202305818

(19)中華民國智慧財產局

(12)發明說明書公開本 (11)公開編號：TW 202305818 A

(43)公開日： 中華民國 112 (2023) 年 02 月 01 日

(21)申請案號：111113903 (22)申請日： 中華民國 111 (2022) 年 04 月 12 日

(51)Int. Cl.： *G16B30/00 (2019.01)* *C12Q1/68 (2018.01)*
G16B40/10 (2019.01) *G16B40/20 (2019.01)*

(30)優先權：2021/04/12 美國 63/173,728

(71)申請人：香港中文大學 (香港地區) THE CHINESE UNIVERSITY OF HONG KONG (HK)
香港

(72)發明人：盧 煜明 LO, YUK-MING DENNIS (GB)；趙 慧君 CHIU, ROSSA WAI KWUN (AU)；陳 君賜 CHAN, KWAN CHEE (HK)；江培勇 JIANG, PEIYONG (CN)；鄭 淑恒 CHENG, SUK HANG (HK)；鄧佳恩 DENG, JIAEN (CN)

(74)代理人：陳長文

申請實體審查：無 申請專利範圍項數：50 項 圖式數：23 共 94 頁

(54)名稱

使用電信號之鹼基修飾分析

(57)摘要

本文中描述使用電信號及其他資料判定鹼基修飾之系統及方法。實施例可利用自與測序相關之電信號獲得之特徵，諸如使用奈米孔獲得之彼等特徵，該等特徵受各種鹼基修飾影響，以及判定甲基化狀態之目標位置之周圍窗口中核苷酸之標識。其他特徵可包含對應於該核苷酸之區段電信號的統計值的向量及核酸分子之區域中之窗口中的電信號之統計值。所偵測之鹼基修飾可用於對生物樣本進行額外分析。

Systems and methods for determining base modifications using electrical signals and other data is described herein. Embodiments can make use of features derived from electrical signals related to sequencing, such as those acquired from using a nanopore, that are affected by the various base modifications, as well as an identity of nucleotides in a window around a target position whose methylation status is determined. Other features may include a vector of statistical values of a segment of the electrical signal corresponding to the nucleotide and a statistical value of the electrical signal in a window in a region of the nucleic acid molecule. The detected base modifications can be used for additional analysis of a biological sample.

指定代表圖：

符號簡單說明：

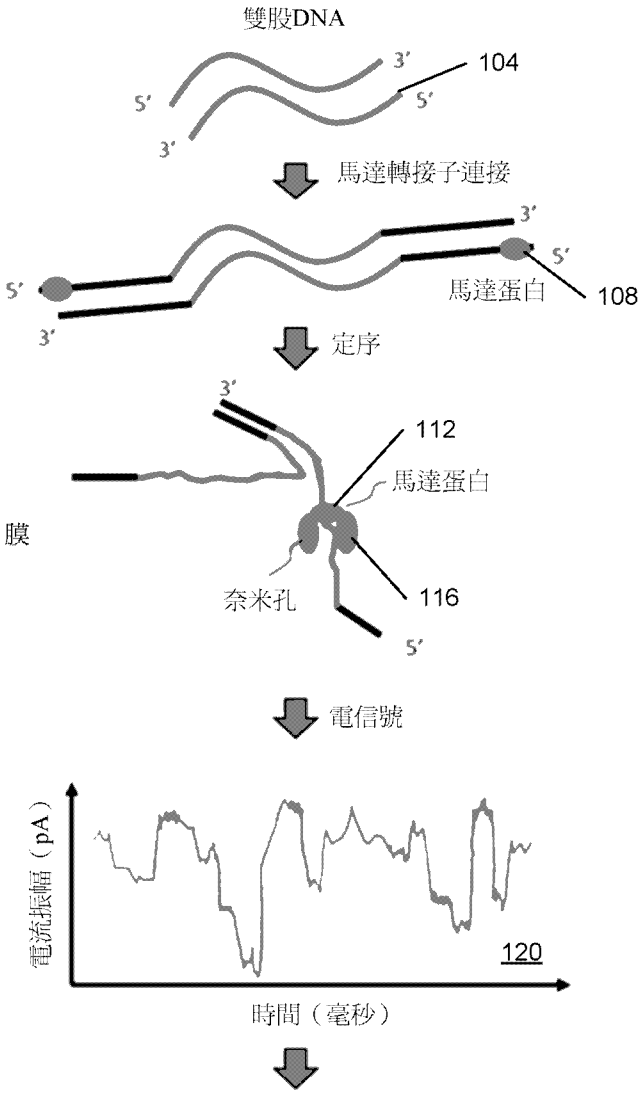
104:DNA 分子

108:馬達蛋白

112:馬達蛋白

116:奈米孔

120:曲線圖



1）鹼基識別

2）鹼基修飾分析

【圖1】

【發明摘要】

【中文發明名稱】

使用電信號之鹼基修飾分析

【英文發明名稱】

BASE MODIFICATION ANALYSIS USING ELECTRICAL SIGNALS

【中文】

本文中描述使用電信號及其他資料判定鹼基修飾之系統及方法。實施例可利用自與測序相關之電信號獲得之特徵，諸如使用奈米孔獲得之彼等特徵，該等特徵受各種鹼基修飾影響，以及判定甲基化狀態之目標位置之周圍窗口中核苷酸之標識。其他特徵可包含對應於該核苷酸之區段電信號的統計值的向量及核酸分子之區域中之窗口中的電信號之統計值。所偵測之鹼基修飾可用於對生物樣本進行額外分析。

【英文】

Systems and methods for determining base modifications using electrical signals and other data is described herein. Embodiments can make use of features derived from electrical signals related to sequencing, such as those acquired from using a nanopore, that are affected by the various base modifications, as well as an identity of nucleotides in a window around a target position whose methylation status is determined. Other features may include a vector of statistical values of a segment of the electrical signal corresponding to the nucleotide and a statistical value of the electrical signal in a window in a region of the nucleic acid molecule. The detected base modifications can be used for additional

analysis of a biological sample.

【指定代表圖】

圖1

【代表圖之符號簡單說明】

104: DNA分子

108: 馬達蛋白

112: 馬達蛋白

116: 奈米孔

120: 曲線圖

【發明說明書】

【中文發明名稱】

使用電信號之鹼基修飾分析

【英文發明名稱】

BASE MODIFICATION ANALYSIS USING ELECTRICAL SIGNALS

【技術領域】

【先前技術】

【0001】核酸中鹼基修飾之存在在包含病毒、細菌、植物、真菌、線蟲、昆蟲及脊椎動物（例如人類）等的不同生物體中各不相同。最常見的鹼基修飾為將甲基添加至不同位置的不同 DNA 鹼基，亦即所謂的甲基化。在胞嘧啶、腺嘌呤、胸腺嘧啶及鳥嘌呤上均已發現甲基化，諸如 5mC（5-甲基胞嘧啶）、4mC（N4-甲基胞嘧啶）、5hmC（5-羥甲基胞嘧啶）、5fC（5-甲醯基胞嘧啶）、5caC（5-羧基胞嘧啶）、1mA（N1-甲基腺嘌呤）、3mA（N3-甲基腺嘌呤）、N6-甲基腺嘌呤（6mA）、7mA（N7-甲基腺嘌呤）、3mC（N3-甲基胞嘧啶）、2mG（N2-甲基鳥嘌呤）、6mG（O6-甲基鳥嘌呤）、7mG（N7-甲基鳥嘌呤）、3mT（N3-甲基胸腺嘧啶）及 4mT（O4-甲基胸腺嘧啶）。在脊椎動物基因體中，5mC 為最常見的鹼基甲基化類型，其次為鳥嘌呤（亦即在 CpG 情況下）。

【0002】DNA 甲基化對哺乳動物的發育至關重要，且在基因表現及沉默、胚胎發育、轉錄、染色質結構、X 染色體失活、防止重複元件的活性、維持有絲分裂過程中基因體的穩定性及調控親源基因體印記方面具有顯著作用。

【0003】DNA 甲基化在啟動子及強化子的沉默中以協調的方式發揮著許

多重要作用（Robertson, 2005；Smith 及 Meissner, 2013）。已發現許多人類疾病與 DNA 甲基化之畸變有關，包含但不限於印記病症（例如貝克威思-威德曼症候群（Beckwith-Wiedemann syndrome）及普瑞德威利症候群（Prader-Willi syndrome））、重複不穩定性疾病（例如 X 脆折症候群）、自體免疫性病症（例如全身性紅斑狼瘡）、代謝障礙（例如 I 型及 II 型糖尿病）、神經病症、衰老等。

【0004】準確量測 DNA 分子上之甲基化修飾將具有許多臨床意義。一種廣泛使用的量測 DNA 甲基化之方法為經由使用亞硫酸氫鹽測序（BS-seq）（Lister 等人, 2009；Frommer 等人, 1992）。在此方法中，DNA 樣本首先用亞硫酸氫鹽處理，將未甲基化之胞嘧啶（亦即 C）轉化為尿嘧啶。相反，甲基化之胞嘧啶保持不變。隨後藉由 DNA 測序分析亞硫酸氫鹽修飾之 DNA。在另一種方法中，在亞硫酸氫鹽轉化之後，接著使用可區分具有不同甲基化譜之經亞硫酸氫鹽轉化之 DNA 的引子對經修飾之 DNA 進行聚合酶鏈反應（PCR）擴增（Herman 等人, 1996）。後一種方法稱為甲基化特異性 PCR。

【0005】此類基於亞硫酸氫鹽之方法的一個缺點為，據報導亞硫酸氫鹽轉化步驟會顯著降解大多數經處理之 DNA（Grunau, 2001）。另一個缺點為亞硫酸氫鹽轉化步驟會產生強烈的 CG 偏差（Olova 等人, 2018），導致具有異質甲基化狀態之 DNA 混合物典型的信雜比降低。此外，由於在亞硫酸氫鹽處理期間 DNA 之降解，亞硫酸氫鹽測序將不係對長 DNA 分子進行測序的理想方法。

【0006】正在持續的努力實現核酸之鹼基修飾的無亞硫酸氫鹽測定。然而，很少有商業上可行的工具能夠達到與亞硫酸氫鹽測序相當的敏感度及特異度程度。奈米孔測序為一種不需要對樣本進行化學標記的具有吸引力的測序類型。用奈米孔測序偵測鹼基修飾可為成本相對較低的且高效的。

【0007】因此，需要用奈米孔測序來判定鹼基修飾。在本揭示案中，吾等描述了處理藉由具有高靈敏度及特異性之奈米孔測序所產生之電流信號以用於

鹼基修飾測定的新穎方法及系統。

【發明內容】

【0008】 所描述之實施例允許在沒有模板 DNA 預處理（諸如酶促及/或化學轉化，或蛋白質及/或抗體結合）之情況下測定核酸中之鹼基修飾，諸如 5mC。本揭示案中存在之實施例可用於偵測不同類型之鹼基修飾，例如，包含但不限於 4mC、5hmC、5fC、5caC、1mA、3mA、6mA、7mA、3mC、2mG、6mG、7mG、3mT、4mT 等。此類實施例可利用自與測序相關之電信號獲得之特徵(諸如使用奈米孔獲得之彼等特徵，該等特徵受各種鹼基修飾影響)，以及判定甲基化狀態之目標位置之周圍窗口中核苷酸之標識。核苷酸之原始電信號亦可與核苷酸上游或下游之核苷酸有關。可以使用合適的技術將原始電信號分配給不同的核苷酸。

【0009】 本發明之實施例可與奈米孔測序一起使用。奈米孔測序系統之一個實例為由牛津奈米孔科技有限公司（Oxford Nanopore Technologies）商業化之系統。方法可使用使用奈米孔測量之電信號。方法可使用核苷酸之標識、核苷酸相對於目標位置之位置、包含對應於該核苷酸之區段電信號的統計值的向量及核酸分子之區域中之窗口中的電信號之統計值。

【0010】 吾等開發之方法可充當偵測生物樣本中鹼基修飾之工具，以評定樣本中之甲基化譜，用於各種目的，包含但不限於研究及診斷目的。偵測到的甲基化譜可用於不同的分析。甲基化譜可用於偵測細胞 DNA 之來源（例如母體或胎兒、組織、細菌）。偵測組織中之異常甲基化譜有助於鑑別個體之發育病症及其他病症。

【0011】 可參考以下詳細描述及隨附圖式來獲得對本發明之實施例之性質及優勢的較佳理解。

【圖式簡單說明】

【0012】 **圖 1** 示出奈米孔測序。

【0013】 **圖 2** 示出根據本發明之實施例的不同信號特徵。

【0014】 **圖 3** 示出根據本發明之實施例的電流信號分段及信號特徵向量之建構。

【0015】 **圖 4** 為根據本發明之實施例的每個核苷酸穿過奈米孔之事件長度（亦即，持續時間）的分佈圖。

【0016】 **圖 5** 示出根據本發明之實施例的使用包括電流模式、測序位置及測序背景（sequencing context）之整合式表示矩陣的 5mC 偵測之原理。

【0017】 **圖 6** 示出根據本發明之實施例的使用包括電流模式、測序位置及基於雙股 DNA 之兩個股的測序背景之整合式表示矩陣的鹼基修飾偵測之原理。

【0018】 **圖 7** 展示根據本發明之實施例的核尺寸對鹼基修飾分析之效能的影響。

【0019】 **圖 8** 展示根據本發明之實施例的關於甲基化偵測之用於訓練及測試之測序分子數目。

【0020】 **圖 9A 至圖 9D** 為根據本發明之實施例的使用 IPM-CNN 及 IPM-RNN 方法的 WGA DNA 與經 M.SssI 處理之 DNA 資料集之間的 CpG 甲基化概率的盒狀圖。

【0021】 **圖 10A 及圖 10B** 展示根據本發明之實施例的訓練資料集及測試資料集之接受者操作特徵（ROC）曲線。

【0022】 **圖 11** 為根據本發明之實施例的用於甲基化分析之不同工具之效能的表。

【0023】 **圖 12** 為根據本發明之實施例的偵測核酸分子中核苷酸之修飾的方法之流程圖。

【0024】 **圖 13** 為根據本發明之實施例的偵測核酸分子中核苷酸之修飾的方法之流程圖。

【0025】 **圖 14** 示出根據本發明之實施例的量測系統。

【0026】 **圖 15** 展示可與根據本發明之實施例的系統及方法一起使用的實例電腦系統之方塊圖。

【0027】 **圖 16** 展示根據本發明之實施例的不同參數組合對 ROC 曲線下面積（AUC）之影響的圖。

【0028】 **圖 17** 展示根據本發明之實施例的窗口大小對 AUC 之影響的圖。

【0029】 **圖 18** 示出根據本發明之實施例的使用包括電流模式、測序位置及測序背景之整合式表示矩陣的 6mA 偵測之原理。

【0030】 **圖 19** 展示根據本發明之實施例的 6mA 偵測之 AUC 的圖。

【0031】 **圖 20** 為針對根據本發明之實施例的源自白血球層（buffy coat）及 NPC 腫瘤樣本之 DNA，藉由 IPM-RNN 模型測定之單分子甲基化程度的比較。

【0032】 **圖 21** 展示根據本發明之實施例的單分子甲基化模式之實例。

【0033】 **圖 22** 為根據本發明之實施例的母體特異性及胎兒特異性游離 DNA 分子之單分子甲基化程度的圖。

【0034】 **圖 23** 為根據本發明之實施例的使用由 IPM-CNN 模型判定之甲基化模式判定游離 DNA 分子之胎兒及母體來源的 ROC 曲線。

【實施方式】

相關申請案之交叉參考

【0035】本申請案主張 2021 年 4 月 12 日申請之美國臨時專利申請案 63/173,728 之優先權益，其以全文引用之方式併入本文中且用於所有目的。

術語

【0036】「*組織*」對應於一組細胞，其共同歸類為一個功能單元。可在單一組織中找到超過一種類型之細胞。不同類型的組織可由不同類型的細胞（例如肝細胞、肺泡細胞或血細胞）組成，但亦可對應於來自不同生物體之組織（母親與胎兒；接受移植之個體的組織；經微生物或病毒感染之生物體的組織）或健康細胞與腫瘤細胞。「參考組織」可對應於用於判定組織特異性甲基化程度之組織。來自不同個體之相同組織類型之多個樣本可用於測定該組織類型之組織特異性甲基化程度。

【0037】「*生物樣本*」係指取自人類個體之任何細胞樣本。生物樣本可為組織生檢、細針抽吸物或血細胞。樣品亦可為獲自孕婦之游離樣本，例如血漿或血清或尿液。在各種實施例中，已富集游離 DNA 的來自孕婦之生物樣本（例如經由離心方案獲得之血漿樣本）中的大多數 DNA 可為游離的，例如大於 50%、60%、70%、80%、90%、95%或 99%之 DNA 可為游離的。離心方案可包含例如 3,000 g × 10 分鐘獲得流體部分，及以例如 30,000 g 再離心 10 分鐘以移除殘餘細胞。在某些實施例中，在 3,000 g 離心步驟之後，吾人可接著對流體部分進行過濾（例如使用孔徑（直徑）為 5 μm 或更小的過濾器）。

【0038】「*序列讀數*」係指自核酸分子之任何部分或全部測序的一串核苷酸。舉例而言，序列讀數可為自核酸片段測序之短核苷酸串（例如 20 至 150 個）、在核酸片段之一端或兩端之短核苷酸串或存在於生物樣本中之整個核酸片段的

測序。序列讀數可以多種方式獲得，例如使用測序技術或使用探針，例如雜交陣列或捕獲探針；或擴增技術，諸如聚合酶鏈反應（PCR）或使用單一引子的線性擴增或等溫擴增。

【0039】「*位點*」（亦稱作「基因體位點」）對應於單一位點，其可為單一鹼基位置或相關鹼基位置群，例如 CpG 位點或相關鹼基位置之較大群。「基因座」可對應於包含多個位點之區域。基因座可僅包含一個位點，此將使得基因座在彼情形下等效於一個位點。

【0040】「*甲基化狀態*」係指既定位點處之甲基化狀態。舉例而言，位點可為甲基化的、未甲基化的或在一些情況下不能判定。

【0041】 各基因體位點（例如 CpG 位點）之「*甲基化指數*」可指在該位點處顯示甲基化之 DNA 片段（例如，如由序列讀數或探針判定）相對於涵蓋彼位點之讀數總數的比例。「讀數」可對應於獲自 DNA 片段之資訊（例如，位點處之甲基化狀態）。讀數可使用優先雜交至在一或多個位點處具有特定甲基化狀態之 DNA 片段的試劑（例如引子或探針）來獲得。通常，該等試劑係在用視 DNA 分子之甲基化狀態而有差異地修飾或有差異地辨識 DNA 分子之方法處理後施用，該方法例如為亞硫酸氫鹽轉化、或甲基化敏感限制酶、或甲基化結合蛋白、或抗甲基胞嘧啶抗體、或辨識甲基胞嘧啶及羥甲基胞嘧啶之單分子測序技術（例如單分子即時測序（例如，來自美國太平洋生物科學公司（Pacific Biosciences））及奈米孔測序（例如來自牛津奈米孔科技有限公司））。

【0042】 區域之「*甲基化密度*」可指顯示甲基化之區域內之位點處之讀數數目除以覆蓋該區域中之位點之讀數總數。該等位點可具有特定特性，例如為 CpG 位點。因此，區域之「CpG 甲基化密度」可指顯示 CpG 甲基化之讀數數目除以覆蓋該區域中之 CpG 位點（例如特定 CpG 位點、CpG 島或較大區域內之 CpG 位點）之讀數總數。例如，人類基因體中各 100 kb 區段（bin）之甲基化密

度可自亞硫酸氫鹽處理之後在 CpG 位點處未轉化之胞嘧啶（其對應於甲基化胞嘧啶）的總數測定為映射至 100 kb 區域之序列讀數所覆蓋之所有 CpG 位點的比例。亦可針對其他區段大小，例如 500 bp、5 kb、10 kb、50 kb 或 1 Mb 等執行此分析。區域可為整個基因體或染色體或染色體之一部分（例如染色體臂）。替代地，甲基化密度可在無亞硫酸氫鹽轉化之情況下使用奈米孔測序使用本揭示案所描述之實施例來測定。當區域僅包含 CpG 位點時，CpG 位點之甲基化指數與區域之甲基化密度相同。「甲基化胞嘧啶之比例」可指相對於所分析之胞嘧啶殘基，亦即包含該區域中除 CpG 情形之外的胞嘧啶的總數而言顯示為甲基化（例如在亞硫酸氫鹽轉化之後未經轉化）的胞嘧啶位點「C's」數目。甲基化指數、甲基化密度、在一或多個位點處甲基化之分子計數及在一或多個位點處甲基化之分子（例如胞嘧啶）比例為「*甲基化程度*」之實例。除亞硫酸氫鹽轉化以外，可使用本領域中熟習此項技術者已知之其他方法來查詢 DNA 分子之甲基化狀態，包含但不限於對甲基化狀態敏感的酶（例如甲基化敏感限制酶）、甲基化結合蛋白、使用對甲基化狀態敏感之平台進行的單分子測序（例如奈米孔測序（Schreiber 等人，《國家科學院院刊（Proc Natl Acad Sci）》 2013; 110: 18910-18915）及藉由單分子即時測序（例如來自美國太平洋生物科學公司的單分子即時測序）（Flusberg 等人 《自然-方法（Nat Methods）》 2010; 7: 461-465））。

【0043】「*甲基化組*」提供基因體中之複數個位點或基因座處之 DNA 甲基化之量的量度。甲基化組可對應於所有基因體、基因體之相當大部分或基因體之一或多個相對小的部分。

【0044】「*妊娠血漿甲基化組*」為自妊娠動物（例如人類）之血漿或血清測定的甲基化組。妊娠血漿甲基化組為游離甲基化組之實例，因為血漿及血清包含游離 DNA。妊娠血漿甲基化組亦為混合甲基化組之實例，因為其為來自體內不同器官或組織或細胞之 DNA 的混合物。在一個實施例中，此類細胞為造血細胞，

包含但不限於紅血球系（亦即紅血球）、骨髓系（例如嗜中性白血球及其前驅體）及巨核細胞系之細胞。在妊娠期，血漿甲基化組可含有來自胎兒及母親之甲基化組資訊。「細胞甲基化組」對應於自患者之細胞（例如血球）測定之甲基化組。血細胞之甲基化組稱為血球甲基化組。

【0045】「*甲基化譜*」包含與多個位點或區域之 DNA 或 RNA 甲基化相關的資訊。與 DNA 甲基化相關之資訊可包含但不限於 CpG 位點之甲基化指數、區域中之 CpG 位點之甲基化密度（簡稱 MD）、CpG 位點在相連區域上之分佈、含有超過一個 CpG 位點之區域內的各個別 CpG 位點之甲基化模式或程度，及非 CpG 甲基化。在一個實施例中，甲基化譜可包含超過一種類型之鹼基（例如胞嘧啶或腺嘌呤）之甲基化或非甲基化模式。基因體之相當大部分之甲基化譜可視為等效於甲基化組。哺乳動物基因體中之「DNA 甲基化」通常指將甲基添加至 CpG 二核苷酸當中之胞嘧啶殘基的 5'碳（亦即 5-甲基胞嘧啶）。DNA 甲基化可在例如 CHG 及 CHH 之其他情形下發生於胞嘧啶中，其中 H 為腺嘌呤、胞嘧啶或胸腺嘧啶。胞嘧啶甲基化亦可呈 5-羥甲基胞嘧啶形式。亦已報導非胞嘧啶甲基化，諸如 N^6-甲基腺嘌呤。

【0046】「*甲基化模式*」係指甲基化及非甲基化鹼基之次序。舉例而言，甲基化模式可為單個 DNA 股、單個雙股 DNA 分子或另一類型之核酸分子上之甲基化鹼基之次序。作為一實例，三個連續 CpG 位點可具有以下甲基化模式中之任一者：UUU、MMM、UMM、UMU、UUM、MUM、MUU 或 MMU，其中「U」指示未甲基化位點且「M」指示甲基化位點。當吾人將此概念擴展至包含但不限於甲基化之鹼基修飾時，吾人將使用術語「*修飾模式*」，其係指經修飾及未經修飾鹼基之次序。舉例而言，修飾模式可為單個 DNA 股、單個雙股 DNA 分子或另一類型之核酸分子上之經修飾鹼基之次序。作為一實例，三個連續潛在地可修飾位點可具有以下修飾模式中之任一者：UUU、MMM、UMM、UMU、UUM、

MUM、MUU 或 MMU，其中「U」指示未經修飾位點且「M」指示經修飾位點。不基於甲基化之鹼基修飾之一個實例為諸如於 8-側氧基-鳥嘌呤中之氧化變化。

【0047】 術語「*高甲基化*」及「*低甲基化*」可指單個 DNA 分子之甲基化密度，如藉由其單分子甲基化程度所量測，例如分子內之甲基化鹼基或核苷酸之數目除以彼分子內之可甲基化鹼基或核苷酸之總數。高甲基化分子為其中單分子甲基化程度等於或高於臨限值之分子，該臨限值可根據不同應用而界定。臨限值可為 5%、10%、20%、30%、40%、50%、60%、70%、80%、90%或 95%。低甲基化分子為其中單分子甲基化程度等於或低於臨限值之分子，該臨限值可根據不同應用而界定且可根據不同應用而變化。臨限值可為 5%、10%、20%、30%、40%、50%、60%、70%、80%、90%或 95%。

【0048】 術語「高甲基化」及「低甲基化」亦可指 DNA 分子群體之甲基化程度，如藉由此等分子之多分子甲基化程度所量測。高甲基化分子群體為其中多分子甲基化程度等於或高於臨限值之分子群體，該臨限值可根據不同應用而界定且可根據不同應用而變化。臨限值可為 5%、10%、20%、30%、40%、50%、60%、70%、80%、90%或 95%。低甲基化分子群體為其中多分子甲基化程度等於或低於臨限值之分子群體，該臨限值可根據不同應用而界定。臨限值可為 5%、10%、20%、30%、40%、50%、60%、70%、80%、90%及 95%。在一個實施例中，可將分子群體與一或多個經選擇之基因體區域進行比對。在一個實施例中，一或多個經選擇之基因體區域可與諸如遺傳病症、印記病症、表觀遺傳病症、代謝病症或神經病症之疾病相關。一或多個經選擇之基因體區域之長度可為 50 個核苷酸（nt）、100 nt、200 nt、300 nt、500 nt、1000 nt、2 knt、5 knt、10 knt、20 knt、30 knt、40 knt、50 knt、60 knt、70 knt、80 knt、90 knt、100 knt、200 knt、300 knt、400 knt、500 knt 或 1 Mnt。

【0049】 如本文所使用之術語「分類」係指與樣本之特定特性相關之任何

數字或其他字符。舉例而言，「+」符號（或字組「陽性」）可表示將樣本分類為具有缺失或擴增。分類可為二元的(例如陽性或陰性)或具有更多分類水準(例如 1 至 10 或 0 至 1 之標度）。

【0050】術語「*閾值*」及「*臨限值*」係指操作中所使用之預定數值。舉例而言，截止大小可指一種大小，大於此大小則排除片段。臨限值可為高於或低於特定分類適用之值。在此等情形中之任一者下均可使用此等術語中之任一者。閾值或臨限值可為表示特定分類或在兩種或更多種分類之間進行辨別的「參考值」或源自該參考值。如技術人員應瞭解，此類參考值可以各種方式測定。例如，可針對具有不同已知分類的兩個不同個體群組測定度量，且可選擇參考值作為一個分類的代表（例如平均值）或介於度量的兩個集群之間的值（例如經選擇以獲得所需的靈敏度及特異性）。作為另一實例，參考值可基於樣本之統計分析或模擬來測定。

【0051】「*病理等級*」（或病症等級）可指與生物體相關之病理的量、程度或嚴重性，其可經由對其細胞之分析來量測。病理之另一實例為移植器官之排斥。其他例示性病理可包含基因體印記病症、自體免疫攻擊（例如損害腎臟之狼瘡性腎炎或損害神經系統之多發性硬化症）、發炎性疾病（例如肝炎）、纖維化過程（例如肝硬化）、脂肪浸潤（例如脂肪性肝病）、退行性過程（例如阿茲海默氏病（Alzheimer's disease））及缺血性組織損傷（例如心肌梗塞或中風）。個體之健康狀態可視為無病理之分類。

【0052】「*妊娠相關病症*」包含以母體及/或胎兒組織中基因之相對表現水準異常為特徵的任何病症。此等病症包含但不限於子癇前症、宮內發育遲緩、侵入性胎盤形成、早產、新生兒溶血性疾病、胎盤功能不全、胎兒水腫、胎兒畸形、HELLP（溶血、肝酵素升高及血小板計數低）症候群、全身性紅斑狼瘡（SLE）及母親之其他免疫性疾病。在一些實施例中，妊娠相關病症為與妊娠期間的生理

或形態異常相關的任何病狀。

【0053】 縮寫「*bp*」係指鹼基對。在一些情況下，「bp」可用於表示 DNA 片段之長度，即使 DNA 片段可為單股的且不包含鹼基對。在單股 DNA 之情形下，「bp」可解釋為提供核苷酸之長度。

【0054】 縮寫「*nt*」係指核苷酸。在一些情況下，「nt」可用於表示以鹼基為單位之單股 DNA 長度。此外，「nt」可用於表示相對位置，諸如所分析之基因座之上游或下游。在關於技術概念化、資料顯示、處理及分析之一些情形下，「nt」及「bp」可互換使用。

【0055】 術語「*序列上下文（sequence context）*」可指一段 DNA 中之鹼基組成（A、C、G 或 T）及鹼基順序。此段 DNA 可圍繞進行鹼基修飾分析或作為鹼基修飾分析之目標的鹼基。舉例而言，序列上下文可指進行鹼基修飾分析之鹼基的上游及/或下游的鹼基。

【0056】 術語「*機器學習模型*」可包含基於使用樣本資料（例如訓練資料）對測試資料作出預測之模型，且因此可包含監督式學習。機器學習模型常常使用電腦或處理器來研發。機器學習模型可包含統計模型。

【0057】 術語「*資料分析框架*」可包含可將資料視為輸入且隨後輸出所預測結果之演算法及/或模型。「資料分析框架」之實例包含統計模型、數學模型、機器學習模型、其他人工智慧模型及其組合。

【0058】 術語「*即時測序*」可指涉及在測序所涉及之過程期間進行資料收集或監測的技術。舉例而言，即時測序可涉及當核苷酸股易位奈米孔時對通過該奈米孔之離子電流進行電信號監測。

【0059】 術語「*電信號*」可指傳達資訊之電壓或電流。電信號可以多種規律及/或不規律的信號波形類型及/或形狀，諸如方形波、矩形波、三角形波、鋸齒形波形，或多種脈衝及尖峰來表示。電信號可包含電壓或電流隨時間推移之變

化的視覺表示。可在特定時間（例如，毫秒）對電信號之量測進行採樣。舉例而言，以 1 kHz、2 kHz、3 kHz、4 kHz、5 kHz、10 kHz、20 kHz、30 kHz、40 kHz、50 kHz、100 kHz 等之頻率對電流進行採樣。

【0060】 術語「信號區段」或「區段」可指與對特定核苷酸進行測序相關之電信號之跡線的一部分。該區段可對應於由奈米孔測序中之鹼基識別判定的核苷酸。該區段可涵蓋跡線之某一持續時間。不同區段可具有不同的持續時間。各區段可不重疊。在一些實施例中，電信號幅度可在區段中具有一定的變化。舉例而言，電信號幅度可在該區段中之電信號幅度平均值或中值之 5%、10%、20%、30%或 40%內。

【0061】 術語「*約（about/approximately）*」可意謂在如藉由本領域中一般熟習此項技術者所測定之特定值之可接受誤差範圍內，其將部分地視該值如何經量測或測定，亦即量測系統之限制而定。舉例而言，根據本領域中之實踐，「約」可意謂在 1 或大於 1 個標準差內。可替代地，「約」可意謂既定值之至多 20%、至多 10%、至多 5%或至多 1%之範圍。可替代地，尤其關於生物系統或方法，術語「約」可意謂在值之一定數量級內、在 5 倍內且更佳地在 2 倍內。當特定值描述於本申請案及申請專利範圍中時，除非另外說明，否則應假定術語「約」意謂在特定值之可接受誤差範圍內。術語「約」可具有如本領域中一般熟習此項技術者通常所理解之含義。術語「約」可指±10%。術語「約」可指±5%。

【0062】 需要使用奈米孔測序偵測鹼基修飾（例如甲基化）的準確及有效的方法。調研性研究已研究使用由奈米孔測序產生之電信號分析 DNA 甲基化之可行性（Simpson 等人，《自然方法學（Nat Methods）》2017;14:407-410；Liu 等人，《自然通訊（Nat Commun.）》2019;10:2449；Ni 等人，《生物資訊（Bioinformatics）》2019;35:4586-4595）。5-甲基胞嘧啶（5mC）之報導效能在許多驗證研究中為次佳的。舉例而言，當基於樣本 NA12878 分析 *H. sapiens* R9.4

1D 資料時，使用名為 DeepSignal 之計算工具進行 5mC 偵測之靈敏度據報導為 79%，特異性為 88%，(Ni 等人，《生物資訊》2019;35:4586-4595）。若吾人旨在實現較高特異性（例如>95%），則預期靈敏度將進一步惡化。對於稱為 nanopolish 之另一工具（Liu 等人，《自然通訊》2019；10:2449），當分析相同的資料集時，靈敏度僅為 0.61，特異性為 0.46。nanopolish 軟體係基於具有以下假設之隱藏式馬可夫模型（hidden Markov model）：（1）DNA 序列中之 6-核苷酸寡聚物（亦即 6-單體單元）之電信號遵循高斯分佈（Gaussian distributions）；（2）特定鹼基之甲基化狀態（甲基化或未甲基化）僅取決於前一鹼基之甲基化狀態的概率；（3）輸出僅取決於產生電流信號之甲基化狀態而不取決於任何其他甲基化狀態或任何其他電流信號之特定電流水準的概率。彼等假設在奈米孔測序期間產生之真實電流信號中可能不正確，因此會導致較低敏感度及特異性。

【0063】 用於基於牛津奈米孔測序進行 DNA 甲基化分析的名為 DeepMod 之最新計算工具嘗試使用雙向遞迴神經網路（RNN）。然而，此類方法之設計旨在藉由利用電信號合計測序讀數之預測結果來量測基因體位置中之甲基化程度，因此不具有分析單分子水準下之甲基化模式的能力。另外，整個資料集（包含大腸桿菌（*Escherichia coli*）、萊茵衣藻（*Chlamydomonas reinhardtii*）及智人（*Homo sapiens*））之中值測序深度為約 33×。在許多商業應用中，將需要較低的測序深度以節省經濟成本及分析時間。尚不清楚 DeepMod 軟體是否能夠以實際上有意義的準確性分析單分子水準下之甲基化模式。

【0064】 在一項研究中，Yuen 等人系統地衡量用於由奈米孔測序進行 CpG 甲基化偵測之工具，且得出結論：大多數工具展示高分散性及與每個 CpG 位點之預期甲基化百分比的低一致性（Yuen 等人，bioRxiv.2020; doi: doi.org/10.1101/2020.10.14.340315）。

【0065】 Tse 等人使用來自太平洋生物科學公司（PacBio）之單分子即時測

序（SMRT-seq）報導了 DNA 聚合酶之動力學特徵，包含藉由在 DNA 聚合期間併入經螢光團標記之核苷酸所產生之光信號，諸如脈衝間隔持續時間（IPD）及脈波寬度（PW），該等經螢光團標記之核苷酸可用於基於使用卷積類神經網路分析由超過一個鹼基組成之量測窗口來區分甲基化及未甲基化 CpG 位點（Tse 等人，《美國國家科學院院刊》2021;118: e2019768118；美國專利第 11,091,794 號）。此類量測窗口將 IPD 及 PW 分組成不同的測序背景及測序位置。然而，奈米孔測序使用完全不同的測序機制，視由穿過奈米孔之雙股 DNA 之一個股所引起之電流信號而定。此類原始電信號視穿過奈米孔之不同核苷酸而變化，且特定核苷酸之電信號將受該核苷酸附近之上游及下游核苷酸影響。因此，不同核苷酸將具有偵測到的不同長度的電信號跡線，且甚至相同的核苷酸將具有不同長度的電信號跡線。當分析與特定核苷酸或超過一個穿過奈米孔之核苷酸相關之電信號時，在各鹼基上偵測到的電信號跡線之長度隨時間推移為不固定的。相比之下，使用 PacBio SMRT-seq 進行 5mC 偵測之前述研究係基於兩個與各核苷酸之光信號相關之固定量測，亦即 IPD 及 PW（Tse 等人，《美國國家科學院院刊》2021; 118:e2019768118）。因此，Tse 等人之研究中提出之訓練模型（Tse 等人，《美國國家科學院院刊》2021; 118:e2019768118）不適用於此類藉由奈米孔測序產生之電信號。

【0066】本文所描述之實施例使用自奈米孔測序獲得之電信號來偵測核苷酸修飾。核苷酸修飾可包含本文所描述之任何甲基化。自奈米孔測序獲得之資訊可包含核苷酸之標識、核苷酸相對於目標位置之位置、包含對應於該核苷酸之區段電信號的統計值的向量及核酸分子之區域中之窗口中的電信號之統計值。

【0067】本揭示案中提供之實施例可用於自獲自生物體之細胞樣本（例如，細胞株、實體器官、實體組織、經由內窺鏡檢獲得之樣本、絨毛膜樣本）獲得的 DNA。本揭示案中之實施例亦可用於自環境（例如，細菌、細胞污染物）、

食品（例如肉）獲得的細胞樣本。本揭示案提供之實施例亦可用於自孕婦獲得之血漿或血清。在一些實施例中，本揭示案中提供之方法亦可在首先例如使用雜交探針（Albert 等人, 2007；Okou 等人, 2007；Lee 等人, 2011），或基於物理分離（例如基於大小等）之方法或在限制酶消化（例如 MspI）後，或基於 Cas9 之富集（Watson 等人, 2019）富集基因體碎片的步驟之後應用。儘管本發明不需要酶促或化學轉化來起作用，但在某些實施例中，可包含此類轉化步驟以進一步增強本發明之效能。

【0068】 本揭示案之實施例改良奈米孔測序以能夠準確且有效地偵測經修飾之鹼基。可直接偵測鹼基修飾。實施例可避免可能無法保留所有修飾資訊以供偵測之酶促或化學轉化。另外，某些酶促或化學轉化可能與某些類型之修飾不相容。本揭示案之實施例亦可避免藉由 PCR 擴增，其可能不會將鹼基修飾資訊轉移至 PCR 產物。另外，DNA 之兩個股可一起測序，從而使一個股之序列與其互補序列配對至另一個股。相比之下，PCR 擴增會分開雙股 DNA 之兩個股，因此難以對兩個組成股之序列進行此類組合分析。

【0069】 此外，相比於其他測序技術，奈米孔測序更具有成本效益及便攜性。舉例而言，奈米孔測序系統 Oxford Nanopore Technologies MinION™ 為約 5,000 USD，而基於光信號之測序系統 PacBio SMRT™ Sequel II 系統為約 500,000 至 700,000 USD。奈米孔測序速度為約 450 個核苷酸/秒，而 PacBio SMRT™ 測序為約 5 個核苷酸/秒。因此，在相同的時間段內，奈米孔測序可獲得比基於光信號之測序系統更多的資料。

【0070】 在有或沒有酶促或化學轉化之情況下測定的甲基化譜可用於分析生物樣本。在一個實施例中，甲基化譜可用於偵測細胞 DNA 之來源（例如母體或胎兒、組織或病毒）。偵測組織中之異常甲基化譜有助於鑑別個體之發育障礙。單分子中之甲基化模式可鑑別嵌合（例如在病毒與人類之間）及雜合 DNA

（例如在天然基因體中正常未融合之兩個基因之間）；或在兩個物種之間（例如經由基因或基因體操縱）。

I. 奈米孔測序原理

【0071】 單分子測序技術之一實例為奈米孔測序（牛津奈米孔科技有限公司）。圖 1 展示用於 DNA 分子（例如 DNA 分子 104）之奈米孔測序的原理。當單 DNA 分子穿過具有奈米大小之孔隙時，由離子電流流動跨過膜所引起之電信號模式用於測定核酸之序列。此類孔隙可例如但不限於由蛋白質（例如 α 溶血素、氣單胞菌溶素（aerolysin）及包皮垢分枝桿菌孔蛋白 A（Mycobacterium smegmatis porin A，MspA））或合成材料（諸如矽或石墨烯）產生（Magi 等人，《生物諮詢學簡報（Brief Bioinform.）》2018;19:1256-1272）。

【0072】 在一個實施例中，雙股 DNA 分子會經歷末端修復過程。此過程將 DNA 轉化至鈍端 DNA，接著添加促進測序轉接子連接之 A 尾端。各自攜載馬達蛋白之測序轉接子（亦即，馬達轉接子）（例如，馬達蛋白 108）連接至 DNA 分子之兩端。測序過程係在馬達蛋白（例如，馬達蛋白 112）鬆解雙股 DNA 時開始，使得第一股能夠穿過奈米孔。當 DNA 股穿過奈米孔 116 時，感測器（例如電極）根據序列上下文以及相關鹼基修飾（稱作一維（1D）讀數））量測隨時間推移（毫秒，ms）之離子電流變化（以皮安（pA）為單位）。曲線圖 120 展示實例電流信號與時間。在另一實施例中，髮夾序列轉接子將用於將第一股及其互補股共價栓繫在一起以形成雙股 DNA 分子。因此，在測序期間，測序雙股 DNA 分子之一個股，接著測序互補股（稱作 $1D^2$ 或二維（2D）讀數），此可潛在地改良測序準確性。在又一實施例中，藉由蛋白質栓繫之雙股 DNA 分子之一個末端將增加在完成測序同一分子之第一股之後測序互補股之可能性，從而產生 $1D^2$ 讀數。

【0073】 原始信號（例如曲線圖 120 中之電流）用於鹼基識別及鹼基修飾

分析。在一些實施例中，藉助於機器學習方法，例如但不限於遞迴神經網路（RNN）、卷積類神經網路（CNN）、隱藏式馬可夫模型（HMM）或其一或多個組合實施鹼基識別及鹼基修飾分析。

【0074】在一個實施例中，吾等研發出一種處理藉由奈米孔測序產生之電流信號的新穎方法，且分析經處理之信號以基於卷積類神經網路（CNN）或遞迴神經網路（RNN）判定單分子水準下之 DNA 甲基化。

II. 電流信號分析

【0075】可分析來自奈米孔測序之電流信號以鑑別鹼基修飾。然而，圖 1 中描述之機器學習方法不僅僅使用使用奈米孔獲得之原始電流之輸入。本文所描述之實施例使用電流之部分之一或多個統計值。此等一或多個統計值之向量可與對應於核苷酸之窗口的其他資訊（包含核苷酸之標識及核苷酸之位置）組合。核苷酸之位置可相對於窗口內之目標位置，其中目標位置為偵測到修飾或缺失的位置。可包含核苷酸之窗口之資訊以及核酸分子之區域中之電信號之統計值以形成輸入資料結構。在此等輸入資料結構上訓練之模型可用於偵測鹼基修飾。

A. 電流向量參數

【0076】對於穿過奈米孔之核苷酸股，吾人將偵測 ***N*** 個事件（亦即，與鑑別出之不同核苷酸相關之信號區段）。在一個實施例中，一個事件對應於在鹼基識別期間鑑別出的一個核苷酸，其中在特定單位時間（例如，毫秒）採樣一系列電信號。在一個實例中，以 4 kHz 之頻率對電流進行採樣（Rang 等人，《基因體生物學（Genome Biol.）》 2018;19:90）。在另一實施例中，一個事件對應於在鹼基識別期間鑑別出的超過一個核苷酸，其中以特定時間速率採樣一系列電信號。

【0077】**圖 2** 展示電流信號之曲線圖。y 軸為以皮安為單位之電流振幅。x

軸為以毫秒為單位之時間。圓點（例如圓點 204）展示個別信號量測。通過相鄰圓點之線（例如線 208）指示具有與核苷酸相關之信號量測的信號區段（例如線 208 之 A）。對於事件 i，假設存在 m_i 電流信號，關於事件 i 之電流信號 j 之振幅由 P_{ij} 表示。在一個實施例中，對於一個核苷酸，包含 $X1$、$X2$、$X3$、$X4$ 及 $X5$ 之信號特徵向量用於表徵與該核苷酸相關之電信號之模式。$X1$、$X2$ 及 $X3$ 之定義在圖 2 中示出。$X1$ 為 P_{ij} 之平均值。$X2$ 為 P_{ij} 之標準差。$X3$ 為 P_{ij} 之中值。$X4$ 為電流相對於 $X3$ 之絕對偏差的中值（在圖 2 中僅標記一個絕對偏差）。$X5$ 為 $X1$ 相對於電流信號之平均值的差除以標準差。$X5$ 可視為一個區段之電流信號之 z 評分。

[0078] 在一個實施例中，P_{ij} 可為正規化信號。正規化可涉及藉由使用與部分或整個核苷酸股相關的最小值及最大值自初始範圍重新調整電流信號以使得經正規化信號值在 0 與 1 之範圍內。正規化可涉及重新調整電流信號以使得經正規化信號值之平均值為 0 且標準差為 1。正規化可涉及藉由使用與部分或整個核苷酸股相關的中位值及偏差來重新調整電流信號。

[0079] $X1$ 及 $X2$ 表示與事件 i 相關之 P_{ij} 之平均值及標準差。

[0080] $X1$ 藉由以下定義：

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

[0081] $X2$ 藉由以下定義：

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

[0082] $X3$ 藉由以下界定：

$$X3 = \text{中值}(P_{ij}),$$

其中 i 在 l 至 r 範圍內，包含含測驗基修飾分析之驗基周圍的事件（例如 CpG 位點處之甲基化）。變數 l 及 r 表示一系列事件（對應於核苷酸序列）之窗

1 的左右側。*l* 與 *r* 之間的核苷酸序列應通常比下文論述之電流信號模式之整合式表示短（稱為 IPM）長。對於既定事件 *i*，*j* 在 1 至 m_i 之範圍內。*X3* 可為用於判定所有區段之中值電流信號。*X3* 對於所有區段可為相同的值，因為 *X3* 係使用不止一個區段之電流測定的。在一些實施例中，*X3* 可用於特定窗口。在其他實施例中，*X3* 可為跨越多個窗口之中值。

[0083] *X4* 藉由以下定義：

$$X4 = \text{中值}(|P_{ij} - X3|),$$

其中中表示絕對值；且 *i* 在 *l* 至 *r* 範圍內，包含含胞嘧啶修飾分析之胞嘧啶周圍的事件（例如，CpG 位點處之甲基化）。對於既定 *i*，*j* 在 1 至 m_i 之範圍內。*X4* 可為用於判定所有區段之電流信號之絕對偏差的中值。*X4* 可使用不止一個區段之電流（例如使用所有採樣之電流值）來計算且因此對於所有區段可為相同的值。

[0084] *X5* 藉由以下定義：

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

其中 $\mu = \frac{\sum_{M=1}^{r} m_r p_M}{M-1}$ 且 $\sigma = \sqrt{\frac{\sum_{M=1}^{r} m_r (p_M - \mu)^2}{M-1}}$

其中 i 在 1 至 r 範圍內，包含含胞嘧啶修飾分析之胞嘧啶周圍的事件（例如 CpG 位點處之甲基化）。對於既定 i，j 在 1 至 m_i 之範圍內。M 為針對在 1 至 r 範圍內之事件採樣之電流信號的總數目。與複數個電流信號相關且用於測定 X3 之區域大小可為 DNA 片段之大小。舉例而言，若 DNA 片段為 500 bp，則區域之大小為 500。若片段為 300 bp，則區域之大小為 300。在一些實施例中，將 DNA 片段進一步劃分成較小子片段以測定 X3 可能為有用的。用於測定 X3 之區域大小可為 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 等。

【0085】 ***X1*** 及 ***X2*** 可用於反映在事件 ***i*** 內的信號變化，表示各核苷酸之電信號之局部模式。***X3***、***X4*** 及 ***X5*** 可用於反映事件 ***i*** 相對於在 ***l*** 至 ***r*** 範圍內之其他周圍事件的信號變化。在一些實施例中，周圍事件可為查詢鹼基修飾分析之鹼基之 X-nt 上游及 Y-nt 下游。X 可包含但不限於 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 10000；Y 可包含但不限於 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 10000。在一個實施例中，周圍事件可為穿過奈米孔之整個核苷酸股。

B. 單股分析

【0086】 **圖 3** 展示電流信號之曲線圖。y 軸為以皮安為單位之電流振幅。x 軸為以毫秒為單位之時間。跡線 304 為隨時間推移之電流振幅。信號區段（例如區段 308）為跡線 304 之與核苷酸相關的部分。電流變化將視穿過奈米孔之不同核苷酸而變化。奈米孔測序中之鹼基識別通常依賴於將電流信號轉化成不同的局部靜止狀態（亦即，事件）。將電流信號轉化成不同事件的過程稱為電信號分段。離子電流變化包含但不限於對應於信號區段中之一或多個核苷酸的事件振幅（例如以皮安，pA 為單位量測）、離子電流之方向、對應於信號區段中之一或多個核苷酸的電流事件之持續時間、離子電流之變化率及不同信號區段之間的相對振幅。振幅可指電流之強度或量值且不一定暗示交流電。使用例如命名 Tombo 之軟體將彼等電流事件分配至不同鹼基（Stoiber 等人，bioRxiv.2016;

doi.org/10.1101/094672）。一個核苷酸將與一系列具有不同振幅之事件相關。此類工具（Tombo）試圖測試分配至兩個樣本之間的基因體鹼基的奈米孔信號的差異以基於曼-惠特尼 U 測試（Mann-Whitney U test）來推斷此類鹼基經修飾抑或未經修飾（Stoiber 等人，bioRxiv.2016; doi.org/10.1101/094672）。此工具（Tombo）不考慮上游及下游信號以及序列上下文，且不能分析單分子水準下之甲基化模式，因為來自不同序列讀數之所有信號會彙聚至基因體鹼基中。已比較 Tombo 之效能與諸如 Nanopolish 及 DeepSginal 之其他工具之彼等效能（Yuen 等人，bioRxiv.2020; doi: doi.org/10.1101/2020.10.14.340315）。

【0087】在一個實施例中，為表徵信號區段內與核苷酸相關之電流模式，計算該信號區段內之事件之彼等電流振幅的平均值（X1）及標準差（X2）。測定與整個分子相關之事件之電流振幅的中值（X3）及與整個分子相關之事件之電流振幅之中值絕對偏差（X4）藉由下式測定信號區段之正規化信號（X5）：

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

其中 X1 為該信號區段內與所討論之核苷酸相關之事件之彼等電流振幅的平均值；μ 為所研究之整個分子內的事件之彼等電流振幅的平均值；σ 為所研究之整個分子內的事件之彼等電流振幅的標準差。在一個實施例中，可在移除最大值及最小之指定小百分比之後得到平均值及標準差。

【0088】對於一個核苷酸，信號特徵向量，包含 X1、X2、X3、X4 及 X5，用於反映與該核苷酸相關的電信號之模式。舉例而言，區段 308 可具有[X1, X2, X3, X4, X5]之信號特徵向量。

【0089】X1 及 X2 表示在信號區段 *i* 內的事件之電流振幅之平均值及標準差。X3 表示與整個分子相關之事件之電流振幅的中值。X4 表示與整個分子相關之事件之電流振幅的中值絕對偏差。X5 表示信號區段 *i* 之正規化信號。

【0090】圖 4 為信號區段之長度之頻率的曲線圖。與核苷酸相關之電流事

件之長度（亦即，以毫秒為單位之持續時間）係在 x 軸上。長度之頻率展示於 y 軸上。圖 4 展示與核苷酸相關之各信號區段之長度為可變的，其中中值為 9（範圍：1 至 3540）。

【0091】 鹼基修飾將影響與其上游及下游核苷酸相關之電信號。在本揭示案中，吾等共同地利用與用於鹼基修飾分析之核苷酸相關之電流信號、與所關注之核苷酸附近的核苷酸相關之電流信號以及測序背景，以便改良效能。CpG 位點處之 DNA 甲基化（亦即，胞嘧啶之第 5 碳處之甲基化）為脊椎動物基因體中最常見的鹼基甲基化類型。對 CpG 位點處之 DNA 甲基化的分析用作本揭示案之說明性實例。

【0092】 **圖 5** 展示使用一個股之電流信號經由奈米孔測序來判定甲基化的方法。在方塊 504 處，提供雙股 DNA 分子。在方塊 508 處，使雙股 DNA 分子與適用於奈米孔測序之測序轉接子連接。在方塊 512 處，進行奈米孔測序。單雙股分子之一個股移動通過內嵌於膜中之孔隙，從而改變流動通過奈米孔之離子電流信號。在方塊 516 處，獲得電流信號。可例如藉由跨電極來量測離子電流信號。

【0093】 將藉由使用例如 Tombo 之分段步驟處理電流信號（Stoiber 等人，bioRxiv.2016; doi.org/10.1101/094672）。此等分段式電事件將分配至不同核苷酸。在方塊 520 處，建構整合式表示矩陣（IPM）。IPM 為電流信號模式之矩陣，其包含每個鹼基之電流信號、測序背景及跨越用於鹼基修飾分析之基因座附近或周圍的一系列核苷酸的測序位置資訊。在一個實施例中，與核苷酸相關之分段式電事件藉由信號特徵向量，亦即，[X1, X2, X3, X4, X5]描述。CpG 位點內之胞嘧啶及例如該胞嘧啶之上游及下游 10-nt（亦即，例如總共 21 nt）以及多個信號特徵向量用於形成電流信號模式之 IPM。出於說明之目的，5'-T[CCATGC]CAT**C**GTC[GATGCA]G-3'之 21-nt 序列用作一實例，得到 IPM 524。

為簡單起見，省略括號中之鹼基（由「…」表示）。對於與腺嘌呤之鹼基（「A」）對應的-2 位置，與「A」相關之信號特徵向量，[X1=1.7，X2=0.29, X3=24.2, X4=436, X5=-0.3]填寫在「-2」行與「A」列之間的對應單元格中。相同行中之其他單元格填寫為「0」。使用相同的規則填寫與 21-nt 序列上下文相關之每個核苷酸的剩餘信號特徵向量，由此形成 21-nt IPM。因此，此 IPM 將同時對電流信號模式、測序背景、測序位置以及隨時間推移而改變之模式進行編碼。源自甲基化及未甲基化 DNA 資料集之多個 IPM 用於訓練 CNN 或 RNN 模型，該模型隨後將用於判定測試樣本中 CpG 位點處之甲基化狀態。

【0094】 方塊 528 展示 CNN 分析。對於 CNN 分析，將 IPM 饋入輸入層中，接著為卷積層及輸出層之處理。CPG 之甲基化概率（亦即，輸出甲基化評分，在 0 至 1 之範圍內）係基於輸出層中之 sigmoid 函數來測定。此方法稱為 IPM-CNN。在一個實施例中，甲基化 CpG 位點（經 M.SssI 處理之 DNA）及未甲基化 CpG 位點（全基因體擴增（WGA）之 DNA）之 IPM 用於訓練 CNN 模型。自經 M.Sss 處理之 DNA 獲得的資料集中 CpG 位點之甲基化目標值定義為「1」，而自 WGA DNA 獲得的資料集中 CpG 位點之甲基化目標值定義為「0」。藉由經由迭代地更新模型參數使由 sigmoid 函數計算之輸出評分與所要目標輸出（二進位值：0 或 1）之間的總預測誤差最小化來獲得 IPM-CNN 之最佳參數。總預測誤差係藉由深度學習演算法（keras.io/）中之 sigmoid 交叉熵損失函數來測定。自訓練資料集得知之模型參數用於分析測試資料集中之甲基化狀態，輸出表明 CpG 位點被甲基化之可能性的概率性評分（亦即甲基化概率）。在一個實施例中，CNN 模型利用四個二維（2D）卷積層，各自具有 32、64、128、256 個核尺寸為 25 的過濾器。彼等卷積層使用矯正線性單元（ReLU）之激活函數。隨後應用批次正規化層。進一步增加一個扁平化層，接著丟棄速率為 0.5 之丟棄層，且接著為全連接層，該全連接層包括 200 個使用 ReLU 激活函數之神經元。最終

應用具有一個神經元之輸出層，利用 sigmoid 激活函數得到 CpG 位點甲基化之概率評分（亦即甲基化概率）。CNN 模型之程式係基於 Keras 深度學習框架（https://keras.io/）實施。

【0095】 方塊 532 展示 RNN 分析。對於 RNN 分析，將 IPM 饋入輸入層中，接著為長短期記憶（LSTM）層及輸出層之處理。CpG 之甲基化概率（在 0 至 1 之範圍內）係基於輸出層中之 sigmoid 函數來測定。此方法稱為 IPM-RNN。使用與 IPM-RNN 中使用之訓練程序類似的訓練程序，藉由經由迭代地更新模型參數而使由 sigmoid 函數計算之輸出評分與所要目標輸出（二進位值：0 或 1）之間的總預測誤差最小化來獲得 IPM-RNN 之最佳參數。自訓練資料集得知之模型參數用於分析測試資料集中之甲基化狀態，輸出表明 CpG 位點被甲基化之可能性的概率性評分（亦即甲基化概率）。在一個實施例中，將具有 LSTM 單元之 RNN 模型與兩個全連接隱藏層一起使用，該兩個全連接隱藏層各自具有 256 個隱藏節點。最後一層之後為具有丟棄速率 0.2 之丟棄層。最終應用具有一個神經元之輸出層，利用 sigmoid 激活函數得到 CpG 位點甲基化之概率性評分（亦即甲基化概率）。CNN 模型之程式係基於 Keras 深度學習框架（keras.io/）實施。

C. 雙股分析

【0096】 **圖 6** 展示使用兩個 DNA 股之電流信號經由奈米孔測序判定甲基化的方法。在一個實施例中，當雙股 DNA 分子以第二核苷酸股（稱為互補股或克里克股（Crick strand））將在第一核苷酸股（稱為沃森股（Watson strand））完成穿過奈米孔之後緊接著穿過同一奈米孔的方式測序時，可獲得此類雙股 DNA 分子之兩個核苷酸股的電流信號。用於在同一奈米孔對雙股 DNA 之兩個核苷酸股進行依序測序的此類技術稱為 $1D^2$ 或 2D 測序。在方塊 604 處，提供雙股 DNA 分子。在方塊 608 處，使雙股 DNA 分子與適用於奈米孔測序之測序轉接子連接。在方塊 612 處，使單雙股分子之一個股移動通過內嵌於膜中之孔隙，

接著使互補股移動通過該孔隙。在方塊 616 處，獲得每個雙股 DNA 分子之兩個股的電流信號。可藉由跨電極來量測離子電流信號。所獲得的電流信號用於推導 DNA 分子之核苷酸資訊，該資訊係使用 Guppy（牛津奈米孔科技有限公司（Oxford Nanopore Technologies Ltd））進行測序（亦即，鹼基識別)。在一些實施例中，可使用其他鹼基識別工具，包含但不限於 Albacore（nanoporetech.com/）、WaveNano（Wang 等人，《定量生物學（Quantitative Biology.）》，2018;6:359-368）、Chiron（Teng 等人，《大數據科學（GigaScience.）》2018;7:giy037）、Flappie（github.com/nanoporetech/flappie）、Scrappie（github.com/nanoporetech/scrappie）等。

【0097】以特定時間速率（例如毫秒）採樣點電流信號將分配至所偵測的不同核苷酸用於鹼基修飾分析。將藉由使用例如 Tombo 之分段步驟處理電流信號（Stoiber 等人，bioRxiv.2016; doi.org/10.1101/094672）。此等分段式電事件將分配至不同核苷酸。在方塊 620 處，建構包含每個雙股 DNA 分子之兩個股的整合式表示矩陣（IPM）。在一個實施例中，與核苷酸相關之分段式電事件藉由信號特徵向量，亦即，[***X1, X2, X3, X4, X5***]描述。獲得互補股之對應鹼基的信號特徵向量，亦即[***X1', X2', X3', X4', X5'***]。CpG 位點內之胞嘧啶及例如該胞嘧啶之上游及下游 10-nt（亦即，例如總共 21 nt）以及多個信號特徵向量用於形成電流信號模式之 IPM。獲得相同的雙股 DNA 分子之互補股中的對應鹼基之 IPM。合併自沃森股及克里克股得到的 IPM，進而形成具有較高維度之新穎 IPM 矩陣以用於鹼基修飾分析。

【0098】在一些實施例中，可使用其他計算工具將電流信號指配至不同核苷酸，包含 NanoMod（Liu 等人，《英國醫學委員會基因體學（BMC Genomics.）》2019;20:78）、Albacore（nanoporetech.com/）、Chiron（Teng 等人，《大數據科學（GigaScience.）》2018;7:giy037）、Nanopolish（Simpson 等人，《自然方法

學（Nat Methods.）》2017;13:407-410）、Scrappie（https://github.com/nanoporetech/scrappie）、UNCALLED（Kovaka 等人，《自然生物技術（Nat Biotechnol.）》2020; doi:10.1038/s41587-020-0731-9）等。此等計算工具及其他描述用於雙股分析之技術可用於單股分析。

【0099】 出於說明之目的，5'-T[CCATGC]CATCGTC[GATGCA]G-3'之 21-nt 序列作為一實例用作 IPM 624 之基礎。IPM 624 可類似於 IPM 524，但包含沃森股及克里克股兩者。為簡單起見，省略括號中之鹼基（由「…」表示）。對於與沃森股中之腺嘌呤之鹼基（「A」）對應的-2 位置，與「A」相關之信號特徵向量，亦即[X1 = 1.7, X2 = 0.29, X3 = 436, X4 = 24.2, X5 = -0.3]填寫在由「沃森股」指示之區域中的「-2」行與「A」列之間的對應單元格中。對於其在互補股（亦即克里克股）中之對應鹼基「T」，與「T」相關之信號特徵向量，[X1' = -1.9, X2' = 0.23, X3' = 24.2, X4' = 436, X5'= -1.4]填寫在由「克里克股」指示之區域中的「-2」行與「T」列之間的對應單元格中。相同行中之其他單元格填寫為「0」。在一些實施例中，可改變信號特徵向量中要素之次序。舉例而言，可使用[X2, X1, X3, X4, X5]、[X2, X3, X4, X5, X1]、[X1, X3, X5, X4, X2]或其他組合。在一些實施例中，信號特徵向量之大小可不侷限於 5。舉例而言，藉由增加更多處理之電信號特徵或原始電信號，信號特徵向量之大小可包含但不限於 6、7、8、9、10、15、20、30、40、50、100 等。藉由編輯或刪除信號特徵向量中之一些特徵，信號特徵向量之大小可包含但不限於 1、2、3、4。

【0100】 使用相同的規則填寫與 21-nt 序列上下文相關之每個核苷酸之剩餘信號特徵向量，由此形成 21-nt IPM。因此，此 IPM 將同時對電流信號模式、測序背景、測序位置以及隨時間推移而改變之模式進行編碼。源自甲基化及未甲基化 DNA 資料集之多個 IPM 用於訓練 CNN 或 RNN 模型，該模型隨後將用於判定測試樣本中 CpG 位點處之甲基化狀態。

〔0101〕方塊 628 展示 CNN 分析。在實施例中，CNN 模型利用四個二維（2D）卷積層，各自具有 32、64、128、256 個核大小為 1×25 的過濾器。後等卷積層使用矯正線性單元（ReLU）之激活函數。隨後應用批次正規化層。進一步增加一個最大池化層，接著去乘速率為 0.5 之去乘層，且接著為全連接層，該全連接層包括 200 個使用 ReLU 激活函數之神經元。最終應用具有一個神經元之輸出層，利用 sigmoid 激活函數得到 CpG 位點甲基化之概率性評分（亦即甲基化概率）。CNN 模型之程式係基於 Keras 深度學習框架（keras.io/）實施。在一典實施例中，各人可改變核大小 $m \times m$，其中「m」可包含但不限於 1、2、3、4、5、10、15、20、30、35、40、45、50、100 等，且「m」可包含但不限於 1、2、3、4、5、10、15、20、30、35、40、45、50、100 等。

〔0102〕圖 7 為核大小對驗基修飾分析之效能的影響的表。第一行展示不同的核大小。第二行展示訓練資料集之 AUC（ROC 接受者操作特徵曲線下面積）。第三行展示測試資料集之 AUC。圖 7 展示一系列核大小，諸如 1×5、1×10、1×15、1×20 及 1×25 將在區分甲基化 CpG 位點及未甲基化 CpG 位點中提供相當的效能，如藉由分別為 0.96、0.96、0.97、0.96 及 0.96 之 AUC 所指示。

〔0103〕方塊 632 展示 RNN 分析。在實施例中，將具有 LSTM 單元之 RNN 模型與兩個全連接隱藏層一起使用，該兩個全連接隱藏層各自具有 256 個隱藏節點。LSTM 隱藏單元之生流輸出係藉由生流輸入及儲存於 LSTM 單元中之先前資訊來測定。作為一個實例，與 21 nt IPM 之第一列指示的位置相關之生物特徵向量 $|X1, X2, X3, X4, X5|$ 被視為在特定時間步長下的 LSTM 單元之輸入 X_t。正向 LSTM RNN 將基於如下之運算根據時間步長遞迴地計算隱藏層 $\overrightarrow{h}$（Gers 等人，《IEEE 神經網路彙刊（IEEE Transactions on Neural Networks）》2001;12:1333-1340）：

$$A_{t,r} = sigmoid\left(W_{xa,r}X_{t,r} + W_{ha,r}H_{t-1,r} + W_{ca,r} \odot C_{t-1,r} + b_{a,r}\right),$$

$$I_{t,r} = sigmoid\left(W_{xf,r}X_{t,r} + W_{hf,r}H_{t-1,r} + W_{cf,r}\odot C_{t-1,r} + b_{f,r}\right),$$

$$C_{tr} = I_{t,r}\odot C_{t-1,r} + A_{t,r}\odot tanh\left(W_{xc,r}X_{tr} + W_{hc,r}\odot C_{tr} + b_{c,r}\right),$$

$$O_{t,r} = sigmoid\left(W_{xo,r}X_{t,r} + W_{ho,r}H_{t-1,r} + W_{co,r}\odot C_{t,r} + b_{o,r}\right),$$

$$H_{t,r} = O_{t,r}\odot tanh\left(C_{t,r}\right)。$$

[0104] 反向 LSTM RNN 將其於如下之運算根據時間步長遞迴地計算隱藏向 H（Gers 等人，《IEEE 神經網路彙刊》2001;12:1333-1340）：

$$A_{t,B} = sigmoid\left(W_{xa,B}X_{t,B} + W_{ha,B}H_{t-1,B} + W_{ca,B}\odot C_{t-1,B} + b_{a,B}\right),$$

$$I_{t,B} = sigmoid\left(W_{xf,B}X_{t,B} + W_{hf,B}H_{t-1,B} + W_{cf,B}\odot C_{t-1,B} + b_{f,B}\right),$$

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

$$O_{t,B} = sigmoid\left(W_{xo,B}X_{t,B} + W_{ho,B}H_{t-1,B} + W_{co,B}\odot C_{t,B} + b_{o,B}\right),$$

$$H_{t,B} = O_{t,B}\odot tanh\left(C_{t,B}\right)。$$

其中 W 及 b 為權重及偏差；X is 輸入向量；A 為輸入門之激活向量；F 為遺忘門之 sigmoid 函數；C 為單元狀態；O 為輸出門之 sigmoid 函數且 H 為 LSTM 隱藏單元之輸出。

[0105] 將正向及反向 LSTM RNN 單元之輸出合併。

$$Z_t = H_{t,F}\oplus H_{t,B}。$$

[0106] LSTM RNN 輸出之最後一層之後為具有去棄速率 0.2 之去棄層。最終應用具有一個神經元之輸出層，利用 sigmoid 激活函數得到 CpG 位點甲基化之概率性評分（亦即甲基化概率）。CNN 模型之程式係基於 Keras 深度學習框架（keras.io/）實施。

D. 參數分析

[0107] 分析不同雪流向量參數及不同窗口大小對 AUC（ROC 接受者操作特徵曲線下面積）之影響。下等根據本揭示案提供之實施例基於 IPM CNN 模型分析在使用 IPM 中之不同參數的情況下的區分能力。為此目的，分別分析來

自 WGA DNA 及經 M.SssI 處理之 DNA 資料集的 8,282 個分子（38,238 個 CpG 位點）及 8,247 個分子（39,708 個 CpG 位點）。

【0108】 **圖 16** 展示不同參數組合對 AUC 之影響的曲線圖。電流向量參數之不同組合係在 x 軸上，且 AUC 係在 y 軸上。圖 16 展示使用 IPM 中但不限於 X1、X2、X3、X4 及 X5 之不同參數組合會產生 CpG 甲基化分析之不同效能。舉例而言，使用 IPM 中之 X1 產生 0.954 之 AUC，而 IPM 中 X1 及 X2 之組合產生 0.893 之 AUC。IPM 中 X1、X2 及 X3 之組合使 AUC 提高至 0.963。IPM 中 X1、X2、X3 及 X4 之組合使 AUC 進一步提高至 0.978，接著在此實例中使用 X1、X2、X3、X4 及 X5 的情況下使效能平穩在 0.977 之 AUC。因此，在一些實施例中，IPM 中之不同參數組合將允許吾人測定在區分甲基化及未甲基化 CpG 位點中之所需效能。

【0109】 測試單獨地而非組合地使用 X1、X2、X3、X4 及 X5。單獨地使用 X1、X2、X3、X4 及 X5 之結果分別為 0.95、0.92、0.98、0.88 及 0.95 之 AUC。X3（亦即，區域中之 P_{ij} 之中值）得到 0.98 之高 AUC。高 AUC 可至少部分為完整片段水準上之甲基化差異的結果。所使用之資料集涉及 WGA（完全未甲基化）及 M.Sssl（完全甲基化）。然而，實際上片段將不為完全甲基化或完全未甲基化的。對於並非完全甲基化或完全未甲基化之樣本，單獨使用 X3 可不會產生高 AUC。

【0110】 **圖 17** 展示窗口大小對 AUC 之影響的曲線圖。x 軸展示以核苷酸為單位之窗口大小。y 軸展示 AUC。IPM 中使用之核苷酸數目（又稱窗口大小）將捕獲在奈米孔測序期間產生之電流信號的不同資訊內容，且可影響甲基化分析之效能。圖 17 展示使用 IPM-CNN 模型區分甲基化及未甲基化 CpG 位點的效能呈現：隨著 IPM 中使用之核苷酸數目自 1 nt 增加至 10 nt，AUC 自 0.715 逐步增加至 0.969。在此實例中，在 7 nt 之窗口大小處達到效能平穩。因此，在一些

實施例中，調節 IPM 之窗口大小將允許吾人測定在區分甲基化及未甲基化 CpG 位點中之所需效能。

【0111】實施例可不需要使用產生最高 AUC 的電流向量參數或窗口大小之組合。較低 AUC 對於某些用途可能足夠，或較高 AUC 可不值得與額外參數相關的額外計算及儲存成本。此外，可調節不同參數以實現期望 AUC、特異性及/或靈敏度。舉例而言，較大窗口大小可用於補償較少使用 X1、X2、X3、X4 及 X5 中之參數。

E. 6mA 修飾之偵測

【0112】為測定電流信號分析對除 5mC 以外之修飾的適用性，使用電流信號分析來偵測 N6-甲基腺嘌呤（6mA）。

【0113】**圖 18** 展示使用一個股之電流信號經由奈米孔測序來判定 6mA 甲基化的方法。圖 18 類似於展示用於判定 5mC 甲基化之方法的圖 5。在方塊 1804 處，提供雙股 DNA 分子。在方塊 1808 處，使雙股 DNA 分子與適用於奈米孔測序之測序轉接子連接。在方塊 1812 處，進行奈米孔測序。在方塊 1816 處，獲得電流信號。在方塊 1820 處，建構整合式表示矩陣（IPM）。方塊 1804 至 1820 可與方塊 504 至 520 相同。

【0114】出於判定 6mA 甲基化之說明目的，5'-G[TACCCG]GGT**A**CTG[TCTAGA]G-3'之 21-nt 序列作為一實例用作 IPM 之基礎，以進行甲基化分析之核苷酸 A（例如對應於 0 位置）為中心。IPM 1824 展示使用 21-nt 序列之結果。為簡單起見，省略括號中之鹼基（由「…」表示）。對於與一個股中之腺嘌呤之鹼基（「A」）對應的 0 位置，與「A」相關之信號特徵向量（亦即，[X1 = 0.39, X2 = 0.04, X3 = 389, X4 = 46.3, X5 = 0.32]）填寫在矩陣之「0」行與「A」列之間的對應單元格。相同行中之其他單元格填寫為「0」。在一些實施例中，可改變信號特徵向量中要素之次序。舉例而言，可使用[X2, X1,

X3, X4, X5]、[X2, X3, X4, X5, X1]、[X1, X3, X5, X4, X2]或其他組合。在一些實施例中，信號特徵向量之大小可不僅為 5。舉例而言，藉由增加更多處理之電信號特徵或原始電信號，信號特徵向量之大小可包含但不限於 6、7、8、9、10、15、20、30、40、50、100 等。藉由編輯或刪除信號特徵向量中之一些特徵，信號特徵向量之大小可包含但不限於 1、2、3 或 4。

【0115】 使用相同的規則填寫與 21-nt 序列上下文相關之每個核苷酸之剩餘信號特徵向量，由此形成 21-nt IPM。因此，此 IPM 將同時對電流信號模式、測序背景、測序位置以及隨時間推移而改變之模式進行編碼。源自與核苷酸 A 相關聯之甲基化及未甲基化 DNA 資料集之多個 IPM 用於訓練 CNN 或 RNN 模型，該模型隨後將用於判定測試樣本中 A 位點處之甲基化狀態。方塊 1828 展示 CNN 分析，且方塊 1832 展示 RNN 分析。此等方塊可與方塊 528 及 532 相同。

【0116】 為測試上文示出之吾等方法（IPM-CNN 或 IPM-RNN）是否能夠判定腺嘌呤甲基化（6mA），吾等下載包括來自先前研究（Rand 等人，《自然方法》2017；14:411-413）之 pUC19 質體 DNA 之奈米孔測序結果的兩個公共資料集。第一資料集（6mA 資料集）係由在含有 *dam* 及 *dcm* 甲基轉移酶兩者之大腸桿菌（*E.coli*）中生長之 pUC19 質體 DNA 產生，其中所有 GATC 模體經推測為 A 位點均甲基化。第二資料集（uA 資料集）係由用未經修飾之核苷酸進行 PCR 擴增的 DNA 產生，其中所有 A 位點經推測為未甲基化。在訓練程序中，吾等使用 IPM-CNN 模型分析來自 6mA 資料集的 2052 個含有 GATC 模體之分子及來自 uA 資料集的 2081 個分子。

【0117】 **圖 19** 展示使用 IPM-CNN 模型得到的 AUC。x 軸展示特異性。y 軸展示靈敏度。線 1904 展示訓練資料集的結果。訓練資料集之 AUC 為 0.94。在訓練程序中，吾等將訓練之 IPM-CNN 模型應用於來自 6mA 資料集的 522 個含有 GATC 模體之分子及來自 uA 資料集的 481 個分子。測試資料集之 AUC 為

0.92。另外，當使用 IPM-RNN 模型時，對於訓練及測試資料集兩者均得到 0.89 之 AUC。此等資料表明 IPM-CNN 及 IPM-RNN 可允許區分 6mA 位點與未甲基化 A 位點。

【0118】在實施例中，用於人類或非人類 DNA 之 6mA 判定的訓練資料集可基於分別使用 6mA 核苷酸及未甲基化 A 核苷酸進行 PCR 擴增來建構。在幾個 PCR 週期之後，大部分 DNA 分子將攜載 6mA 核苷酸以用於由 6mA 核苷酸進行擴增之 DNA 產生之資料集，而大部分 DNA 分子將攜載未甲基化 A 核苷酸以用於由未甲基化 A 核苷酸進行擴增之 DNA 產生之資料集。此兩種類型之資料集可用於訓練 CNN 及/或 RNN 模型以判定測試樣本中 A 核苷酸之甲基化狀態。

【0119】使用電流信號分析偵測除 5mC 之外的 6mA 證實此分析適用於其他甲基化類型。因此，此等方法應準確地偵測本文所描述之其他甲基化。

F. 人類個體之非腫瘤與腫瘤組織之間的 CpG 甲基化分析

【0120】藉由使用本文所描述之實施例判定的位點之甲基化可用於區分不同類型的組織。使用根據本揭示案之實施例的 IPM-RNN 模型，吾等分析源自鼻咽癌（NPC）腫瘤及白血球層樣本之細胞 DNA 分子的甲基化模式。為此目的，吾等使用來自 NPC 腫瘤之 147 個分子，其中中值大小為 4,406 bp（四分位數範圍（IQR）：1,962 至 8,128 bp）且中值為 32 個 CpG/分子（IQR：13 至 61）。吾等分析來自白血球層之另外 147 個分子，其中中值大小為 6,823 bp（四分位數範圍（IQR）：2,515 至 9,304 bp）且中值為 49 個 CpG/分子（IQR：23 至 118）。

【0121】**圖 20** 展示來自白血球層樣本及 NPC 腫瘤組織樣本之 DNA 分子的比較圖。x 軸展示組織類型。y 軸展示呈百分比形式的甲基化程度。發現白血球層中之單分子甲基化程度（亦即，分子中判定為甲基化之 CpG 位點的百分比）（中值：74.8%；IQR：71.1%至 80.1%）顯著高於 NPC 腫瘤中之單分子甲基化程度（中值：50；IQR：45.7 至 53.1）（*P* 值<0.0001，威爾卡森秩和檢定（Wilcoxon

rank-sum test））。源自腫瘤組織之 DNA 分子呈現為低甲基化，其與基於短讀數亞硫酸氫鹽測序之先前結論一致（Chan 等人，《美國國家科學院院刊》 2013; 110:18761-8）。然而，本文所述之新穎奈米孔測序技術允許對幾乎整個長 DNA 分子進行測序，且分析 DNA 分子之甲基化模式。舉例而言，奈米孔測序可分析大小大於 600 bp 之 DNA 分子，其不能藉由短讀數測序平台（例如 Illumina）進行查詢。

【0122】 **圖 21** 示出腫瘤 DNA 分子及白血球層 DNA 分子中之甲基化模式。實心黑色圓（例如圓 2104）指示甲基化 CpG 位點。空心圓（例如圓 2108）指示未甲基化 CpG 位點。圓展示 CpG 位點相對於所分析之 DNA 分子之 5'端的相對位置（亦即，圖中 DNA 分子之左側更接近 5'端）。如圖 21 中所示，相比於源自白血球層樣本之彼等 DNA 分子，源自腫瘤組織之 DNA 分子傾向於在分子中攜載更多未甲基化 CpG 位點。僅 5.4%的來自白血球層樣本之分子具有＜50%之單分子甲基化程度及 2,091 bp 之中值長度。相比之下，39.5%的來自 NPC 腫瘤組織之分子具有＜50%之單分子甲基化程度及 2,924 bp 之中值長度。DNA 分子之長度在 897 bp 至 10,424 bp 範圍內。

【0123】 此等資料展示本文所描述之用於偵測甲基化之奈米孔測序技術可用於單分子甲基化模式分析以區分來自組織活檢體樣本的各 DNA 分子之組織來源（例如非腫瘤 DNA 與腫瘤 DNA 分子）。組織活檢體之單分子甲基化模式分析將允許檢查腫瘤級別或亞型、監測癌症或其他疾病之治療、評估器官異常（例如腎臟衰竭）等。

G. 胎兒與母體 DNA 分子之間的分析

【0124】 藉由使用本文所描述之實施例判定的位點之甲基化可用於區分胎兒與母體 DNA 分子。根據 IPM-CNN 模型，吾等藉由 1,262 個胎兒特異性游離 DNA 分子（中值大小：530 bp；IQR：361 至 779 bp）及 6,108 個母體特異性

游離 DNA 分子（中值大小：668 bp；IQR：448 至 1,089 bp）之至少 5 個 CpG 位點，利用母體白血球層與胎盤組織之間的 SNP 資訊來判定單分子甲基化模式，該等分子係獲自妊娠三個月之孕婦。此孕婦之血漿 DNA 中之胎兒 DNA 分數為 26.0%。

【0125】 圖 **22** 展示母體特異性與胎兒特異性 DNA 分子之間的單分子甲基化程度。x 軸展示游離 DNA 分子之類別：母體特異性或胎兒特異性。y 軸展示呈百分比形式的單分子甲基化程度。單血漿 DNA 分子之中值甲基化程度（亦即，分子中判定為甲基化之 CpG 位點的百分比）對於胎兒特異性游離 DNA 分子為 66.6%（IQR：28.5%至 86.6%），其顯著低於母體特異性游離 DNA 分子之中值甲基化程度（中值：78.5%；IQR：50%至 93.7%）（*P* 值：<0.0001，曼-惠特尼 *U 測試*）。結果表明使用游離 DNA 分子之甲基化資訊允許區分各血漿 DNA 分子之母體及胎兒來源。

【0126】 另外，藉由比較由 IPM-CNN 模型判定之甲基化模式與如 2021 年 2 月 5 日申請之美國專利申請案第 17/168,950 號中所描述之白血球層及胎盤組織之各別參考甲基化模式，吾人可得到 0.87 之 AUC，以用於區分孕婦中胎兒及母體來源之血漿 DNA 分子。

【0127】 **圖 23** 展示基於由 IPM-CNN 模型判定之甲基化模式對孕婦中之游離 DNA 分子進行胎兒及母體來源分析的 ROC 曲線。x 軸為特異性，且 y 軸為靈敏度。

III. 用於評估基於 IPM 之甲基化測定的資料集

【0128】 未甲基化資料集含有經由全基因體擴增（WGA）製備的經擴增 DNA 之測序結果（表示為 WGA DNA 資料集）。使用 WGA 中之未經修飾之核苷酸得到幾乎不含鹼基修飾之擴增 DNA（除了少量輸入基因體 DNA 以外）。甲基化資料集含有在測序之前藉由 M.SssI（CpG 甲基轉移酶，自含有來自螺原體

屬菌株 MQ1 之甲基轉移酶基因的大腸桿菌菌株分離，將使雙股 DNA 中之所有 CpG 位點甲基化）處理之 DNA 的測序結果（表示為經 M.SssI 處理之 DNA 資料集）。M.SssI 甲基轉移酶致使 CpG 位點甲基化。

【0129】 為製備 WGA DNA 資料集，藉由將反應混合物（含有 phi29 反應緩衝液及 dNTP）在 95℃下之加熱塊中培育 5 分鐘接著冷卻至 4℃，將核酸外切酶抗性隨機引子預退火為 DNA 模板之 1 ng。接著將 phi29 聚合酶添加至反應混合物中且在 30℃下培育 4 小時。DNA 用 Ampure XP 珠粒純化且用 Qubit 螢光計定量。通常，200 ng DNA 可獲自 20 μl 反應物。

【0130】 為製備經 M.SssI 處理之 DNA 資料集，在 WGA 之後，將一半 DNA 用 M.SssI 酶處理。將甲基轉移酶反應緩衝液、S-腺苷甲硫胺酸（SAM）及 M.SssI 與 DNA 混合，且在 37℃下培育 2 小時。藉由在 65℃下加熱 20 分鐘使反應停止。連接測序套組（SQK-LSK109）(牛津奈米孔)用於庫製備.用 NEBNext FFPE DNA 修復混合物以及 NEBNext Ultra II 末端修復/dA-加尾模組處理 DNA。在 Ampure XP 珠粒清除之後，藉由添加轉接子混合物、連接緩衝液及 NEBNext Quick T4 DNA 連接酶將測序轉接子連接至經修復之 DNA。經連接之 DNA 用 Ampure XP 珠粒清潔且用短片段緩衝液洗滌。將庫再懸浮於溶離緩衝液中。R9.4.1 流通池用於對 WGA（樣本_01）及經 M.SssI 處理（樣本_02）庫中之每一者進行測序。流通池首先用含有沖洗繫鏈液（Flush Tether）及沖洗緩衝液之流通池預處理混合物進行預處理（primed）。接著藉由混合測序緩衝液、負載珠粒及 DNA 庫來製備負載庫之混合物。以逐滴方式將負載庫之混合物添加至流通池樣本口中。將負載之流通池插入 PromethION 中之狹槽中且使用默認參數測序 64 小時。

【0131】 針對樣本_01 及樣本_02，吾等分別獲得 15.6 及 15.3 百萬奈米孔測序讀數，其中 13.8（88.7%）及 13.8（90.7%）百萬讀數可藉由使用 Minimap2

（Li H，《生物資訊（Bioinformatics）》2018;34(18):3094-3100）與人類參考基因體（UCSC hg19）對準。樣本_01 及樣本_02 之中值讀數長度分別為 510 nt（四分位數範圍（IQR）：333 至 778 nt）及 606 nt（IQR：382 至 911 nt）。在一些實施例中，BLASR（Mark J Chaisson 等人,《BMC 生物資訊（BMC Bioinformatics）》2012; 13: 238）、BLAST（Altschul SF 等人,《分子生物學期刊（J Mol Biol.）》1990;215(3):403-410）、BLAT（Kent WJ,《基因體研究》2002;12(4):656-664）、BWA (Li H 等人,《生物資訊》2010;26(5):589-595）、NGMLR（Sedlazeck FJ 等人,《自然方法》2018;15(6):461-468）及 LAST（Kielbasa SM 等人,《基因體研究》 2011;21(3):487-493）可用於將經測序讀數與參考基因體進行比對。

【0132】 **圖 8** 為展示基於 IPM 用於訓練及測試 CNN 及 RNN 模型之測序分子之數目的表。第一行為資料集。經 M.SssI 處理之 DNA 為甲基化 DNA 資料集，且 WGA DNA 為未甲基化 DNA 資料集。第二行為用於訓練之分子數目及 CpG 位點數目。第三行為用於測試之分子數目及 CpG 位點數目。對於訓練資料集，吾等隨機使用分別來自經 M.SssI 處理之 DNA（甲基化 DNA）及 WGA DNA（未甲基化 DNA）的 7,989 及 8,052 個測序分子。此訓練資料集包括 38,470 個甲基化 CpG 位點及 37,150 個未甲基化 CpG 位點。對於測試資料集，吾等隨機使用分別來自經 M.SssI 處理之 DNA（甲基化 DNA）及 WGA DNA（未甲基化 DNA）的 4、826 及 5,041 個測序分子。此訓練資料集包括 9,716 個甲基化 CpG 位點及 11,444 個未甲基化 CpG 位點。

【0133】 **圖 9A 至圖 9D** 為使用 IPM-CNN 及 IPM-RNN 方法的 WGA DNA 與經 M.SssI 處理之 DNA 資料集之間的 CpG 之甲基化概率的盒狀圖。圖具有在 x 軸上之資料集。甲基化概率係在 y 軸上。圖 9A 及圖 9B 展示使用 IPM-CNN 分析之結果。圖 9A 展示對訓練資料集之 IPM-CNN 分析，其中經 M.SssI 處理之 DNA 資料集中 CpG 之甲基化概率（中值：0.99；IQR：0.987 至 0.999）顯著高

於 WGA DNA 資料集中之甲基化概率（中值：0.03；IQR：0.001 至 0.15）（*P* 值＜0.0001，曼-惠特尼 U 測試）。圖 9B 展示對測試資料集之 IPM-CNN 分析，其亦展示 WGA（中值：0.4；IQR：0.002 至 0.18）與經 M.SssI 處理之 DNA 資料集（中值：0.99；IQR：0.980 至 0.999）之間的 CpG 之甲基化概率的顯著差異（*P* 值＜0.0001，曼-惠特尼 U 測試）。

【0134】圖 9C 及圖 9D 展示使用 IPM-RNN 分析之結果。圖 9C 展示對訓練資料集之 IPM-RNN 分析，其中經 M.SssI 處理之 DNA 資料集中 CpG 之甲基化概率（中值：0.994；IQR：0.92 至 0.99 ）顯著高於 WGA DNA 資料集中之甲基化概率（中值：0.079；IQR：0.059 至 0.118）（P 值＜0.0001，曼-惠特尼 U 測試）。圖 9D 展示對測試資料集之 IPM-RNN 分析，其亦展示 WGA（中值：0.077；IQR：0.057 至 0.115）與經 M.SssI 處理之 DNA 資料集（中值：0.994；IQR：0.919 至 0.999）之間的 CpG 之甲基化概率的顯著差異（P 值＜0.0001，曼-惠特尼 U 測試）。此等結果表明，根據本揭示案提供之實施例用途由奈米孔測序產生之電信號判定 CpG 位點之甲基化狀態係可行的。在一個實施例中，0.5 之甲基化概率閾值可用於判定 CpG 位點之甲基化狀態。在使用此閾值的情況下，對於 IPM-CNN 分析，DNA 甲基化偵測之特異性及靈敏度對於訓練資料集分別為 96%及 91%，且對於測試資料集分別為 93%及 88%。對於 IPM-RNN 分析，DNA 甲基化偵測之特異性及靈敏度對於訓練及測試資料集兩者分別為 97%及 88%。在一些實施例中，可根據各種應用調節甲基化概率之閾值。

【0135】**圖 10A** 及**圖 10B** 展示接受者操作特徵（ROC）曲線分析。特異性展示於 x 軸上。靈敏度展示於 y 軸上。圖 10A 展示訓練資料集之結果。圖 10B 展示測試資料集之結果。IPM-CNN 結果以線 1004 及 1008 展示。IPM-RNN 結果以線 1012 及 1016 展示。DeepMod（Liu 等人，《自然通訊》2019; 10:2449）結果以線 1020 及 1024 展示。Nanopolish（Liu 等人，《自然通訊》2019; 10:2449）

結果以線 1028 及 1032 展示。基於 IPM 之 CNN 及 RNN 分析為訓練及測試資料集兩者供應良好效能，其中ROC曲線下面積(AUC)不小於0.95。相比於DeepMod（0.83）及 nanopolish（0.91），基於 IPM 之 CNN 及 RNN 模型在測試資料集中產生 ROC 曲線下面積（AUC）為 0.95 及 0.97 的更佳效能。發現基於 IPM 之 RNN 或 CNN 與其他包含 DeepMod 及 nanopolish 之工具的所有比較的 *P* 值（DeLong 測試）＜0.0001。此等結果表明 IPM-CNN 及 IPM-RNN 在 DNA 甲基化分析方面優於其他工具。

【0136】 **圖 11** 為針對不同分析之既定特異性之靈敏度的表。第一行展示分析類型。第二行展示靈敏度。第三行展示特異性。圖 11 展示在既定特異性下，IPM-CNN 及 IPM-RNN 分析實現高得多的靈敏度。舉例而言，在 90%之特異性下，IPM-CNN 及 IPM-RNN 分別分析實現 90%及 93%之靈敏度，而 DeepMod 及 nanopolish 方法分別實現僅 53%及 74%之靈敏度。在 95%之特異性下，IPM-CNN 及 IPM-RNN 分析分別實現 86%及 90%之靈敏度，而 DeepMod 及 nanopolish 方法僅分別實現 38%及 55%之靈敏度。在 99%之特異性下，IPM-CNN 及 IPM-RNN 分析分別實現 70%及 83%之靈敏度，而 DeepMod 及 nanopolish 分別實現僅 13%及 16%之靈敏度。此等結果進一步證實，序列區段之電流信號模式之整合式表示矩陣將大大地提高 DNA 甲基化測定之準確性。特定言之，IPM-RNN 在彼等方法中產生最佳的效能。

【0137】 在一些實施例中，對於 IPM，經歷鹼基修飾分析之鹼基周圍的 DNA 鏈段之長度可為對稱或不對稱的。舉例而言，該鹼基之上游 X-nt 及下游 Y-nt 可用於鹼基修飾分析。X 可包含但不限於 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

10000；Y 可包含但不限於 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 10000。X 及 Y 可相同或不同。

【0138】在一些實施例中，核酸中之鹼基修飾將根據本揭示案中之實施例在不同生物體中進行分析，該等生物體包含病毒、細菌、植物、真菌、線蟲、昆蟲及脊椎動物（例如人類）等。最常見的鹼基修飾為將甲基添加至不同位置之不同 DNA 鹼基中，亦即所謂的甲基化。在胞嘧啶、腺嘌呤、胸腺嘧啶及鳥嘌呤上均已發現甲基化，諸如 5mC（5-甲基胞嘧啶）、4mC（N4-甲基胞嘧啶）、5hmC（5-羥甲基胞嘧啶）、5fC（5-甲醯基胞嘧啶）、5caC（5-羧基胞嘧啶）、1mA（N1-甲基腺嘌呤）、3mA（N3-甲基腺嘌呤）、6mA（N6-甲基腺嘌呤）、7mA（N7-甲基腺嘌呤）、3mC（N3-甲基胞嘧啶）、2mG（N2-甲基鳥嘌呤）、6mG（O6-甲基鳥嘌呤）、7mG（N7-甲基鳥嘌呤）、3mT（N3-甲基胸腺嘧啶）及 4mT（O4-甲基胸腺嘧啶）。

【0139】在一些實施例中，可藉由不同的統計及/或數學模型分析電流信號模式之整合式表示矩陣，該等模型包含但不限於線性回歸、邏輯回歸、深度遞迴神經網路（例如長短期記憶，LSTM）、貝氏分類（Bayes classifier）、隱藏式馬可夫模型（HMM）、線性判別分析（LDA）、k 均值聚類、具有雜訊的基於密度之空間聚類應用（DBSCAN）、隨機森林演算法及支持向量機（SVM）。在又一實施例中，自然語言處理將應用於電信號分析以進行鹼基修飾分析。

【0140】在一些實施例中，可使用不同類型的奈米孔，包含但不限於生物奈米孔，諸如蛋白質 α-溶血素及其藉由蛋白質工程化技術之變化形式、由程式化細菌產生之孔蛋白、由合成材料、、石墨烯製成之固態奈米孔等。

【0141】在實施例中，此等方法可用於藉由參考諸如人類參考基因體（hg19）之參考基因體設計引導 RNA，例如長散佈核元件（LINE）重複序列來靶向大量共用同源序列之長 DNA 分子。在一個實例中，此類分析可用於分析孕婦之母體血漿中之循環游離 DNA，以偵測胎兒非整倍體（Kinde 等人《公共科學圖書館·綜合（PLOS One）》2012;7(7):e41162。在實施例中，經去活化或『死亡』Cas9（dCas9）及其相關單引導 RNA（sgRNA）可用於在不切割雙股 DNA 分子之情況下富集經靶向長 DNA。舉例而言，sgRNA 之 3'端可經設計以攜帶額外通用短序列。吾人可使用與彼通用短序列互補之經生物素標記之單股寡核苷酸以捕獲 dCas9 所結合的彼等目標長 DNA 分子。在另一實施例中，吾人可使用經生物素標記之 dCas9 蛋白或 sgRNA 或兩者以促進富集。

【0142】在實施例中，吾人可執行尺寸選擇以在對所關注之一或多個特定基因體區域無限制之情況下使用包含但不限於化學方法、物理方法、酶促方法、基於凝膠之方法及基於磁珠之方法或合併遠不止該等途徑的方法的途徑富集長 DNA 片段。

IV. 實例方法

【0143】此部分展示使用機器學習模型偵測鹼基修飾及訓練用於偵測鹼基修飾之機器學習模型的實例方法。

A. 修飾之偵測

【0144】**圖 12** 為與偵測核酸分子中核苷酸之修飾相關的例示性方法 1200 之流程圖。修飾可包含本文所描述之任何甲基化或任何氧化。氧化可為 8-側氧基-鳥嘌呤。在一些實施方案中，圖 12 之一或多個程序方塊可藉由系統（例如量測系統 1400）執行。在一些實施方案中，圖 12 之一或多個程序方塊可由與系統分離或包含該系統之另一裝置或裝置群組執行。另外或可替代地，圖 12 之一或多個程序方塊可藉由量測系統 1400 之一或多個組件執行，諸如偵測器 1420、邏

輯系統 1430、局部記憶體 1435、外部記憶體 1440、儲存裝置 1445 及/或處理器 1450。

【0145】 在方塊 1210 處，接收輸入資料結構。輸入資料結構可對應於樣本核酸分子中測序之核苷酸的窗口。藉由量測對應於核苷酸之電信號來對樣本核酸分子進行測序。電信號可為電流、電壓、電阻、電感、電容或阻抗。可藉由使用奈米孔進行測序。方法 1200 可進一步包含使用奈米孔對樣本核酸進行測序。奈米孔可為本文所描述之任何奈米孔。

【0146】 輸入資料結構可包含若干特性之值。針對窗口內之每個核苷酸的特性可包含核苷酸之標識、核苷酸相對於各個窗口內之目標位置的位置及包含電信號之對應於核苷酸之區段之第一區段統計值的向量。特性可包含核酸分子之等於或大於窗口之區域中電信號之第一區統計值。舉例而言，輸入資料結構可包含整合式表示矩陣[IPM]。

【0147】 核苷酸之標識可為鹼基（例如 A、T、C 或 G）。可經由利用奈米孔測序之鹼基識別技術來判定鹼基。鹼基識別技術可使電信號之區段與核苷酸相關聯。核苷酸之位置可為相對於目標位置之核苷酸距離。舉例而言，當核苷酸在一個方向上距離目標位置一個核苷酸時，位置可為+1，且當核苷酸在相反方向上距離目標位置一個核苷酸時，位置可為-1。

【0148】 第一區段統計值可表示電信號之對應於核苷酸之區段的平均值。在一些實施例中，第一區段統計值可表示電信號之對應於核苷酸之區段的電信號變化（例如標準差）。在實施例中，第一區段統計值可表示電信號之對應於核苷酸之區段的平均值的正規化值。正規化可包含重新調整以使得第一區段統計值在某一範圍（例如 0 至 1 之範圍）內。正規化可包含使用部分或所有核苷酸股之中值、平均值及/或偏差。正規化可為本文所描述之任何正規化，包含 z-評分（例如 X5）。

【0149】向量可包含第二區段統計值，其表示電信號之對應於核苷酸之區段的變化。向量可包含第三區段統計值，其表示第一區段統計值之正規化值。向量可包含本文所描述之變數 X1、X2 及 X5 的任何組合。

【0150】第一區統計值可表示該區域中電信號之平均值或中值。舉例而言，第一區統計值可為 X3。在實施例中，第一區統計值可表示電信號相對於該區域中之電信號之平均值或中值的變化的絕對值之中值或平均值。變化可為標準差。舉例而言，第一區統計值可為 X4。在一些實施例中，第一區統計值可為可選的。

【0151】輸入資料結構可進一步包含第二區統計值，其表示電信號相對於該區域中之電信號之平均值或中值的變化的絕對值的中值或平均值。舉例而言，第二區統計值可為 X4。

【0152】對於窗口內之不同核苷酸，第一區統計值可為相同值。對於窗口內之不同核苷酸，第二區統計值可為相同值。因此，第一區統計值及第二區統計值可視為與具有第一區段統計值及/或第二區段統計值之向量不同。替代地，對於每個核苷酸，向量亦可包含第一區統計值及/或第二區統計值可包含在向量中，即使該等值在核苷酸之間為相同的。在 IPM 524 及 IPM 624 中示出重複該等區域統計值的途徑。

【0153】該區域可在樣本核酸分子之一個股上。在一些實施例中，該區域可在樣本核酸分子之兩個股上。窗口可包含樣本核酸分子之兩個股上的核苷酸。該區域可為樣本核酸分子。該區域可包含至少 5、10、15、20、25、30、50、100、200、300、400、500、1k、5k、10k、50k 或 1M 個核苷酸。在一些實施例中，該區域可少於 50、100、200、300、400、500、1k、5k、10k、50k 或 1M 個核苷酸。該區域可以目標位置處之核苷酸為中心。

【0154】核苷酸之窗口可以目標位置處之核苷酸為中心。在一些實施例

中，窗口可不以目標位置處之核苷酸為中心。窗口可包含目標位置處之核苷酸的上游 X-nt 及下游 Y-nt。X 可包含但不限於 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 10000；Y 可包含但不限於 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 10000。窗口中核苷酸之最小數目可為 2、3、4、5、6、7、8、9、10、20、30、40、50、100、200，或大於目標位置之上游及下游之核苷酸數目中任一者之和的數目。窗口可類似於圖 5 中展示及描述的窗口。

【0155】 窗口可包含核酸分子之兩個股，類似於圖 6 所描述之技術。

【0156】 在方塊 1220 處，將輸入資料結構輸入至模型中。藉由接收第一複數個第一資料結構來訓練模型。第一複數個資料結構之每個第一資料結構對應於複數個第一核酸分子之各個核酸分子中測序之核苷酸的各個窗口。藉由量測對應於核苷酸之電信號來對第一核酸分子中之每一者進行測序。修飾在每個第一核酸分子之每個窗口中目標位置處的核苷酸中具有已知的第一狀態。每個第一資料結構包含與輸入資料結構相同之特性的值。模型可為本文所描述之任何機器學習模型。

【0157】 藉由儲存複數個第一訓練樣本來進一步訓練模型。每個第一訓練樣本包含第一複數個第一資料結構中之一者及指示目標位置處之核苷酸之第一狀態的第一標記。另外，當將第一複數個第一資料結構輸入至模型時，藉由基於模型之匹配或不匹配第一標記之相應標記的輸出使用複數個第一訓練樣本使模

型之參數最佳化而訓練模型。模型之輸出指定在各個窗口中目標位置處之核苷酸是否具有修飾。訓練可如稍後圖 13 所描述來進行。

【0158】 在方塊 1230 處，使用該模型判定修飾是否存在於輸入資料結構中窗口內之目標位置處的核苷酸中。

【0159】 修飾狀態可用於進一步分析中。在自孕婦獲得之樣本中，在本揭示案中之實施例可用於基於甲基化狀態判定血漿 DNA 分子之胎兒或母體來源。可藉由具有比參考值更高或更低的甲基化程度之基因體區域判定母體或胎兒來源。在實施例中，自孕婦獲得之樣本可為游離的，例如血漿或血清。在一些實施例中，樣本核酸分子可鑑別為與預定基因體區域對準。可已知預定基因體區域在胎兒或母體基因體中為高甲基化或低甲基化的。該方法可包含使用目標位置處之核苷酸之修飾狀態及視情況樣本核酸分子之一或多個其他核苷酸之修飾狀態來判定樣本核酸為胎兒來源抑或母體來源。

【0160】 判定樣本核酸分子為胎兒來源抑或母體來源可包含使用一或多個核苷酸之甲基化狀態來判定樣本核酸分子之甲基化程度。可將樣本核酸分子之甲基化程度與參考值進行比較。參考值可由一或多個母體核酸分子之甲基化程度來測定。將樣本核酸分子之甲基化程度與參考值進行比較可包含判定樣本核酸分子之甲基化程度低於參考值。判定樣本核酸分子為胎兒來源抑或母體來源可包含使用該比較判定樣本核酸分子為胎兒來源。

【0161】 在一些實施例中，樣本核酸分子可為複數個樣本核酸分子中之一個樣本核酸分子。該方法可進一步包含使用甲基化狀態判定複數個樣本核酸分子中之每一者為胎兒來源抑或母體來源。可使用對複數個樣本核酸分子之胎兒或母體來源的判定來測定胎兒分數。

【0162】 在一些實施例中，修飾狀態可用於判定拷貝數畸變是否存在於一區域中。修飾可為甲基化。樣本核酸分子可為游離的且獲自懷有胎兒之女性個體

之生物樣本。樣本核酸分子可為複數個樣本核酸分子中之一個樣本核酸分子。該方法可進一步包含將複數個樣本核酸分子鑑別為與胎兒基因體之區域對準。可判定複數個樣本核酸分子中之每個樣本核酸分子之一或多個核苷酸的修飾狀態。可使用複數個樣本核酸分子中之每個樣本核酸分子之一或多個核苷酸之甲基化狀態判定該區域之甲基化程度。該方法可進一步包含使用甲基化程度判定拷貝數畸變是否存在於胎兒基因體之區域中。該區域可為染色體，且該方法可進一步包含判定存在拷貝數畸變及判定胎兒具有染色體非整倍體。

【0163】 可判定修飾存在於一或多個核苷酸處。可使用在一或多個核苷酸處之修飾的存在來判定病症之分類。病症之分類可包含使用修飾之數目。可將修飾之數目與臨限值進行比較。替代或另外地，分類可包含一或多個修飾之位置。一或多個修飾之位置可藉由將核酸分子之序列讀數與參考基因體比對來判定。若已知與病症相關之某些位置顯示為具有修飾，則可判定病症。舉例而言，甲基化位點之模式可與病症之參考模式進行比較，且可基於該比較判定病症。與參考模式之匹配或與參考模式之實質性匹配（例如，80%、90%或 95%或更高）可指示病症或病症之可能性較高。病症可為任何妊娠相關之病症（例如子癇前症、宮內發育遲緩、侵入性胎盤形成及早產）。

【0164】 可分析統計學上顯著數目個核酸分子以便提供對一或多個懷孕個體中之病症、組織來源或臨床相關之 DNA 分數的準確判定。在一些實施例中，分析至少 1,000 個核酸分子。在其他實施例中，可分析至少 10,000 或 50,000 或 100,000 或 500,000 或 1,000,000 或 5,000,000 個核酸分子。作為另一實例，可產生至少 10,000 或 50,000 或 100,000 或 500,000 或 1,000,000 或 5,000,000 個序列讀數。

【0165】 該方法可包含判定病症之分類為個體患有該病症。分類可包含使用修飾之數目及/或修飾之位點的病症等級。

【0166】 可使用一或多個核苷酸處之修飾的存在判定胎兒 DNA 分數、胎兒甲基化譜、母體甲基化譜、印記基因區域之存在。

【0167】 方法 1200 可包含額外實施方案，諸如任何單一實施方案或下文描述及/或結合本文中在別處描述之一或多個其他方法之實施方案的任何組合。

【0168】 儘管圖 12 展示方法 1200 之實例方塊，但在一些實施方案中，相比於圖 12 中所描繪之彼等方塊，方法 1200 可包含額外方塊、更少方塊、不同方塊或以不同方式配置之方塊。另外或替代地，可並行地執行方法 1200 之方塊中之兩者或多於兩者。

B. 模型訓練

【0169】 **圖 13** 展示偵測核酸分子中核苷酸之修飾的例示性方法 1300。例示性方法 1300 可為訓練用於偵測修飾之模型的方法。該修飾可包含甲基化。甲基化可包含本文所描述之任何甲基化。該修飾可具有離散狀態，諸如甲基化及未甲基化，且可能指定甲基化之類型。因此，核苷酸可能有多於兩種狀態（分類）。圖 13 中之訓練可與圖 12 之方法 1200 一起使用。

【0170】 在方塊 1310 處，接收複數個第一資料結構。本文描述資料結構之各種實例，例如在圖 5 及圖 6 中。第一複數個第一資料結構中之每個第一資料結構可對應於複數個第一核酸分子之各個核酸分子中測序之核苷酸的各個窗口。與第一複數個資料結構相關之每個窗口可包含 4 個或更多個連續核苷酸，包含 5、6、7、8、9、10、11、12、13、14、15、16、17、18、19、20、21 或更多個連續核苷酸。每個窗口可具有相同數目之連續核苷酸。窗口可為重疊的。每個窗口可包含第一核酸分子之第一股上的核苷酸及第一核酸分子之第二股上的核苷酸。第一資料結構亦可包含窗口內之每個核苷酸的股特性之值。股特性可指示核苷酸存在於第一股或第二股。窗口可包含第二股中與第一股中對應位置之核苷酸不互補的核苷酸。在一些實施例中，第二股上之所有核苷酸均與第一股上之核

苷酸互補。在一些實施例中，每個窗口可包含第一核酸分子之僅一股上的核苷酸。

【0171】第一複數個第一資料結構可包含 5,000 至 10,000、10,000 至 50,000、50,000 至 100,000、100,000 至 200,000、200,000 至 500,000、500,000 至 1,000,000 或 1,000,000 或更多個第一資料結構。複數個第一核酸分子可包含至少 1,000、10,000、50,000、100,000、500,000、1,000,000、5,000,000 或更多個核酸分子。作為另一實例，可產生至少 10,000 或 50,000 或 100,000 或 500,000 或 1,000,000 或 5,000,000 個序列讀數。

【0172】藉由量測與核苷酸對應之電信號來對第一核酸分子中之每一者進行測序。電信號可來自奈米孔測序。

【0173】修飾在每個第一核酸分子之每個窗口中目標位置處的核苷酸中具有已知的第一狀態。第一狀態可為核苷酸中不存在修飾，或可為核苷酸中存在修飾。可已知第一核酸分子中不存在修飾，或可對第一核酸分子進行處理以使得修飾不存在。可已知第一核酸分子中存在修飾，或可對第一核酸分子進行處理以使得修飾存在。若第一狀態為不存在修飾，則修飾可在每個第一核酸分子之每個窗口中不存在，而非僅在目標位置不存在。已知的第一狀態可包含第一資料結構之第一部分的甲基化狀態及第一資料結構之第二部分的未甲基化狀態。可經由使用亞硫酸氫鹽測序或使用單分子即時測序之光信號的技術來判定已知的甲基化第一狀態。

【0174】目標位置可為各個窗口之中心。對於具有跨越偶數個核苷酸之窗口，目標位置可為緊靠窗口中心的上游或緊靠下游的位置。在一些實施例中，目標位置可在各個窗口之任何其他位置，包含第一位置或最後位置。舉例而言，若窗口跨越一個股之 n 個核苷酸，自第 1 位至第 n 位（上游或下游），則目標位置可在第 1 位至第 n 位的任何位置。

【0175】每個第一資料結構包含窗口內之特性的值。特性可為方塊 1210 處描述之特性中之任一者。

【0176】在方塊 1320 處，儲存複數個第一訓練樣本。每個第一訓練樣本包含第一複數個第一資料結構中之一者及指示目標位置處之核苷酸之修飾的第一狀態的第一標記。

【0177】在方塊 1330 處，接收第二複數個第二資料結構。方塊 1330 為情況選用的。第二複數個第二資料結構中之每個第二資料結構對應於複數個第二核酸分子中之各個核酸分子中測序之核苷酸的各個窗口。第二複數個核酸分子可與複數個第一核酸分子相同或不同。修飾在每個第二核酸分子之每個窗口內的目標位置處的核苷酸中具有已知的第二狀態。第二狀態為與第一狀態不同的狀態。舉例而言，若第一狀態為存在修飾，則第二狀態為不存在修飾，反之亦然。每個第二資料結構包含與第一複數個第一資料結構相同之特性的值。

【0178】在方塊 1340 處，儲存複數個第二訓練樣本。方塊 1340 為視情況選用的。每個第二訓練樣本包含第二複數個第二資料結構中之一者及指示目標位置處之核苷酸之修飾的第二狀態的第二標記。

【0179】在方塊 1350 處，使用複數個第一訓練樣本及視情況選用之複數個第二訓練樣本訓練模型。當將第一複數個第一資料結構及視情況選用之第二複數個第二資料結構輸入至模型時，藉由基於模型之匹配或不匹配第一標記及視情況選用之第二標記的相應標記的輸出使模型之參數最佳化來進行訓練。模型之輸出指定在各個窗口中目標位置處之核苷酸是否具有修飾。該方法可僅包含複數個第一訓練樣本，因為模型可將離群值鑑別為與第一狀態不同的狀態。該模型可為統計模型，亦稱為機器學習模型。

【0180】在一些實施例中，模型之輸出可包含處於複數個狀態中之每一者的概率。可將具有最高概率之狀態視為狀態。

【0181】該模型可包含卷積神經網路（CNN）。CNN 可包含一組卷積過濾器，其經組態以過濾第一複數個資料結構及視情況選用之第二複數個資料結構。過濾器可為本文所描述之任何過濾器。每層之過濾器的數目可為 10 至 20、20 至 30、30 至 40、40 至 50、50 至 60、60 至 70、70 至 80、80 至 90、90 至 100、100 至 150、150 至 200 或更多。過濾器之內核尺寸可為 2、3、4、5、6、7、8、9、10、11、12、13、14、15、15 至 20、20 至 30、30 至 40 或更多。CNN 可包含經組態以接收經過濾之第一複數個資料結構及視情況選用之經過濾之第二複數個資料結構的輸入層。CNN 亦可包含複數個隱藏層，其包含複數個節點。複數個隱藏層中之第一層耦合至輸入層。CNN 可進一步包含輸出層，其耦合至複數個隱藏層之最後一層且經組態以輸出輸出資料結構。輸出資料結構可包含該等特性。

【0182】該模型可包含遞迴類神經網路（RNN）。RNN 模型包含多個與量測窗口中之複數個核苷酸相關聯的長短期記憶（LSTM）單元。LSTM 單元之數目可等於量測窗口中核苷酸之數目。在一些實施例中，LSTM 單元之數目可少於量測窗口中核苷酸之數目。LSTM 單元之數目可為但不限於 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 等。一個 LSTM 單元可將與電流信號特徵相關之資訊傳輸至下一個 LSTM 單元，該資訊將經歷多輪線性或非線性變換。此類跨越 LSTM 單元之資訊傳輸通常以順序方式（例如根據時間步長）組構。此類跨越 LSTM 單元之資訊傳輸可為雙向的（亦即，包含時間順序及備用時間順序）。每個 LSTM 單元包含可程式化運算，諸如遺忘門、輸入門、單元狀態及輸出門。經由彼等運算，一個 LSTM 可判定來自前一時間步長之電流信號資訊是否為記住的或不相關的且可被遺忘（遺忘門）。一個 LSTM 單元嘗試學習自輸入達至此單元（輸入門）

的新資訊。該單元將更新的資訊自當前時間步長傳遞至下一個時間步長（輸出門）。本文中之單元狀態攜載該資訊以及所有時間步長。可使用 LSTM 單元之多個層。LSTM 層之數目可為 1、2、3、4、5、6、7、8、9、10、15、20、30 等。可使用各層之間的全連接。sigmoid 函數通常用作輸入門、輸出門及遺忘門之門函數（gating function）。sigmoid 函數之輸出值可在 0 與 1 之間，從而判定沒有資訊流動通過該等門或資訊完全流動通過該等門。雙曲正切激活函數（又稱 Tanh）可用作輸出激活函數，其處理來自輸出門之資訊值以形成值在-1 與 1 之間的新資訊，該資訊可傳遞至下一個 LSTM 單元。在一些實施例中，吾人可使用其他激活函數，包含但不限於二進制階梯函數、線性激活函數、sigmoid 函數、矯正線性單元等。由 LSTM 之最終層產生的值可傳遞至輸出層（亦即，密集層，具有一定數目之神經元）上，其中每個神經元均為全連接。密集層中之神經元數目可為但不限於 2、3、4、5、6、7、8、9、10、20、30、40、50、100、200、300、400、500、1000、2000 個等。吾人可使用多個密集層，包含但不限於 1、2、3、4、5、6、7、8、9、10、20、30、40、50、100、5000、1000 個等。輸出層可輸出甲基化評分，例如基於 sigmoid 激活函數或 SoftMax 激活函數，其可用於對甲基化狀態進行分類。舉例而言，若甲基化評分大於 0.5，則判定鹼基為甲基化。否則，判定鹼基為未甲基化。在一些實施例中，用於對甲基化狀態進行分類之臨限值可為但不限於至少 0.1、0.2、0.3、0.4、0.6、0.7、0.8、0.9 等。在一些實施例中，可丟棄模型中之一些神經元以使過度擬合問題最小化。丟棄之神經元百分比可為但不限於 1%、5%、10%、15%、20%、25%、30%、40%、50%、60%、70%等，其可根據不同層而不同。

【0183】該模型可包含監督式學習模型。監督式學習模型可包含不同的方法及演算法，包含分析學習、人工神經網路、後向傳播、提昇（boosting）（元演算法）、貝氏統計、案例式推理、決策樹學習、歸納邏輯程式設計、高斯過程

回歸（Gaussian process regression）、基因程式設計、資料分組處理方法、核估計法（kernel estimator）、學習自動機、學習分類系統、最小訊息長度（決策樹、決策圖等）、多線性子空間學習、單純貝氏分類（naive Bayes classifier）、最大熵分類、條件隨機場、最近相鄰演算法、可能近似正確學習（PAC）學習、漣波下降規則（ripple down rule）、知識獲取方法、符號機器學習演算法、子符號機器學習演算法、支持向量機、最小複雜度機器（MCM）、隨機森林、分類集成、有序分類、資料預處理、處理不平衡資料集、統計關係學習或 Proaftn（一種多準則分類演算法）。模型可線性回歸、邏輯回歸、深度遞迴類神經網路（例如長短期記憶體，LSTM）、貝氏分類器、隱藏式馬可夫模型（HMM）、線性判別分析（LDA）、k 均值聚類、具有雜訊的基於密度之空間聚類應用（DBSCAN）、隨機森林演算法、支持向量機（SVM）或本文所描述之任何模型。

【0184】作為訓練機器學習模型之一部分，機器學習模型之參數（諸如權重、臨限值，例如可用於神經網路中之激活函數等）可基於訓練樣本（訓練集）而經最佳化，以提供對目標位置處之核苷酸的修飾進行分類的最佳化準確度。可進行各種形式之最佳化，例如反向傳播、經驗風險最小化及結構風險最小化。可使用驗證樣本集（資料結構及標記）來驗證模型之準確度。可使用訓練集中用於訓練及驗證之各個部分來進行交叉驗證。該模型可包括複數個子模型，從而提供集合模型。子模型可為較弱的模型，一旦組合就提供更準確的最終模型。

V. 例示性系統

【0185】**圖 14** 示出根據本發明之一實施例的量測系統 1400。如圖所示之系統包含在樣本架 1410 內之樣本 1405，諸如 DNA 分子，其中樣本 1405 可與檢定 1408 接觸，以提供物理特徵 1415 的信號。樣本架之一實例可為包含檢定之探針及/或引子的流通池或液滴藉以移動之套管（在包含液滴之檢定的情況下）。藉由偵測器 1420 偵測樣本之物理特徵 1415（例如，螢光強度、電壓或電流）。

偵測器 1420 可按時間間隔（例如，週期性間隔）進行量測，以獲得構成資料信號之資料點。在一個實施例中，類比數位轉換器在複數個時間將來自偵測器之類比信號轉換成數位形式。樣品架 1410 及偵測器 1420 可形成檢定裝置，例如根據本文所描述之實施例進行測序之測序裝置。資料信號 1425 自偵測器 1420 發送至邏輯系統 1430。資料信號 1425 可儲存於局部記憶體 1435、外部記憶體 1440 或儲存裝置 1445 中。

【0186】 邏輯系統 1430 可為或可包含電腦系統、ASIC、微處理器等。其亦可包含顯示器（例如監測器、LED 顯示器等）及使用者輸入裝置（例如滑鼠、鍵盤、按鈕等）。邏輯系統 1430 及其他組件可為獨立的或網路連接之電腦系統的一部分，或其可直接連接至或併入包含偵測器 1420 及/或樣品架 1410 之裝置（例如測序裝置）中。邏輯系統 1430 亦可包含在處理器 1450 中執行之軟體。邏輯系統 1430 可包含電腦可讀媒體，其儲存用於控制系統 1400 執行本文所描述之方法中之任一者的指令。舉例而言，邏輯系統 1430 可向包含樣品架 1410 之系統提供命令，使得執行測序或其他物理操作。此類物理操作可以特定次序進行，例如在試劑以特定次序添加及移除之情況下。此類物理操作可由可用於獲得樣本且執行分析之例如包含機械臂之機器人系統執行。

【0187】 本文所提及之任一種電腦系統可利用任何適合數目個子系統。此類子系統之實例展示於**圖 15** 中之電腦系統 10 中。在一些實施例中，電腦系統包含單一電腦設備，其中子系統可為電腦設備之組件。在其他實施例中，電腦系統可包含具有內部組件之多個電腦設備，其各自為一個子系統。電腦系統可包含桌上型及膝上型電腦、平板電腦、行動電話、其他行動裝置及基於雲端之系統。

【0188】 圖 15 中所示之子系統經由系統匯流排 75 互連。展示額外子系統，諸如列印機 74、鍵盤 78、一或多個儲存裝置 79、與顯示適配器 82 耦接之監測器 76（例如顯示屏幕，諸如 LED）及其他裝置。耦接至輸入/輸出（I/O）控制器

71 之周邊裝置及 I/O 裝置可藉由此項技術中已知之多種手段（諸如輸入/輸出（I/O）埠 77（例如，USB、Lightning、Thunderbolt™））連接至電腦系統。舉例而言，I/O 埠 77 或外部介面 81（例如乙太網路（Ethernet）、Wi-Fi 等）可用於將電腦系統 10 連接至廣域網路（諸如網際網路）、滑鼠輸入裝置或掃描儀。經由系統匯流排 75 互連允許中央處理器 73 與各子系統通信且控制系統記憶體 72 或儲存裝置 79（例如，固定磁碟，諸如硬碟機，或光碟）執行複數個指令，以及子系統之間的資訊交換。系統記憶體 72 及/或一或多個儲存裝置 79 可實施為電腦可讀媒體。另一子系統為資料收集裝置 85，諸如照相機、麥克風、加速計及其類似物。本文中所提及之資料中之任一者可自一個組件輸出至另一組件且可輸出至使用者。

【0189】 電腦系統可包含複數個相同組件或子系統，例如藉由外部介面 81、藉由內部介面或經由可移式儲存裝置連接在一起，該等可移式儲存裝置可自一個組件連接至另一組件且移除。在一些實施例中，電腦系統、子系統或設備可經由網路通信。在此類情況下，一台電腦可視為用戶端，且另一台電腦視為伺服器，其中各電腦可為同一電腦系統之一部分用戶端及伺服器各自可包含多個系統、子系統或組件。

【0190】 實施例之態樣可以控制邏輯形式使用硬體電路（例如特殊應用積體電路或場域可程式化閘陣列）及/或使用具有大體上可程式化處理器之電腦軟體以模組化或整合式方式來實施。如本文所使用，處理器可包含單核處理器、同一個積體晶片上之多核處理器或單一電路板或網路硬體以及專用硬體上之多個處理單元。基於本文所提供之揭示內容及教示內容，本領域中之一般熟習此項技術者將知道及瞭解使用硬體及硬體與軟體之組合來實施本發明之實施例的其他方式及/或方法。

【0191】 本申請案中所描述之任何軟體組件或功能可使用例如習知或面

向對象技術，以軟體程式碼形式實施，該軟體程式碼係由使用任何適合電腦語言（諸如 Java、C、C++、C#、Objective-C、Swift）或腳本處理語言（諸如 Perl 或 Python）的處理器執行。軟體程式碼可以一系列指令或命令形式儲存於電腦可讀取媒體上以進行儲存及/或傳輸。適合之非暫時性電腦可讀媒體可包含隨機存取記憶體（RAM）、唯讀記憶體（ROM）、磁性媒體（諸如硬碟機或軟碟機）或光學媒體，諸如光碟（CD）或數位化通用光碟（DVD）或藍光碟、快閃記憶體及其類似者。電腦可讀媒體可為此類儲存或傳輸裝置之任何組合。

【0192】 此類程式亦可使用適用於經由符合多種協定之有線、光學及/或無線網路（包含網際網路）傳輸的載波信號來編碼及傳輸。因此，電腦可讀取媒體可使用以此類程式編碼之資料信號建立。以程式碼編碼之電腦可讀媒體可與相容裝置一起封裝或與其他裝置分開提供（例如經由網際網路下載）。任何此類電腦可讀媒體可存在於單一電腦產品（例如硬碟機、CD 或整個電腦系統）上或其內部，且可存在於系統或網路內之不同電腦產品上或其內部。電腦系統可包含用於向使用者提供本文所提及之任何結果的監測器、列印機、或其他適合之顯示器。

【0193】 本文中所描述之方法中之任一者可完全或部分地使用電腦系統來執行，該電腦系統包含可經組態以執行步驟之一或多個處理器。因此，實施例可針對經組態以執行本文所描述之任何方法之步驟的電腦系統，潛在地使用不同組件執行各別步驟或各別步驟群組。儘管以帶編號之步驟形式呈現，但本文中之方法之步驟可同時或在不同時間或以不同順序執行。另外，此等步驟之一部分可與其他方法之其他步驟之部分一起使用。另外，可視情況選用步驟之全部或部分。此外，任何方法之任何步驟可使用用於執行此等步驟之系統的模組、單元、電路或其他構件來執行。

【0194】 可在不脫離本發明之實施例的精神及範疇的情況下以任何合適

方式組合特定實施例之特定細節。然而，本發明之其他實施例可針對與各個別態樣或此等個別態樣之特定組合相關的特定實施例。

【0195】已出於說明及描述之目的呈現本揭示案之例示性實施例的上述描述。其並不意欲為詳盡的或將本揭示案限於所描述之精確形式，且鑒於以上教示，許多修改及變化為可能的。

【0196】除非相反地特定指示，否則「一（a/an）」或「該（the）」之敍述欲意謂「一或多個（種）」。除非相反地特定指示，否則「或」之使用欲意謂「包括性的或」，而非「互斥性的或」。提及「第一」組件不一定需要提供第二組件。此外，除非明確陳述，否則提及「第一」或「第二」組件不會將所提及組件限制於特定位置。術語「基於」意指「至少部分地基於」。

【0197】出於所有目的，本文所提及之所有專利、專利申請案、公開案及描述均以全文引用之方式併入。不承認任一者為先前技術。

【符號說明】

【0198】

10: 電腦系統

71: 輸入/輸出（I/O）控制器

72: 系統記憶體

73: 中央處理器

74: 列印機

75: 系統匯流排

76: 監測器

77: I/O 埠

78: 鍵盤

79: 一或多個儲存裝置

81: 外部介面

82: 顯示適配器

85: 資料收集裝置

104: DNA 分子

108: 馬達蛋白

112: 馬達蛋白

116: 奈米孔

120: 曲線圖

204: 圓點

208: 線

304: 跡線

308: 區段

504: 方塊

508: 方塊

512: 方塊

516: 方塊

520: 方塊

524: 方塊

528: 方塊

532: 方塊

604: 方塊

608: 方塊

612: 方塊

616: 方塊

620: 方塊

624: 方塊

628: 方塊

632: 方塊

1004: 線

1008: 線

1012: 線

1016: 線

1020: 線

1024: 線

1028: 線

1032: 線

1200: 方法

1210: 方塊

1220: 方塊

1230: 方塊

1300: 方法

1310: 方塊

1320: 方塊

1330: 方塊

1340: 方塊

1350: 方塊

1400: 量測系統

1405: 樣本

1408: 檢定

1410: 樣本架

1415: 物理特徵

1420: 偵測器

1425: 資料信號

1430: 邏輯系統

1435: 局部記憶體

1440: 外部記憶體

1445: 儲存裝置

1450: 處理器

1804: 方塊

1808: 方塊

1812: 方塊

1816: 方塊

1820: 方塊

1824: 方塊

1828: 方塊

1832: 方塊

1904: 線

2104: 實心黑色圓

2108: 空心圓

【發明申請專利範圍】

【請求項1】

一種用於偵測核酸分子中核苷酸之修飾的方法，該方法包括：

接收輸入資料結構，該輸入資料結構對應於樣本核酸分子中測序之核苷酸的窗口，其中該樣本核酸分子係藉由量測對應於該等核苷酸之電信號來測序，該輸入資料結構包括以下特性之值：

對於該窗口內之每個核苷酸：

該核苷酸之標識，

該核苷酸相對於各個窗口內之目標位置的位置，及

包括該電信號之對應於該核苷酸之區段的第一區段統計值的向量；

將該輸入資料結構輸入至模型中，該模型藉由以下操作進行訓練：

接收第一複數個第一資料結構，該第一複數個第一資料結構中之每個第一資料結構對應於複數個第一核酸分子之各個核酸分子中測序之核苷酸的各個窗口，其中該等第一核酸分子中之每一者係藉由量測對應於該等核苷酸之該電信號來測序，其中該修飾在每個第一核酸分子之每個窗口中目標位置處之核苷酸中具有已知的第一狀態，每個第一資料結構包括與該輸入資料結構相同特性之值，

儲存複數個第一訓練樣本，每個樣本包含該第一複數個第一資料結構中之一者及指示該目標位置處之核苷酸之第一狀態的第一標記，及

當將該第一複數個第一資料結構輸入至該模型時，基於該模型之匹配或不匹配該等第一標記之相應標記的輸出，使用該複數個第一訓練樣本來使該模型之參數最佳化，其中該模型之輸出指定該各個窗口中該目標位置處之核苷酸是否具有該修飾，

使用該模型判定該修飾是否存在於該輸入資料結構中之該窗口內之該目標位置處的核苷酸中。

【請求項2】

如請求項1之方法，其中該第一區段統計值表示該電信號之對應於該核苷酸之該區段的平均值。

【請求項3】

如請求項1之方法，其中該第一區段統計值表示該電信號之對應於該核苷酸之該區段的該電信號之變化。

【請求項4】

如請求項1之方法，其中該第一區段統計值表示該電信號之對應於該核苷酸之該區段的平均值的正規化值。

【請求項5】

如請求項1、2或4中任一項之方法，其中該向量包括表示該電信號之對應於該核苷酸之該區段之變化的第二區段統計值。

【請求項6】

如請求項1、2或3中任一項之方法，其中該向量包括表示該電信號之對應於該核苷酸之該區段的平均值之正規化值的第二區段統計值。

【請求項7】

如請求項2之方法，其中：

該向量包括表示該電信號之對應於該核苷酸之該區段的變化的第二區段統計值，且

該向量包括表示該第一區段統計值之正規化值的第三區段統計值。

【請求項8】

如前述請求項中任一項之方法，其中該輸入資料結構包括該核酸分子

之區域中的該電信號之第一區統計值等於或大於該窗口的值。

【請求項9】

如請求項8之方法，其中該第一區統計值表示該區域中的該電信號之平均值或中值。

【請求項10】

如請求項8之方法，其中該第一區統計值表示相對於該區域中之該電信號之該平均值或中值的該電信號之變化絕對值的中值或平均值。

【請求項11】

如請求項9之方法，其中該輸入資料結構進一步包括第二區統計值，該第二區統計值表示相對於該區域中之該電信號之該平均值或中值的該電信號之變化絕對值的中值或平均值。

【請求項12】

如請求項8至11中任一項之方法，其中該區域係在該樣本核酸分子之一個股上。

【請求項13】

如請求項8至12中任一項之方法，其中該區域為該樣本核酸分子或包括至少5、10、15、20、25、30、50、100、200、300、400、500或1k、5k、10k、50k或1M個核苷酸。

【請求項14】

如請求項8至13中任一項之方法，其中該區域係以該核苷酸為中心。

【請求項15】

如前述請求項中任一項之方法，其中該窗口包括該樣本核酸分子之兩個股上的核苷酸。

【請求項16】

如前述請求項中任一項之方法，其中該修飾為甲基化或氧化。

【請求項17】

如前述請求項中任一項之方法，其中該電信號為電流、電壓、電阻、電感、電容或阻抗。

【請求項18】

如前述請求項中任一項之方法，其進一步包括使用奈米孔對該樣本核酸分子進行測序。

【請求項19】

如請求項1之方法，其中：

該修飾為甲基化，且

該樣本核酸分子為游離的且獲自懷有胎兒之女性個體之生物樣本，

該方法進一步包括：

使用該目標位置處之核苷酸之修飾狀態判定該樣本核酸分子為胎兒來源抑或母體來源，其中該修飾狀態為該修飾是否存在，且視情況判定該樣本核酸分子之一或多個其他核苷酸之修飾狀態。

【請求項20】

如請求項19之方法，其中判定該樣本核酸分子為胎兒來源抑或母體來源包括：

使用該一或多個核苷酸之該等修飾狀態判定該樣本核酸分子之甲基化程度；及

將該樣本核酸分子之該甲基化程度與參考值進行比較。

【請求項21】

如請求項20之方法，其中該參考值係由一或多個母體核酸分子之甲基化程度測定。

【請求項22】

如請求項20之方法，其中：

將該樣本核酸分子之該甲基化程度與該參考值進行比較包括判定該樣本核酸分子之該甲基化程度低於該參考值，及

判定該樣本核酸分子為胎兒來源抑或母體來源包括使用該比較判定該樣本核酸分子為胎兒來源。

【請求項23】

如請求項19之方法，其進一步包括：

將該樣本核酸分子鑑別為與預定基因體區域對齊。

【請求項24】

如請求項19之方法，其中：

該樣本核酸分子為複數個樣本核酸分子中之一個樣本核酸分子，

該方法進一步包括：

使用該等修飾狀態判定該複數個樣本核酸分子中之每一者為胎兒來源抑或母體來源，及

使用對該複數個樣本核酸分子之胎兒或母體來源的判定來測定胎兒分數。

【請求項25】

如請求項1之方法，其中：

該修飾為甲基化，

該樣本核酸分子為游離的且獲自懷有胎兒之女性個體之生物樣本，且

該樣本核酸分子為複數個樣本核酸分子中之一個樣本核酸分子，

該方法進一步包括：

將該複數個樣本核酸分子鑑別為與胎兒基因體之區域對齊，

判定該複數個樣本核酸分子中之每一個樣本核酸分子之一或多個核苷酸的修飾狀態，

使用該複數個樣本核酸分子中之每一個樣本核酸分子的該一或多個核苷酸之該等修飾狀態來判定該區域之甲基化程度，及

使用該甲基化程度判定該胎兒基因體之該區域中是否存在拷貝數畸變。

【請求項26】

一種用於偵測核酸分子中核苷酸之修飾的方法，該方法包括：

接收第一複數個第一資料結構，該第一複數個第一資料結構中之每個第一資料結構對應於複數個第一核酸分子中之各個核酸分子中測序的核苷酸之各個窗口，其中該等第一核酸分子中之每一者係藉由量測對應於該等核苷酸之電信號來測序，其中該修飾在每個第一核酸分子之每個窗口中目標位置處之核苷酸中具有已知的第一狀態，每個第一資料結構包括以下特性之值：

對於該窗口內之每個核苷酸：

該核苷酸之標識，

該核苷酸相對於各個窗口內之目標位置的位置，及

包括該電信號之對應於該核苷酸之區段的第一區段統計值的向量；

儲存複數個第一訓練樣本，每個樣本包含該第一複數個第一資料結構中之一者及指示該目標位置處之核苷酸之修飾之第一狀態的第一標記；及

當將該第一複數個第一資料結構輸入至模型時，基於該模型之匹配或不匹配該等第一標記之相應標記的輸出，使用該複數個第一訓練樣

本使該模型之參數最佳化而訓練該模型，其中該模型之輸出指定該各個窗口中該目標位置處之核苷酸是否具有該修飾。

【請求項27】

如請求項26之方法，其進一步包括：

接收第二複數個第二資料結構，該第二複數個第二資料結構中之每個第二資料結構對應於複數個第二核酸分子之各個核酸分子中測序之核苷酸的各個窗口，其中該修飾在每個第二核酸分子之每個窗口內之目標位置處之核苷酸中具有已知的第二狀態，每個第二資料結構包括與該第一複數個第一資料結構相同的特性之值；

儲存複數個第二訓練樣本，每個樣本包含該第二複數個第二資料結構中之一者及指示該目標位置處之核苷酸之第二狀態的第二標記；

其中訓練：

該第一狀態或該第二狀態為存在該修飾，且另一狀態為不存在該修飾，

該模型進一步包括當將該第二複數個第二資料結構輸入至該模型時，基於該模型之匹配或不匹配該等第二標記之相應標記的輸出，使用該複數個第二訓練樣本使該模型之參數最佳化。

【請求項28】

如請求項27之方法，其中該複數個第一核酸分子與該複數個第二核酸分子相同。

【請求項29】

如請求項26之方法，其中：

與該第一複數個第一資料結構相關之每個窗口包括該第一核酸分子之第一股上的核苷酸及該第一核酸分子之第二股上的核苷酸，及

每個第一資料結構進一步包括對於該窗口內之每個核苷酸之股特性的值，該股特性指示該核苷酸存在於該第一股或該第二股上。

【請求項30】

如請求項26之方法，其中該修飾包括該目標位置處之核苷酸的甲基化。

【請求項31】

如請求項30之方法，其中該等已知的第一狀態包含該等第一資料結構之第一部分的甲基化狀態及該等第一資料結構之第二部分的未甲基化狀態。

【請求項32】

如請求項26之方法，其中該第一區段統計值表示該電信號之對應於該核苷酸之該區段的平均值。

【請求項33】

如請求項26之方法，其中該第一區段統計值表示該電信號之對應於該核苷酸之該區段的該電信號之變化。

【請求項34】

如請求項26之方法，其中該第一區段統計值表示該電信號之對應於該核苷酸之該區段的平均值的正規化值。

【請求項35】

如請求項26、32或34中任一項之方法，其中該向量包括表示該電信號之對應於該核苷酸之該區段之變化的第二區段統計值。

【請求項36】

如請求項26、32或33中任一項之方法，其中該向量包括表示該電信號之對應於該核苷酸之該區段的平均值之正規化值的第二區段統計值。

【請求項37】

如請求項32之方法，其中：

該向量包括表示該電信號之對應於該核苷酸之該區段的變化的第二區段統計值，且

該向量包括表示該第一區段統計值之正規化值的第三區段統計值。

【請求項38】

如請求項26至37中任一項之方法，其中每個第一資料結構包括該各個核酸分子之區域中的該電信號之第一區統計值等於或大於該窗口的值。

【請求項39】

如請求項38之方法，其中該第一區統計值表示該區域中的該電信號之平均值或中值。

【請求項40】

如請求項38之方法，其中該第一區統計值表示相對於該區域中之該電信號之該平均值或中值的該電信號之變化絕對值的中值或平均值。

【請求項41】

如請求項39之方法，其中該第一資料結構進一步包括第二區統計值，該第二區統計值表示相對於該區域中之該電信號之該平均值或中值的該電信號之變化絕對值的中值或平均值。

【請求項42】

如請求項38至41中任一項之方法，其中該區域係在該各個核酸分子之一個股上。

【請求項43】

如請求項38至45中任一項之方法，其中該區域為該各個核酸分子或包括至少5、10、15、20、25、30、50、100、200、300、400、500或1k、5k、

10k、50k或1M個核苷酸。

【請求項44】

如請求項38至43中任一項之方法，其中該區域係以該核苷酸為中心。

【請求項45】

如請求項26至44中任一項之方法，其中該窗口包括該各個核酸分子之兩個股上的核苷酸。

【請求項46】

一種電腦產品，其包括儲存複數個指令之非暫時性電腦可讀媒體，該複數個指令在執行時控制電腦系統以執行如前述請求項中任一項之方法。

【請求項47】

一種系統，其包括：

如請求項46之電腦產品；及

一或多個處理器，其用於執行儲存於該電腦可讀媒體上之指令。

【請求項48】

一種系統，其包括用於執行上述方法中之任一者的構件。

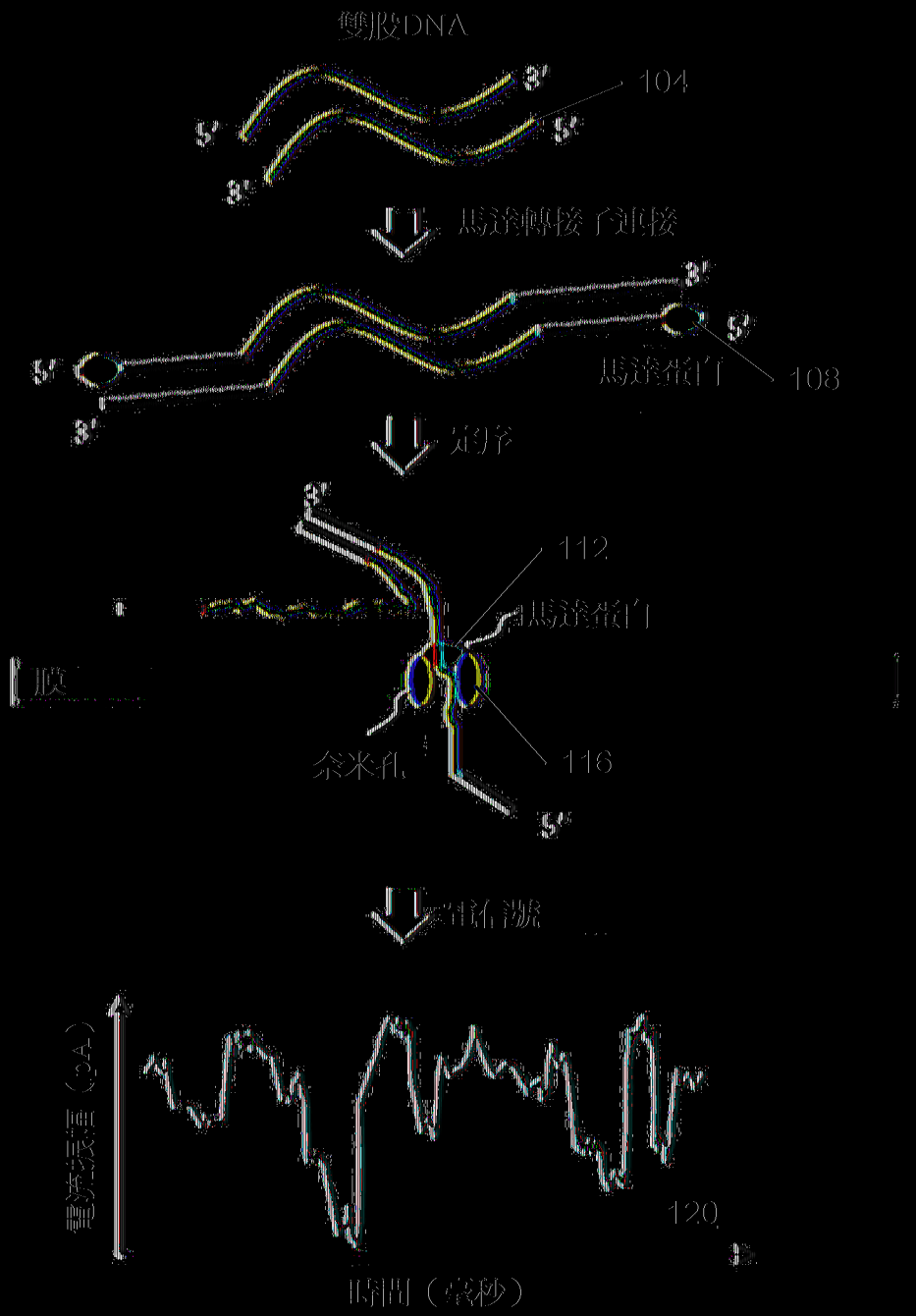
【請求項49】

一種系統，其包括經組態以執行上述方法中之任一者的一或多個處理器。

【請求項50】

一種系統，其包括分別執行上述方法中之任一者的步驟的模組。

[(發明圖式)]

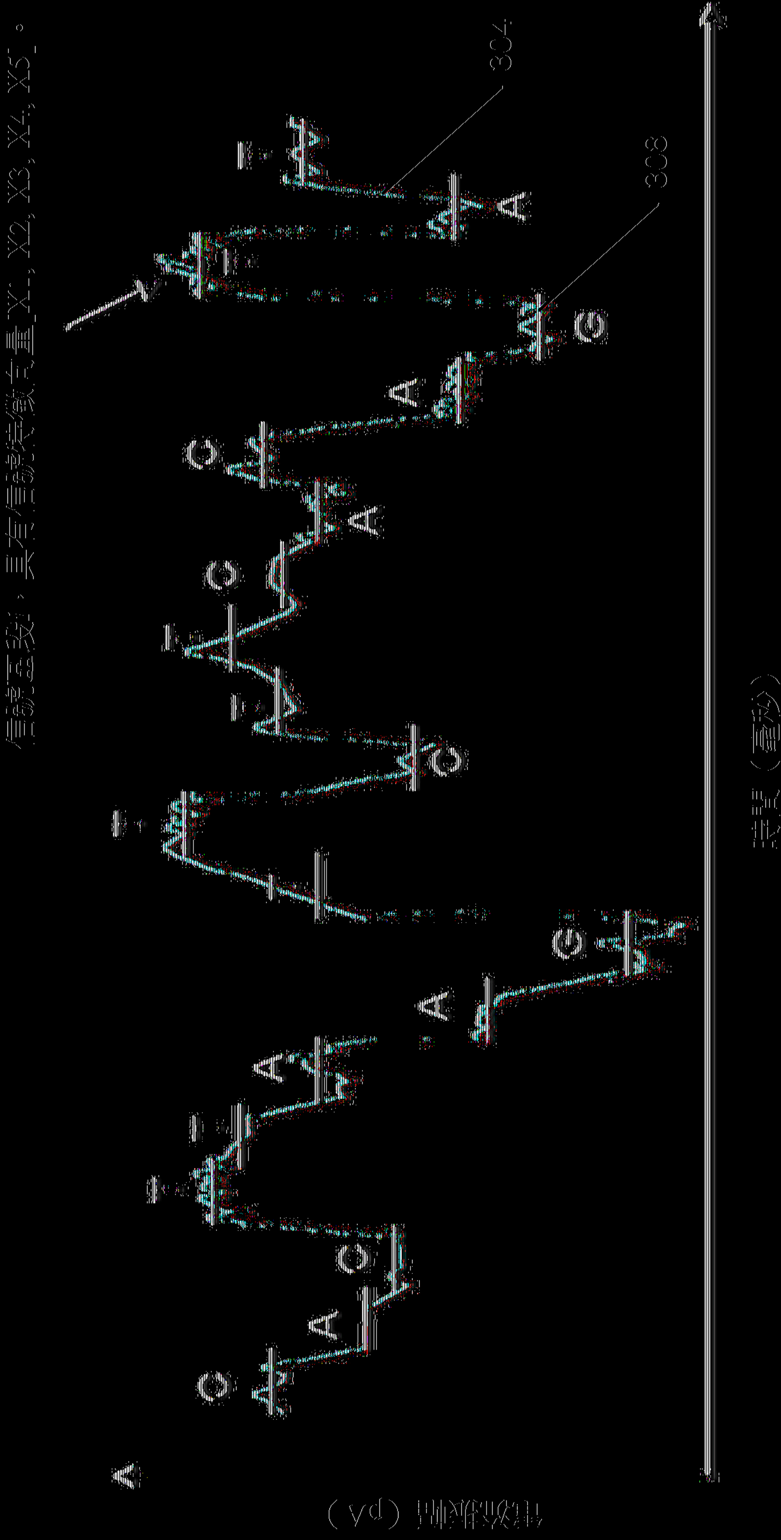


1）檢體識別

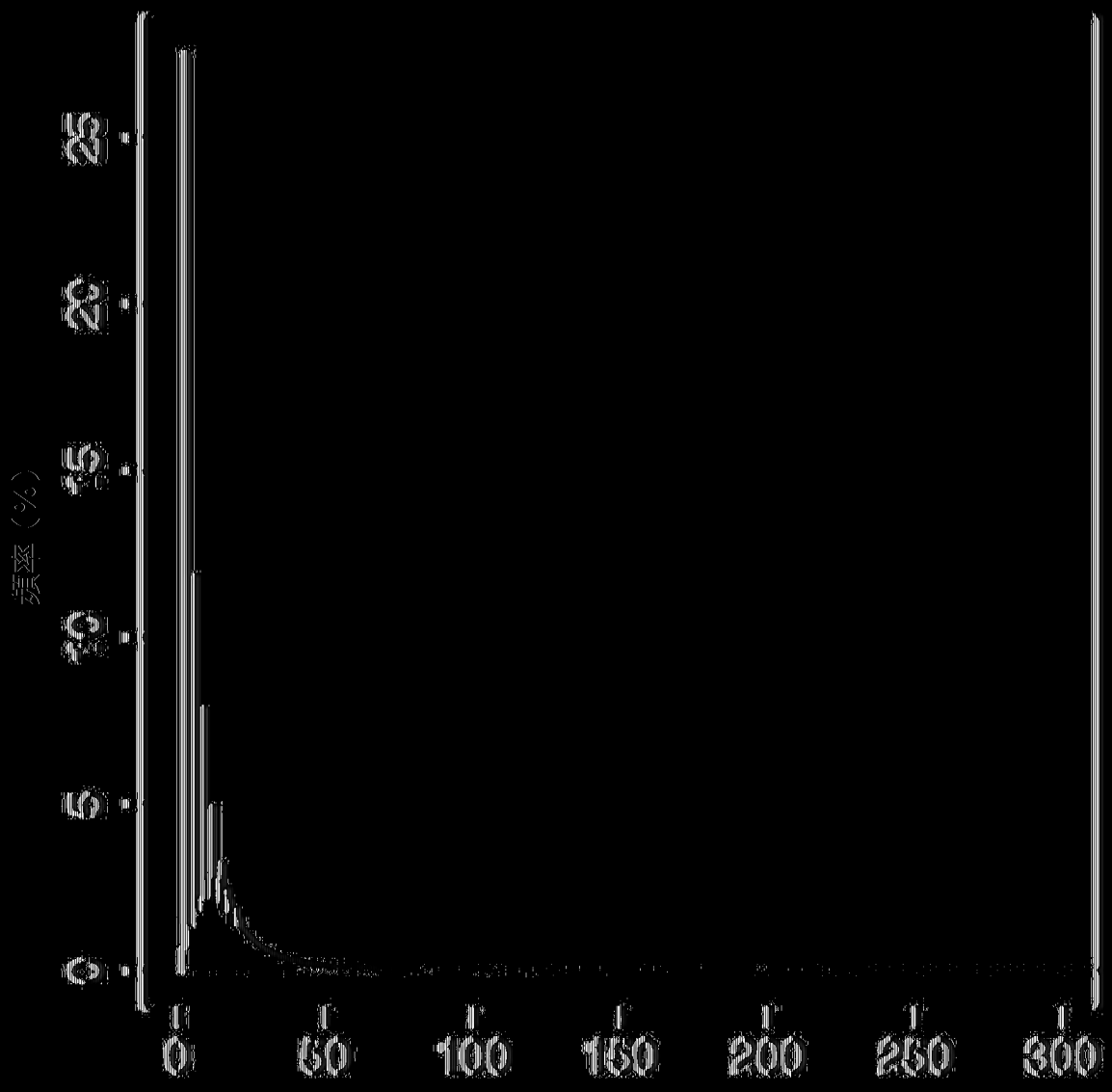
2）檢體修飾分析

（圖1）

電流值（pA）
208
204



【圖3】



（圖4）

雙股DNA分子 · 504

⇓

用於奈米孔定序之定序轉接子連接 · 508

⇓

奈米孔定序 · 512

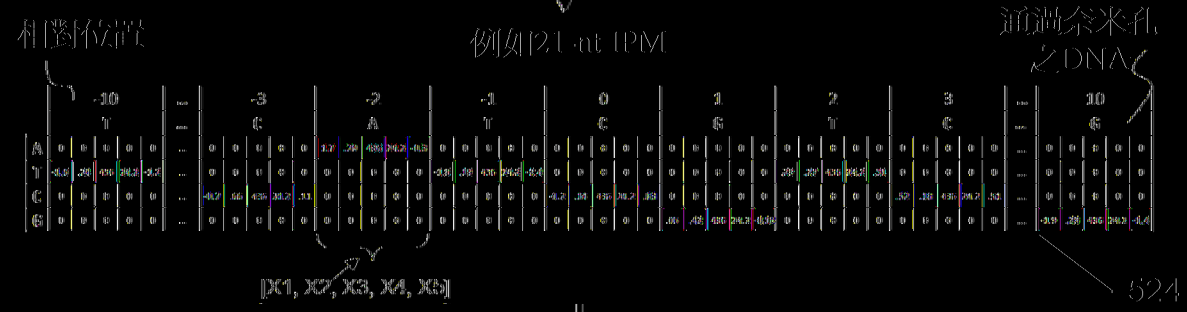
⇓

獲得電流信號 · 516

⇓

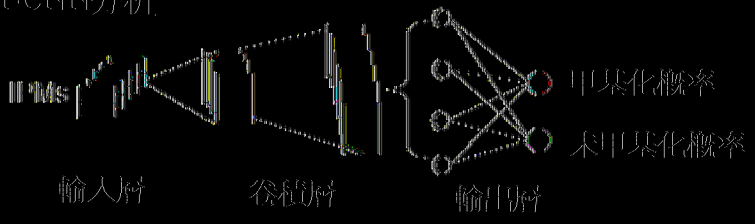
建構電流信號模式之整合式表示矩陣（IPM），包括鹼基之電流信號、定序背景及路徑用於鹼基修飾分析之其中所述或周圍的一系列核苷酸的定序位置資訊 · 520

⇓

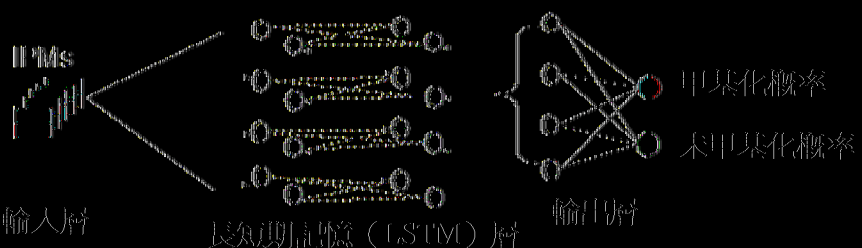


⇓

（a）IPM-CNN分析



（b）IPM-RNN分析



（圖5）

雙股DNA分子 · 604

⇓

用於奈米孔定序之定序轉接子連接 · 608

⇓

奈米孔定序，雙股DNA之第一股在第一股穿過奈米孔之後穿過可奈米孔（1D²或2D）。 · 612

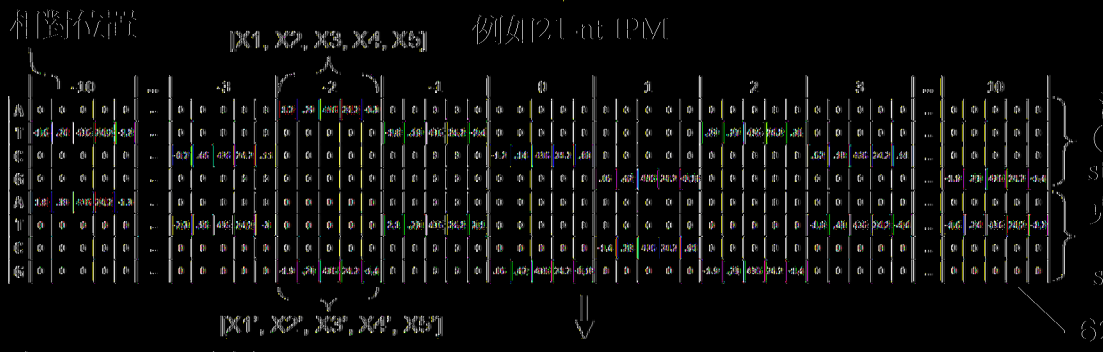
⇓

獲得每個雙股DNA分子之兩個股的電流信號 · 616

⇓

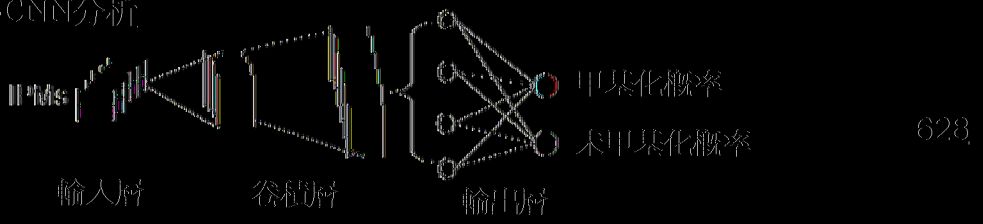
矩陣化每個雙股DNA分子之兩個股的整合式表示矩陣（IPM） · 620

⇓



⇓

（a）IPM-CNN分析



（b）IPM-RNN分析

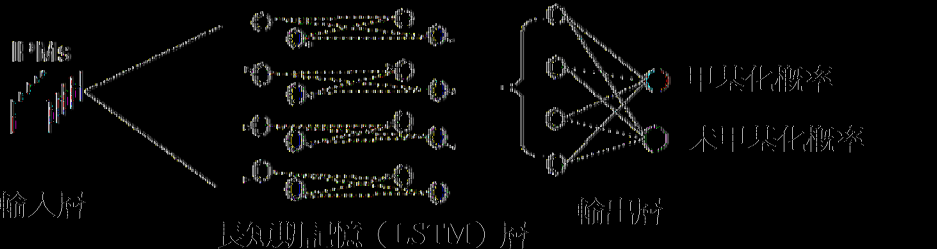


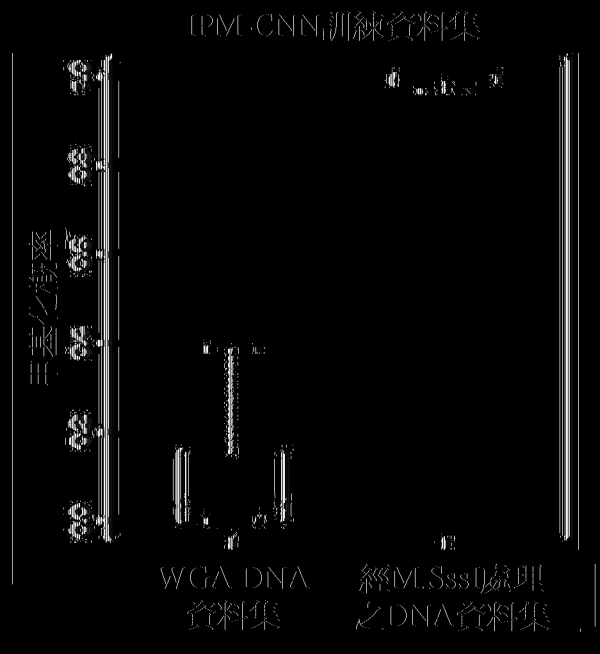
圖6

核大小	AUC	
	訓練資料集	測試資料集
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

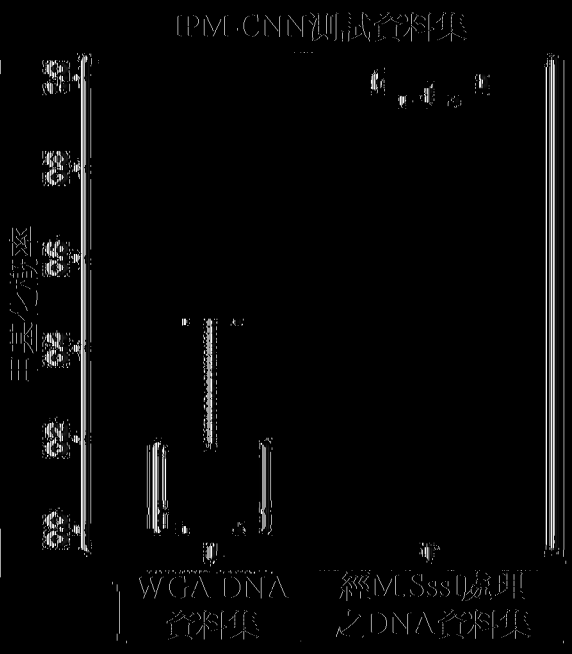
（續7）

資料集	分子數目（CpG位點數目）	
	訓練	測試
經M.SssI處理之DNA	7,989 (38,470)	4,826 (9,716)
WGADNA	8,052 (37,150)	5,041 (11,444)

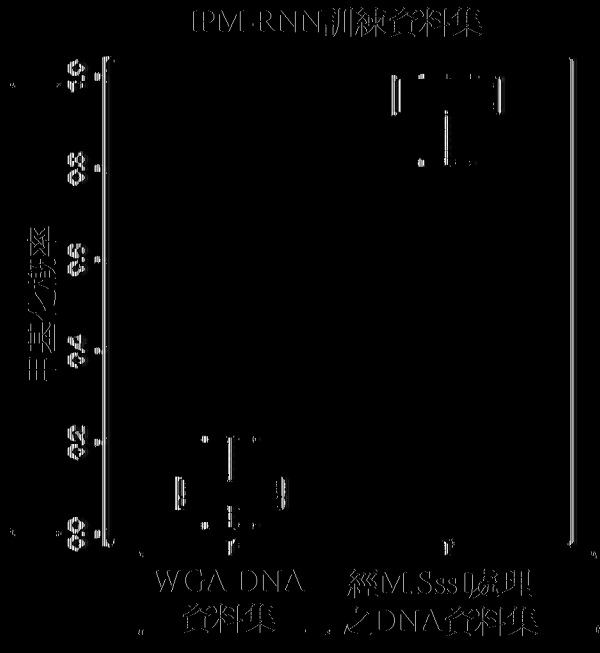
（圖8）



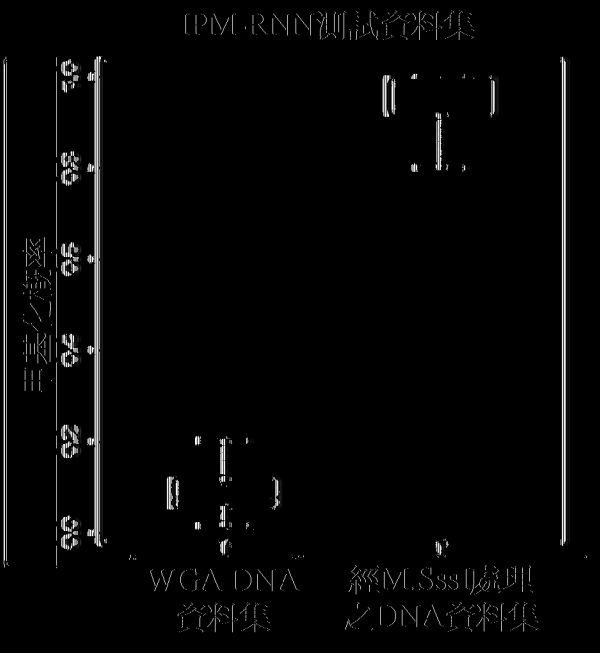
（圖9A）



（圖9B）



（圖9C）

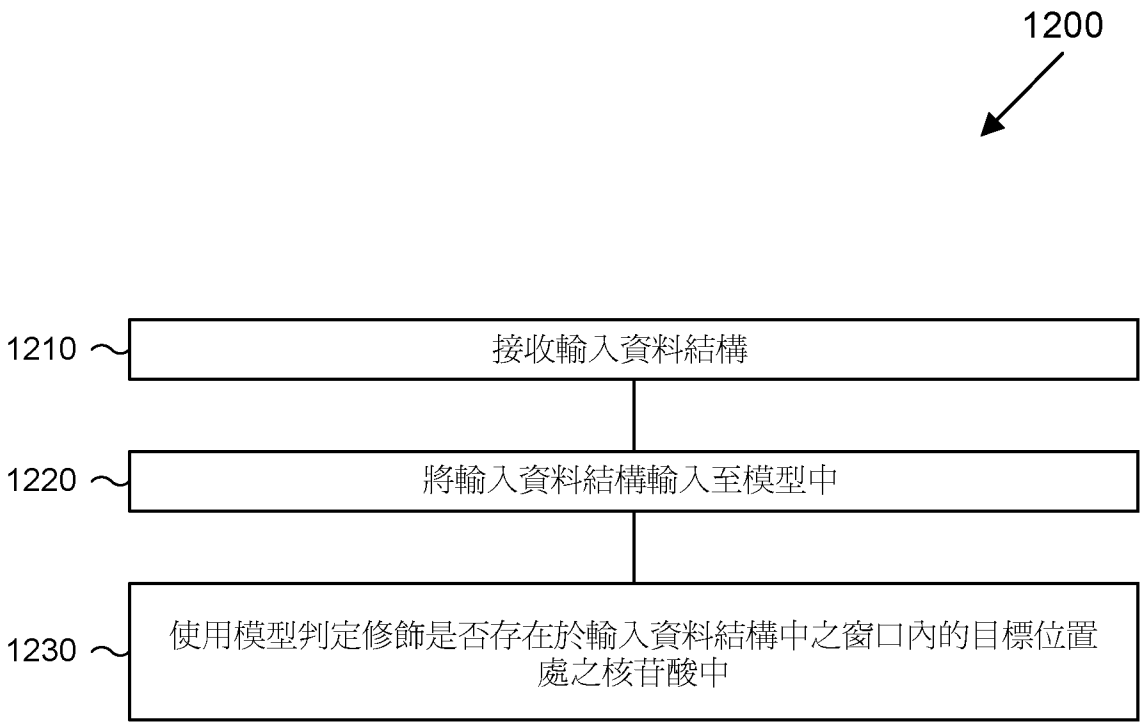


（圖9D）

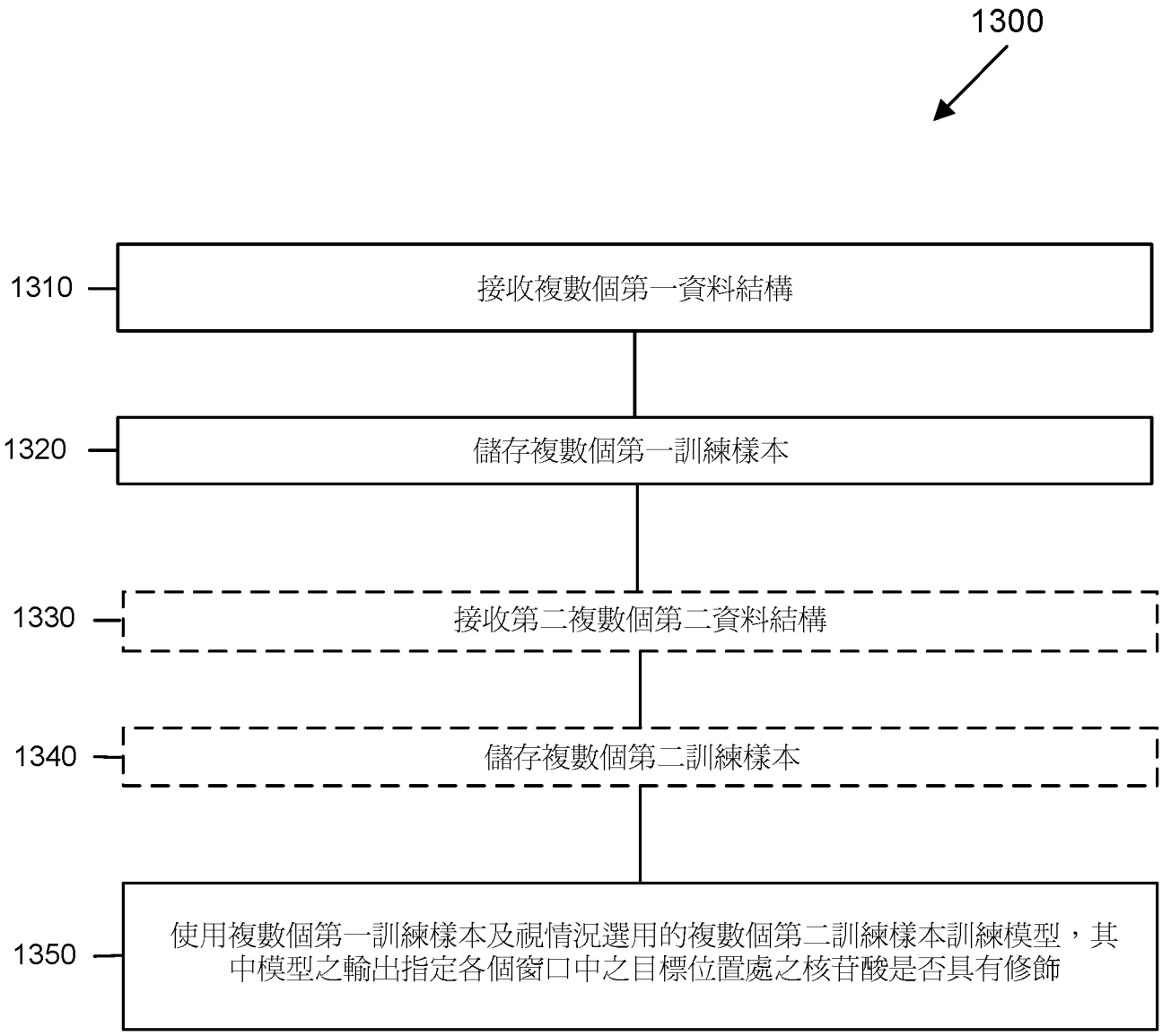
1.0
0.8
0.6
0.4
0.2
0.0
1.0
0.8
0.6
0.4
0.2
0.0
1016
1008
1032
1024
1004
1028
1020
1012

	感度	特異性
IPM-CNN	90%	
IPM-RNN	93%	90%
DeepMod	58%	
nanopolish	74%	
IPM-CNN	86%	
IPM-RNN	90%	95%
DeepMod	38%	
nanopolish	55%	
IPM-CNN	70%	
IPM-RNN	89%	99%
DeepMod	13%	
nanopolish	16%	

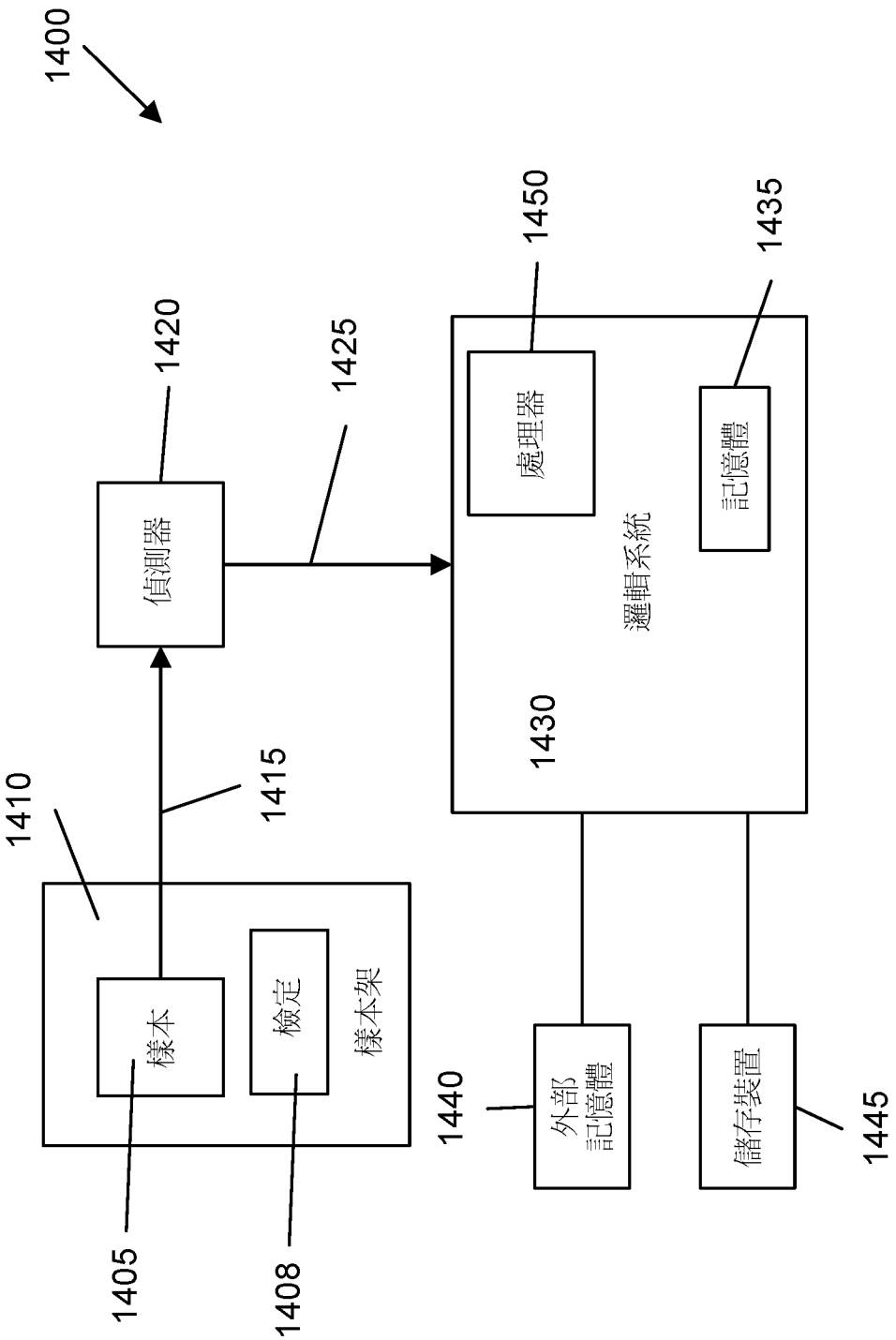
（附圖）


1200
1210
接收輸入資料結構
1220
將輸入資料結構輸入至模型中
1230
使用模型判定修飾是否存在於輸入資料結構中之窗口內的目標位置處之核苷酸中

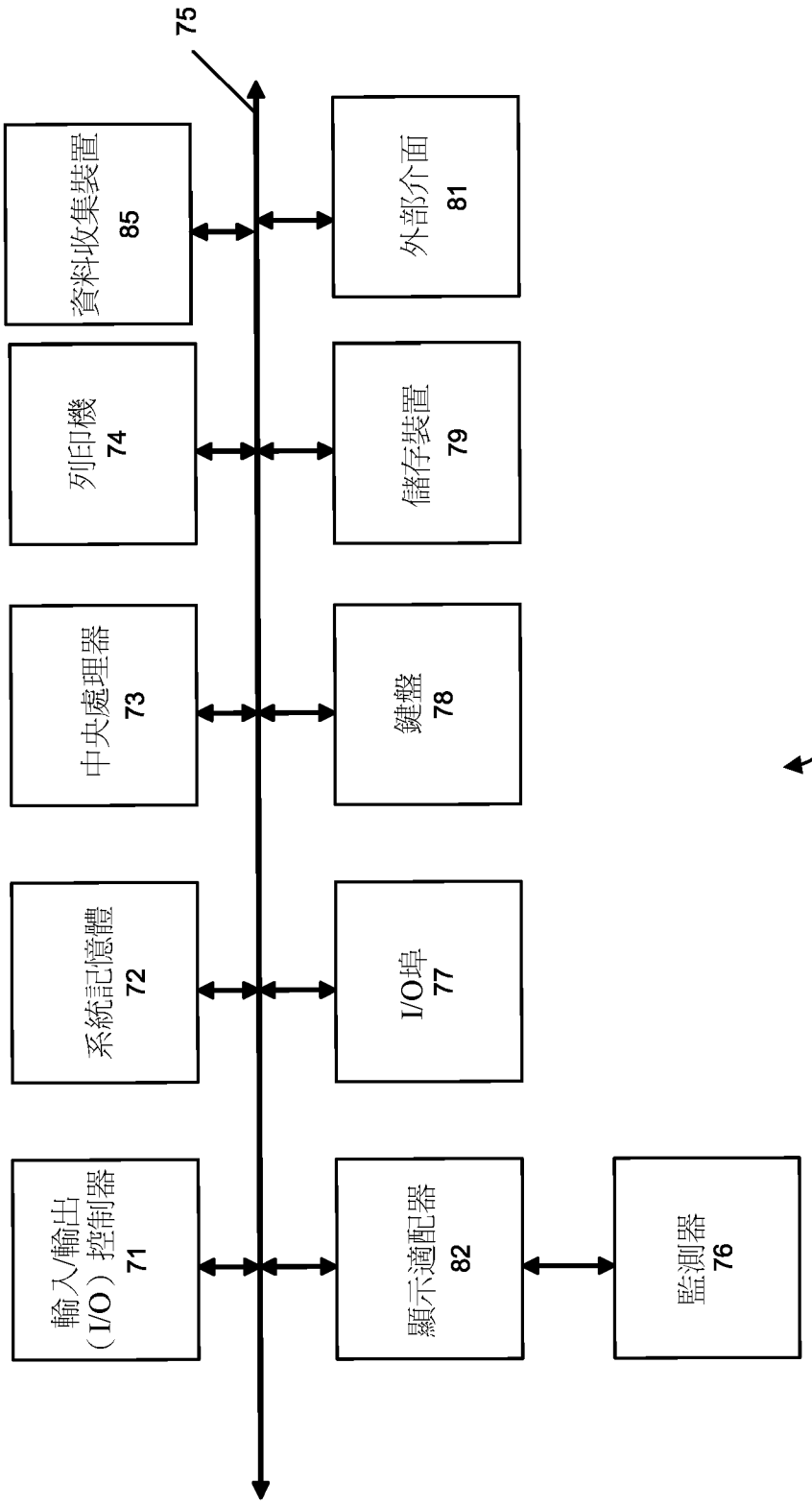
【圖12】


1300
1310
接收複數個第一資料結構
1320
儲存複數個第一訓練樣本
1330
接收第二複數個第二資料結構
1340
儲存複數個第二訓練樣本
1350
使用複數個第一訓練樣本及視情況選用的複數個第二訓練樣本訓練模型，其中模型之輸出指定各個窗口中之目標位置處之核苷酸是否具有修飾

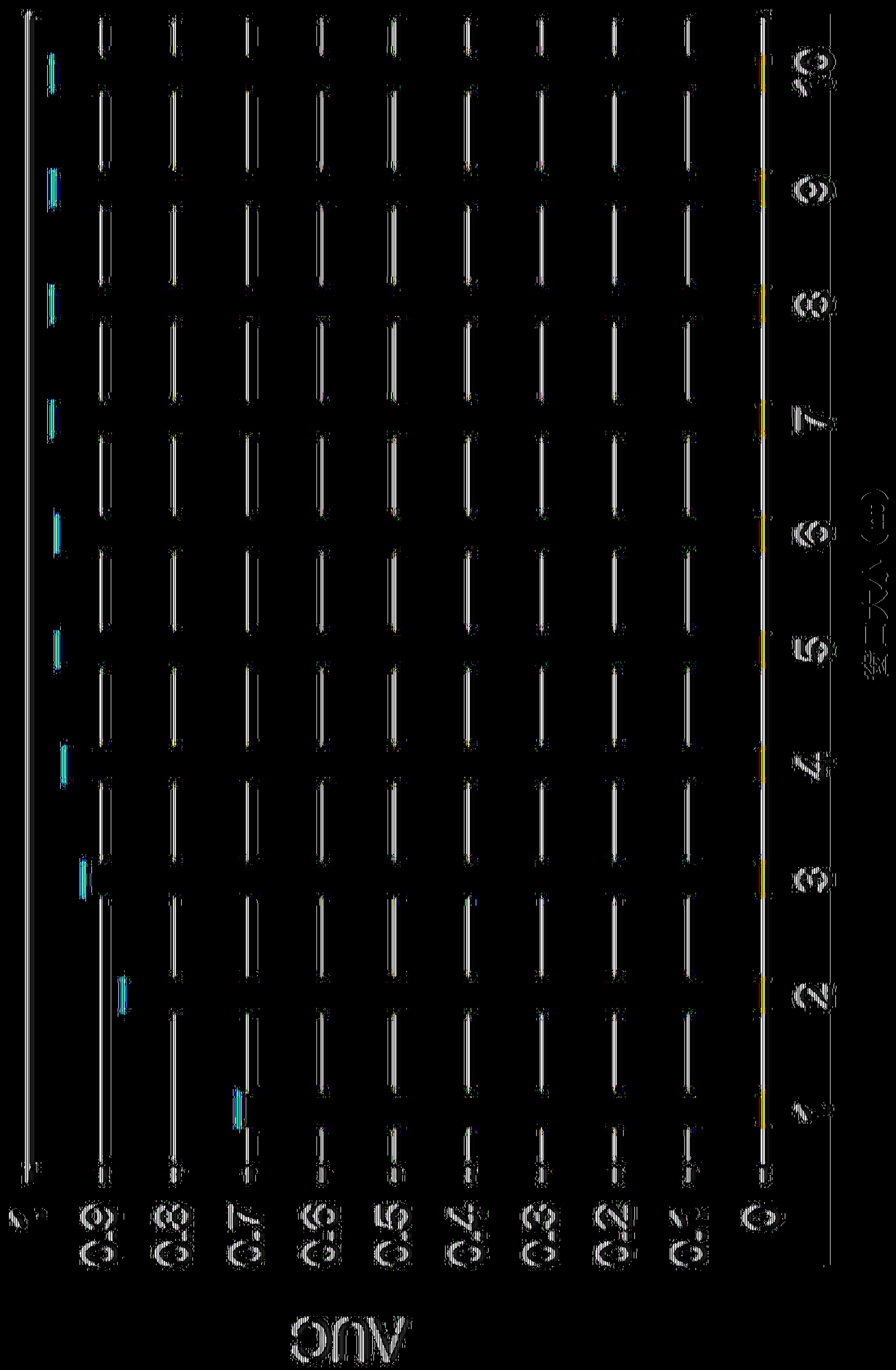
【圖13】


1400
1420
1425
1450
1435
偵測器
處理器
記憶體
邏輯系統
1430
1410
1415
樣本
檢定
樣本架
1405
1408
1440
外部
記憶體
儲存裝置
1445

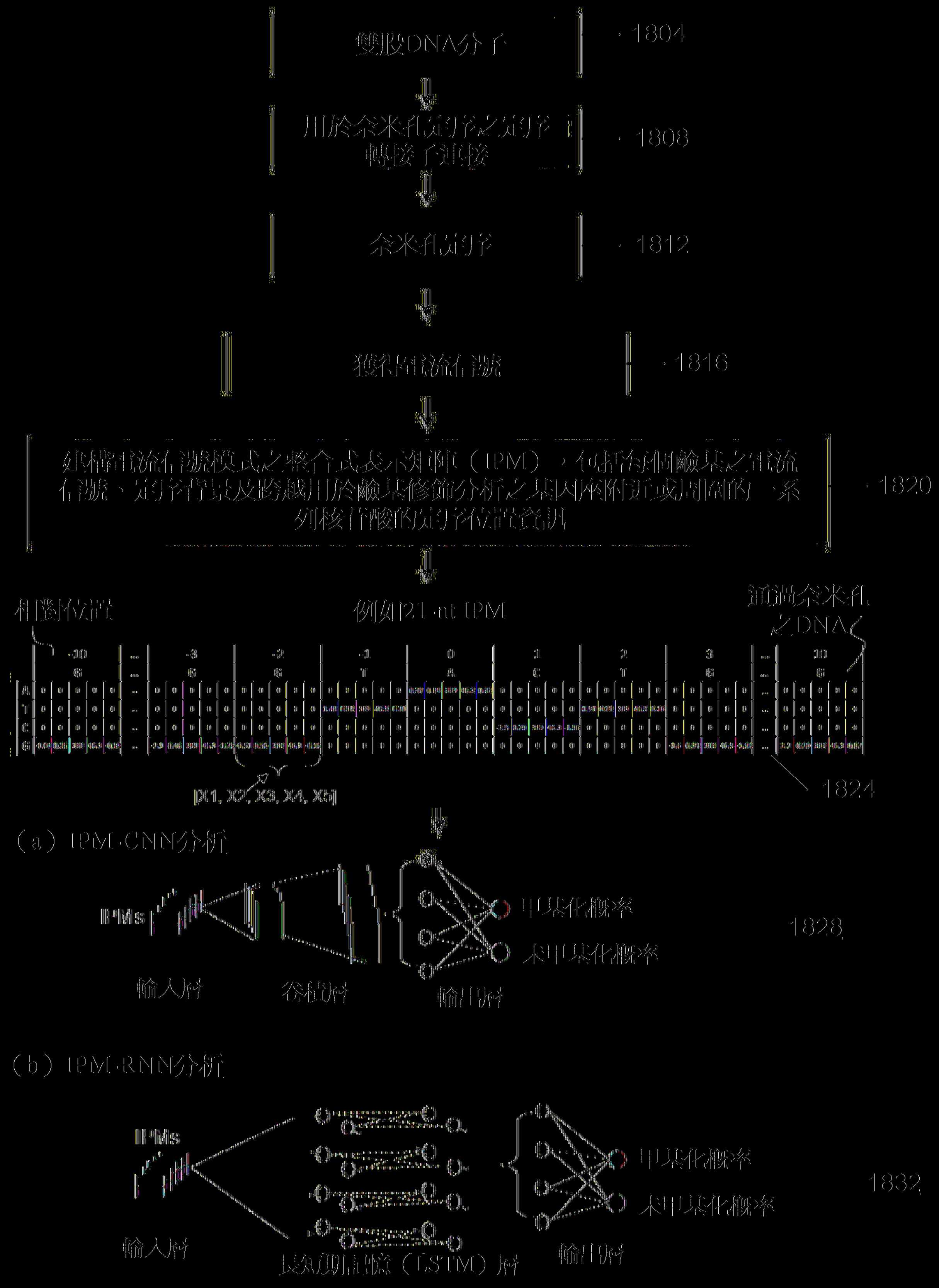
【圖14】


75
資料收集裝置
85
外部介面
81
列印機
74
儲存裝置
79
中央處理器
73
鍵盤
78
系統記憶體
72
I/O埠
77
輸入/輸出（I/O）控制器
71
顯示適配器
82
監測器
76
10

【圖15】

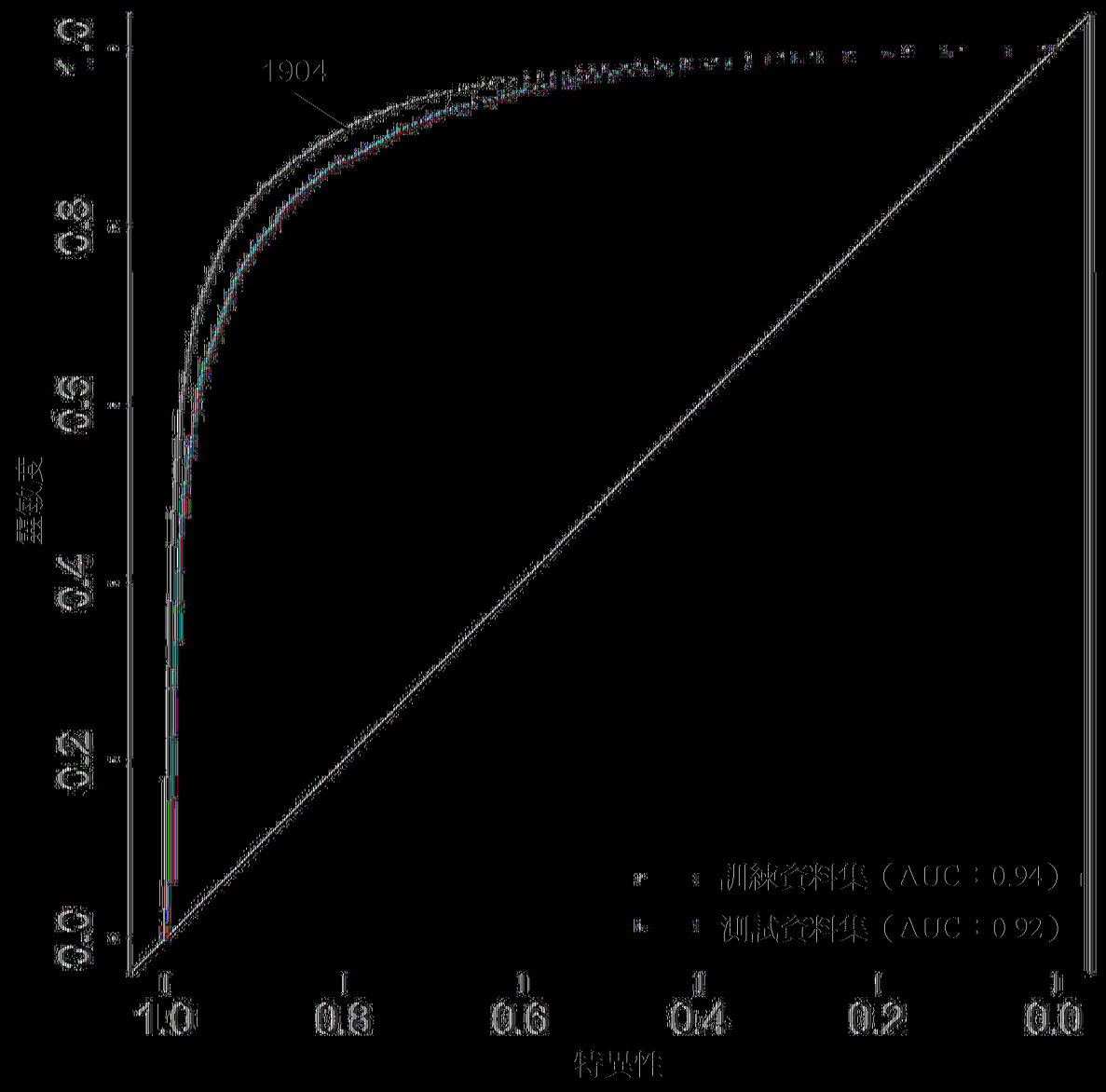


【圖 7】

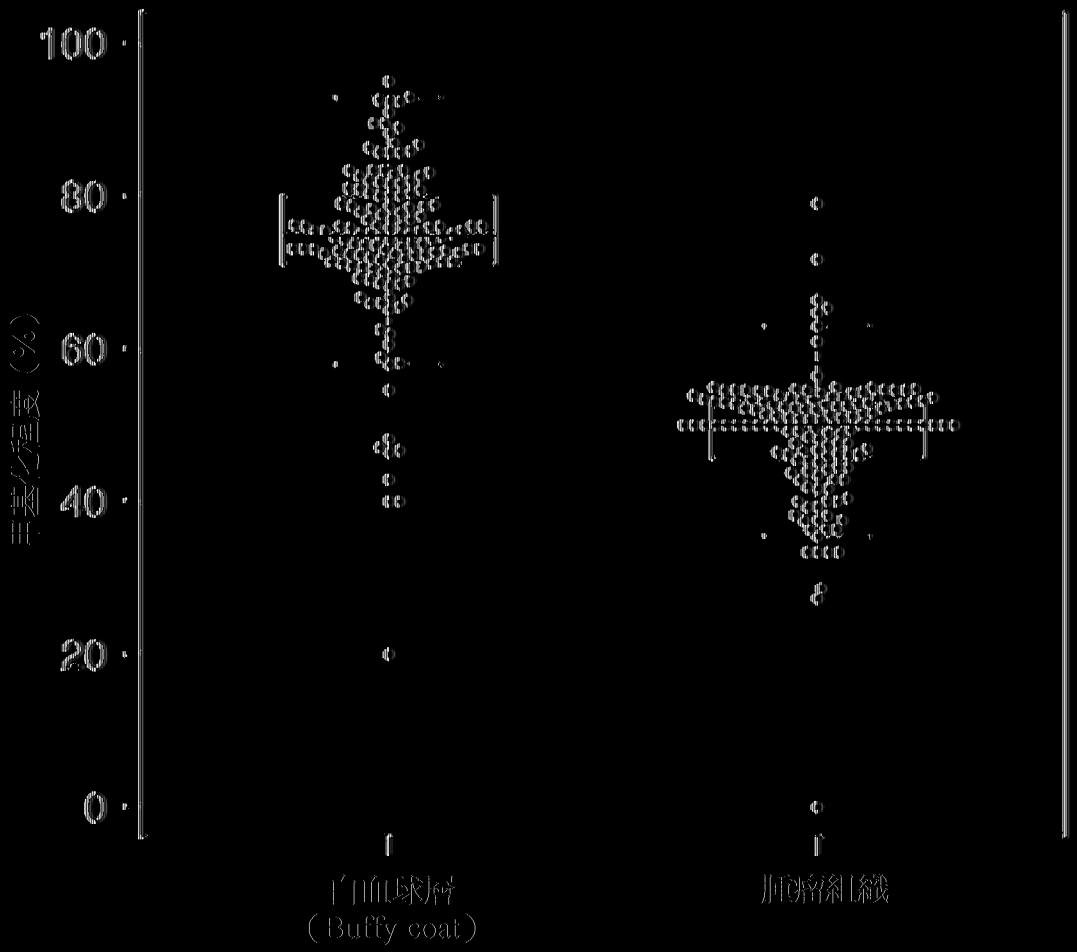


雙股DNA分子
1804
用於奈米孔定序之定序轉接子連接
1808
奈米孔定序
1812
獲得電流信號
1816
建構電流信號模式之整合式表示矩陣（IPM），包括各個鹼基之電流信號、定序背景及跨越用於鹼基修飾分析之基因座附近或周圍的一系列核苷酸的定序位置資訊
1820
相對位置
例如21-nt IPM
通過奈米孔之DNA
[X1, X2, X3, X4, X5]
1824
（a）IPM-CNN分析
IPMs
甲基化機率
未甲基化機率
1828
輸入層
卷積層
輸出層
（b）IPM-RNN分析
IPMs
甲基化機率
未甲基化機率
1832
輸入層
長短期記憶（LSTM）層
輸出層

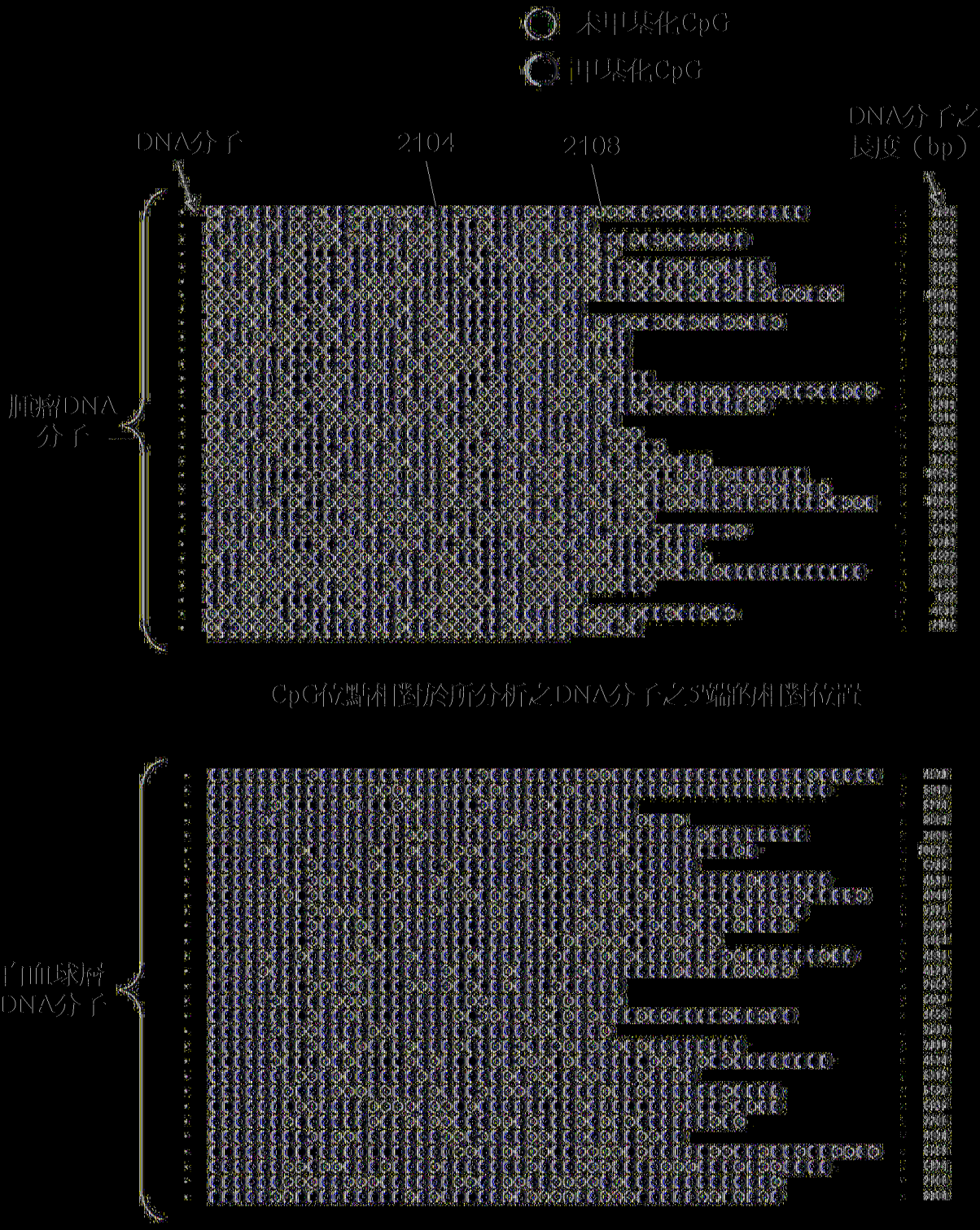
【圖18】


1904
1.0
0.8
0.6
0.4
0.2
0.0
1.0
0.8
0.6
0.4
0.2
0.0
特異性
訓練資料集（AUC：0.94）
測試資料集（AUC：0.92）

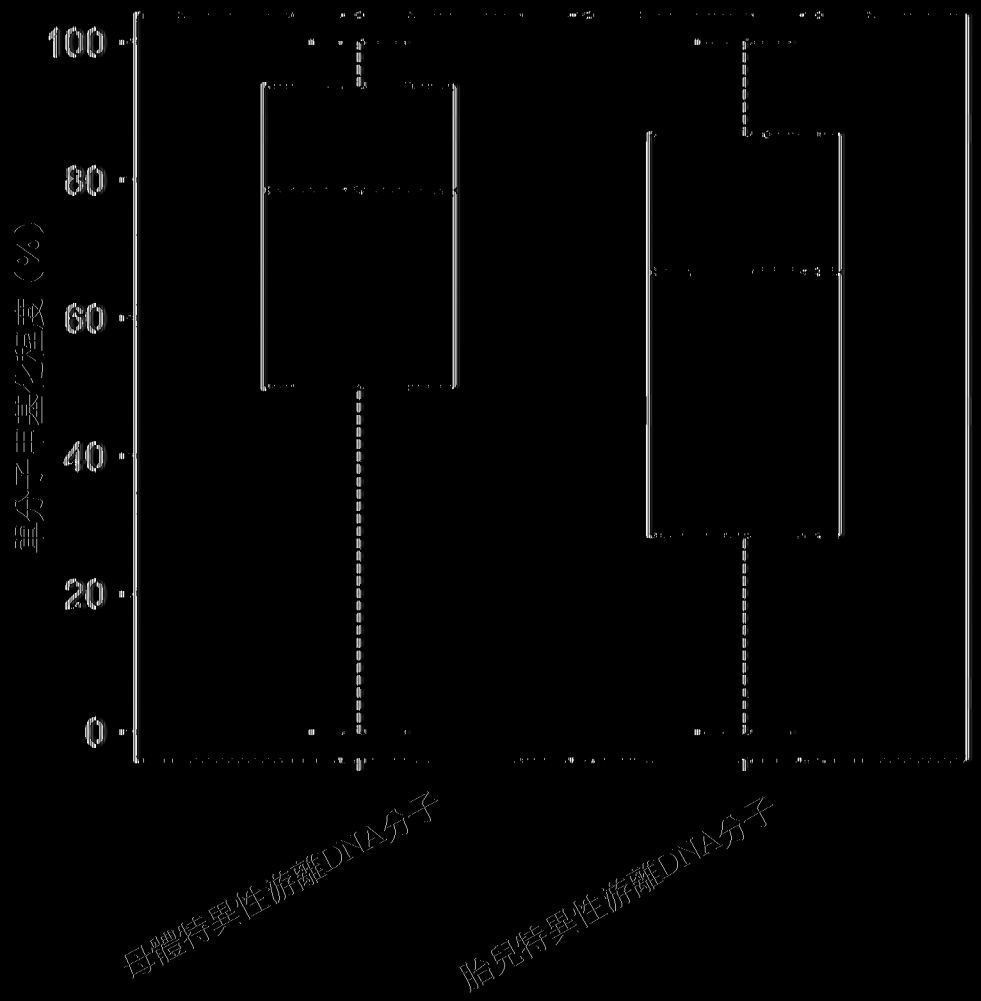
（圖19）



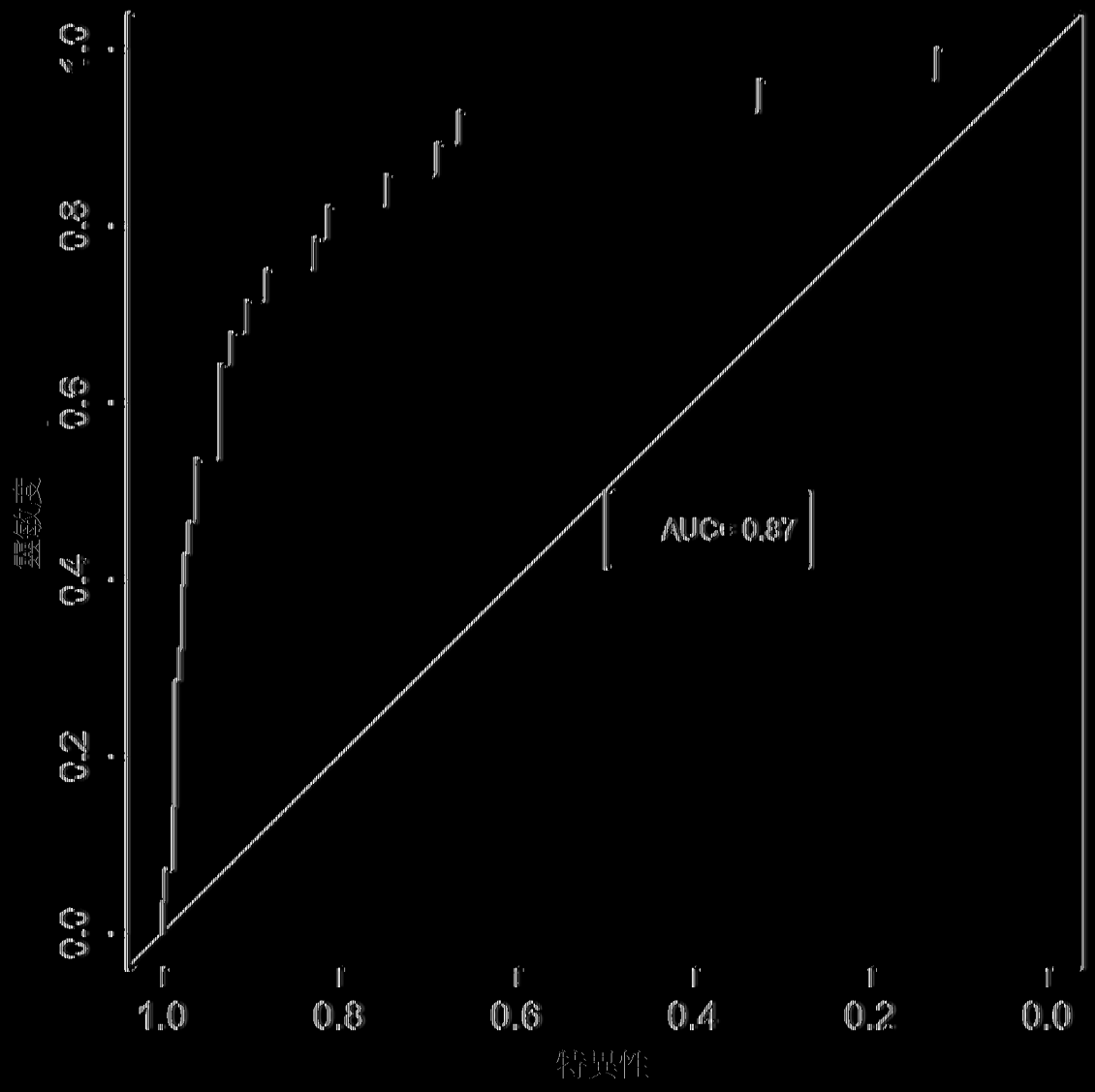
(圖20)


未甲基化CpG
甲基化CpG
DNA分子
2104
2108
DNA分子之長度（bp）
腫瘤DNA分子
CpG位點相對於所分析之DNA分子之5'端的相對位置
游離DNA分子
CpG位點相對於所分析之DNA分子之5'端的相對位置

（圖21）



（圖22.）



（圖23）