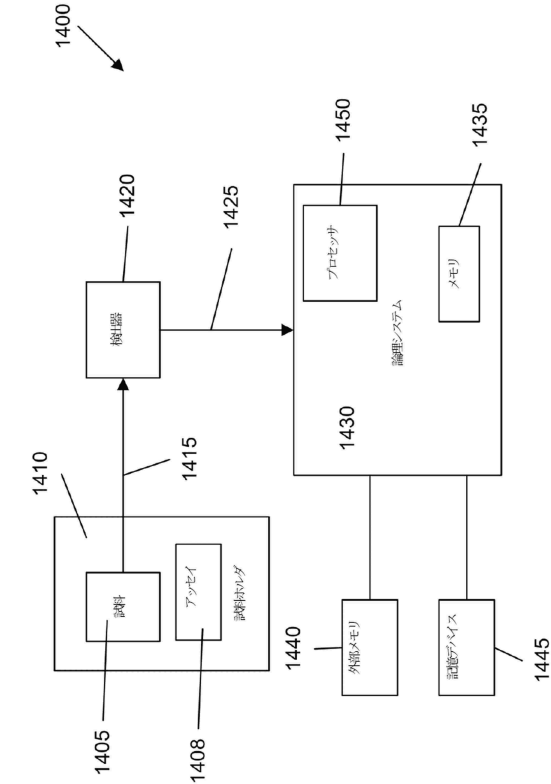
【図 1 2】



【図 1 3】



【図 1 4】



10

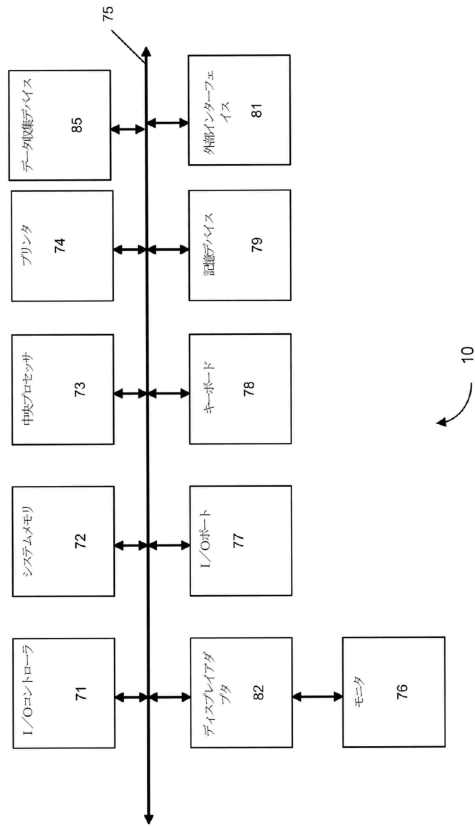
20

30

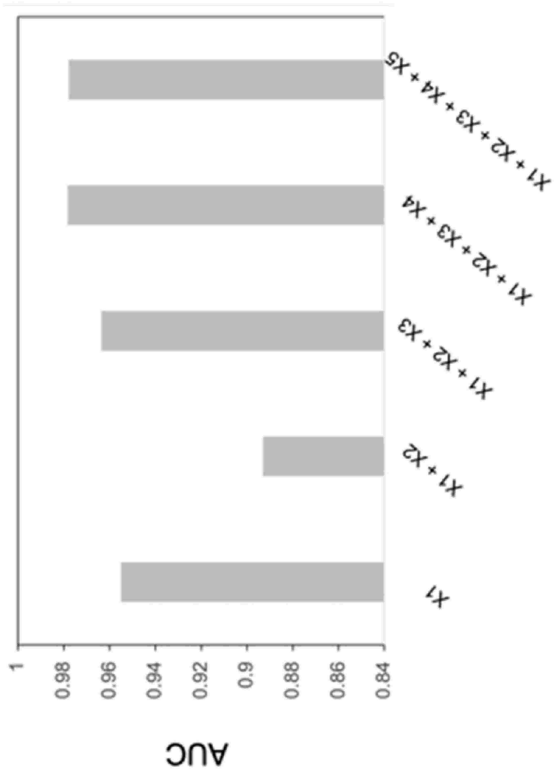
40

50

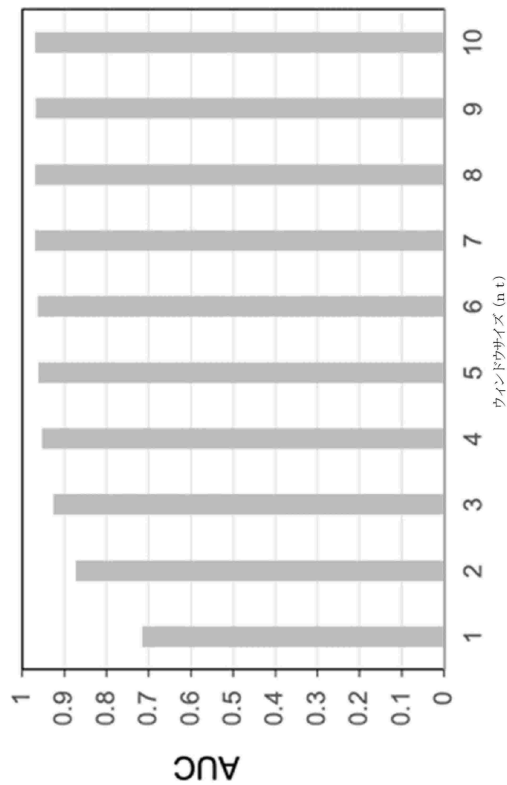
【図 15】



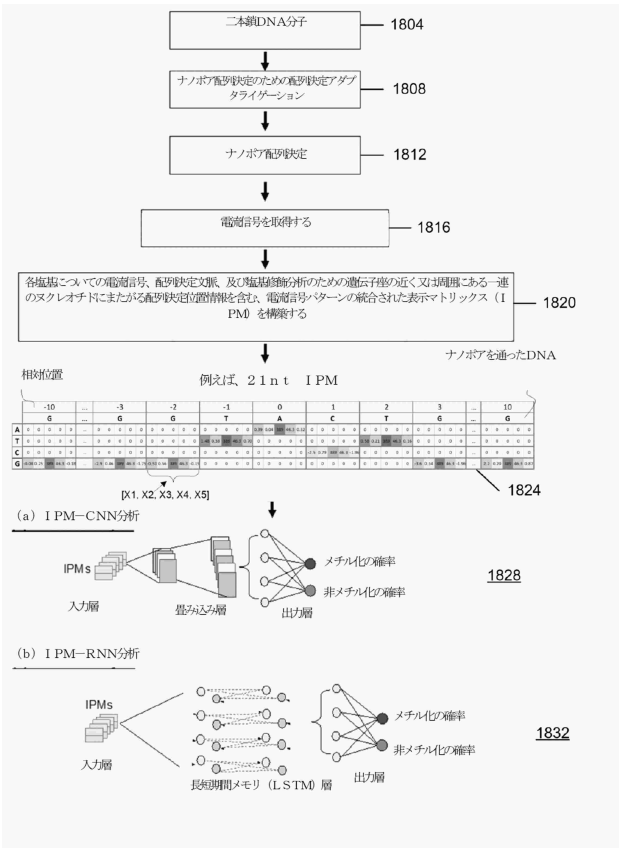
【図 16】



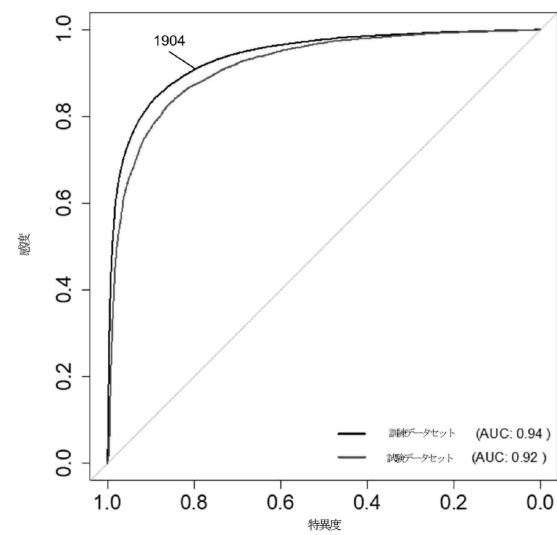
【図 17】



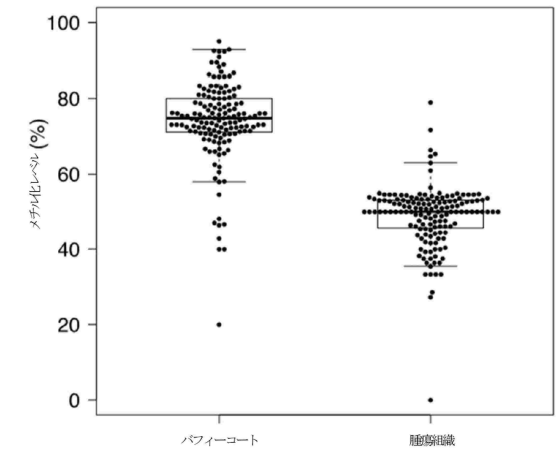
【図 18】



【図 19】

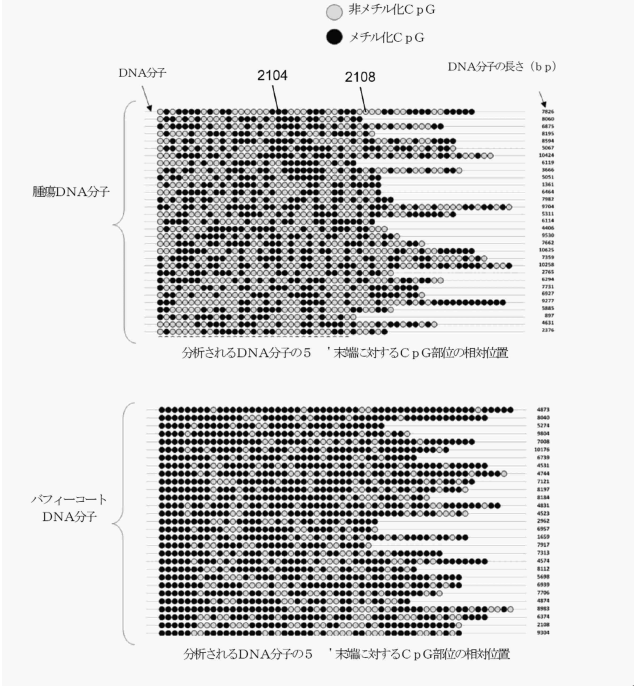


【図 20】

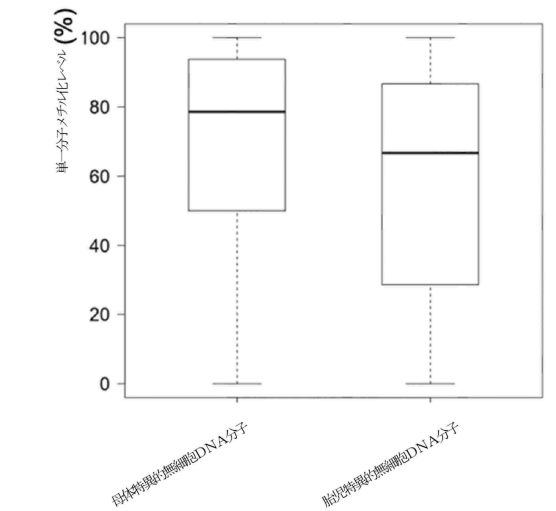


10

【図 21】



【図 22】



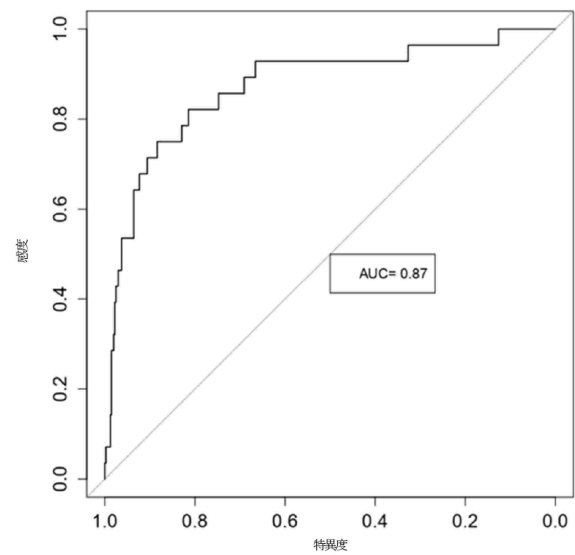
20

30

40

50

【図 2 3】



10

20

30

40

50

【国際調査報告】

INTERNATIONAL SEARCH REPORT		International application No. PCT/CN2022/086260
A. CLASSIFICATION OF SUBJECT MATTER C12Q 1/68(2018.01)i According to International Patent Classification (IPC) or to both national classification and IPC		
B. FIELDS SEARCHED Minimum documentation searched (classification system followed by classification symbols) C12Q Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched Electronic data base consulted during the international search (name of data base and, where practicable, search terms used) WPI:EPODOC;CNKI;CNPAT: sequencing, detect+, base, modification, nucleotide, electrical, signal, nanopore, methyl group, data structure, model, optimizing, training, current, voltage, resistance, inductance, capacitance, impedance		
C. DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	US 2021047679 A1 (The Chinese University of Hong Kong) 18 February 2021 (2021-02-18) claims 1-17, description, paragraphs [0670]-[0677]	1-50
Y	CN 107849607 A (The Chinese University of Hong Kong) 27 March 2018 (2018-03-27) abstract, description, paragraphs [0074]-[0084]	1-50
A	CN 104662165 A (PACIFIC BIOSCIENCES OF CALIFORNIA INC.) 27 May 2015 (2015-05-27) the whole document	1-50
A	US 2002197639 A1 (SHIA, Michael A et al.) 26 December 2002 (2002-12-26) the whole document	1-50
A	US 2011014607 A1 (JIRTLE, Randy L et al.) 20 January 2011 (2011-01-20) the whole document	1-50
A	US 2010028877 A1 (SCHATZ, Philipp) 04 February 2010 (2010-02-04) the whole document	1-50
☐ Further documents are listed in the continuation of Box C. ☒ See patent family annex.		
* Special categories of cited documents: “A” document defining the general state of the art which is not considered to be of particular relevance “E” earlier application or patent but published on or after the international filing date “L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified) “O” document referring to an oral disclosure, use, exhibition or other means “P” document published prior to the international filing date but later than the priority date claimed “T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention “X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone “Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art “&” document member of the same patent family		
Date of the actual completion of the international search 02 June 2022		Date of mailing of the international search report 05 July 2022
Name and mailing address of the ISA/CN National Intellectual Property Administration, PRC 6, Xitucheng Rd., Jimen Bridge, Haidian District, Beijing 100088, China Facsimile No. (86-10)62019451		Authorized officer XU,Jing Telephone No. 86- (10) -53962627

Form PCT/ISA/210 (second sheet) (January 2015)

10

20

30

40

50

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2022/086260

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
US	2021047679	A1	18 February 2021	EP	3827092	A1	02 June 2021
				CA	3110884	A1	25 February 2021
				WO	2021032060	A1	25 February 2021
				TW	202124728	A	01 July 2021
				CN	112752853	A	04 May 2021
				US	2021363571	A1	25 November 2021
				KR	20210068396	A	09 June 2021
				AU	2020323958	A1	04 March 2021
				US	11091794	B2	17 August 2021
				GB	2590573	A	30 June 2021
				GB	2590573	B	22 November 2021
				GB	2590032	A	16 June 2021
				GB	2590032	B	08 December 2021
				GB	2590854	A	07 July 2021
				GB	2590854	B	30 March 2022
CN	107849607	A	27 March 2018	US	2017044606	A1	16 February 2017
				TW	202130815	A	16 August 2021
				EP	3334843	A1	20 June 2018
				JP	2021118691	A	12 August 2021
				WO	2017025059	A1	16 February 2017
				AU	2020202153	A1	16 April 2020
				TW	201718874	A	01 June 2017
				JP	2018531583	A	01 November 2018
				CA	2995422	A1	16 February 2017
				AU	2016305103	A1	22 March 2018
				JP	6867045	B2	28 April 2021
				TW	728994	B	01 June 2021
CN	104662165	A	27 May 2015	US	2016130646	A1	12 May 2016
				US	2013303385	A1	14 November 2013
				US	2019024162	A1	24 January 2019
				WO	2013148400	A1	03 October 2013
				US	2014179564	A1	26 June 2014
				US	2014295498	A1	02 October 2014
				US	9238836	B2	19 January 2016
				US	9290800	B2	22 March 2016
				US	10590484	B2	17 March 2020
				EP	2831283	A1	04 February 2015
US	2002197639	A1	26 December 2002	JP	2004536596	A	09 December 2004
				EP	1402071	A2	31 March 2004
				AU	2002318336	A1	23 December 2002
				WO	02101353	A2	19 December 2002
US	2011014607	A1	20 January 2011	WO	2008070144	A2	12 June 2008
US	2010028877	A1	04 February 2010	WO	2008047234	A1	24 April 2008
				EP	2087137	A1	12 August 2009
				EP	2087137	B1	30 November 2011

Form PCT/ISA/210 (patent family annex) (January 2015)

10

20

30

40

50

フロントページの続き

MK,MT,NL,NO,PL,PT,RO,RS,SE,SI,SK,SM,TR),OA(BF,BJ,CF,CG,CI,CM,GA,GN,GQ,GW,KM,ML,MR,N
E,SN,TD,TG),AE,AG,AL,AM,AO,AT,AU,AZ,BA,BB,BG,BH,BN,BR,BW,BY,BZ,CA,CH,CL,CN,CO,CR,CU,
CZ,DE,DJ,DK,DM,DO,DZ,EC,EE,EG,ES,FI,GB,GD,GE,GH,GM,GT,HN,HR,HU,ID,IL,IN,IR,IS,IT,JM,JO,J
P,KE,KG,KH,KN,KP,KR,KW,KZ,LA,LC,LK,LR,LS,LU,LY,MA,MD,ME,MG,MK,MN,MW,MX,MY,MZ,N
A,NG,NI,NO,NZ,OM,PA,PE,PG,PH,PL,PT,QA,RO,RS,RU,RW,SA,SC,SD,SE,SG,SK,SL,ST,SV,SY,TH,TJ,
TM,TN,TR,TT,TZ,UA,UG,US,UZ,VC,VN,WS,ZA,ZM,ZW

(特許庁注：以下のものは登録商標)

1. S W I F T

2. P Y T H O N

弁理士 石川 大輔

(74)代理人 230113332

弁護士 山本 健策

(72)発明者 ロー, ユクーミン デニス

中華人民共和国 香港, ニュー テリトリーズ, シャーティン, ピ チウ ビルディング, ルー
ム 3 0 1, オーアールケーティーエス, ザ チャイニーズ ユニバーシティ オブ ホンコン 気付

(72)発明者 チウ, ロッサ ワイ クン

中華人民共和国 香港, ニュー テリトリーズ, シャーティン, ピ チウ ビルディング, ルー
ム 3 0 1, オーアールケーティーエス, ザ チャイニーズ ユニバーシティ オブ ホンコン 気付

(72)発明者 チャン, クワン チー

中華人民共和国 香港, ニュー テリトリーズ, シャーティン, ピ チウ ビルディング, ルー
ム 3 0 1, オーアールケーティーエス, ザ チャイニーズ ユニバーシティ オブ ホンコン 気付

(72)発明者 チアン, ペイヨン

中華人民共和国 香港, ニュー テリトリーズ, シャーティン, ピ チウ ビルディング, ルー
ム 3 0 1, オーアールケーティーエス, ザ チャイニーズ ユニバーシティ オブ ホンコン 気付

(72)発明者 チェン, スク ハン

中華人民共和国 香港, ニュー テリトリーズ, シャーティン, ピ チウ ビルディング, ルー
ム 3 0 1, オーアールケーティーエス, ザ チャイニーズ ユニバーシティ オブ ホンコン 気付

(72)発明者 デン, ジアエン

中華人民共和国 香港, ニュー テリトリーズ, シャーティン, ピ チウ ビルディング, ルー
ム 3 0 1, オーアールケーティーエス, ザ チャイニーズ ユニバーシティ オブ ホンコン 気付

F ターム (参考) 4B029 AA07 AA23 BB11 FA03

4B063 QA01 QA13 QA18 QA19 QQ02 QQ03 QQ42 QX05